



# **Community Analysis Package 6.0**

Copyright 2019, PISCES Conservation Ltd

# Community Analysis Package 6.0

## Searching for structure in community data

---

*by Richard Seaby and Peter Henderson*

*Community Analysis Package is designed to help you identify the pattern and structure that often lies at the heart of ecological communities.*

*It offers a wide range of tried and tested techniques to handle and analyse the complex data that ecological sampling often generates.*

*Using the latest data handling techniques, CAP 6.0 can give you an insight into your data in seconds, and with vivid graphics, enable you to print and publish your results simply, quickly and easily.*

*CAP 6 now integrates with R, to allow you to explore your data further.*

# Community Analysis Package 6.0

**Copyright 2019, PISCES Conservation Ltd**

All rights reserved. No parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - without the written permission of the publisher.

Products that are referred to in this document may be either trademarks and/or registered trademarks of the respective owners. The publisher and the author make no claim to these trademarks.

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

Printed: September 2019 in England

**Publisher**

*Pisces Conservation Ltd*

# Table of Contents

Foreword	I
<b>Part I Introduction to CAP</b>	<b>2</b>
1 Improvements implemented in CAP 5.0	3
Improvements to previous versions of CAP	3
2 Methods offered by CAP	4
3 Installation and System Requirements	4
4 General instructions	5
5 User preferences	5
6 Quick guide to running a data set	6
7 Video guides	8
8 Choose demo data set	8
9 Large data files	9
10 Maximum size of the data set and computation speed	9
11 Common errors and problems	10
<b>Part II Comparing and classifying communities</b>	<b>14</b>
1 Searching for similarity	14
2 Cluster analysis	14
3 Multivariate analysis	15
4 TWINSpan	16
5 Principal Component Analysis (PCA)	16
6 Reciprocal Averaging (RA)	17
7 Discriminant Analysis	17
<b>Part III Demonstration data sets</b>	<b>19</b>
1 Worked Example - Stream Invertebrates	20
2 Worked Example - Japanese Pottery	22
3 Worked Example - Martinsville Igneous	26
4 Worked Example - Cicada Song	30
<b>Part IV Raw Data</b>	<b>34</b>
1 Creating and editing a data set	34
Organising your data for analysis	35
Preparing large data sets in a spreadsheet program	36
Import from Excel	37
Data entry from within CAP	38
Creating a data grid of a certain size	41
Copying and pasting data	42
Editing existing data	44
Saving new or edited data	45
<b>Part V Working Data</b>	<b>48</b>



1 Data transformations .....	49
2 Relativisations .....	49
3 Dealing with zeros or sparse data .....	50
4 Transposing data .....	51
<b>Part VI Grouping</b> .....	<b>53</b>
1 Changing group properties .....	56
<b>Part VII Summary tab</b> .....	<b>59</b>
1 Data set statistics .....	59
<b>Part VIII Ordination</b> .....	<b>63</b>
1 Principal Component Analysis - PCA .....	63
Variance .....	63
Scores .....	64
Eigenvectors .....	65
Viewing the correlation between variables (species) .....	65
Cross products .....	66
PCA plot .....	66
Principal Axis vs Variable Plot .....	69
Scree Plot .....	70
2 Detrended Correspondence Analysis - DECORANA .....	70
Computations - DECORANA .....	71
Species Scores - DECORANA .....	72
Sample Scores - DECORANA .....	73
DECORANA plot .....	74
3 Non-metric Multi-Dimensional Scaling .....	75
Starting Configuration - MDS .....	76
Site coordinates .....	77
Stress .....	78
MDS plots .....	80
4 Reciprocal Averaging - RA .....	82
Computations - Reciprocal Averaging .....	83
Species Scores - Reciprocal Averaging .....	83
Sample Scores - Reciprocal Averaging .....	84
RA plot .....	85
<b>Part IX TWINSpan</b> .....	<b>88</b>
1 Setup window - TWINSpan .....	88
2 TWINSpan Text .....	89
3 Site summary .....	90
4 Variables summary .....	91
5 Dendrogram sites .....	92
6 Dendrogram species .....	93
7 Editing TWINSpan Dendrograms .....	94
8 TWINSpan Out .....	96
<b>Part X Clustering</b> .....	<b>98</b>
1 Agglomerative cluster analysis .....	99
Ward's .....	100

Single linkage .....	100
Complete linkage .....	100
Average linkage .....	100
McQuitty's .....	101
Gower's .....	101
Centroid .....	101
Cluster groups .....	101
Dendrogram - Cluster analysis .....	102
Edit Dendrogram.....	103
Cluster summary .....	105
<b>2 Divisive cluster analysis .....</b>	<b>106</b>
Plot Clusters .....	107
Cluster Groups .....	108
Cluster Summary .....	108

## **Part XI Similarity and Distance Measures 110**

<b>1 Simple matching .....</b>	<b>111</b>
<b>2 S3 .....</b>	<b>111</b>
<b>3 Rogers_Tanimoto .....</b>	<b>112</b>
<b>4 S4 .....</b>	<b>112</b>
<b>5 S5 .....</b>	<b>113</b>
<b>6 S6 .....</b>	<b>113</b>
<b>7 Jaccards .....</b>	<b>114</b>
<b>8 Sørensen .....</b>	<b>114</b>
<b>9 S9 .....</b>	<b>115</b>
<b>10 S10 .....</b>	<b>115</b>
<b>11 Russell &amp; Rao .....</b>	<b>116</b>
<b>12 Kulczynski .....</b>	<b>116</b>
<b>13 S13 .....</b>	<b>116</b>
<b>14 Ochiai .....</b>	<b>117</b>
<b>15 Q1 .....</b>	<b>117</b>
<b>16 Q2 .....</b>	<b>118</b>
<b>17 Kulczynski-Quantitative .....</b>	<b>118</b>
<b>18 Steinhaus .....</b>	<b>119</b>
<b>19 Euclidean .....</b>	<b>119</b>
<b>20 Average .....</b>	<b>119</b>
<b>21 Chord .....</b>	<b>119</b>
<b>22 Geodesic .....</b>	<b>120</b>
<b>23 Manhattan .....</b>	<b>120</b>
<b>24 Mean Character Difference (Czekanowski) .....</b>	<b>120</b>
<b>25 Whittaker .....</b>	<b>121</b>
<b>26 Canberra .....</b>	<b>121</b>
<b>27 Bray-Curtis .....</b>	<b>121</b>
<b>28 Squared Chord Distance .....</b>	<b>121</b>
<b>29 Mahalanobis distance .....</b>	<b>122</b>
<b>30 Renkonen .....</b>	<b>122</b>

<b>Part XII Association analysis</b>	<b>124</b>
<b>Part XIII Group Tests</b>	<b>127</b>
1 Analysis of Similarity (ANOSIM) .....	128
2 Similarity Percentages (SIMPER) .....	129
3 Discriminant Analysis .....	131
Eigenvalues - DA .....	132
Discriminant Function Coefficients .....	133
Fisher's Discriminant Functions .....	134
Group Centroids - DA .....	135
Significance tests - DA .....	136
Discriminant analysis plot .....	137
Dispersion matrices - DA .....	137
Site Coordinates - DA .....	138
Predictive Validation .....	138
<b>Part XIV Variable Filtering</b>	<b>141</b>
1 Variable Filtering - Setup .....	143
<b>Part XV Compare</b>	<b>146</b>
1 Compare samples .....	146
2 Profile Plot .....	147
3 Scatter Plot .....	148
4 Matrix Plot .....	149
<b>Part XVI Printing, editing and saving results</b>	<b>153</b>
1 Exporting and copying charts .....	153
2 Exporting dendrograms .....	154
3 Printing charts and dendrograms .....	155
4 Editing charts .....	157
Zooming on charts .....	158
The Editing Chart dialog .....	160
Drawing a perimeter .....	161
Preparing charts for output - Chart Tools .....	162
Themes for charts .....	167
5 Printing and exporting grid and text output .....	168
<b>Part XVII Obtaining help</b>	<b>172</b>
1 References .....	172
2 Citation .....	173
<b>Part XVIII Run R code</b>	<b>175</b>
1 Setting up R .....	175
Upgrading R .....	177
2 Reset R exe Path .....	177
3 Installing R packages .....	178
4 PCA - Correlation in R .....	179

5	PCA - Covariance in R .....	180
6	DECORANA R .....	181
7	Reciprocal Averaging .....	182
8	MDS - Bray Curtis R .....	183
9	MDS - Jaccard R .....	184
10	Clustering R .....	185
11	ANOSIM R .....	186
12	SIMPER .....	187
13	PCA - Cor - Outlier R .....	188
14	PCA - Covar - Outlier R .....	190

<b>Index</b>	<b>192</b>
--------------	------------

# Licence Agreement

## PISCES LICENSE AGREEMENT

This is a legal agreement between you the end user and PISCES Conservation Ltd. Lymington (PISCES). BY OPENING THIS PACKAGE YOU ARE AGREEING TO BE BOUND BY THE TERMS OF THIS AGREEMENT. IF YOU DO NOT AGREE TO THE TERMS OF THIS AGREEMENT PROMPTLY RETURN THE UNOPENED PACKAGE AND ALL ACCOMPANYING ITEMS (including written material) TO THE PLACE YOU OBTAINED THEM FOR A FULL REFUND.

1. GRANT OF LICENSE - This PISCES License Agreement ('License') permits you to use one copy of the PISCES software product acquired with this License (SOFTWARE) on any single computer, provided the SOFTWARE is in use on only one computer at any time. If you have multiple Licenses for the SOFTWARE then at any time, you may have as many copies of the SOFTWARE in use as you have Licenses. The SOFTWARE is 'in use' on a computer when it is loaded into the temporary memory (i.e. RAM) or installed into the permanent memory (e.g. hard disk, CD ROM, or other storage device) of that computer, except that a copy installed on a network server for the sole purpose of distribution to other computers is not 'in use'. If the anticipated number of users of the SOFTWARE will exceed the number of applicable Licenses then you must have a reasonable mechanism or process in place to assure that the number of persons using the SOFTWARE concurrently does not exceed the number of Licenses. If the SOFTWARE is permanently installed on the hard disk or other storage device of a computer (other than network server) and one person uses that computer more than 80% of the time it is in use then that person may also use the SOFTWARE on a portable or home computer.

2. COPYRIGHT - The SOFTWARE is owned by PISCES or its suppliers and is protected by all applicable national laws. Therefore, you must treat the SOFTWARE like any other copyrighted material (e.g. a book) except that if the software is not copy protected you may either (a) make one copy of the SOFTWARE solely for backup or archival purposes, or (b) transfer the SOFTWARE to a single hard disk provided you keep the original solely for backup or archival purpose. You may not copy the Product manual(s) or written materials accompanying the SOFTWARE.

3. OTHER RESTRICTIONS - You may not rent or lease the SOFTWARE, but you may transfer your rights under this PISCES License Agreement on a permanent basis provided you transfer all copies of the SOFTWARE and all written materials, and the recipient agrees to the terms of this Agreement. You may not reverse engineer, decompile or disassemble the SOFTWARE. Any transfer must include the most recent update and all prior versions.

LIMITED WARRANTY - PISCES warrants that (a) the SOFTWARE will perform substantially in accordance with the accompanying Product Manual(s) for a period of 90 days from the date of receipt; and (b) any PISCES supplied hardware accompanying the SOFTWARE will be free from defects in materials and workmanship under nominal use and service for a period of one year from the date of receipt. Any implied warranties on the SOFTWARE and hardware are limited to 90 days and one (1) year respectively or the shortest period permitted by applicable law, whichever is greater.

CUSTOMER REMEDIES - PISCES'S entire liability and your exclusive remedy shall be, at PISCES option, either (a) return of the price paid or (b) repair or replacement of the SOFTWARE or hardware that does not meet PISCES'S Limited Warranty, and which is returned to PISCES with a copy of your receipt. This Limited Warranty is void if failure of the SOFTWARE or hardware has resulted from accident, abuse or misapplication. Any replacement SOFTWARE will be warranted for the remainder of the original warranty period or 30 days, whichever is longer.

NO OTHER WARRANTIES - TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, PISCES DISCLAIMS ALL OTHER WARRANTIES, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WITH RESPECT TO THE SOFTWARE, THE ACCOMPANYING PRODUCT MANUAL(S) AND WRITTEN MATERIALS, AND ANY ACCOMPANYING HARDWARE. THE LIMITED WARRANTY CONTAINED HEREIN GIVES YOU SPECIFIC LEGAL RIGHTS.

NO LIABILITY FOR CONSEQUENTIAL DAMAGES - TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW PISCES AND ITS SUPPLIERS SHALL NOT BE LIABLE FOR ANY OTHER DAMAGES WHATSOEVER (INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF BUSINESS PROFITS, BUSINESS INTERRUPTION, LOSS OF BUSINESS INFORMATION, OR OTHER PECUNIARY LOSS) ARISING OUT OF THE USE OF OR INABILITY TO USE THIS PISCES PRODUCT, EVEN IF PISCES HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. IN ANY CASE, PISCES'S ENTIRE LIABILITY UNDER ANY PROVISION OF THIS AGREEMENT SHALL BE LIMITED TO THE AMOUNT ACTUALLY PAID BY YOU FOR THE SOFTWARE.

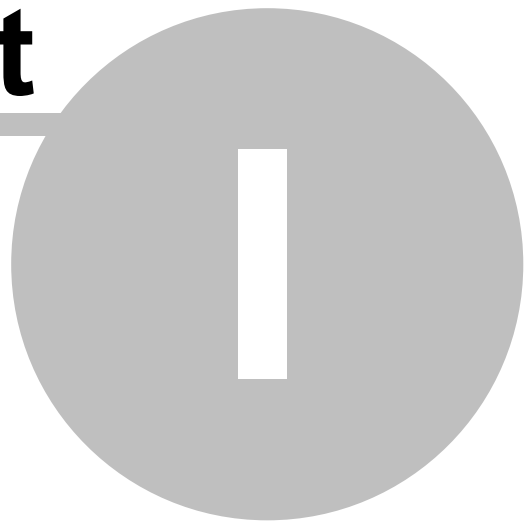
This Agreement is governed by the laws of England.

Should you have any questions concerning this Agreement, or if you desire to contact PISCES for any reason, please use the address information enclosed in this product to contact PISCES or write:

PISCES Conservation Ltd  
IRC House, The Square  
Pennington, Lymington  
Hampshire, England  
SO41 8GN  
Tel 01590 676622

# Part

---



# 1 Introduction to CAP

CAP (Community Analysis Package) is a Windows program that offers a range of analytical techniques commonly used by researchers in fields such as biology, geology, palaeontology, archaeology and the social sciences. CAP is also tested to work on Apple machines using a PC emulator.

Programs to carry out many of these techniques have long been available, but, they are often difficult to use and they frequently have little or limited graphical output. CAP has been designed for ease of use on PCs running under Windows. Data can be organised using standard Windows programs such as Excel and the output from CAP is displayed, exported and printed using standard Windows techniques. This results in a program that is easily used by both students and professionals.

CAP is particularly useful for teaching because it allows students to quickly enter data, try different transformations and explore a range of methods within a familiar Windows setting.

The input data set is arranged as a two-dimensional array. In many scientific disciplines the samples, which are normally collected from set localities and may be called, for example, quadrats, individuals or stations, form the columns. The variables for each sample are the rows, and may comprise for example the numbers of each species or other taxon observed, a score in a test, the frequency of a particular type of object or a chemical concentration.

Throughout this Help system, we have often referred to the contents of the rows as species, for the sake of simplicity, but it should be borne in mind that they may comprise other definitions.

CAP does not include multivariate methods which seek to understand the relationship between two sets of data, as occurs when both physical and biological data have been recorded for a number of samples. However, two other Pisces programs, [Ecom](#) and [Fuzzy Grouping](#), are designed for this type of data.

The methods on offer in CAP include both ordination methods such as Principal Component Analysis ([PCA](#)<sup>[63]</sup>) and Reciprocal Averaging ([RA](#)<sup>[82]</sup>), and classification methods such as [Discriminant Analysis](#)<sup>[131]</sup> and [TWINSpan](#)<sup>[88]</sup>, plus a wide range of clustering procedures. Taken together they provide a powerful suite of methods with which to explore, compare and analyse community structure. For all methods CAP offers high quality graphical and tabulated output which is organised in a tabbed notebook style. The program will run on Apple machines with PC emulation software.

CAP uses the same data structure as Species Diversity and Richness IV which calculates a wide range of diversity and species richness measures. Together with Ecom and Fuzzy Grouping, the programs offer a very extensive range of methods for the analysis of ecological communities and multivariate relationships in general.

CAP 6 introduces the ability to run R code directly from within the program. There have been many other smaller improvements in the program.

CAP 6.0 was developed and produced by Richard Seaby, Peter Henderson and Robin Somes, and was released in August 2019.

[List of methods offered by CAP](#)<sup>[4]</sup>

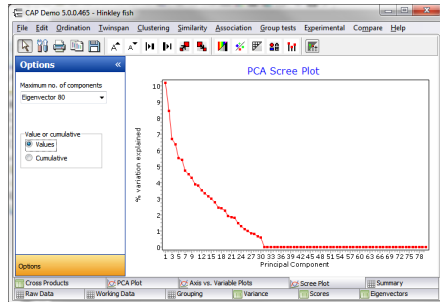
[Improvements implemented in CAP 5.0 \(2014\)](#)<sup>[3]</sup>

[Improvements implemented in older versions of CAP](#)<sup>[3]</sup>

## 1.1 Improvements implemented in CAP 5.0

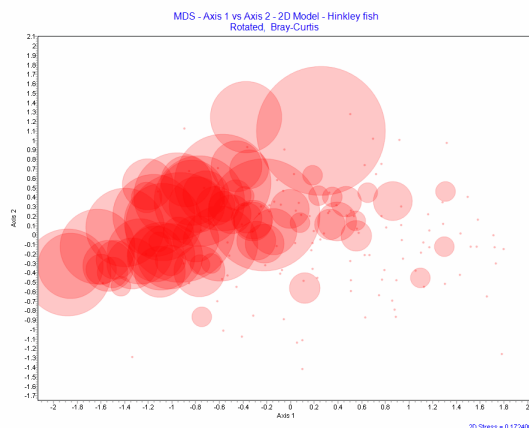
A wide range of new features were added for most of the methods.

1. The methods for selecting and [editing group membership](#)<sup>[53]</sup> were extensively developed.
2. New [dendrogram plotting options](#)<sup>[103]</sup> were implemented.
3. In PCA: [a scree plot](#)<sup>[70]</sup> showing how the fraction of total variance in the data is explained or represented by each Principal Component.



A scree plot offered by PCA

4. [Multidimensional scaling plots](#)<sup>[80]</sup> were improved so that a bubble plot showing the relative abundance of a selected variable in each sample can be displayed. This shows the contribution of the species to the ordination.



An example of a bubble plot generated using the Hinkley fish demo data set.

5. New data exploration options, [Profile Plot](#)<sup>[147]</sup>, [Scatter Plot](#)<sup>[148]</sup> and [Matrix Plot](#)<sup>[149]</sup>.
6. Completely new [Print Preview](#)<sup>[155]</sup> for dendrograms and charts.
7. The video guides were removed from the program, and are now hosted on YouTube.

### 1.1.1 Improvements to previous versions of CAP

The complete list of improvements and changes made to CAP over the years can be found on our website on the [Community Analysis Package](#) page - look for the link called **Update History/Upgrades**.

We are continually striving to improve CAP, and all our other software products, if you have any suggestions or ideas we are always keen to hear them. Please [contact us](#)<sup>[172]</sup>.



## 1.2 Methods offered by CAP

[Data set statistics](#)<sup>[59]</sup>  
[Species correlations](#)<sup>[65]</sup>  
[Data transformations](#)<sup>[49]</sup>  
[Changing the data to relative magnitudes](#)<sup>[49]</sup>  
[Dealing with zeros](#)<sup>[50]</sup>  
[Transposing data](#)<sup>[51]</sup>  
[Analysis of Similarity - ANOSIM](#)<sup>[128]</sup>  
[Similarity percentages -SIMPER](#)<sup>[129]</sup>  
[Discriminant Analysis - Canonical Variate Analysis](#)<sup>[131]</sup>  
[Principal Component Analysis- PCA](#)<sup>[63]</sup>  
[Non-Metric Multidimensional scaling](#)<sup>[75]</sup>  
[Reciprocal Averaging-RA](#)<sup>[82]</sup>  
[Detrended Correspondence Analysis-DECORANA](#)<sup>[70]</sup>  
[Two-way indicator species analysis-TWINSPAN](#)<sup>[88]</sup>  
[Agglomerative cluster analysis](#)<sup>[99]</sup>  
[Divisive cluster analysis](#)<sup>[106]</sup>  
[Similarity measures](#)<sup>[110]</sup>  
[Association analysis](#)<sup>[124]</sup>

### Experimental methods:

[Variable Filtering](#)<sup>[141]</sup>

## 1.3 Installation and System Requirements

### System requirements:

1. A PC running Windows XP or later (It is likely that CAP will run under older versions of Windows, but we regret that we cannot guarantee this).
2. 50 MB of spare hard disk space
3. INTERNET CONNECTION FOR THE VIDEO GUIDES

CAP does not limit the size of your data set, however your hardware will. To use CAP with very large data sets (1000 or more species or samples) you will need a fast modern machine with 1 GB memory or more to perform such calculations quickly. For more information, see [Maximum size of the data set](#)<sup>[9]</sup>.

### Download Installation:

1. Run the downloaded exe file: the installation process should begin automatically - follow the on-screen instructions.
2. When installation is complete, there will be a CAP entry under Start: Programs. An uninstall facility will also be created, in case you wish to remove the program, and a Desktop shortcut.

### Older versions of CAP:

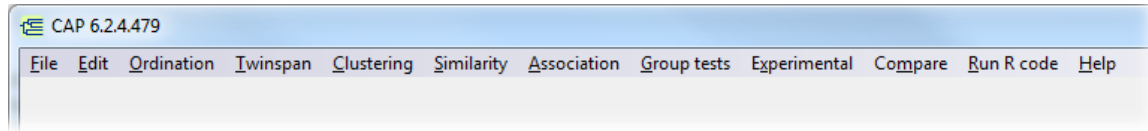
If you already have an old version of CAP installed on your computer, it is not necessary to remove it before installing CAP, since the new version will be installed to a different location.

**Uninstalling CAP:** Note that the CAP folder in My Documents, and its sub-folder Rtemp, may need to be deleted manually after uninstalling.

## 1.4 General instructions

Start CAP in the normal Windows fashion either by clicking on the Desktop icon or from Start: All Programs: CAP 6.

Along the top bar are a number of drop-down menus. These work in the same way as other standard Windows programs.



**File:** To open, re-open, create, export, print and save data sets.

**Edit:** To cut copy and paste to and from the active window.

**Ordination:** To select an ordination method to apply to the data.

**Twinspan:** To initiate a Twinspan analysis.

**Clustering:** To select a cluster analysis.

**Similarity:** To display similarity measures between samples.

**Association:** To perform a Chi squared association test.

**Group tests:** To assign samples to groups and look for similarities between the groups.

**Experimental:** To use the experimental [Species Filtering](#)<sup>[141]</sup> method.

**Compare:** To compare samples in terms of the variables present.

**Run R Code:** To use R code snippets natively within the CAP program.

**Help:** to enter the Help system.

When the program is started, you will be presented with a blank working area. Use File: Open or Reopen to select and open your data file, or File: New to create a new data set. The raw data will be displayed on the [Raw data](#)<sup>[34]</sup> page; there is a corresponding [Working data](#)<sup>[48]</sup> page where you can manipulate and transform the data. Initially, the Raw Data and Working Data will display the same values – except that CAP will remove any zero-sum rows or columns from the Working Data array. Once a transformation or some other adjustment has been applied to the data (for details see below) the Working Data form will display the adjusted data. The analyses will only use the Working Data set; your original (Raw) data will be unchanged, unless you choose to save the transformed data set over the existing file.

[Obtaining help](#)<sup>[172]</sup>

[Quick guide to running a data set](#)<sup>[6]</sup>

[Demonstration data sets](#)<sup>[19]</sup>

[Creating and editing a data set](#)<sup>[34]</sup>

[Maximum size of the data set and computation speed](#)<sup>[9]</sup>

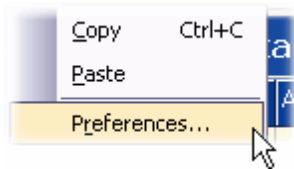
[Printing and exporting your results](#)<sup>[153]</sup>

[Zooming on charts](#)<sup>[158]</sup>

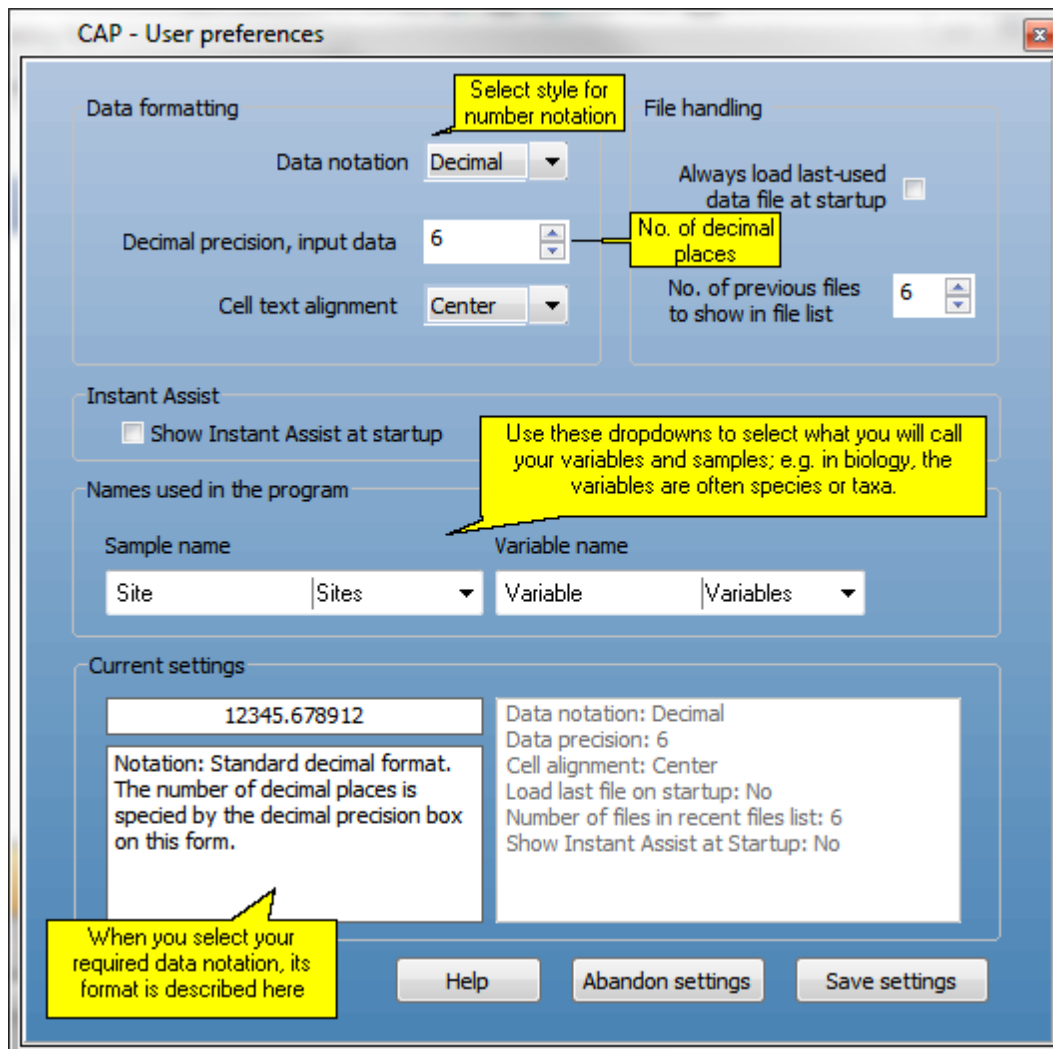
For more information about using CAP and a [video demonstration](#)<sup>[8]</sup> see the **Help: Guides: Introduction**.

## 1.5 User preferences

CAP's user preferences offer considerable control over the display and handling of your data. From the **Edit** menu, choose Preferences.



The following screen will be displayed:



Settings for data notation, precision and alignment can be set in the Data formatting group. The Data notation drop-down allows you to select between Decimal, Scientific, General and Number; these choices are explained in the 'Current settings' text box at bottom left.

In the File handling group, the options to reload the last file you were using whenever the program starts up, and also to set the number of files visible in the file list, can be set. You can also choose whether or not to show the Instant Assist feature on program startup.

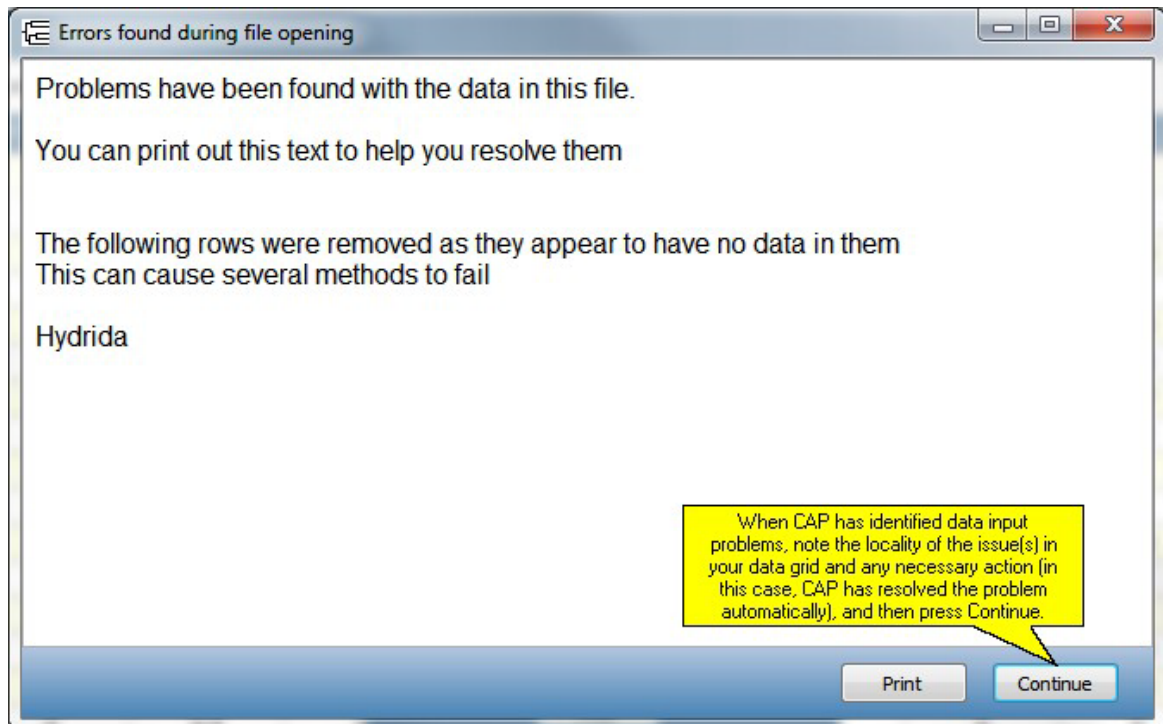
## 1.6 Quick guide to running a data set

With CAP data can be imported directly from an Excel spreadsheet as a native Excel file (.xls), or as a Comma-separated value (.csv) or text (.txt) file - see [Preparing large data sets](#)<sup>[36]</sup>, and [Import from Excel](#)<sup>[37]</sup> for further details.

1. Start CAP and open your data file from the **File** menu. The default file type is .csv; use the drop-down menu to select either .xls or .txt if required.

2. The data will now open in the Raw Data window. CAP will alert you if there are any problems - such as rows or columns that sum to zero - which might cause the calculations to malfunction. Sometimes it will require you to take some action, such as replacing a cell's contents, otherwise it will make the required changes automatically. CAP will only make these changes to the [Working Data](#)<sup>[48]</sup> grid; the Raw Data will always contain the complete original data array. The stored data file will not be altered unless you use Save or Save As to save the Working Data over the existing data file.

The alert screen is shown below. When you press Continue, the data will automatically be transferred to the Working Data tab.



3. If necessary, edit the [Raw Data grid](#)<sup>[34]</sup> to correct any problems highlighted during the file opening.
4. On the [Working Data tab](#)<sup>[48]</sup>, undertake any transformations or relativisations you require. Note that any changes undertaken on the working data will not change the raw data. Once an adjustment has been made, remember to click on the Submit button to create adjusted data sets for analysis.
5. Select a method from the menu toolbar, choose the setup options and run the analysis. Use the tabbed windows to examine the output. All aspects of the charts and plots can be altered using the buttons on the toolbar above the chart (shown below). For more instructions on editing charts, see [Printing and saving results](#)<sup>[153]</sup>.



6. Use **File: Print** or the **Printer** button on the toolbar to print the charts, and **Edit: Copy** or the **Copy** button to copy it to the clipboard for pasting into another application.
7. Use the **Save** button on the toolbar to save the chart as an image file for use in another

application, or as a TeeChart Pro (\*.tee) file. See [exporting charts](#)<sup>[153]</sup> for further details.

For more information and a video demonstration see the **Help: Guides: Data In**.

## 1.7 Video guides

Previous versions of CAP included video guides installed with the program. While the guides are very useful, the installation of the videos occasionally caused anti-virus software to flag one of them as being infected with a trojan. It was proven that there was no such infection, but the problem periodically resurfaced.

Instead, we now host all the instructional videos on YouTube. This has several advantages:

1. The size of the installation file is less than 50% of the old version, making for quicker and easier downloading when buying digitally. This is especially useful for users in remote locations with slow internet connections.
2. Faster and simpler installation.
3. No more false warnings from anti-virus software about trojan infection.
4. We can easily issue updated versions of the video guides.

You can find our YouTube channel [here](#).

The individual guides featured are:

[General introduction to CAP](#)

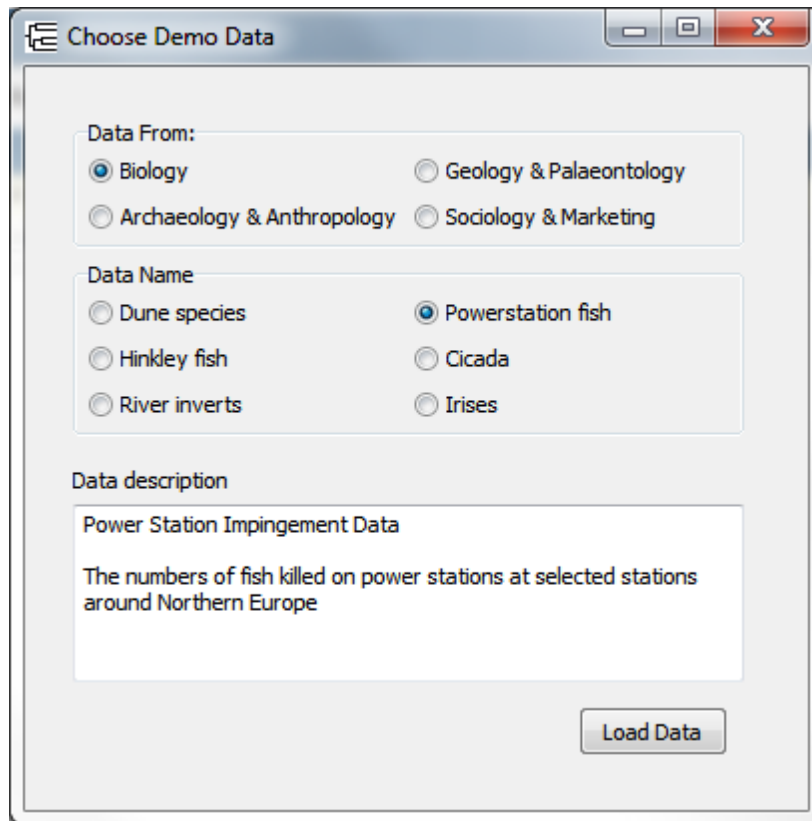
[Input & output](#)

[Creating groups](#)

## 1.8 Choose demo data set

CAP comes with 16 demo data sets, which have been chosen to demonstrate the full range of capabilities of the program; most of the examples shown in this Help system use one or another of these data sets. They fall in to 4 categories; Biology, Geology & Palaeontology, Archaeology & Anthropology, and Sociology & Marketing. They are all described, and 4 detailed worked examples are provided, in the [Demonstration Data Sets section](#)<sup>[19]</sup>.

To open a demo data set, click **File: Open a demo data set**, and select the one you require from the dialog:



## 1.9 Large data files

If you are planning to use large datasets with CAP then you may find the PISCES LIST COMBINER very useful. Old versions of Excel have a limit of 255 columns per worksheet. LIST COMBINER will take all your worksheets and combine them all into one large csv file (summing and sorting the values for each variable along the way) ready to load directly into CAP. See [www.pisces-conservation.com/softutils.html](http://www.pisces-conservation.com/softutils.html) for more details.

See also [Maximum size of the data set](#)<sup>9</sup>.

## 1.10 Maximum size of the data set and computation speed

In CAP, array handling technology is used to enable the analysis of very large data sets.

Theoretically, the size of the input data matrix is unlimited, although, in practice, the memory resources of the PC are usually the limiting factor. In tests, we have run CAP with a data array of 5000 variables x 5000 samples without difficulty.

Because of its unusual computational requirements, which involves the generation of new data items in the form of pseudospecies, TWINSpan is more demanding in memory usage. The maximum size of the data set whilst running TWINSpan is therefore likely to be smaller than for the other multivariate methods. It should still easily be sufficient for most users' needs, however.

Our approach to data handling means that on a reasonably modern PC, most of the computation, even on very large data sets, can be completed within a few seconds. All computations with data sets of up to 100 samples by 100 variables are usually almost instantaneous.

If you do run into speed problems with very large data sets, it will generally be improved if there are no other Windows applications running at the time.

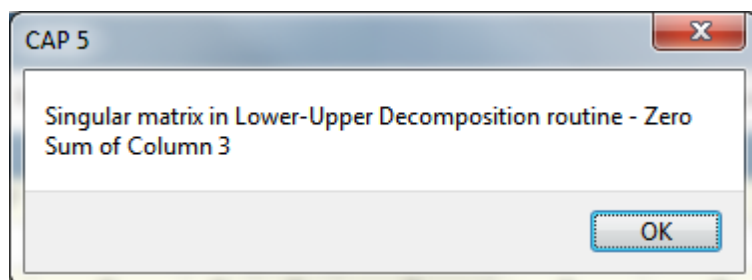
## 1.11 Common errors and problems

### Common errors and problems

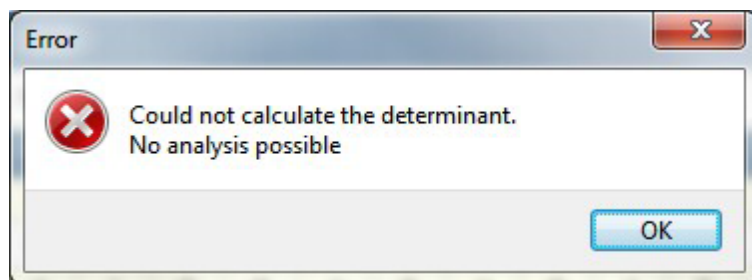
No matter how much care is taken with data preparation, it is not unusual to encounter problems when initially loading and running analyses on a data set. These are usually simple to rectify. Common error messages and their solutions are shown below.

#### **1. Singular Matrix in Lower-Upper Decomposition routine.**

Some data sets simply will not work with a Discriminant Analysis, if there is low (or zero) variability in the samples/sites, or if there are close correlations between sites or variables. The program will show two error messages; the first shows where in the data set the problem lies:

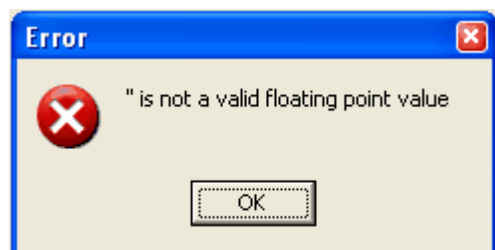


while the second states that no analysis is possible:



#### **2. " is not a valid floating point value.**

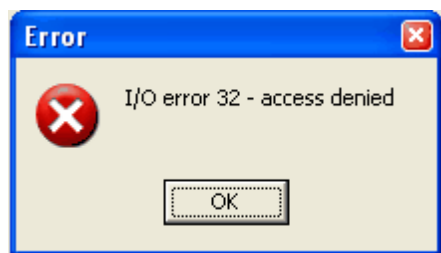
This will occur if the raw data set holds blank columns or rows - ones that sum to zero. Remove blank columns and rows by using '[Handling zeros](#)'<sup>[50]</sup> in the Working Data tab. It may also occur if the raw data holds a blank cell. In some cases CAP will identify the problem cell which should be edited. Normally it is because the data has been prepared in a spreadsheet using blanks to represent zero values.



Occasionally, this error can occur because a blank space or a character has been accidentally entered into a cell outside the data matrix when it was being prepared in a spreadsheet program. To prevent this happening, it is good practice, before saving your data set as a .csv file, to highlight the first 10 or so blank rows and columns below and to the right of the data matrix, and press 'Delete'. This will clear the cells of any accidentally-entered contents.

### 3. I/O error 32 - access denied.

This will occur if the data file you are trying to open is currently being used by another program - normally the spreadsheet which was used to organise the data.



Close the file in other programs and try again.

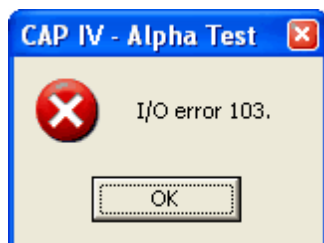
### 4. When a data file is opened, all the data are dumped into the first cell, rather than opening in the grid properly.

This is usually because there are one or more blank cells in the second row of data. If the blanks are replaced by zeros, or another row with no blanks is put in the second position, the problem should disappear. Row 2 is the crucial one; blank cells are tolerated elsewhere.

### 5. I entered the data in a spreadsheet program, and saved it as a csv file. When I try to open it in CAP, the numbers are separated by ';' not ',' so the analysis will not run.

This is because some non-British systems use ; instead of , to separate values in a csv file. Open Windows Explorer, and change the file extension from csv to txt (i.e. Filename.csv becomes Filename.txt). Open the txt file in Word or another word-processing program. Use the Find/Replace function to replace every ; with a , then save the file again. Change the txt file extension back to csv. It should now open and run perfectly in CAP.

### 6. I attempted to run TWINSpan using the default settings but got an I/O error 103.



If you have large numbers in your data set this may occur. The reason is that, as part of its normal operation, TWINSpan adds some very large numbers to each data value, in order to increase the discrimination between them. If you already have large numbers the resulting values may be too large for the data types used within the program. The solution is to transform your data. Usually a square root transformation will resolve the matter.

### 7. Interpreting TWINSpan output.

In TWINSpan OUT: concerning the list of Indicators and their sign (i.e. Species 1 (+), Species 2 (-)), does the (+) represent species present in the class and (-) represent species NOT present in the class? If so, then are the only sample points (quadrats) representing the TWINSpan classes those listed under the positive group?

In the TWINSpan output Species1 (+) indicates that species1 is characteristic of the quadrats classified to the right of the centroid of the primary axis of the ordination. It does not mean that the Species1 is absent in all the quadrats to the left of the centroid. An output of, say, Species2 (-) would mean that Species2 is characteristic of quadrats to the left of the centroid. Remember that if TWINSpan is not undertaken on presence/absence data; it uses pseudospecies so the same



species can be an indicator at different levels of abundance.

In the case of presence/absence data there will certainly be a strong tendency for a positive species only to be present in the group of quadrats to the right. However, this is not necessarily always the case.

#### **8. Invalid floating point operation.**

This often indicates that there are rows or columns with no data in them. Use the Handling Zeros button on the working data tab to find and remove the offending row or column.

#### **9. My files do not appear to be saved.**

This is often caused by the lack of the correct file extension. By default, the Open File dialog shows .csv files. To see other files in the Open File dialog, select .xls, .txt or All Files from the Files of type: drop-down menu. WARNING: the latter option will show all the files in the directory - whether they are compatible with the program or not. If you have saved a file without an extension either add the extension outside the program (open the folder in Windows Explorer or My Computer, single-click the file and press F2 to edit the file name), or open the file and save with the correct extension using 'Save as...'

#### **10. When I press F1, no Help screen appears.**

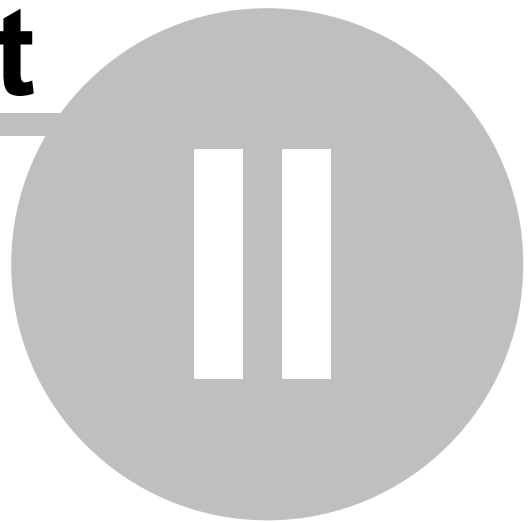
Ensure that the window on which you are seeking help is the active one.

#### **11. I want to use a similarity or distance measure that is not offered by CAP.**

[Contact PISCES](#)<sup>[172]</sup> and we will try to implement the measure for you.

# Part

---



## 2 Comparing and classifying communities

Even a quite modest field survey can produce a bewildering amount of information on the presence and abundance of species, pottery or fossils, and it is frequently difficult to identify and summarise the main features and inter-relationships between communities. This section describes the different approaches and techniques you can use. The aim is to describe the ways in which your data can be presented and explored, and little attention is given to the theoretical background of the techniques.

Useful sources for information about the mathematical background to these techniques are [Legendre & Legendre \(1998\)](#)<sup>[172]</sup> and [Kent & Coker \(1992\)](#)<sup>[172]</sup>.

[Organising your data for analysis](#)<sup>[35]</sup>

[Searching for similarity](#)<sup>[14]</sup>

[Multivariate analysis](#)<sup>[15]</sup>

[Cluster analysis](#)<sup>[14]</sup>

[TWINSpan](#)<sup>[16]</sup>

[Principal Component Analysis \(PCA\)](#)<sup>[16]</sup>

[Non-metric Multi-Dimensional Scaling \(n-MDS\)](#)<sup>[75]</sup>

[Reciprocal averaging \(RA & DECORANA\)](#)<sup>[17]</sup>

[Discriminant Analysis](#)<sup>[17]</sup>

### 2.1 Searching for similarity

When we compare the flora or fauna sampled at different localities, we often wish to know how similar are their species assemblages. Numerous methods have been devised for the measurement of similarity, the most successful of which are described below. [Legendre & Legendre \(1998\)](#)<sup>[172]</sup> give a more complete account of similarity and distance measures.

[Similarity indices](#)<sup>[110]</sup> are simple measures of either the extent to which two habitats have variables in common (Q analysis), or variables have habitats in common (R analysis). Binary similarity coefficients use presence-absence data, and more complex quantitative coefficients can be used if you have data on species abundance. When comparing the variables at two localities, indices can be divided into those that take account of the absence of a variable from both communities (double zero methods), and those that do not. In most ecological applications it is unwise to use double-zero methods as they assign a high level of similarity to localities which both lack many species. We would not normally consider two sites highly similar because their only common feature was the joint lack of a group of species, which could occur because of sampling errors or because both sites were unsuitable.

### 2.2 Cluster analysis

When numbers of sites or habitats are to be compared, the [similarity measures](#)<sup>[110]</sup> offered by CAP can form the basis of cluster analysis, which seeks to identify groups of sites, or stations that are similar in their species composition.

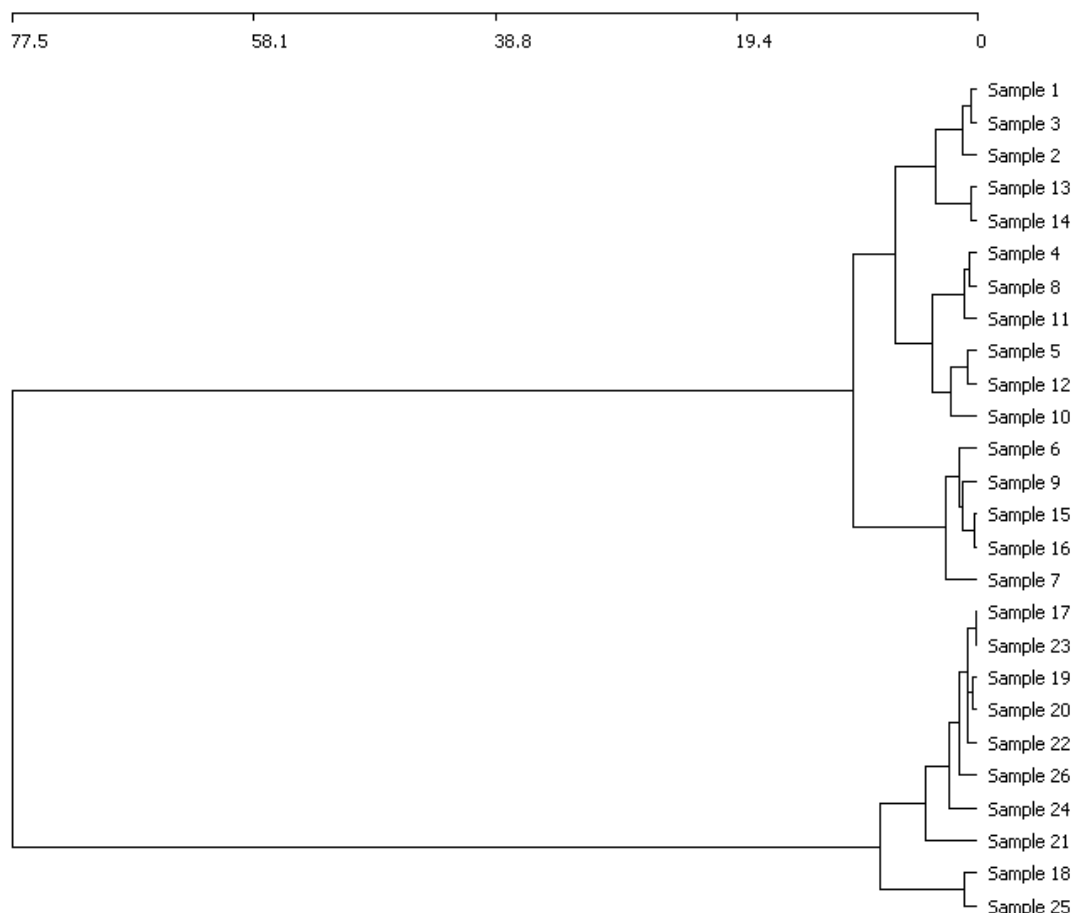
Classification methods comprise two principal types, hierarchical, where objects are assigned to groups that are themselves arranged into groups as in a dendrogram, and non-hierarchical, where the objects are simply assigned to groups. The methods are further classified as either [agglomerative](#)<sup>[99]</sup>, where the analysis proceeds from the objects by sequentially uniting them, or [divisive](#)<sup>[106]</sup>, where all the objects start as members of a single group which is repeatedly divided. For computational and presentational reasons hierarchical-agglomerative methods are the most popular.

The basic computational scheme used in cluster analysis can be illustrated using single linkage

cluster analysis as an example. This is the simplest procedure and consists of the following steps.

1. Start with  $n$  groups each containing a single object (sites or variables).
2. Calculate, using the similarity measure of choice, the array of between-object similarities.
3. Find the two objects with the greatest similarity, and group them into a single object.
4. Assign similarities between this group and each of the other objects using the rule that the new similarity will be the greater of the two similarities prior to the join.
5. Continue steps 3 and 4 until only one object is left.

The results from a cluster analysis are usually presented in the form of a dendrogram:



The problem with all classification methods is that there can be no objective criteria of the best classification; indeed even randomly-generated data can produce a pleasing and convincing dendrogram. Always consider carefully whether the groupings identified seem to make sense and reflect some feature of the natural world.

If you wish to use R see [Run R code](#)<sup>175</sup>

## 2.3 Multivariate analysis

Multivariate analysis is used when the objective is to search for relationships between, or classify objects that are defined by, a number of attributes. Generally, we seek to show the relationship between sites (or samples) using the measured variables (eg species) as the attributes. Data sets can be large, for example marine benthic or forest beetle faunal studies can easily require analysis of a matrix of 100 samples (stations) by 350 species, and thus multivariate analysis requires a computer. If the objective is to assign objects to a number of discrete groups then cluster analysis

should be considered. If there is no *a priori* reason to believe the objects will or could naturally fall into groups, then an ordination technique may be more suitable. Ordination assumes the objects form a continuum of variation and the objective is often to generate hypotheses about the environmental factor(s) that mould community structure.

There is a considerable literature on multivariate techniques. Useful texts for ecologists are [Legendre & Legendre \(1998\)](#)<sup>[172]</sup>, [Digby & Kempton \(1987\)](#)<sup>[172]</sup> and [Kent and Coker \(1992\)](#)<sup>[172]</sup>.

## 2.4 TWINSpan

[TWINSpan](#)<sup>[88]</sup> was developed for, and is still mostly used by, botanists. However, the method can be used by other disciplines, although in fields such as archeology the concept of a pseudospecies may be difficult to present. This method can be used with presence-absence, % cover and quantitative data. The ability to effectively handle % cover makes the method particularly attractive to botanists. The Two-Way INdicator SPecies ANalysis procedure (Hill *et al.* 1975) also produces dendrograms of the relationship between species and samples, but uses the Reciprocal Averaging ordination method to order the species and samples. Thus the method is something of a hybrid between classificatory and ordination methods. It is particularly attractive in studies where the objective is to classify communities so that field workers can quickly assign an area to a community type, and is much favoured by botanists. This is because it identifies indicator species characteristic of each group identified.

TWINSpan is a useful technique when you are seeking to identify species that can be used to characterise particular communities. It is, however, not always an easy method to understand. A particular oddity of the method is the concept of 'pseudospecies'. Each species is divided into a number of pseudospecies which represent the different abundance levels at which it was found.

## 2.5 Principal Component Analysis (PCA)

[Principal Component Analysis](#)<sup>[63]</sup> (PCA) is the oldest and still one of the most frequently-used ordination techniques in community ecology. It is most appropriate for full quantitative data, but can be used if abundance is classified into a number of abundance classes. The objective of the method is to express the relationship between the samples in a 2- or 3-dimensional space that can be plotted and usefully visualised. This can only be achieved if many of the variables are positively or negatively correlated. Normally this will be so for a number of reasons. First, there is the interdependence between organisms in an ecosystem, and second, because many variables respond similarly to environmental variables such as temperature and water.

General descriptions of the procedure for biologists are given [Legendre & Legendre \(1998\)](#)<sup>[172]</sup>, [Digby & Kempton \(1987\)](#)<sup>[172]</sup>, [Kent & Coker \(1992\)](#)<sup>[172]</sup>.

The analysis is undertaken on either the between-sample variance-covariance matrix, or the correlation matrix. If the variables vary greatly in abundance you will probably need to transform the data by taking logarithms or using a square-root transformation. Logarithmic transformations would be excellent if it were not for the fact that zeros cannot be handled. A frequently-used procedure is to add 1 to all the observations. This can distort the output, and so it is probably more appropriate to use a square-root transformation.

If you undertake a PCA on the correlation matrix you will be giving all variables, irrespective of abundance, equal weighting, whereas the analysis undertaken on the variance-covariance matrix will reflect differences in abundance, but can result in the numerically-dominant variables determining the output. When successful, PCA will present major features of a complex community in only 2 or 3 dimensions and the ordination of samples (sites) along these new axes can be related to underlying environmental factors that are moulding community structure. PCA can be judged a success when the first two or three principal axes explain an appreciable proportion of the total variability in the data set. For large ecological data sets with > 20 species, if the three largest axes can explain more than 30% of the variance, this would generally be considered satisfactory.

If you wish to use R see [Run R code](#)<sup>[175]</sup>

## 2.6 Reciprocal Averaging (RA)

[Reciprocal Averaging](#)<sup>[82]</sup> (also termed Correspondence Analysis) and an adjusted version called Detrended Correspondence Analysis ([DECORANA](#)<sup>[70]</sup>) are the final types of ordination method that are frequently used. They are best used on quantitative data, although they can give good results with classed abundance data. Both of these methods are particularly effective when it is suspected that the sites (samples) can be arranged along an environmental gradient. Further, the method allows the site and variable (e.g. species) ordinations to be plotted on the same figure, which allows the influence of the variables in determining the ordination of the samples to be uncovered.

If you wish to use R see [Run R code](#)<sup>[175]</sup>

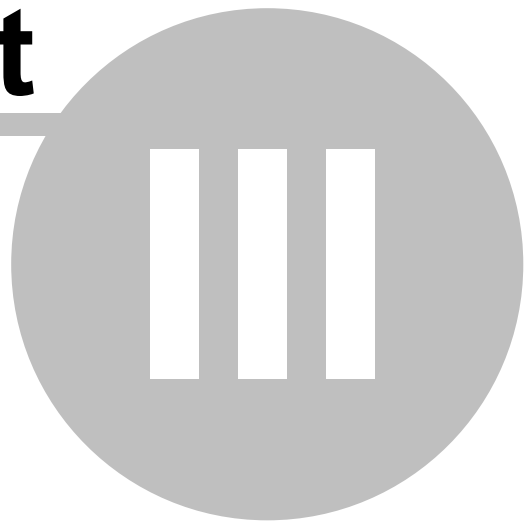
## 2.7 Discriminant Analysis

[Discriminant analysis](#)<sup>[131]</sup> (also called canonical variate analysis) is a standard method for testing the significance of previously-defined groups, identifying and describing which variables distinguish between groups, and producing a model to allocate new samples to a group. DA allows the relationship between groups of samples to be displayed graphically.

A goal of a discriminant analysis is to produce a simple function that, given the measurements of a number of variables, will classify an object into a known group. The point to note is that for DA you require pre-defined groups.

# Part

---



### 3 Demonstration data sets

Installing CAP also installs a selection of demonstration data sets of widely differing size and type. These allow you to experiment with the program and by opening these files in Excel or another spreadsheet, see how the data are organised. Files are embedded within the program.

#### From biology

Aldabra Atoll vegetation survey data - *aldabra veg.csv*

Coastal dune vegetation survey - *dune species.csv*

Hinkley Point fish survey - *Hinkley fish.csv*

Southern English chalk stream invertebrate survey - *River inverts.csv* - see [Worked Example Stream Invertebrates](#) <sup>[20]</sup>

Data on fish caught on power stations in N. Europe - *Powerstation fish.csv*

Data on the songs of different cicada species - *cicada.csv* - see [Worked Example Cicada Song](#) <sup>[30]</sup>

Data on the morphology of 3 species of iris flower - *irises.csv*

The *Hinkley fish.csv* file holds the actual monthly captures of fish collected at Hinkley Point in the Severn Estuary over a 14 year period. You can find more information on this exceptional time series, which is now in its 34th year, [on our website](#).

#### From archeology and anthropology

Measurements on Egyptian skulls through time - *Egyptian skulls.csv*

Analysis of Japanese potsherds - *Jomon Hall.csv* - see [Worked Example Japanese Pottery](#) <sup>[22]</sup>

Seriation of Ancient Greek finds - *Melos.csv*

Analysis of Nigerian potsherds - *Nigerian pottery.csv*

Comparison of chemistry of Romano-British pottery - *Romano British pottery.csv*

East Anglian employment in the middle ages - *16 century east anglia.csv*

#### From geology and palaeontology

Analysis of Ordovician molluscs - *Ordovician fossils.csv*

Petrology of igneous rock Martinsville Virginia - *Petrology.csv* - see [Worked Example Martinsville Igneous](#) <sup>[26]</sup>

#### Sociology and marketing

Soft drink preferences - *beverages.csv*

Lawnmower use - *lawnmower.csv*



### 3.1 Worked Example - Stream Invertebrates

To help you understand the full capabilities of the program and its methods, we present below a worked example, using one of the data files installed with the program: *River inverts.csv*.

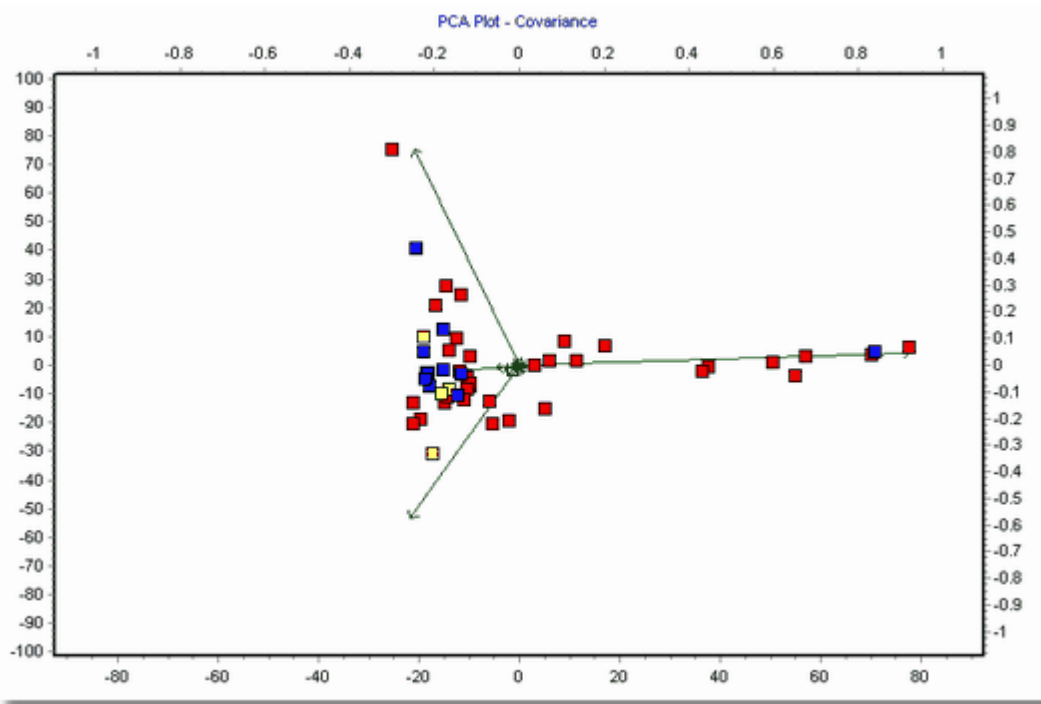
The data were collected in 2000 as part of an investigation into the effects of river restoration on biodiversity in several chalk streams in southern England. The report on the investigation, *Effects of physical restructuring of channels on the flora and fauna of three Wessex rivers*, can be downloaded in Adobe Acrobat PDF format from our website at [consult.pisces-conservation.com/latestreports.html](http://consult.pisces-conservation.com/latestreports.html)

Concentrating on macroinvertebrates, the data comprise percentage composition by family from restored and unrestored reaches of the 3 rivers and 2 tributaries. There are 61 families (rows) across 49 sample sites (columns).

The site names are coded with the initial letter of the river, the site name, and 'U' or 'R' for unrestored or restored. The rivers are the Wylye, its tributary the Till, the Piddle, its tributary the Devil's Brook, and the River Avon. So for instance, the unrestored stretch at Hyams Farm on the Avon is coded Ahyam U, and the restored site at Great Wishford on the Wylye is coded Wgwish R.

To begin the example, open the data file from the File: Open a demo data set menu. On opening the file, a warning box pops up, to inform you that the row titled 'Hydrida' has no data - i.e. all the values are zero. The program will automatically remove this row from the [Working Data](#) set.

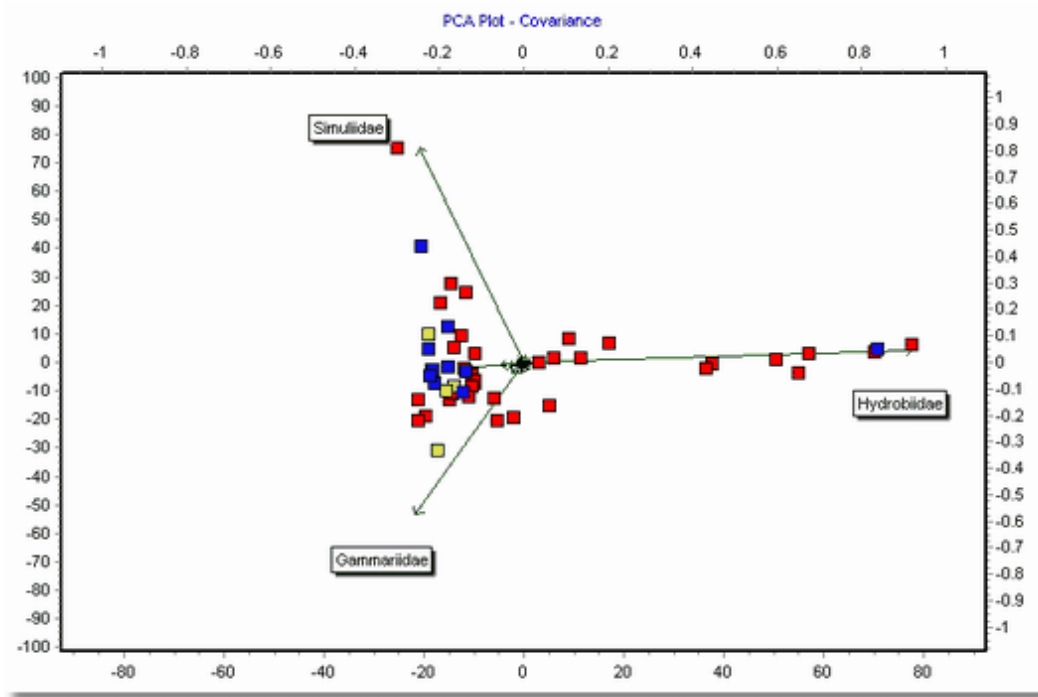
Let us first perform a Principal Component Analysis; from the Ordination menu, choose PCA - Covariance. Click on the PCA Plot tab which appears at the bottom, to view the resulting chart (shown below with all labels removed, for clarity). Note that, because the PCA depends on a randomisation of the data, the chart may appear as a mirror image of the one below. This is unimportant.



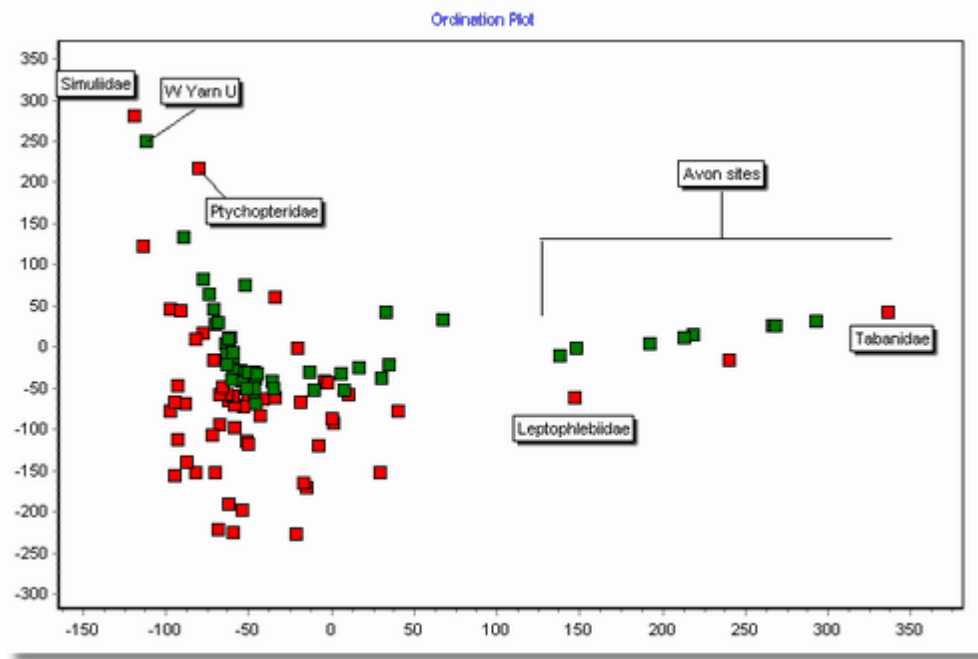
Immediately, we can see that most of the species vectors are tightly grouped in the centre of the plot, but there are 3 strong vectors and a number of sites associated with them. By experimenting

with the labelling, we can see that the three families are Simuliidae, Hydrobiidae and Gammaridae. By looking at the site labels, it becomes obvious that the 8 sites plotted in a line to the right are the A sites, from the River Avon. The vector suggests that these sites are strongly correlated with the Hydrobiidae. Similarly, the site at the top left, Wyarn U, is strongly correlated with the Simuliidae. We can confirm this from the working data; over 90% of the individuals in the Wyarn U sample were Simuliidae.

To give the chart more impact, we should now label the chart to show the vector titles etc; follow the instructions for [adding annotations](#)<sup>[164]</sup> under [Preparing Charts for Output](#)<sup>[162]</sup>.



Similar conclusions can be reached by performing Reciprocal Averaging; choose Ordination : Reciprocal Averaging to perform the analysis. After adding [annotations](#)<sup>[164]</sup> and [lines](#)<sup>[164]</sup>, the plot appears as below:



W Yarn U is correlated with the Simuliidae and Ptychopteridae, whereas the group of 8 Avon sites is strongly associated with Hydrobiidae, Tabanidae and Leptophlebiidae.

## 3.2 Worked Example - Japanese Pottery

Demonstration data set: *Jomon Hall.csv*.

Reference: Hall, M. E., 2001. *Pottery styles during the early Jomon period: geochemical perspectives on the Moroiso and Ukishima pottery styles*. *Archaeometry* 43, 59-75.

This example is based on the study by Hall (2001) of pottery shards from the early Jomon period (c. 5000±2500 BC). Energy-dispersive X-ray fluorescence was used to determine the concentration of 15 minor and trace elements in 92 pottery sherds. The sherds came from four sites in the Kanto region and belong to either the Moroiso or Ukishima style of pottery.

The author reasoned that if the pottery were locally produced, we should expect to find statistically significant differences in the chemical composition between potsherds from different sites. If there are no differences between sites, then we can assume that the Jomon potters utilized raw materials that were geochemically similar, or that the pottery was part of a trade/exchange/redistribution network between settlements. For each sherd the elemental composition of barium (Ba), copper (Cu), gallium (Ga), iron (Fe), lead (Pb), manganese (Mn), nickel (Ni), niobium (Nb), rubidium (Rb), strontium (Sr), thorium (Th), titanium (Ti), yttrium (Y), zinc (Zn) and zirconium (Zr) were measured.

### Preliminary data examination and transformation

The concentration of the elements present varied greatly from about 105 ppm for iron (Fe) to around 10 ppm for yttrium (Y). The author therefore undertook a log 10 transformation on the data to reduce the dominance of Fe and Ti in the analysis. Given the 5 orders of magnitude difference in concentrations and the fact that the data set holds no zeros, a log transformation is a good choice. Niobium was removed from the data set prior to analysis as it was generally below the detectable limit.

### The use of the correlation matrix

PCA was done on the correlation matrix of the log-transformed data. By using the correlation matrix the author was giving all the elements the same influence on the final ordination. This is the correct choice if it is believed that all elements can potentially equally contribute to the identification of

similarities between sherds, irrespective of concentration. In fact for these data the author would have reached substantially similar conclusions if the variance-covariance matrix of log-transformed data had been used instead.

### Results

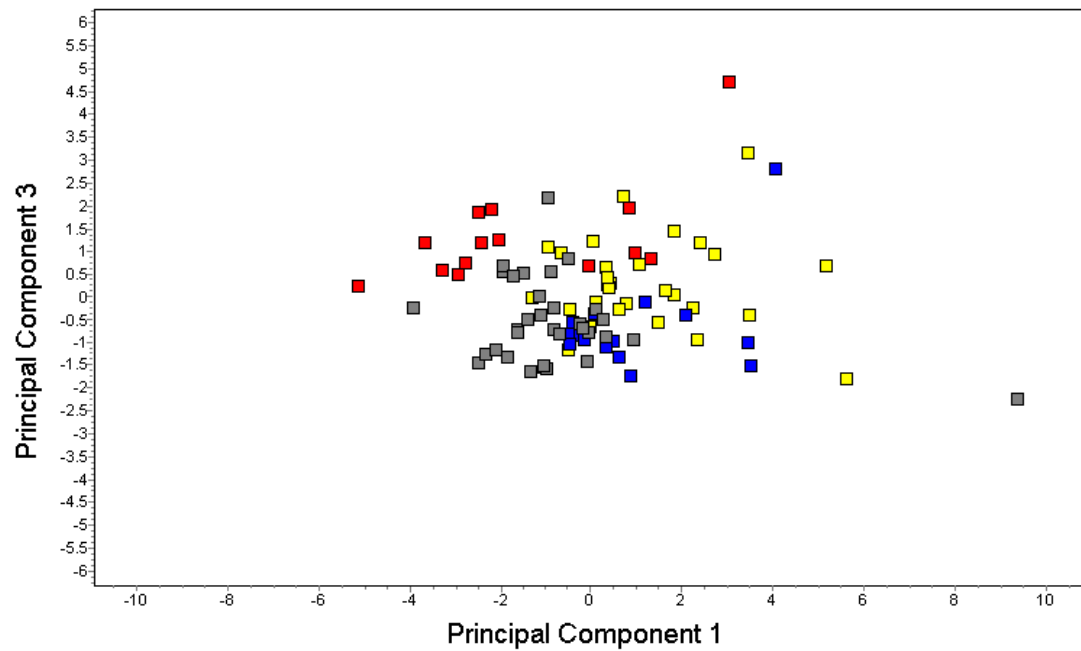
As shown in the table below, the first 3 axes explained about 57.89 % of the total variability in the data set. The sum of all the eigenvalues, which is a measure of the total variability, is 14, which is simply the sum of the number of variables used in the analysis. This summation is always true when a correlation matrix is used. Therefore the percentage variability explained by the largest eigenvalue is  $4.713/14 \times 100 = 33.66\%$ .

	Eigenvalues	Cumulative percentage of the total variance
1	4.713	33.66
2	1.954	47.62
3	1.437	57.89
4	1.228	66.66
5	0.9771	73.64

These results suggest that much of the variability in elemental composition can be expressed in 3 dimensions. The first 4 dimensions are probably meaningful (eigenvalues > 1).

An examination of the 3 2D plots possible for the 3 largest components showed that the position of the sherds in the 2 dimensional space defined by the 1st and 3rd principal components separated the sherds into the 4 localities (Fig 1). Four sample outliers in the PCA were a:mb:002, k:uk2:008, n:uk:137 and s:mb:007. For example, the sherd s:mb:007 is represented by the grey square on the far right of the plot. By repeating the analysis with the outliers removed (Fig 3) we can see more clearly the grouping of the shards between the 4 sites. In the figure, each of the sites is coded as a different colour. You will see for example that the blue squares, representing the Narita 60 site are clustered in a single discrete area.

### PCA Plot - Correlation Matrix - Jomon Sherds - all data



### PCA Plot - Correlation Matrix - Outliers Removed

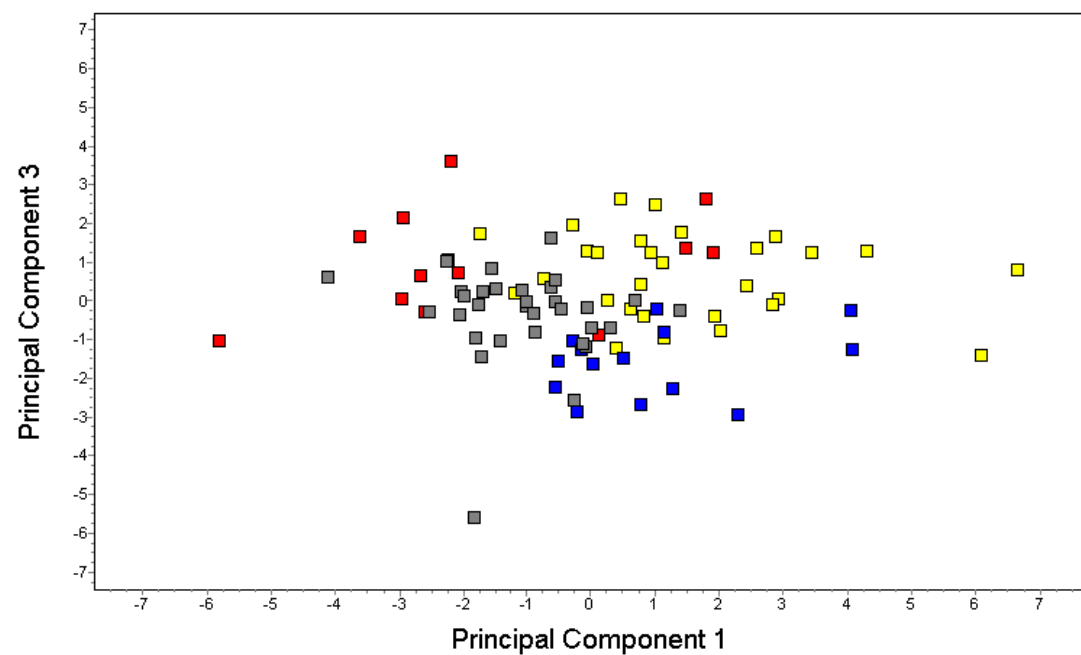
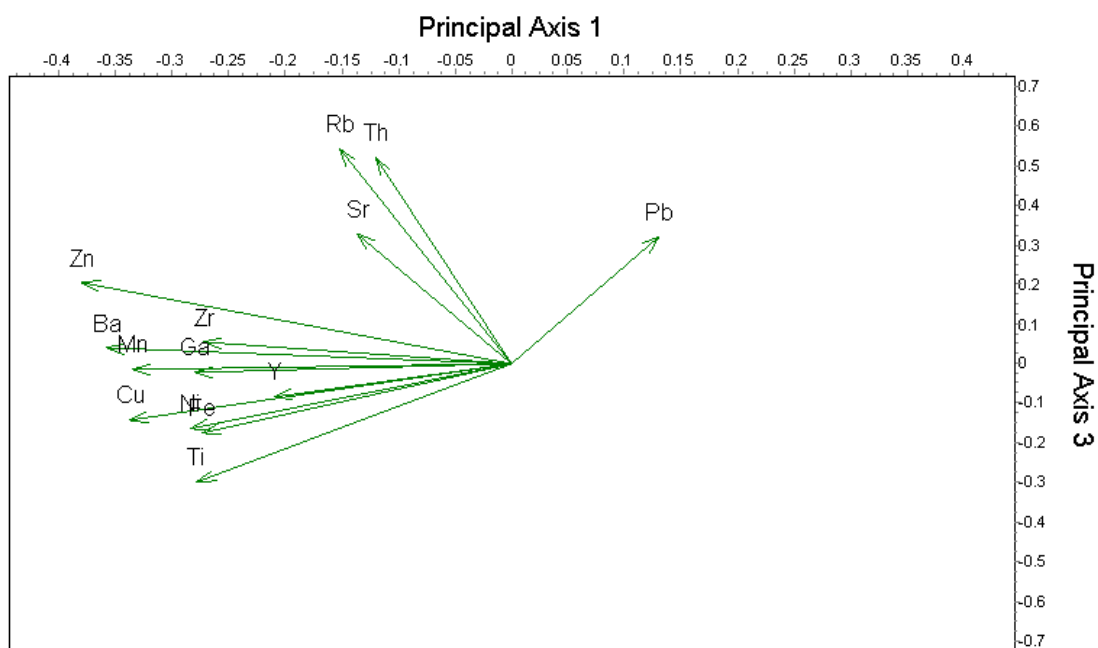


Figure 1: PCA ordination of Jomon potsherds. Red – Aryoshi-kita, Yellow – Kamikaizuka, Blue - Narita 60 and Grey - Shouninzuka.

## PCA Plot - Correlation Matrix - Outliers Removed

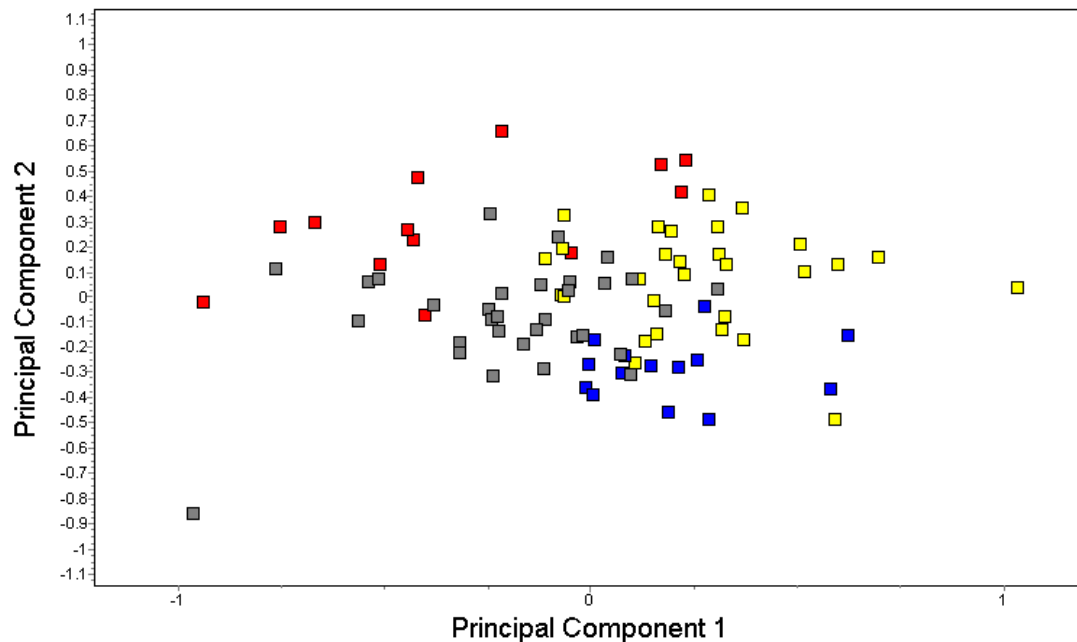


**Figure 2: A plot of the eigenvectors for the 14 elements used for the PCA.**

The plot of the eigenvectors in Fig 2 shows that Principal axis 1 is a measure of the concentration of the elements Zn, Ba, Mn, Zr, Ga, Cu, Ni, Fe, Y and Ti present, with sherds to the left (negative direction) of the axis having the largest concentrations. Axis 3 is a measure of Sr, Rb, Th and Pb concentration with the greatest concentrations at the top (positive direction) of the axis.

The samples were also classified by pottery style – Red is Moroiso A, Yellow Moroiso B and Blue Ukishima. By comparing the plot showing the sites and the styles it is apparent that the Ukishima pottery is found at all the sites. Note that there is some difference in the elemental composition in the pottery styles, with Moroiso sherds having generally higher concentrations of all the elements measured.

### PCA Plot - Variance-Covariance Matrix - Outliers removed



**Figure 3: PCA ordination of Jomon potsherds using the Variance-covariance matrix. Red – Aryoshi-kita, Yellow – Kamikaizuka, Blue - Narita 60 and Grey - Shouninzuka.**

#### Conclusions

Hall (2001) concluded that Principal Component Analysis indicates that there are four major groups in the data set, which correspond to site location. This indicates the majority of Early Jomon pottery found at four sites in Chiba Prefecture was made from locally available raw materials. While the Kamikaizuka and Shouninzuka groups overlap, both sites are less than 10 km apart and their potters could have shared the same raw material sources.

For sites having both Moroiso and Ukishima pottery, both styles of pottery were made from the same or geochemically similar raw materials. This suggests that both styles were probably made at the same site, and indicates that if the different pottery styles are reflecting ethnic identity, then intermarriage between ethnic groups is occurring. Alternatively, the pottery styles could be reflecting some sort of social interaction between groups.

### 3.3 Worked Example - Martinsville Igneous

Demonstration data set: *Petrology.csv*

Reference: P. C. Ragland, J. F. Conley, W. C. Parker, and J. A. Van Orman, 1997, *Use of principal components analysis in petrology: an example from the Martinsville igneous complex, Virginia, U.S.* A. Mineralogy and Petrology 60:165-184.

This example is based on the study by Ragland *et al.* (1997). This paper examined the utility of PCA for the analysis of the relationship between geological structures using a chemical dataset for the Martinsville igneous complex (MIC), Virginia, USA. The study sought to answer 4 main questions:

- Can PCA discern geochemical trends or relationships among lithologic units that have petrogenetic significance? If, so, what are these trends and what do they indicate about the origin of the rocks?
- Can PCA determine which of the original chemical variables are the most meaningful and do these correspond to the traditionally accepted variables, such as  $\text{SiO}_2$  and  $\text{MgO}$ ?
- Are the PCA-generated variables as useful or more useful for petrogenetic purposes than the

original chemical variables?

- Overall, is PCA a useful alternative to the traditional approaches for examining these types of geochemical and petrologic data?

### Preliminary data examination and transformation

The data set comprised data on the percentage weight of the oxides of 10 major elements and the concentration in parts per million of 3 trace elements (Rb, Sr and Zr). The authors checked the variables for normality and noted that MgO was not normal and so log-transformed this variable. This transformation does in fact make no difference to the ordination or the resulting conclusions.

### The use of the correlation matrix

PCA was undertaken on the correlation matrix; this was essential if the ordination was not to be dominated by the three variables with the largest variance because of their magnitude (Zr, Sr and Mg). This is the correct choice if it is believed that all elements can potentially equally contribute to the study of the relationships between the rocks.

### Results

As shown in the table below, the first 2 axes explained about 72.9 % of the total variability in the data set. The sum of all the eigenvalues, which is a measure of the total variability, is 14, which is simply the sum of the number of variables used in the analysis, because the correlation matrix was used. Therefore the percentage variability explained by the largest eigenvalue is  $7.28/14 \times 100 = 52.01\%$ . The first 3 dimensions are probably meaningful (eigenvalues > 1).

	Eigenvalues	Cumulative percentage of the total variance
1	7.282	52.01
2	2.918	72.86
3	1.27	81.93
4	0.8209	87.79
5	0.5165	91.48

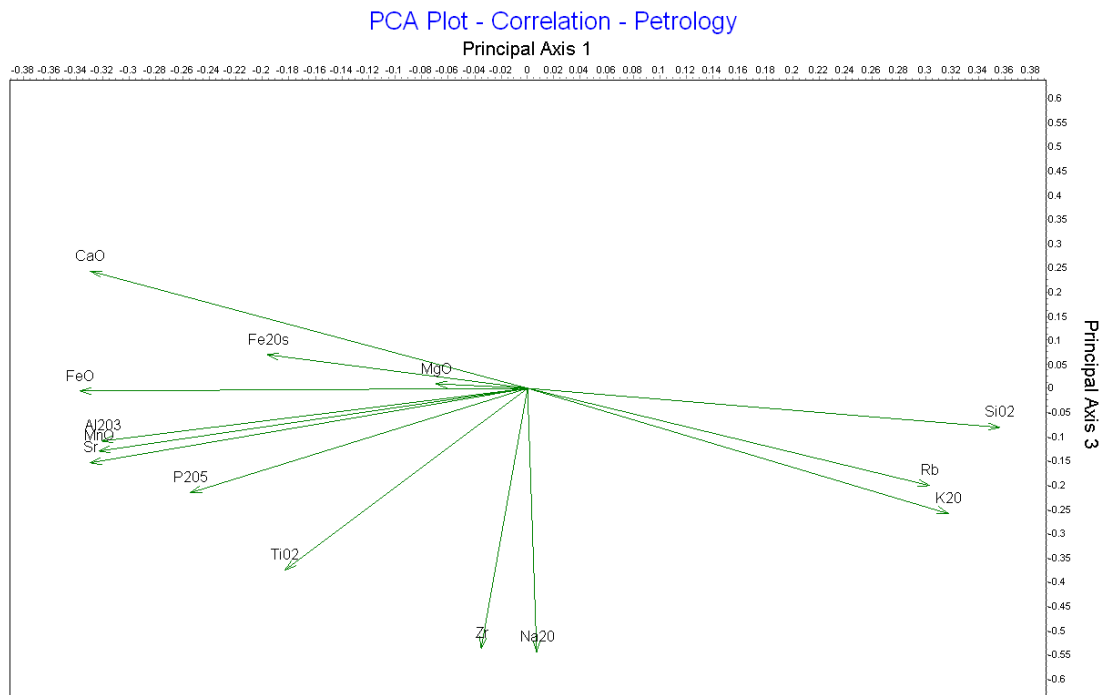
The authors reported a higher percentage of the variability explained by the first two axes probably because they combined the percentage composition for the two iron oxides into a single variable. However, this makes little difference to the ordination produced.

These results show that much of the variability in chemical composition can be expressed in 2 dimensions.

The plot of the eigenvectors (Fig 1) shows that Principal Axis 1 arranges the samples so that those with the highest concentrations of Ca, Fe, Al, Mn, Sr, P and Ti are towards the left (negative) and those with highest concentrations of Si, Rb and K to the right (positive). Axis 2 is a measure of Zr and Na concentration with the greatest concentrations at the bottom (negative direction) of the axis. The authors recognised 4 groups of eigenvectors. 1) Si, Rb, K; 2) Ca, Mg, 3) Fe, Al, Sr, Mn; 4) P, Ti; and 5) Na, Zr. When grouping eigenvectors you must consider the angle between them, not the length of the vector. The present results would suggest that P and Ti make a poor group, but they can be viewed as intermediate between {Zr, Na} and {Fe, Al, Sr, Mn}.



An examination of the 2D plots (Fig 2) shows a clear clustering of the rock samples. When the samples are grouped according to their mineralogy it is clear that the PCA ordination based on chemistry produces a similar classification. For example, that the syenodiorite can best be distinguished by its relatively high Na and Zr contents. The granites are characterized by relatively high Si, Rb, and K, and the Rich Acres gabbros being relatively enriched in Mg, Ca, and Fe. The authors do not consider these finding “particularly a surprise” and if the PCA “only confirmed the mineralogical groupings and chemical differences easily apparent, they would be of limited value.” The particular value of the PCA is in showing relationships and hybrids. For instance, the hybrid Leatherwood rocks (blue squares) are intermediate in composition between the granites (yellow squares) and the diorites (red squares).



**Fig 1: A plot of the eigenvectors.**

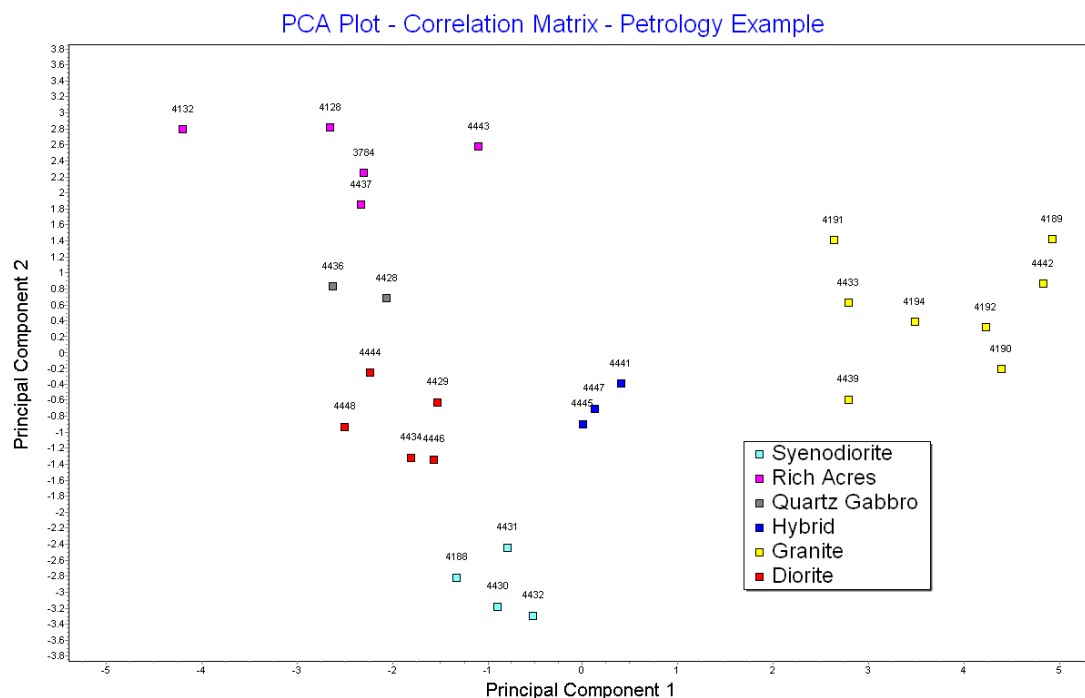


Fig 2: PCA of the rock samples.

## Conclusions

Ragland *et al.* (1997) concluded that Principal Component Analysis was useful: "... PCA is an insightful tool in petrology and geochemistry and is recommended as a first-step, exploratory technique for a dataset of chemical analyses. It allows the researcher to determine which of the original variables may be the most useful in characterizing the dataset." Further, it is capable of identifying possible relationships and hybrids and thus can be used as an aid when generating hypotheses about relationships between and origins of rocks.

## Alternative approaches

Ragland *et al.* (1997) used the correlation matrix for the PCA which had the effect of giving equal weighting to every element. The plot below shows the ordination of the sites using the variance-covariance matrix calculated with all variables log transformed. It is interesting to note that essentially the same clusters are formed but the eigenvectors show a number of tight pairs {Zr, Na}, {K, Rb}, {Mn, Fe} and {Ca, Mg}. This plot also shows that it possible to place the samples and the variable eigenvectors on the same plot. Some authors, including Ragland *et al.* (1997), plot only the apex of the eigenvectors. To avoid confusion, this should be avoided; the relationship between eigenvectors given by their angular difference is more easily studied if they are plotted as vectors.

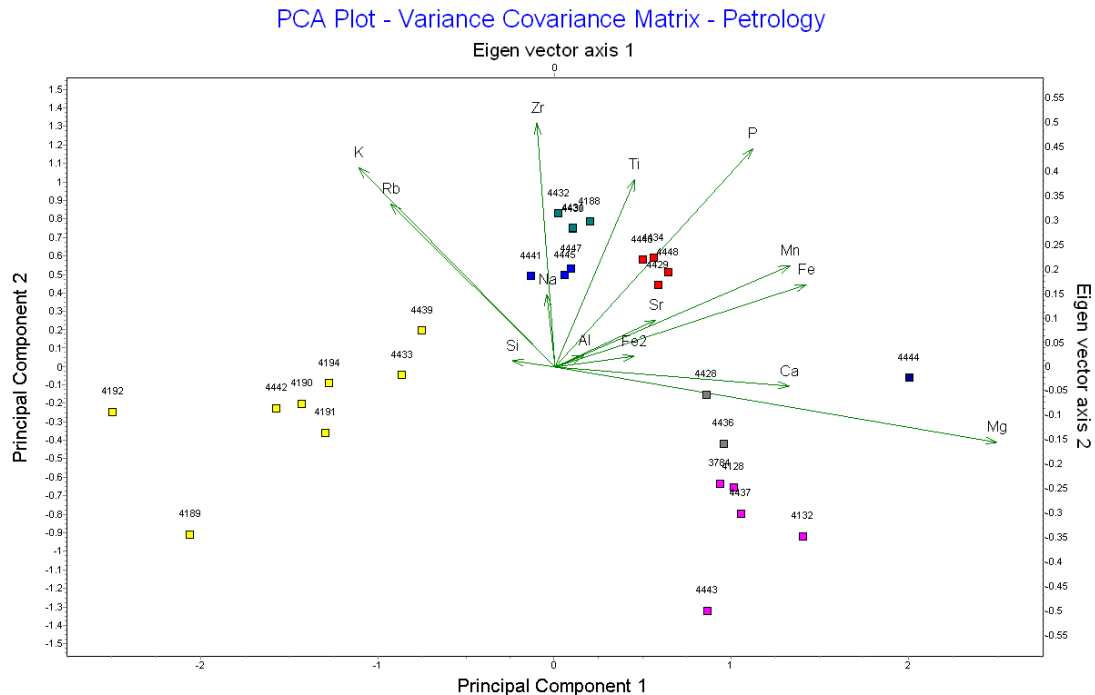


Fig 3: PCA for petrology example based on the covariance matrix.

### 3.4 Worked Example - Cicada Song

Demonstration data set: *cicada.csv*.

Reference: Ohya, E., 2004. *Identification of Tibicen cicada species by a Principal Component Analysis of their songs*. Anais da Academia Brasileira de Ciências 76: 441-444.

Ohya (2004) used recordings of cicadas to demonstrate that the songs of different species could be differentiated using PCA. This example shows the use of PCA to compare the features of time series. It also shows that standard measurements for known types, in this case species, can be included in the data set so that samples can be assigned to groups by their proximity to these standards within the ordination space.

#### Preliminary data examination and transformation

The data set comprised observations of Peak frequency (Hz), Mean Frequency (Hz) and No of pulses per 0.2 s. Recordings were made on 12 individuals of unknown species and 3 standard sets for the species *Tibicen japonicus*, *T. flammatus* and *T. bihamatus*. No transformations were undertaken.

#### The use of the correlation matrix

PCA was undertaken on the correlation matrix; this was essential if the ordination was not to be dominated by frequency measurements, which were between 5000 and 7000, while the pulse rate ranged between 8 and 20. It would also have been possible to have rescaled frequency in KHz and used the variance-covariance matrix.

## Results

As shown in the table below, the first 2 axes explained about 98% of the total variability in the data set, demonstrating that a 2 D graph can show the relationship between the cicada species. The sum of all the eigenvalues, which is a measure of the total variability, is 3, which is simply the sum of the number of variables used in the analysis, because the correlation matrix was used with 3 variables. The first dimension only has one eigenvalue > 1), but the second is required to distinguish between *T. japonicus* and *T. flammatus*.

	Eigenvalues	Cumulative percentage of the total variance
1	2.69	89.62
2	0.26	98.15

Figure 1 is a biplot of the eigenvectors and the sample scores. They can be effectively placed on the same graph because of the small number of variables and samples included in this study. The 3 samples for known species have been marked as large squares and labelled with the species name.

The 2D plots shows a clear separation between the species and the clustering of the unknown samples around the *T. bihamatus* standard indicates that all samples except for S1 can be assigned to this species. S1 has a song very similar to that of *T. japonicus*.

## PCA Plot - Correlation Matrix - Cicada Song

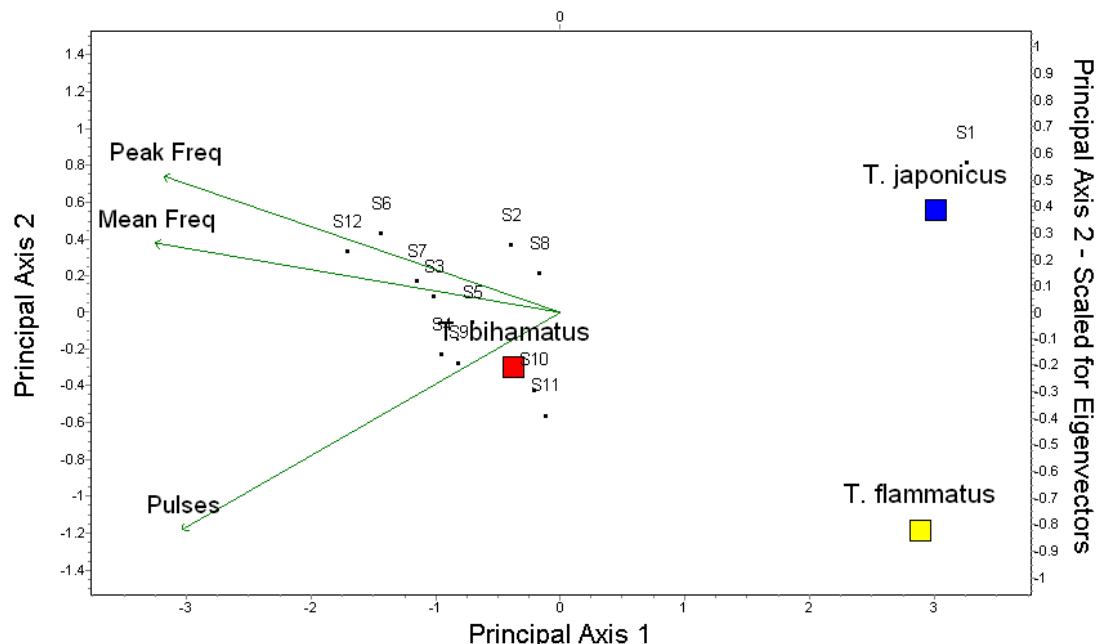


Fig 1: PCA plot of the song of 3 species of cicada.

## Conclusions

As the author stated "The cluster analysis of the PCA scores clearly separated *T. japonicus*, *T.*

*flammatus* and *T. bihamatus* from each other and allocated the samples as expected.“ He did include a warning: “However, one should collect real specimens with each sound recording in order to check the result of this method.”

**Alternative approaches**

There are no other methods that work as well as a PCA using the correlation matrix for data of this type.

**Part**

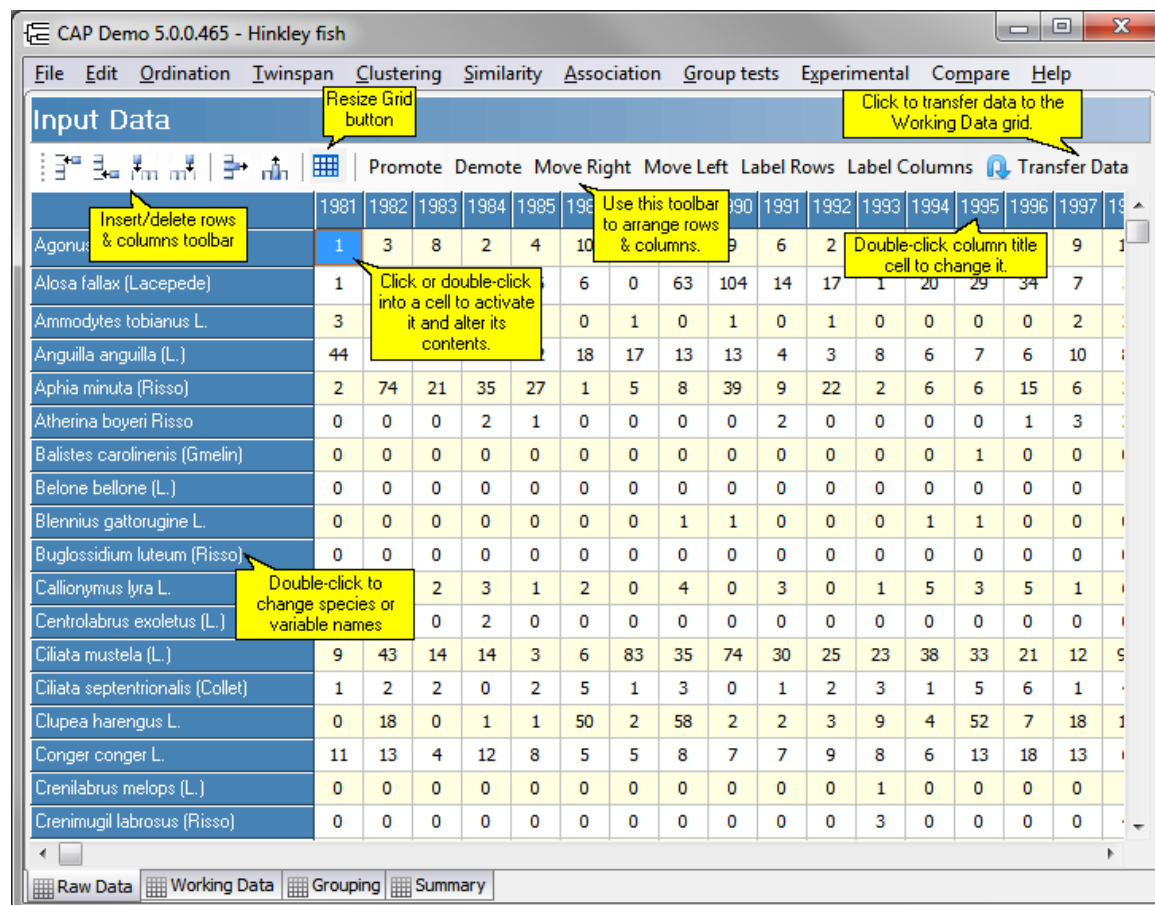
---



**IV**

## 4 Raw Data

The Raw Data tab shows you the data imported or entered into CAP. These data can be edited and saved. The actual data used by CAP to undertake any calculations is shown in the [Working Data](#) <sup>[48]</sup> tab. If you make changes to the Raw Data grid, you must then press the Transfer Data button, or switch to the Working Data tab, and click 'Reload Raw Data', before you can carry out analyses on the changed data.



See [Creating and editing a data set](#) <sup>[34]</sup> for methods for data entry.

### 4.1 Creating and editing a data set

There are a number of ways to enter data into CAP.

[Preparing large data sets in a spreadsheet program](#) <sup>[36]</sup>

[Data entry from within CAP](#) <sup>[38]</sup>

[Copy and Pasting data into CAP](#) <sup>[42]</sup>

[Importing from Excel](#) <sup>[37]</sup>

See also:

[Editing existing data](#) <sup>[44]</sup>

[Saving edited data or creating a new data file](#) <sup>[45]</sup>

[Maximum size of the data set](#) <sup>[9]</sup>

[Organising your data for analysis](#) <sup>[35]</sup>

### 4.1.1 Organising your data for analysis

Multivariate data sets can often be arranged as a two-dimensional matrix, consisting of N samples (or stations) forming the columns, and the S variables (e.g. species or % chemical composition) forming the rows. This matrix or grid can hold either the observed abundance of each variable, an abundance score, or presence-absence information. Presence-absence data are just recorded as 1 if the species was found in the sample, and 0 if not. The most convenient way to store and organise these data is with the use of a spreadsheet program such as Excel. Most statistical and multivariate software can import data from these spreadsheets. If the objective is to search for species inter-relationships then the N column by S row matrix of species correlations or similarities may be formed, and what is termed an R analysis undertaken. If the objective is to identify samples with similar communities then the N column by S row matrix of sample correlations or similarities may be used in a Q analysis.

	site 1	site 2	site 3	site 4	site 5	site 6	site 7	site 8	site 9	site 10	site 11
site 12	site 13	site 14	site 15	site 16	site 17						
Achmil	1	3	0	0	2	2	2	0	0	4	0
0	0	0	0	0	2						
Agrsto	0	0	4	8	0	0	0	4	3	0	0
4	5	4	4	7	0						
Airpra	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	2						
Alogen	0	2	7	2	0	0	0	5	3	0	0
8	5	0	0	4	0						
Antodo	0	0	0	0	4	3	2	0	0	4	0
0	0	0	0	0	4						
Belper	0	3	2	2	2	0	0	0	0	2	0
0	0	0	0	0	0						
Brohor	0	4	0	3	2	0	2	0	0	4	0
0	0	0	0	0	0						
Chealb	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0						
Cirarv	0	0	0	2	0	0	0	0	0	0	0
0	0	0	0	0	0						
Elepal	0	0	0	0	0	0	0	4	0	0	0
0	0	4	5	8	0						
Elyrep	4	4	4	4	4	0	0	0	6	0	0
0	0	0	0	0	0						
Empnig	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0						
Hyprad	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	0	2						
Junart	0	0	0	0	0	0	0	4	4	0	0
0	0	0	3	3	0						
Junbuf	0	0	0	0	0	0	2	0	4	0	0
4	3	0	0	0	0						
Leoaut	0	5	2	2	3	3	3	3	2	3	5
2	2	2	2	0	2						
Lolper	7	5	6	5	2	6	6	4	2	6	7
0	0	0	0	0	0						
Plalan	0	0	0	0	5	5	5	0	0	3	3
0	0	0	0	0	2						
Poapra	4	4	5	4	2	3	4	4	4	4	4
0	2	0	0	0	1						
Poatri	2	7	6	5	6	4	5	4	5	4	0
4	9	0	0	2	0						
Potpal	0	0	0	0	0	0	0	0	0	0	0
0	0	2	2	0	0						
Ranfla	0	0	0	0	0	0	0	2	0	0	0
0	2	2	2	2	0						



Rumace	0	0	0	0	5	6	3	0	2	0	0
2	0	0	0	0	0						
Sagpro	0	0	0	5	0	0	0	2	2	0	2
4	2	0	0	0	0						
Salrep	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0						
Tripra	0	0	0	0	2	5	2	0	0	0	0
0	0	0	0	0	0						
Trirep	0	5	2	1	2	5	2	2	3	6	3
3	2	6	1	0	0						
Viclat	0	0	0	0	0	0	0	0	0	1	2
0	0	0	0	0	0						
Brarut	0	0	2	2	2	6	2	2	2	2	4
4	0	0	4	4	0						
Calculus	0	0	0	0	0	0	0	0	0	0	0
0	0	4	0	3	0						

**Table 1: An example of biological data. This is part of the Dune meadow data used by Jongman *et al.* (1995). The scientific names of the species have been abbreviated.**

### 4.1.2 Preparing large data sets in a spreadsheet program

You do not need to create data sets within CAP; in fact the best way to enter large data sets is to use a spreadsheet such as Excel or Lotus 1-2-3, which will give access to a wide range of sorting and editing procedures to ease your task.

A data set should have the following format:

	Sample 1	Sample 2	Sample 3
Variable 1	21	1	5
Variable 2	15	5	0
Variable 3	0	7	0

The normal arrangement of community data within CAP is to have the samples (quadrats) as columns and the species as the rows. However, older versions of Excel have a maximum number of columns of 255, which can prove difficult if you have a data set with a very large number of sites/samples. If this is the case, the data can be arranged in Excel with the variables forming the columns, as the Transpose option within CAP can be used to switch columns and rows. Numbers can be either integer or real; some methods may require integers, in most such cases the program will run with real data, which will be automatically rounded.

The above table and the image below show you how the data will look in Excel. The samples are arranged in columns. Each sample has a title field. Start the first sample in column 2. The data consists of the number of individuals observed in the sample. Put in zeros rather than leaving cells blank. The species names (variable names) are input from row 2 in column 1.

	A	B	C	D	E	F	G	H
1		Hinkley	W Thurrock	Sizewell	Wylfa	Fawley	Oldbury	Heysham
2	Sprat	9565.911	8102.16	89707.602	62.497	3229.1	599.094	7185.167
3	Whiting	9720.2	255488.96	5164.854	3.812	8.26	13452.23	2406.3
4	Goby, Sand	1848.556	22266.52	3250.585	2.462	811.02	55.132	1341.133
5	Herring	102.617	15982.05	3044.474	16.756	1952.52	27.566	2810
6	Pipefish, Greater	0.728	69.285	1931.988	2.065	121.66	0	27.167
7	Pipefish, Nilsson's	17.467	601.245	1302.485	2.621	38.5	27.566	335.333
8	Sole (Dover sole)	737.239	3326.145	700.029	34.385	25.2	238.906	186.4
9	Flounder	548.017	20711.1	668.275	2.859	207.2	3013.887	260.167
10	Dab	622.25	717.495	616.228	10.8	5.18	0	420.2
11	Pout	1357.306	2285.94	471.287	6.432	706.16	22245.79	4400.467
12	Hooknose (Pogge)	47.306	426.87	292.251	41.453	4.76	0	636
13	Plaice	29.839	307.83	234.737	17.391	5.88	9.189	843.933

When using Excel use the **Save As** function to save your data as a \*.csv file. This will result in a data file in the format used by CAP. Alternatively, you can save as an Excel file (.xls, not .xlsx) which can be [imported into CAP](#)<sup>[37]</sup>. Ensure that the work sheet you are saving only holds the tabulated data for analysis. If your data set has been created using the convention that a blank cell means zero then use the Find and Replace function available in all common spreadsheets to search for blank cells and replace them with 0 (zero).

Occasionally, errors occur because a blank space or a character has been accidentally entered into a cell outside the data matrix. To prevent this happening, it is good practice, before saving your data set as a .csv file, to highlight the first 10 or so blank rows and columns below and to the right of the data matrix, and press 'Delete'. This will clear the cells of any accidentally-entered contents.

The csv file can also be opened and edited in a text editor (e.g. Notepad) or word processor (Microsoft Word), in which case it will appear like this:

,Title1,Title2,Title3

Variable 1,21,1,5

Variable 2,15,5,0

Variable 3,0,7,0

Variable 4,1,9,0

Variable 5,0,0,8

Note the leading comma on the first row which will make the first cell blank.

see also: [Quick guide to running a dataset](#)<sup>[6]</sup>; [Data entry from within CAP](#)<sup>[38]</sup>

### 4.1.3 Import from Excel

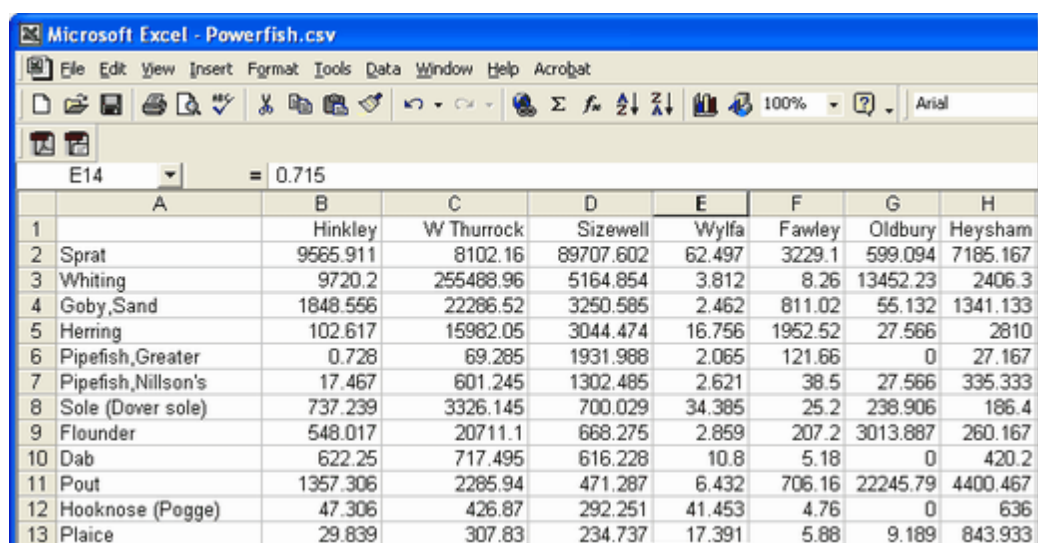
CAP offers the ability to import data directly from a Microsoft Excel spreadsheet. The program can import .csv (Comma-separated values) files, .xls files (Excel 97-2003 format), but notxlsx (Excel 2007/10/13) folders.

It is important, however, that a number of points are observed:

1. If you are using a spreadsheet with multiple worksheets, the data will be imported from the worksheet that was open (visible) when the workbook was last saved.
2. The data should be present as a contiguous rectangular block, starting at Cell A1.
3. Cell A1 itself should be empty, with sample names present in Row 1 and variable (species) names present in Column 1.
4. The import procedure ignores formulae in cells and imports the visible values.
5. Ensure that all cells rightwards and downwards from cell B2 contain numerical data.

6. CAP can import directly from Excel whether Excel is open or not, ***provided the worksheet has been saved.***

The image below shows a typical data set organised within Excel. The columns are the samples and the rows are the values for the variables.



The screenshot shows a Microsoft Excel window titled 'Microsoft Excel - Powerfish.csv'. The spreadsheet contains data for 13 fish species (rows 1-13) across 8 locations (columns A-H). The locations are Hinkley, W Thurrock, Sizewell, Wylfa, Fawley, Oldbury, and Heysham. The fish species are Sprat, Whiting, Goby, Sand, Herring, Pipefish, Greater, Pipefish, Nilsson's, Sole (Dover sole), Flounder, Dab, Pout, Hooknose (Pogge), and Plaice. The data values are numerical, representing some metric for each species at each location.

	A	B	C	D	E	F	G	H
1		Hinkley	W Thurrock	Sizewell	Wylfa	Fawley	Oldbury	Heysham
2	Sprat	9565.911	8102.16	89707.602	62.497	3229.1	599.094	7185.167
3	Whiting	9720.2	255488.96	5164.854	3.812	8.26	13452.23	2406.3
4	Goby, Sand	1848.556	22286.52	3250.585	2.462	811.02	55.132	1341.133
5	Herring	102.617	15982.05	3044.474	16.756	1952.52	27.566	2810
6	Pipefish, Greater	0.728	69.285	1931.988	2.065	121.66	0	27.167
7	Pipefish, Nilsson's	17.467	601.245	1302.485	2.621	38.5	27.566	335.333
8	Sole (Dover sole)	737.239	3326.145	700.029	34.385	25.2	238.906	186.4
9	Flounder	548.017	20711.1	668.275	2.859	207.2	3013.887	260.167
10	Dab	622.25	717.495	616.228	10.8	5.18	0	420.2
11	Pout	1357.306	2285.94	471.287	6.432	706.16	22245.79	4400.467
12	Hooknose (Pogge)	47.306	426.87	292.251	41.453	4.76	0	636
13	Plaice	29.839	307.83	234.737	17.391	5.88	9.189	843.933

Remember that you can also copy from Excel and paste directly into a CAP data grid - see [Copy and pasting data](#)<sup>[42]</sup>.

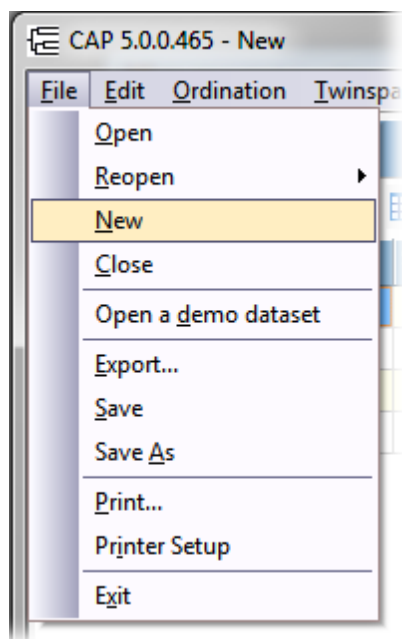
#### 4.1.4 Data entry from within CAP

Data sets can be created and edited within CAP, by creating a blank grid, and either typing in each individual data point by hand, or by [copying/pasting the data](#)<sup>[42]</sup> from elsewhere into the grid. Here, we look at entering data by hand.

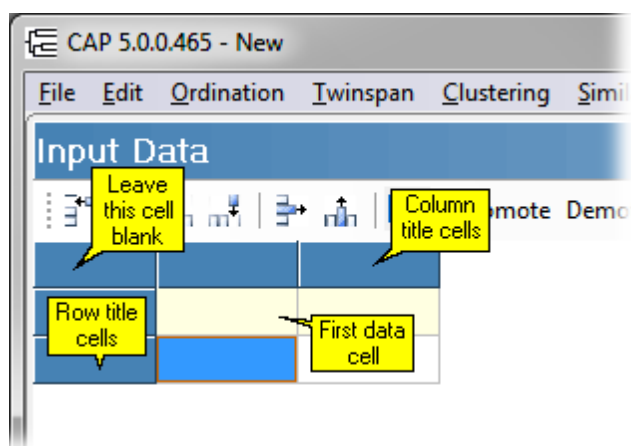
The process of entering data is:

1. Create a new data set
2. Add row and column titles
3. Enter data (before or after creating the row/column titles, it's not really important)
4. Transfer the data set to [Working Data](#)<sup>[48]</sup>
5. Save the data file
6. If you wish, add [grouping information](#)<sup>[53]</sup>.

1. To create a new data set, select File: New from the top menu bar:

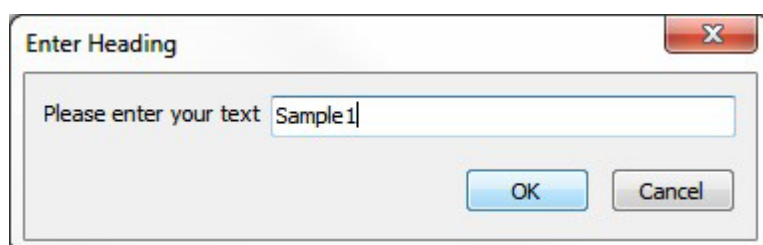


You are presented with a 3 x 3 grid in which the top blue row represents samples, and the left blue column represents variables (e.g. species):



## 2. Row and column titles

To enter column and row titles, double click on the dark blue title cells and enter the cell title into the dialog box that appears:



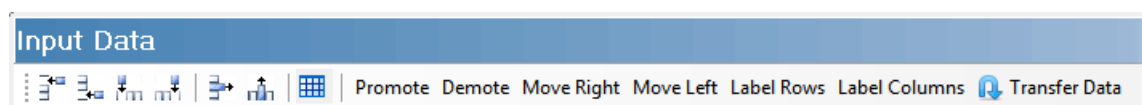
## 3. Entering data

Leave the top left-hand cell empty. To input data, just click into a cell and type a number. Numbers can be either integer or real; some methods may require integers, but in most such cases the program will run with real data which will be automatically rounded.

Pressing the return key moves you sequentially down each column. To type in a column of values

just type a number into the top data cell of the column and then press **Enter** on your keyboard to move into the next cell down.

Above the empty grid is the data input toolbar:



From left to right, the tool bar buttons are:

**Insert row above selected cell**

**Insert row below selected cell**

**Insert column to left of selected cell**

**Insert column to right of selected cell**

**Delete selected row**

**Delete selected column**

**Resize grid**

**Promote** - first data row to title row

**Demote** - title row to first data row

**Move Right** - Insert column to left of entire block of data (i.e. add a column of row header cells).

**Move Left** - Remove row header column from left of entire block of data

**Label Rows** - add labels (Row1, Row2, etc) to the row headers column.

**Label Columns** - add labels (Column1, Column2, etc) to the column headers row.

**Transfer Data** - when you have finished adding your data, you must click this to send the data set to the [Working Data grid](#)<sup>[48]</sup>, before you can carry out any analyses.

#### 4. Transfer the data to the Working Data grid.

When you've finished entering your data, you will have a grid that looks something like this:

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
Row 1	19.6	15.4	22.3	24.3	0	0	7.9
Row 2	5.15	5.75	2.35	7.54	10.21	0.9	9.3
Row 3	0	9.5	3.3	5	6	5	9.1
Row 4	5.5	0	3	4	0	1	2.1

Now, click the Transfer Data button to move the data set to the Working Data grid. Alternatively, click on the [Working Data](#)<sup>[48]</sup> tab at the bottom of the program window, and click on the **Reload Raw Data** button. The data set is now ready for use in the program.

#### 5. Save your data set.

Before you go any further, save your precious data. Click File: Save, choose a name for the data file

#### 6. Add [grouping information](#)<sup>[53]</sup> if necessary.

##### Another way to add rows and columns to the data grid

To add a new row, click into a cell in the bottom row of the grid, and press the Down arrow on your keyboard. To add a new column, click on the bottom right-hand cell in the grid and press Enter on your keyboard.

To remove a row, click on a cell in that row, and press the Delete key on your keyboard. Make sure that the cell itself is selected, not the value it contains:

**Selected cell:**

15.4
5.75
9.5

**Selected value:**

15.4
5.75
9.5

You should note that once a row has been deleted, the Undo function will not restore that row.

See also:

[Creating a data grid of a certain size](#)<sup>[41]</sup>

[Copying and pasting data](#)<sup>[42]</sup>

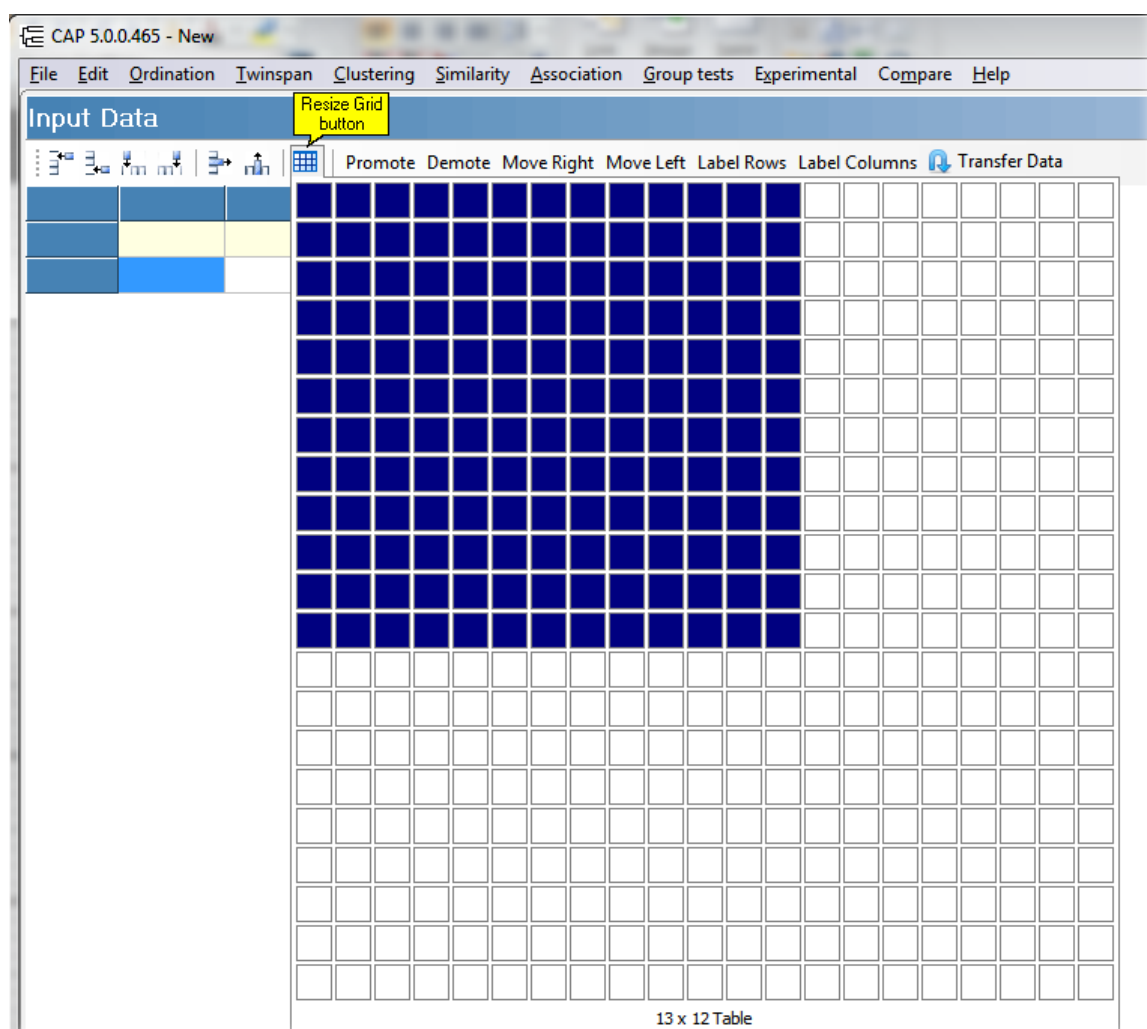
[Importing data from Excel or other spreadsheet programs](#)<sup>[37]</sup>

[Editing existing data](#)<sup>[44]</sup>

#### 4.1.4.1 Creating a data grid of a certain size

##### Creating a data grid of a specific size

If you know the size of the data grid you wish to create (X columns by Y rows), you can use the Resize Grid button (don't forget to add an extra column and row for the header cells); click and drag the cursor down and to the right to select the number of columns/rows:



You can then enter the data by hand, or [copy/paste it from elsewhere](#)<sup>[42]</sup>.

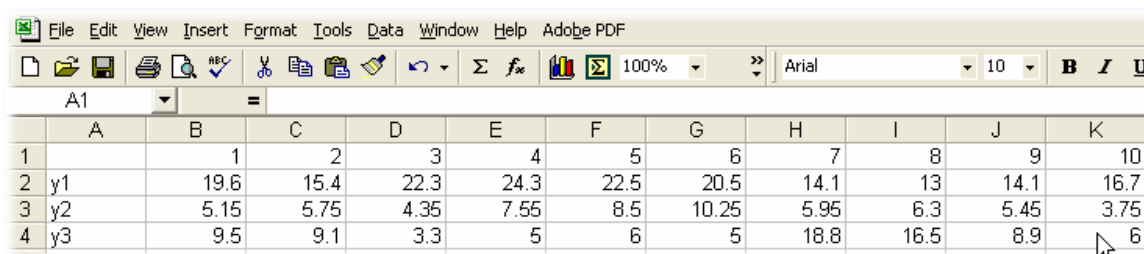
### 4.1.5 Copying and pasting data

You can use standard Copy and Paste Windows techniques to move data into CAP. The following example shows how to copy and paste from Excel.

The process is as follows:

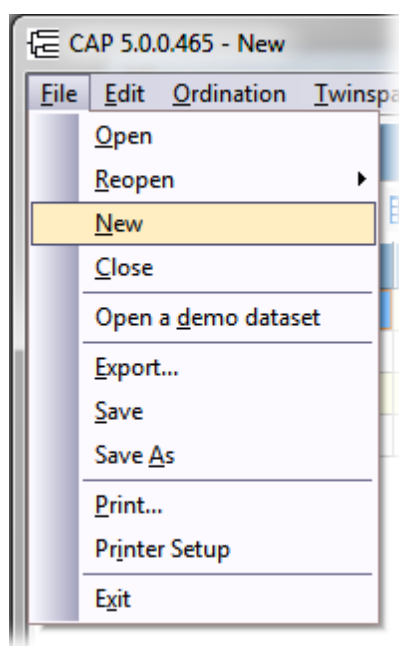
1. **Select and copy the data from its source**
2. **Create an empty data set in CAP**
3. **Paste the data into the grid in CAP, align the row and column titles, or add title information**
4. **Transfer the data set to the Working Data grid**
5. **Save your data set**
6. **Add grouping information (if required).**

1. Select and copy the data in your Excel spreadsheet:

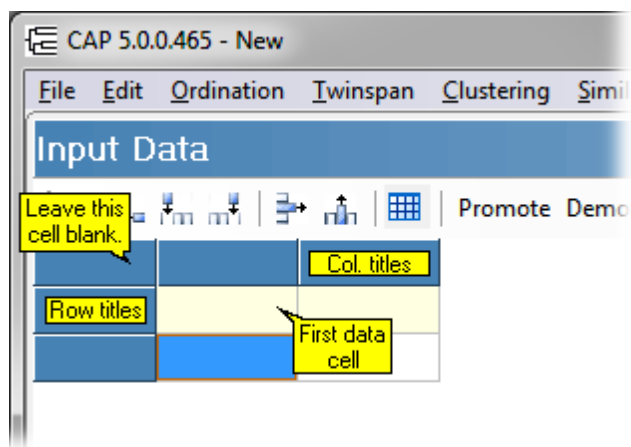


	A	B	C	D	E	F	G	H	I	J	K
1		1	2	3	4	5	6	7	8	9	10
2	y1	19.6	15.4	22.3	24.3	22.5	20.5	14.1	13	14.1	16.7
3	y2	5.15	5.75	4.35	7.55	8.5	10.25	5.95	6.3	5.45	3.75
4	y3	9.5	9.1	3.3	5	6	5	18.8	16.5	8.9	6

2. Open CAP and select **File: New** from the drop-down menus:



You are presented with a 3 x 3 grid in which the top blue row represents sites or samples, and the left blue column represents variables (e.g. species):



3. Click on the first data cell (**not** the top left cell) in the grid.
  - i. If the data you are pasting in already include Row and Column titles, click **Edit: Paste data (with titles)**. The data will be pasted into the grid; the program will automatically add the required number of rows and columns to the grid, and the row/column titles will be placed in the blue title cells:

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
Row 1	19.6	15.4	22.3	24.3	0	0	7.9
Row 2	5.15	5.75	2.35	7.54	10.21	0.9	9.3
Row 3	0	9.5	3.3	5	6	5	9.1
Row 4	5.5	0	3	4	0	1	2.1

- ii. If your data do **not** include row & column titles, click **Edit: Paste data (no titles)**, or press **Ctrl-V** on your keyboard. The data will be pasted into the grid; the program will automatically add the required number of rows and columns to the grid, but the title cells will remain blank:

	19.6	15.4	22.3	24.3	0	
	5.15	5.75	2.35	7.54	10.21	
	0	9.5	3.3	5	6	
	5.5	0	3	4	0	

Now, either click the **Label Rows** and **Label Columns** buttons to add a generic title to each, or double-click into each row and column title cell to add your own titles.

- iii. If by accident you use **Paste data (no title)** or **Ctrl-V** with data containing row and column titles (image below), you then need to move them in to the correct position. Click first the **Promote** and then the **Move left** buttons on the tool bar above the data grid, to move the titles into the correct place:



	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	
Row 1	19.6	15.4	22.3	24.3	0	0	
Row 2	5.15	5.75	2.35	7.54	10.21	0.9	
Row 3	0	9.5	3.3	5	6	5	
Row 4	5.5	0	3	4	0	1	

If you paste data with no title information using **Paste data (with titles)**, so that data appear in the title rows, then click the **Demote** and **Move Right** buttons, to move the data into the correct place.

4. Now, click the Transfer Data button, to move the data set to the [Working Data](#)<sup>[48]</sup> grid. If you prefer, click on the [Working Data](#)<sup>[48]</sup> tab at the bottom of the program window, and click on the **Reload Raw Data** button. The data set is now ready for use in the program.

5. Save your data. Click **File: Save**, and choose a file name and location.

6. If necessary, add [grouping information](#)<sup>[53]</sup> to the data set.

See also:

[Data entry from within CAP](#)<sup>[38]</sup>

[Creating a data grid of a certain size](#)<sup>[41]</sup>

[Importing data from Excel or other spreadsheet programs](#)<sup>[37]</sup>

[Editing existing data](#)<sup>[44]</sup>

### 4.1.6 Editing existing data

You can edit and add data to the Raw Data grid.

**Note:** If you press the Delete key on your keyboard while a **cell**, rather than the **value in the cell**, is selected, the entire row will be deleted (simply re-load the original data file to restore the row). If you do wish to use the Delete key, then make sure that only the **value in the cell** is selected.

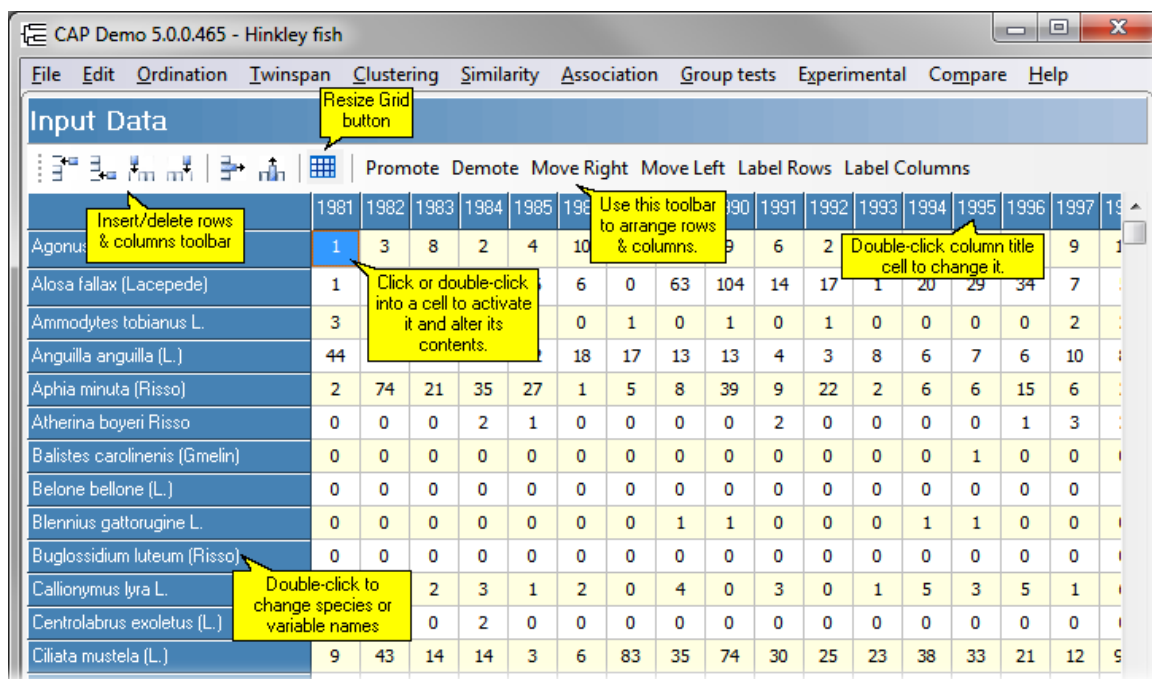
**Selected cell:**

15.4	5.75	9.5
------	------	-----

**Selected value:**

15.4	5.75	9.5
------	------	-----

A single value in the raw data grids can be edited by double-clicking into the cell to select it, and typing in a new value. To add or delete columns and rows, use the tool bar above the raw data grid (shown below):



From left to right, the toolbar buttons are:

**Insert row above selected**

**Insert row below selected**

**Insert column to left of selected**

**Insert column to right of selected**

**Delete selected row**

**Delete selected column**

**Resize grid**

**Promote** first data row to title row

**Demote** title row to first data row

**Move Right** - Insert column to left of entire block of data (i.e. add a column of row header cells).

**Move Left** - Remove row header column from left of entire block of data

**Label Rows** - add labels (Row1, Row2, etc) to the row headers column.

**Label Columns** - add labels (Column1, Column2, etc) to the column headers row.

When you have made changes to the raw data, they must be transferred to the Working Data grid; click on the Working Data tab at the bottom of the program window, and click the Reload Raw Data button. CAP will use the working grid thus created for all subsequent calculations. If you have already done an analysis, for instance a PCA, and then changed the raw data, the PCA results will not be updated immediately; you will need to run the analysis again.

Any changes you make to the raw data will not alter a saved data file until **File: Save** is used, from the Raw Data page.

The Working Data grids can also be [edited](#)<sup>[48]</sup>. Changes made to the working grids will not be transferred to the raw data grid. Transformed or otherwise changed working data can be saved as a new data set using **File: Export** from the Working Data page.

#### 4.1.7 Saving new or edited data

When you have made changes to the Raw Data grid, and/or the grouping information, clicking **File: Save** will save your raw data and any group allocations you have made, **replacing the original data file**. If you wish to preserve the original data file, then your data set can be copied and saved under a different name by selecting **File: Save As**, from the Raw Data page.

To save the [Working Data grid](#)<sup>[48]</sup>, if the dataset has been transformed, transposed or edited in

---

some other way, do the following:

1. Click on the Working Data tab
2. Select **File: Export**, and save as a comma delimited file (\*.csv).

When you select **File: Exit**, or close down the program using the cross at the top right corner of the program window, if the data set has been altered in any way, but not saved, you will be asked if you would like to save the data.

# Part

---

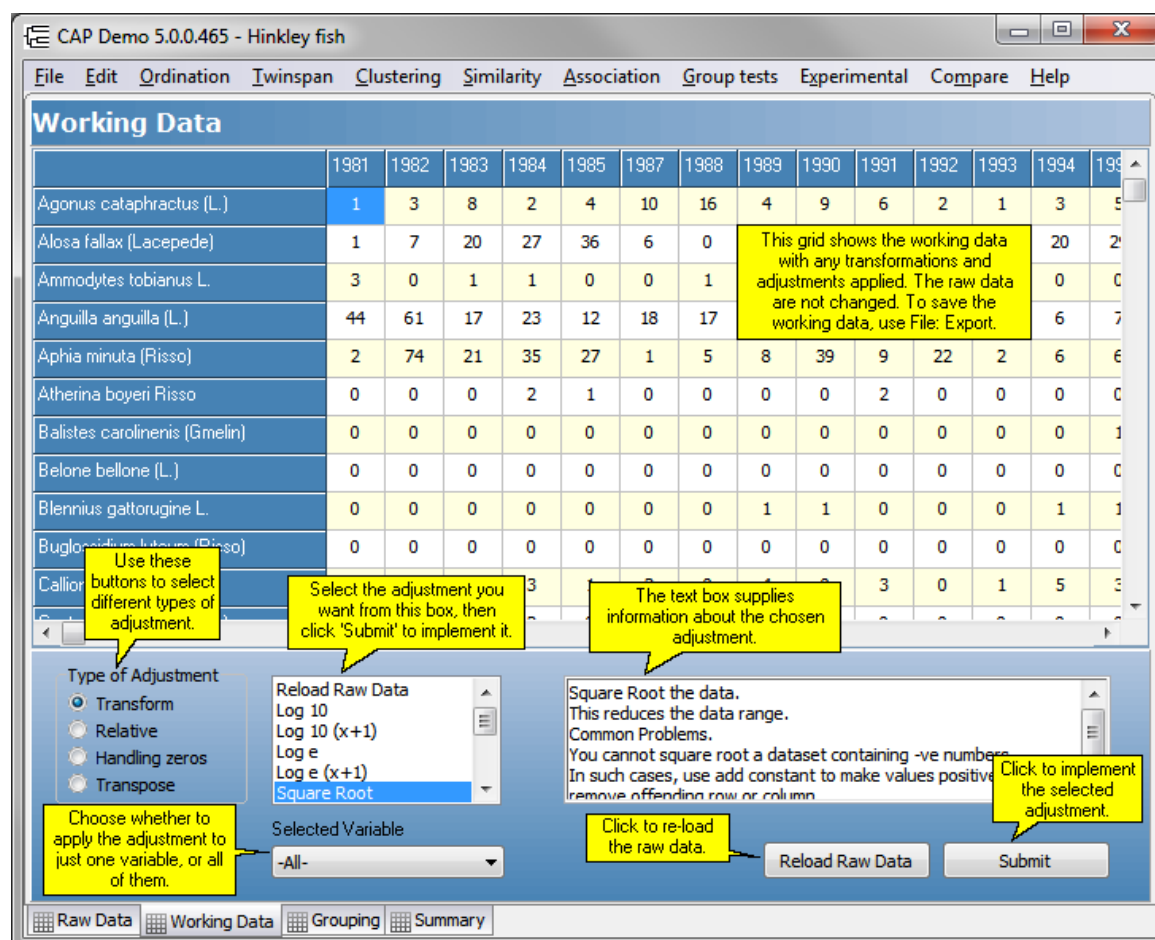


V

## 5 Working Data

The Working Data Tab shows you the data that will be used by any method you select to undertake the calculations. If you have edited the [raw data](#) <sup>[34]</sup>, remember to click on Reload Raw Data to get these changes transferred to the Working Data grid.

CAP is designed to separate the raw and working data so that you can transform and adjust your data in many ways without changing the raw data. The Working Data screen allows you to make a variety of changes to the raw data prior to undertaking an analysis. Initially, you will be presented with a grid filled with the raw data; this can be adjusted using the options in the panel below the data grid. To save the working data as a new file use **File: Export**.



[Data transformations](#) <sup>[49]</sup>

[Relative adjustments](#) <sup>[49]</sup>

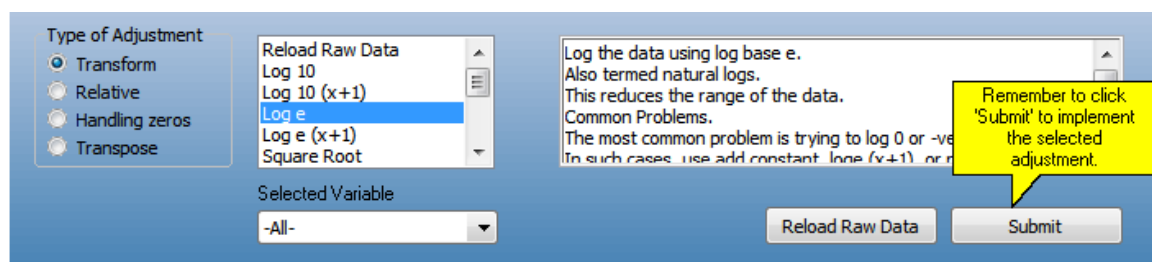
[Dealing with zeros](#) <sup>[50]</sup>

[Transposing data](#) <sup>[51]</sup>

[Saving the working data](#) <sup>[45]</sup>

## 5.1 Data transformations

The values in the [Working Data](#)<sup>[48]</sup> grid can be transformed using a variety of functions. From the Working data window select Transform in the Type of Adjustment panel.



The possible relative measures available within CAP are given below. Select the transformation to be made and click on **Submit** to make the change. The transformation options within CAP are itemised below.

**Reload raw data** - This will cause the working data to revert to the raw data.

**Log(10)** - Each value is transformed to the log to base 10. This cannot be done for numbers  $\leq 0$ .

**Log10(x+1)** - Each value is transformed by adding 1 and then calculating the log to base 10. This is used when the data contains zero values.

**Log e** - Each value is transformed to the log to base e (natural logs). This cannot be done for numbers  $\leq 0$ .

**Log e (x+1)** - Each value is transformed by adding 1 and then calculating the log to base e. This is used when the data contains zero values.

**Square root** - the square root of each number is calculated. This cannot be done for negative numbers.

**Arcsin** - The Arcsin of each value is calculated. A transformation often used for percentage data.

**Arcsin root** - The Arcsin of the square root of each number is calculated.

**Power** - Each value, x, is transformed to  $x^a$ , where a is chosen by the user.

**Add constant** - A constant value, chosen by the user, is added to each value.

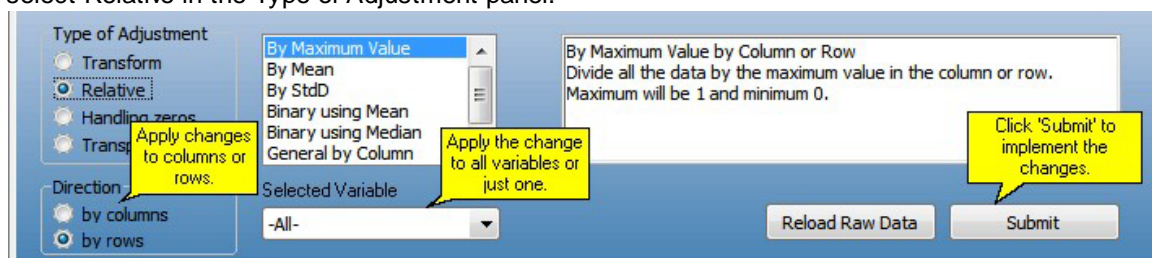
**Subtract constant** - A constant value, chosen by the user is subtracted from each value.

**Multiply by constant** - Each value is multiplied by a constant value chosen by the user.

**Divide by constant** - Each value is divided by a constant value chosen by the user.

## 5.2 Relativisations

The values in each row or column of the [working data](#)<sup>[48]</sup> can be transformed so that their magnitudes are expressed relative to a variety of statistical measures. On the Working Data page select Relative in the Type of Adjustment panel.



The possible relative measures available within CAP are given below. In each case you can select whether the relativisation will be applied to rows or columns. Select the adjustment to be made and click on Submit to make the change.

**By Maximum value** - For each row or column the maximum value is found and all values are divided by the maximum.

**By Mean** - For each row or column the mean value is found and all values are subtracted from the mean.

**By SD** - For each row or column the standard deviation value is found and all values are divided by the standard deviation.

**Binary using Mean** - For each row or column the mean is found and all values above the mean are given the value 1 and all values below the mean zero.

**Binary using Median** - For each row or column the median is found and all values above the median are given the value 1 and all values below the mean zero.

**General by Row** - This allows you to define a general relativisation to be applied to each row. Each value is divided by

$$\sum_{i=1}^{i=x} x \frac{a}{i}$$

where i is the column and a is a user-selected parameter (the default value is 1) entered into the Enter value box displayed in the lower panel.

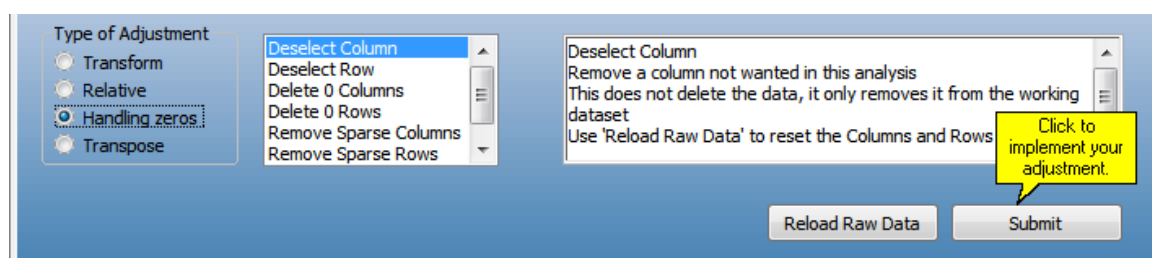
**General by Column** - This allows you to define a general relativisation to be applied to each column. Each value is divided by

$$\sum_{j=1}^{j=x} x \frac{a}{j}$$

where j is the row and a is a user-selected parameter (the default value is 1) entered into the Enter value box displayed in the lower panel.

## 5.3 Dealing with zeros or sparse data

Rows or columns in the working data holding zero values can be removed. On the Working Data page select Handling zeros in the Type of Adjustment panel.



Select the adjustment to be made and click on **Submit** to make the change. The adjustments are not made to the original data (on the Raw Data tab), but to the working data.

The possible options are as follows:

**Deselect Rows:** Removes rows not wanted in this analysis.

**Deselect Columns:** Removes columns not wanted in this analysis.

**Delete zero columns** - Every column in the data set that only contains zeros is removed.

**Delete zero rows** - Every row in the data set that only contains zeros is removed.

**Remove sparse columns** - Every column in the data set which contains < x non-zero elements is removed. The value of x is entered by the user in the At least x non-zero value text box.

**Remove sparse rows** - Every row in the data set which contains < x non-zero elements is removed. The value of x is entered by the user in the At least x non-zero value text box.

**Beals** - Beals smoothing is used on sparse data sets where many of the samples hold a small proportion of the total species list. As the transformation only uses the presence/absence of species, it should not be applied to good quality quantitative data as you will be discarding much of your information. It should only be performed on data arranged with the samples as columns. Described and discussed by Beals (1984) and McCune (1994) respectively, each element,  $e_{ij}$ , in the data array is replaced by:

$$e_{ij} = \frac{1}{N_i \sum_{k=1}^{k=\max S} \frac{Both_{jk}}{Samples_k}}$$

where i is the column (sample) number, j is the row (species) number,  $N_i$  is the number of species in column (sample) i,  $Both_{jk}$  is the number of samples holding both species j and k,  $Samples_k$  is the number of samples holding species k and  $\max S$  is the total number of columns (samples).

## 5.4 Transposing data

Use this option to switch the rows and columns of the data set. Like all the other adjustments it is applied to the working data set. Select **Transpose** in the Type of Adjustment radio box and click on the **Submit** button.

The normal arrangement of community data within CAP is to have the samples (quadrats) as columns and the variables (e.g. species) as the rows. However, old versions of Excel had a maximum number of columns of 255, which can prove difficult if you have a data set with a very large number of sites/samples. If this is the case, the data can be arranged in Excel with the species forming the columns, and the Transpose option within CAP used to switch columns and rows.

**N.B.** If you have allocated your samples to groups, for the purposes of using the ANOSIM or SIMPER methods, you will find that your groups will be lost if you transpose your data. The solution is to save your working data using **File: Export**, then reload the data into CAP; you can then re-assign samples to groups.



# Part

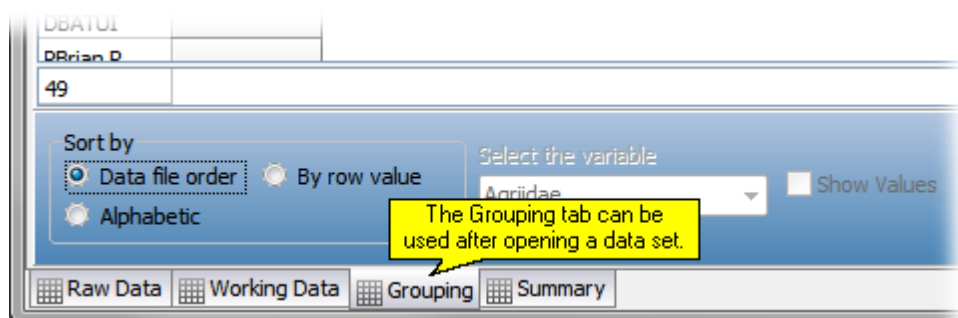
---



VI

## 6 Grouping

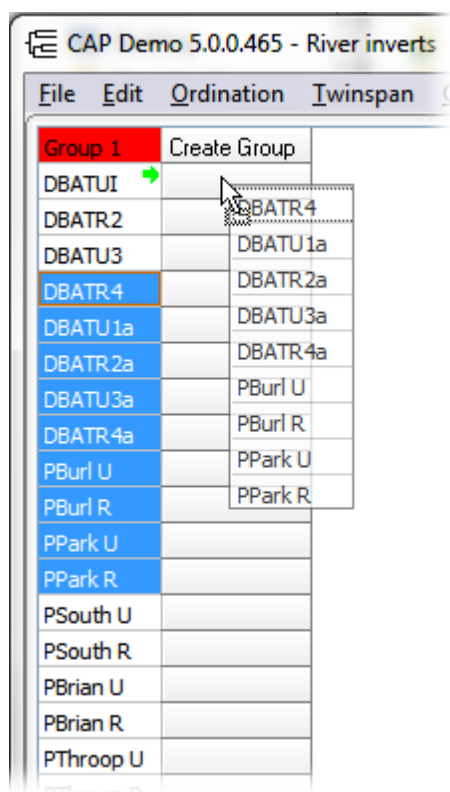
Use the Grouping tab to create groups and assign individual samples to groups:



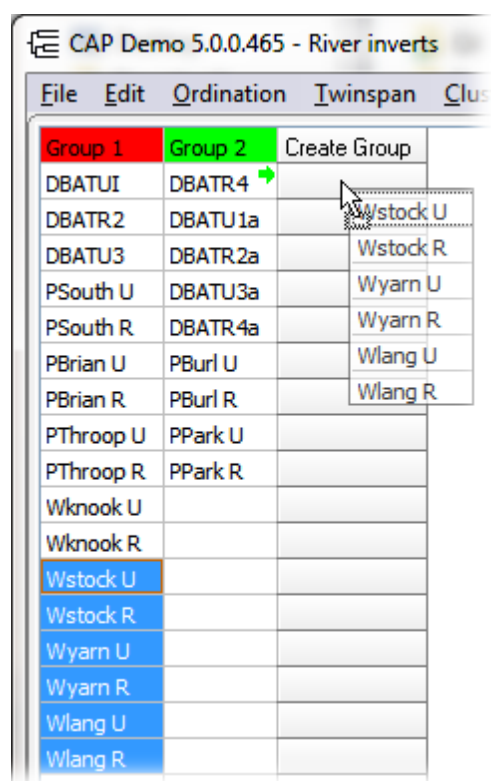
Allocating samples to groups is easily done by either of two methods; **A.**, selecting a sample or block of samples and dragging to another column in the grid, or **B.**, selecting the required samples, and right-clicking to send them to the required group.

### A. Grouping by dragging/dropping

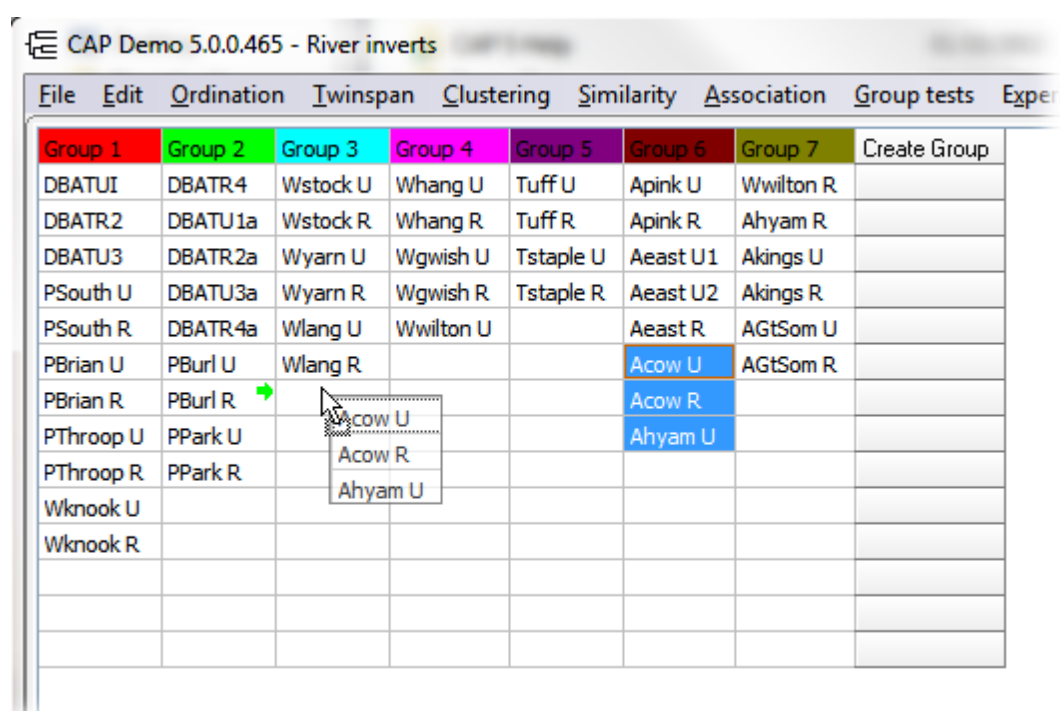
Initially all the samples are in the unassigned column. Click and drag to select one or more samples from Column 1. The selected samples will appear blue. Then simply drag the block of samples to the 'Create Group' column:



You will see a green arrow appear to show you the column the samples will appear in. Release the mouse button to drop the samples into that column; a dialog box will appear to assign a name to the new group. If you do not enter a name a default name will be created by CAP. Further samples can then be dragged from the 1st column into Group 2, or into the 'Create Group' column to create another new group:

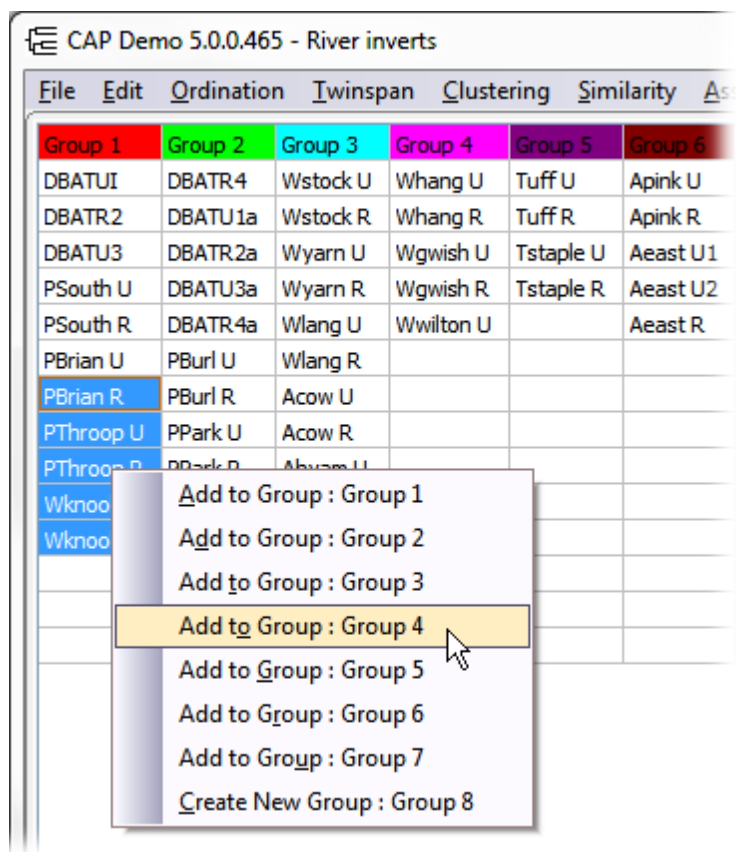


Repeat this process as often as you wish to create further groups; you can also drag and drop samples between groups.

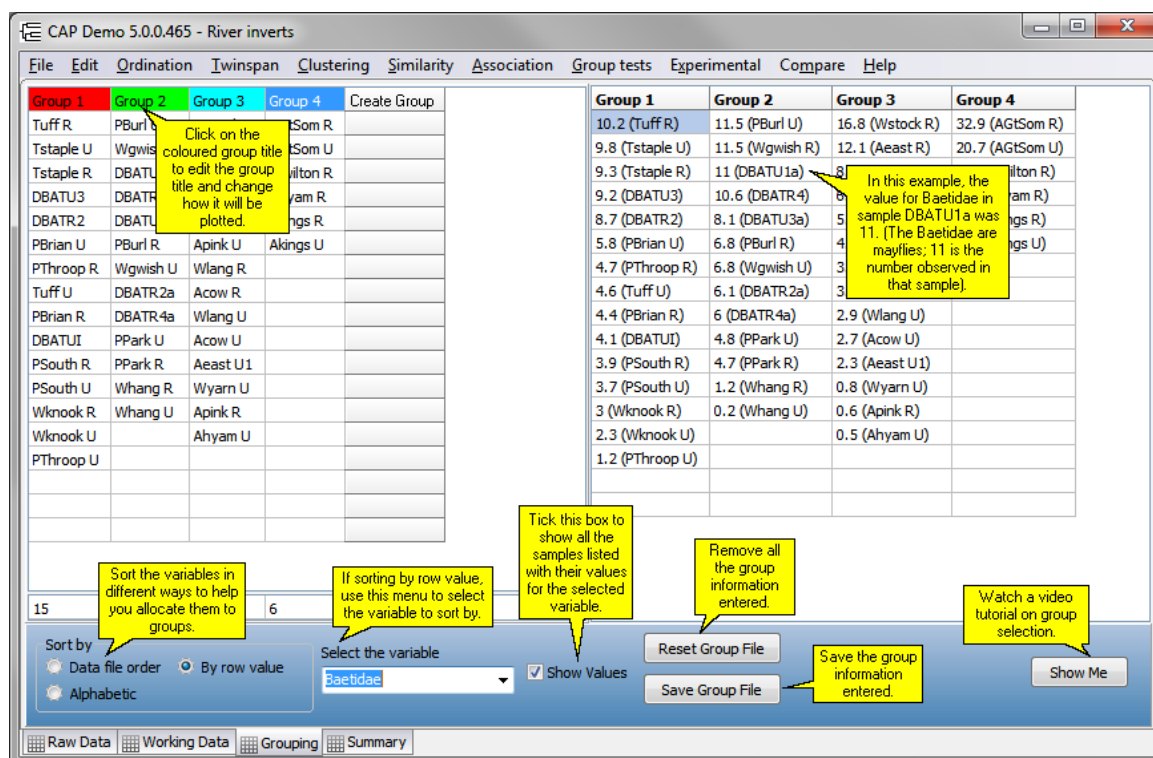


## B. Grouping by right-clicking

Alternatively, you can move samples to a new group, or between existing groups, by clicking and dragging to select the required samples (as above), then right-clicking, and selecting the group to add the samples to, from the pop-up menu:



Once you have created your groups, you can [edit the group names and the colours associated with the groups](#) <sup>[56]</sup>. To help you sort your samples and understand which groups are most appropriate for them, you will find in the bottom panel a number of sorting options.

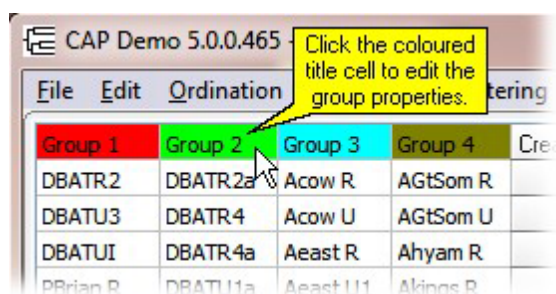


**NOTE:** If a data set has a column with 0 observations in it, that column is always going to be

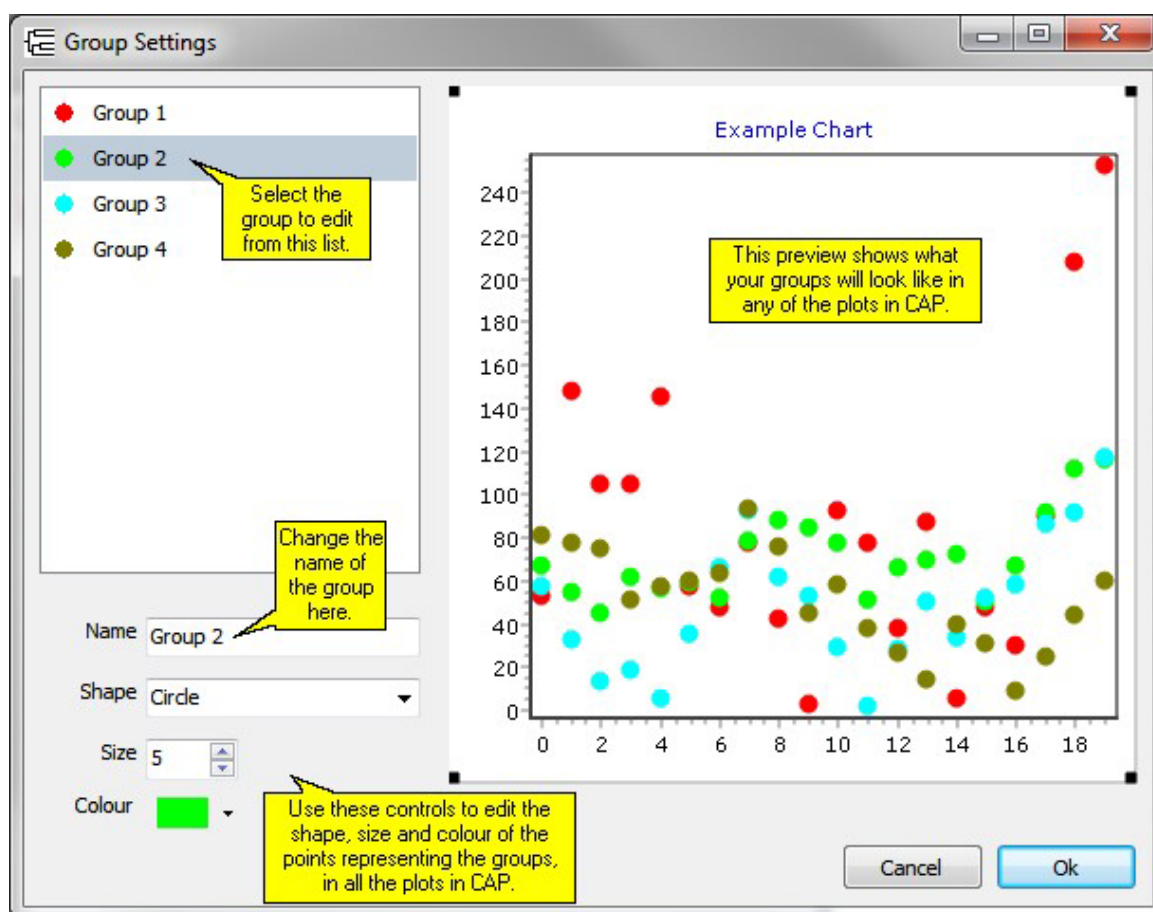
removed on loading the data set, when the data are transferred to the Working Data tab. Even if the raw set is reloaded, the 0 column is again removed. The deselected column is always therefore going to be shown on the Grouping tab, in red italic text.

## 6.1 Changing group properties

You can change how a group will be plotted, and also its name, by left-clicking on the column title cell, on the Grouping page of the program:



This will open the Group Settings dialog:



This dialog will control the way groups are represented throughout every plot produced by CAP. If you wish to change the representation in one particular plot only, you can do so quite easily; on the plot, click the 'Edit' button, then under 'Series', select the group you wish to edit. More information under [Editing charts](#)<sup>[157]</sup>.

On many of the plots, you can also [draw a perimeter line](#)<sup>[161]</sup> around all the members of the defined

groups.

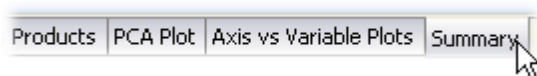
**Part**

---



## 7 Summary tab

After each analysis the summary data are updated and are visible on the Summary tab of the program.

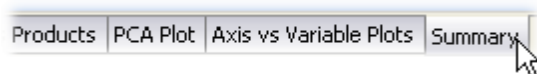


Summary of Data	
General statistics	
No. of rows	61
No. of cols	49
No. of zero cells	1561
No. of non-zero cells	1428
% zero cells	52.224824
Maximum value	90.300000
Minimum value	0.000000
Range	90.300000
Mean	1.631516
Standard deviation	6.133095
Median	0.000000

For details of the statistics on offer, see [Data set statistics](#)<sup>[59]</sup>.

### 7.1 Data set statistics

Summary statistics for both the raw and working data sets are displayed by clicking on the Summary tab.



You can choose summary statistics for either the raw or working data sheets. Statistics can be generated for rows, columns, general summary statistics and the user-defined groups for your data. These statistics are particularly useful prior to undertaking many analyses. For example [PCA](#)<sup>[16]</sup> should not be undertaken on a data set that comprises large numbers of zeros. Use Summary statistics to determine the number of zero elements in your data. The group statistics are also useful when summarising the results of a [Linear Discriminant Analysis](#)<sup>[17]</sup> (also called Canonical Variate Analysis).

When first activated the data grid will display the following general statistics for the working data:



Summary of Data	
General statistics	
No. of rows	61
No. of cols	49
No. of zero cells	1561
No. of non-zero cells	1428
% zero cells	52.224824
Maximum value	90.300000
Minimum value	0.000000
Range	90.300000
Mean	1.631516
Standard deviation	6.133095
Median	0.000000

**No. of Variables (rows)** - This is the number of rows of data in the data set.

**No. of Samples (cols)** - This is the number of columns in the data set.

**No. of zero cells** - This is the number of zero entries in the data matrix.

**No. of non-zero cells** - This is the number of non-zero entries in the data matrix.

**% of zero cells** - This is the number of zero entries divided by the number of cells in the data matrix.

**Maximum value** - This is the maximum value in the data matrix.

**Minimum value** - This is the minimum value in the data matrix.

**Range** - This is the difference between the maximum and minimum values.

**Mean** - This is the mean of all the values in the data matrix.

**Standard deviation** - This is the standard deviation of all the values in the data matrix.

**Median** - This is the median of all the values in the data matrix.

To obtain general statistics on the raw data click the Raw Data radio button in the 'Use' panel situated below the grid to select Raw Data.

Use

☐ Working Data
☒ Raw Data

Statistics

☒ General
☐ Row
☐ Column
☐ Group mean
☐ Group variance

Raw data only

Statistics for the individual rows, columns and groups of either the raw or working data matrix are selected using the Statistics radio buttons situated below the grid.

Values for each row or column are shown in the order they appear in the original data matrix:

Summary of Data												
Row	Mean	Median	Max	Min	Zeros	Non-zeros	% zeros	Sum	Sum Sqr	Variance	Skewness	Kurtosis
Agriidae	0.04	0.00	1.40	0.00	43	6	87.76	2.10	2.07	1.98	0.04	6.49
Ancylidae	0.46	0.20	2.60	0.00	12	37	24.49	22.40	28.70	18.46	0.38	1.91
Asellidae	2.79	1.60	15.20	0.00	6	43	12.24	136.90	881.55	499.07	10.40	1.79

The statistics for rows and columns calculated are as follows:

**Mean** - This is the mean of all the values in the data matrix.

**Median** - This is the median of all the values in the data matrix.

**Max** - This is the maximum value in the data matrix.

**Min** - This is the minimum value in the data matrix.

**Zeros** - This is the number of zero entries in the data matrix.

**Non-zeros** - This is the number of non-zero entries in the data matrix.

**% Zeros** - (Number of zeros/Total number of cells) \* 100

**Sum** - This is the sum of all the values in each row or column in the data matrix.

**Sum Sqr** - This is the sums of squares of all the values in each row or column in the data matrix.

**Variance** - This is the variance of all the values in each row or column in the data matrix.

**Skewness** - This is the skewness of all the values in each row or column in the data matrix.

**Kurtosis** - This is the kurtosis of all the values in each row or column in the data matrix.

For groups, the arithmetic mean and variance for each of the variables in each group are presented.

For example, in the example below a user-defined group of samples called Ashley Rails had a mean Aluminium percentage of 17.32%. Group statistics are only available with Raw Data.

### Summary of Data

	No.	Al	Fe	Mg	Ca	Na
Llanederyn	14	12.56	6.37	4.83	0.20	0.25
Caldicot	2	11.70	5.41	3.86	0.30	0.05
Island Thorns	5	18.18	1.71	0.67	0.03	0.05
Ashley Rails	5	17.32	1.51	0.61	0.05	0.05

**Part**

---



## 8 Ordination

CAP includes 4 well-known and powerful methods of ordination analysis:

[Principal Component Analysis \(PCA\)](#)<sup>[63]</sup>  
[Detrended Correspondence Analysis \(DECORANA\)](#)<sup>[70]</sup>  
[Non-metric Multi-Dimensional Scaling \(NMDS\)](#)<sup>[75]</sup>  
[Reciprocal Averaging \(RA\)](#)<sup>[82]</sup>

### 8.1 Principal Component Analysis - PCA

The relationship between samples (columns) in terms of their variables cannot normally be visualised because this would require a plot with as many axes as there are variables (rows). If your study only includes 3 variables this is possible, but is quite impossible given 4 or more variables or species. PCA is a technique that may summarise the relationship between the samples in a small number of axes that can be plotted. For such a summarisation to work, there must be some degree of correlation between the descriptive variables so that the effect of a number of these variables can be combined into a single axis. For good general introductions to PCA for non-mathematicians see [Kent & Coker](#)<sup>[172]</sup> (1992) and [Legendre & Legendre](#)<sup>[172]</sup> (1983).

From the ordination drop-down menu CAP offers a PCA undertaken on either the correlation or variance-covariance matrix between the descriptors (the variables in the rows). Once either PCA correlation or PCA covariance is selected a PCA on the working data set is undertaken.

Output from a PCA is presented under a number of tabbed components that can each be viewed by clicking on the tab. These are described in turn below:

[Variance-PCA](#)<sup>[63]</sup>  
[Scores-PCA](#)<sup>[64]</sup>  
[Eigenvectors-PCA](#)<sup>[65]</sup>  
[Cross products](#)<sup>[66]</sup>  
[PCA plot](#)<sup>[66]</sup>  
[Principal Axis vs Variable Plot](#)<sup>[69]</sup>  
[Scree Plot](#)<sup>[70]</sup>

If you wish to test if samples in a PCA are outliers using the Mahalanobis distance see [PCA -Cor - Outlier R](#)<sup>[188]</sup> or [PCA -covar -Outlier R](#)<sup>[190]</sup>

If you wish to run a PCA using R see [Run R code](#)<sup>[175]</sup>

#### 8.1.1 Variance

This form presents in the first column the eigenvalues of the dispersion (correlation or variance-covariance) matrix arranged from largest to smallest. In the second column the cumulative total of the eigenvalues is given. The third column, labeled % of Total Variance, gives the cumulative total of the eigenvalues presented as a percentage of the total sum of the eigenvalues. This gives the total variance in the dispersal matrix represented by the cumulative total magnitude of the eigenvalues. If the relationship between the samples (columns) is to be usefully represented by a small number of axes, then the first 3 or 4 eigenvalues should represent a large proportion of the total variance. The amount of the total % variance represented can be seen in the fourth column.

Results - Variance				
	Eigenvalues	Cumulative Total	% of Total Variance	Cum. % of Total Variance
1	7.03	7.03	23.44	23.44
2	5.00	18.23	16.66	40.10
3	3.55		11.85	51.95
4	2.64		8.81	60.76
5	2.14	20.37	4.39	67.89
6	1.76	22.13		73.75
7	1.48	23.60		78.68
8	1.32	24.92	4.39	83.07
9	1.11	26.03	3.69	86.76
10	0.81	26.84	2.70	89.46
11	0.75	27.58	2.48	91.94
12	0.70	28.28	2.32	94.26

See [Printing and exporting text](#)<sup>[66]</sup> to save or print this table.

## 8.1.2 Scores

This table gives the co-ordinates of the different samples (columns of the working data) along each of the axes. These scores are displayed graphically by clicking on the [PCA plot](#)<sup>[66]</sup> tab.

CAP Demo 5.0.0.444 - Hinkley fish								
File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare Help								
Results - Scores								
Comp. Scores	h1	h2	h3	h4	h5	h6	h7	h8
Axis 1	0.86527	0.797676	0.768816	1.58736	1.82331	0.464189	1.95298	0.569
Axis 2	0.90785	1.0906	0.379015	0.565962	-0.733725	-0.835473	-1.21406	0.702
Axis 3	-0.342014	-0.917669	-1.73459	-0.294105	1.44364	0.203005	-0.466905	1.83
Axis 4	-0.18218	-0.0217988	0.512964	-0.154067	1.29685	0.259347	-1.37074	-1.80
Axis 5	0.0228712	0.263171	-0.134391	-0.131014	1.64881	0.0845188	1.67538	-1.28
Axis 6	-0.113379	0.197498	1.16191	-0.516432	-0.0101705	0.238034	0.856435	1.71
Axis 7	-0.0972923	0.486381	0.208919	0.0568635	0.196448	0.156024	-0.898973	0.485
Axis 8	0.0798849	-0.107802	-0.39772	-0.379635	0.759015	-0.260843	0.760929	0.520
Axis 9	-0.153743	0.829893	2.43867	-0.707672	-0.651749	0.0540694	-0.26785	1.90
Axis 10	-0.0556778	-0.245064	-1.63597	0.142262	-0.629275	-0.319547	3.58332	0.190
Axis 11	0.197285	0.387445	0.0776469	-0.402712	2.71843	0.686891	3.21399	0.982
Axis 12	-0.0722087	-0.169923	0.735907	-1.13792	1.66079	-0.00998072	-1.5002	0.743
Axis 13	-0.399794	-0.181702	1.47602	-0.315769	2.80189	0.113015	-5.92979	-7.9967
Axis 14	0.278602	0.360709	1.29388	-2.04687	-0.254076	-0.0538067	-0.0685812	-0.83
Axis 15	-0.330658	-0.0864406	1.56669	1.42773	0.128796	-0.340724	0.884519	-0.13
Axis 16	0.387316	0.955674	1.17735	0.717672	-0.983502	0.290425	-4.26901	0.60
Axis 17	-0.0443264	-0.124094	-1.05004	-1.14322	-0.669943	-0.493028	-1.5197	-0.47

See [Printing and exporting text](#)<sup>[66]</sup> to save or print this table.

### 8.1.3 Eigenvectors

This table gives the eigenvectors associated with each eigenvalue as columns. The eigenvectors are displayed graphically by clicking on the [PCA plot](#)<sup>[66]</sup> tab. Note that you can sort the eigenvectors and the variables by clicking on the title row.

CAP Demo 5.0.0.465 - Hinkley fish

File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental

**Results - Eigenvectors**

Δ	1	2	3	4	5	6
Agonus cataphractus (L.)	0.0005659	-0.00080		0.0028102	-0.0041339	-0.0034021
Alosa fallax (L.)	-0.0029772	-0.00306		0.064558	0.0092315	-0.0660257
Ammodytes n	0.0000051	0.000105		0.000651	0.0000335	0.0001482
Ammodytes tobianus L.	0.0000783	-0.000376		-0.0007483	-0.0014046	-0.0002275
Anguilla anguilla (L.)	-0.0013530	-0.0053685	0.0268276	-0.0085244	0.0039256	0.0138274
Aphia minuta (Risso)	-0.0029187	-0.0044995	0.0261443	-0.0218721	0.0081514	-0.0035041
Atherina boyeri Risso	0.0001144	-0.0005162	-0.0003400	0.0008730	-0.0015584	0.0000411
Balistes carolinensis (Gmelin)	-0.0000242	0.0000380	-0.0000896	-0.0000563	-0.0001712	0.0003414
Belone bellone (L.)	0.0000942	-0.0000484	0.0000057	-0.0000378	-0.0001548	0.0000526
Blennius gattorugine L.	-0.0000410	0.0000044	-0.0001824	-0.0001664	0.0005824	-0.0010306

Click in this title cell to change the alphabetical order of variables.

Click in the eigenvector title cell to change the numerical order of eigenvectors.

See [Printing and exporting text](#)<sup>[68]</sup> to save or print this table.

### 8.1.4 Viewing the correlation between variables (species)

If you wish to view either the Pearson correlation coefficients or the variance - covariance between variables (rows), they can be viewed by selecting Ordination/PCA Correlation or Ordination/PCA Covariance respectively, and clicking on the Cross Products tab.

**Results - Cross Products**

	Agonus cataphractus (L.)	Alosa fallax (Lacepede)	Ammodytes tobianus L.	Anguilla
Agonus cataphractus (L.)	20.2451591	0.4192016	0.3540409	-2.8
Alosa fallax (Lacepede)	0.4192016	165.5797272	-8.6985731	14.0
Ammodytes tobianus L.	0.3540409	-8.6985731	12.9176607	2.6
Anguilla anguilla (L.)	-2.8315649			77.3
Aphia minuta (Risso)	-3.9437773			28.9
Atherina boyeri Risso	0.5965693			-6.5
Balistes carolinensis (Gmelin)	-0.7258010			-1.9
Belone bellone (L.)	0.9696648			0.0
Blennius gattorugine L.	0.4390524	17.1700592	-1.1732386	-1.2
Buglossidium luteum (Risso)	-0.7717339	-5.6673970	-0.0292657	-1.9
Callionymus lyra L.	1.4196661	3.6712954	-0.3301534	-2.8
Centrolabrus exoletus (L.)	-1.9340116	-0.7684171	1.3849481	1.0
Ciliata mustela (L.)	27.7677975	-58.7521324	0.2015389	-84.7
Ciliata septentrionalis (Collet)	-1.0423725	-12.7381201	-3.7925851	-10.8
Clupea harengus L.	21.6786251	-59.9647141	-0.1394902	-158

This element shows that the covariance between the numbers of Ammodytes tobianus and Agonus cataphractus is about 0.354. If a PCA had been undertaken on the correlation matrix, the correlation coefficient would be displayed here.



### 8.1.5 Cross products

This table shows either the variance-covariance matrix or the correlation matrix between the descriptors (variables). The matrix displayed will depend on whether the PCA was undertaken on the correlation or the variance-covariance matrix.

Results - Cross Products				
	Agonus cataphractus (L.)	Alosa fallax (Lacepede)	Ammodytes tobianus L.	Anguilla
Agonus cataphractus (L.)	20.2451591	0.4192016	0.3540409	-2.8
Alosa fallax (Lacepede)	0.4192016	165.5797272	-8.6985731	14.0
Ammodytes tobianus L.	0.3540409	-8.6985731	12.9176607	2.6
Anguilla anguilla (L.)	-2.8315649			77.3
Aphia minuta (Risso)	-3.9437773			28.9
Atherina boyeri Risso	0.5965693			-6.5
Balistes carolinensis (Gmelin)	-0.7258010			-1.9
Belone bellone (L.)	0.9696648			0.0
Blennius gattorugine L.	0.4390524	17.1700592	-1.1732386	-1.2
Buglossidium luteum (Risso)	-0.7717339	-5.6673970	-0.0292657	-1.9
Callionymus lyra L.	1.4196661	3.6712954	-0.3301534	-2.8
Centrolabrus exoletus (L.)	-1.9340116	-0.7684171	1.3849481	1.0
Ciliata mustela (L.)	27.7677975	-58.7521324	0.2015389	-84.7
Ciliata septentrionalis (Collet)	-1.0423725	-12.7381201	-3.7925851	-10.8
Clupea harengus L.	21.6786251	-59.9647141	-0.1394902	-158

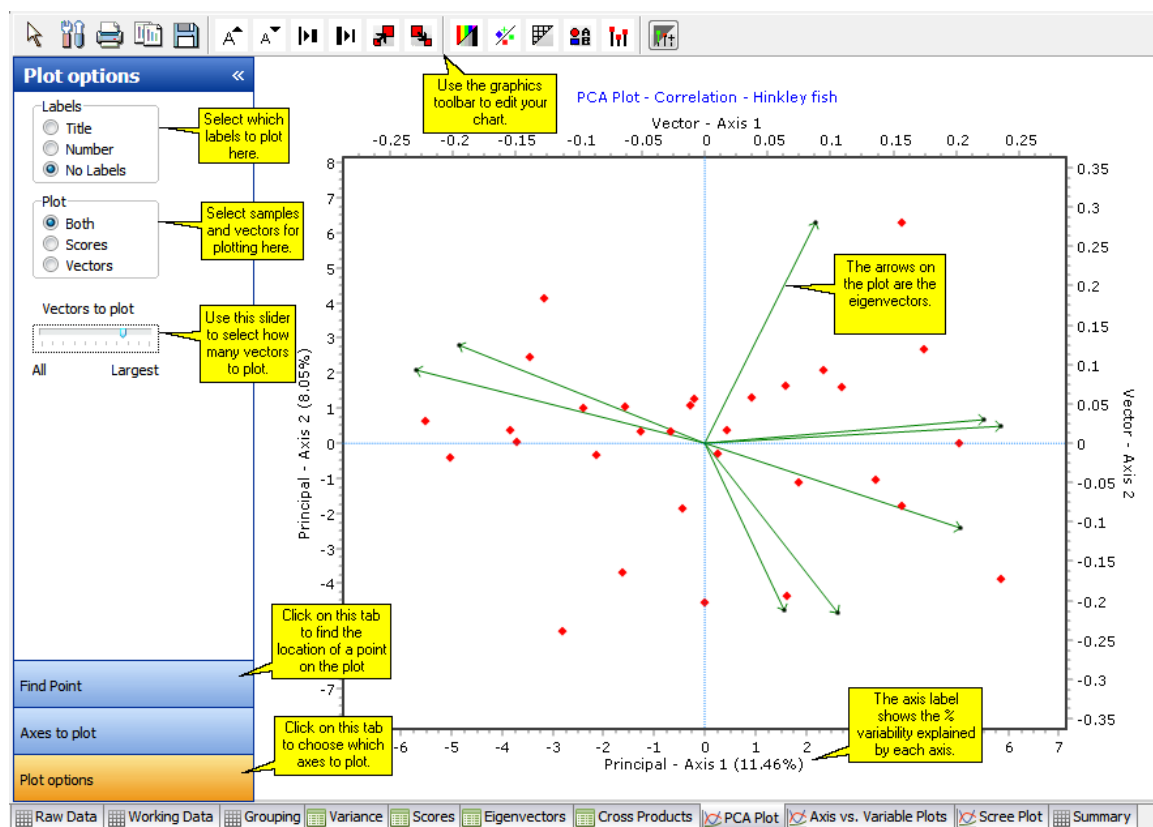
This element shows that the covariance between the numbers of Ammodytes tobianus and Agonus cataphractus is about 0.354. If a PCA had been undertaken on the correlation matrix, the correlation coefficient would be displayed here.

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

### 8.1.6 PCA plot

This window shows plots of the sample [scores](#)<sup>[64]</sup> and the [eigenvectors](#)<sup>[65]</sup>. The graph can be [exported](#)<sup>[153]</sup>, [copied](#)<sup>[153]</sup> and printed.

The display options are selected from the panel to the left of the graph. This panel can be shrunk to the left to make the plot as large as possible. Above the plot there is a toolbar of commonly-used graphics editing tools.



### Plot options tab

The label radio box on the left under Plot options is used to select labelling for the samples.

Select: **Title** to display on the plot the names of the samples.

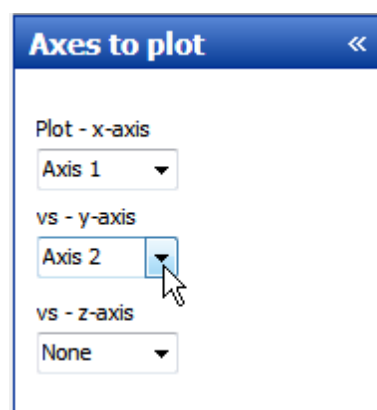
**Number** to show the sequential number of each sample in the working data set.

**No labels** to remove labels from the plot.

The Plot radio box, on the left under Plot options, allows sample scores, the eigenvectors (one for each species or row in the working data ) or both to be plotted. If both are selected then the biplot uses different axes to plot the scores and eigenvectors. The axes for the scores are displayed at the bottom and on the left.

### Axes to plot tab

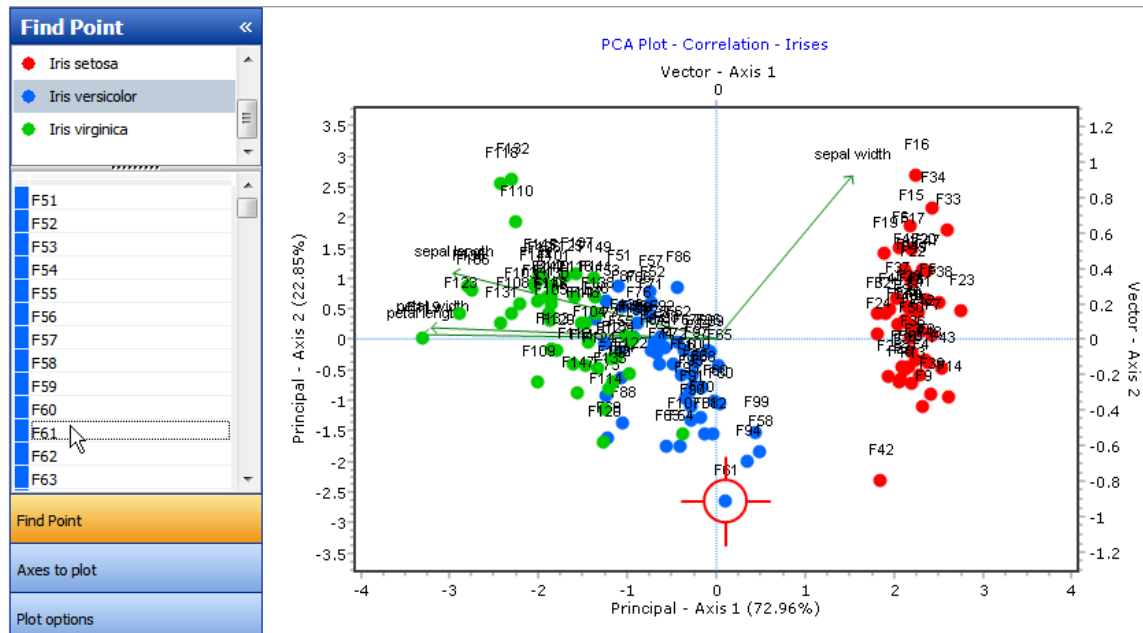
The axes displayed in the plot are selected using the **Axes to plot** tab. The default is axis 1 and axis 2, which will display the relative positions of the samples with respect to the two largest components. A 3D plot is produced if a z variable is selected.



### Find point tab

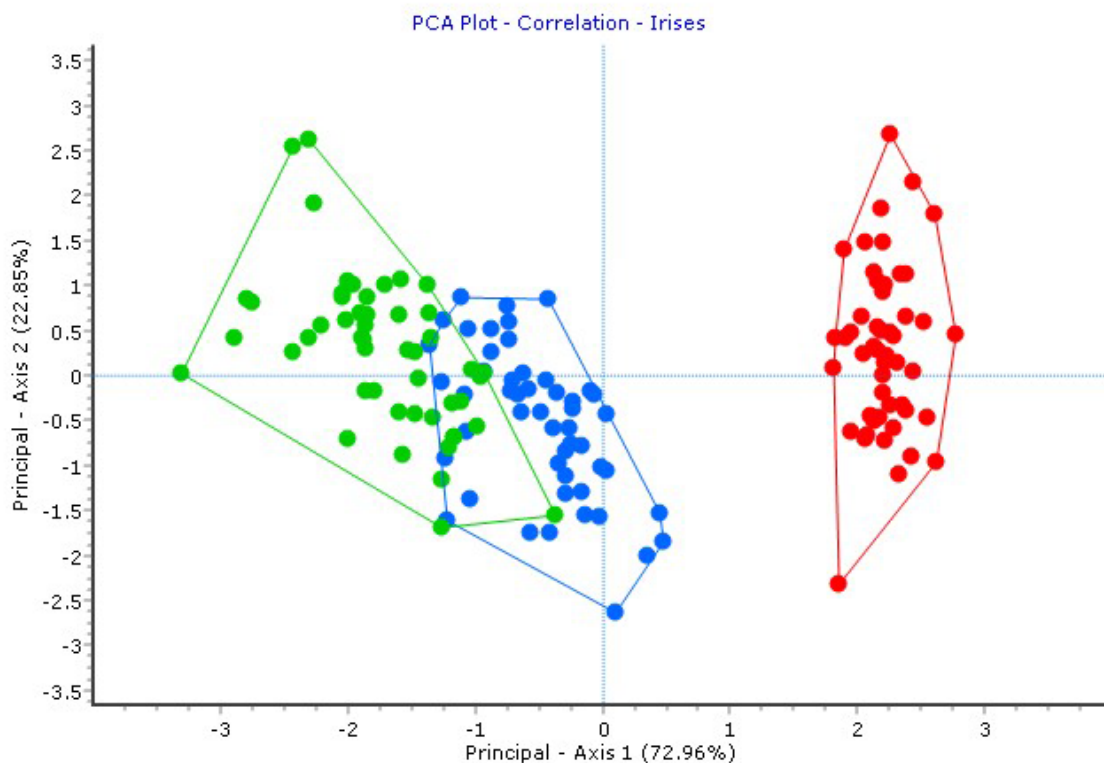


Use the **Find point tab** to locate the position of an individual point on the chart. In the top pane of the tab is a list of the sample groups; click on a group and the individual samples in that group are shown in the lower pane. Double-click and hold down on the sample you wish to locate; the corresponding point on the chart will be marked by a red cross-hairs symbol:



### Setting a perimeter

It is possible to [display the perimeter of each predefined group](#)<sup>[16]</sup>. Below is an example output for Fisher's iris data showing the grouping of the 3 species, with each group outlined:



If you have performed a PCA on an ungrouped data set and, on inspection of the plot, you decide

that a group exists that you wish to emphasise, e.g. for publication, then it will be necessary to scrutinise the plot to identify and take note of all the samples that you wish to encircle. Click on the Grouping tab to show the [group membership editing page](#)<sup>[53]</sup>, assign those samples to a group and then re-run the PCA. Set data point colours as preferred and then [create the perimeter line](#)<sup>[167]</sup>.

See also:

[Editing charts](#)<sup>[157]</sup>

[Drawing a perimeter](#)<sup>[167]</sup>

[Preparing charts for output](#)<sup>[162]</sup>

Printing charts

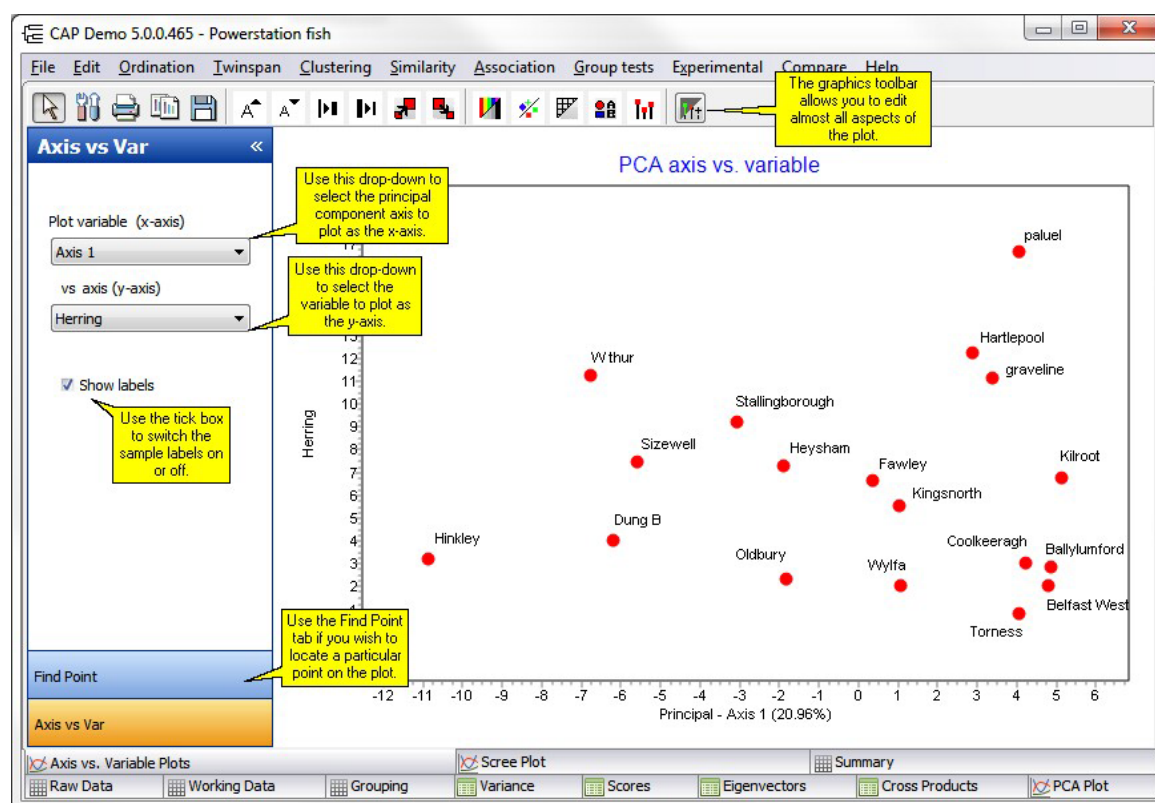
[Exporting charts](#)<sup>[153]</sup>

[Zooming on charts](#)<sup>[158]</sup>

[Themes for graphs](#)<sup>[167]</sup>

### 8.1.7 Principal Axis vs Variable Plot

This window allows for the exploration of the correlation between a selected principal axis produced by PCA and the individual variables. This can be helpful to aid identification of the variables which mostly contribute to the ordination along one axis. The graph can be [exported](#)<sup>[153]</sup>, [copied](#)<sup>[153]</sup> and printed.



See also

[Preparing charts for output](#)<sup>[162]</sup>

Printing charts

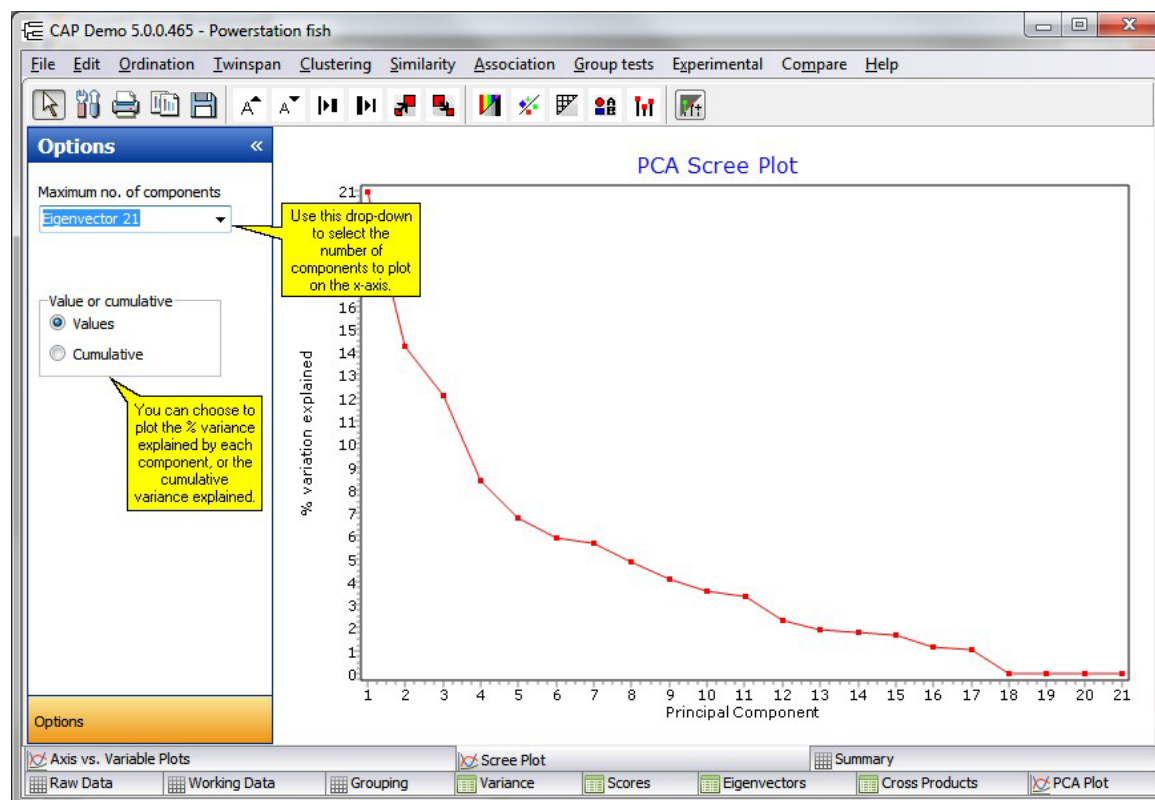
[Exporting charts](#)<sup>[153]</sup>

[Zooming on charts](#)<sup>[158]</sup>

[Themes for graphs](#)<sup>[167]</sup>

### 8.1.8 Scree Plot

A Scree Plot shows the decline in the fraction of total variance in the data explained by each Principal Component. The PCs are plotted in decreasing order of their contribution to the total variance. The display options are selected from the panel to the left of the graph. There is also the option to plot the cumulative variance explained by 1, 2, 3 or more principal components.

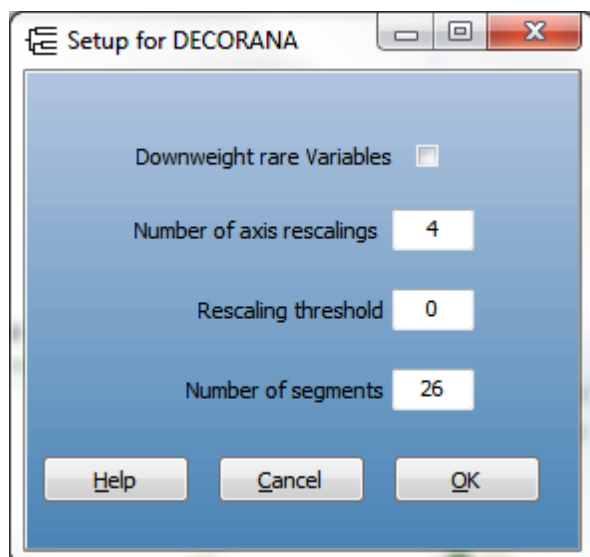


## 8.2 Detrended Correspondence Analysis - DECORANA

Detrended Correspondence Analysis was devised by Hill (1979) as an attempt to improve upon reciprocal averaging (RA). Two problems that occur with reciprocal averaging are termed the 'arch effect' and 'end point compression'. When the first and second axes produced by RA are plotted, it is often observed that the points are arranged in an arch, because of the quadratic relationship between the axes, rather than because of any relationship between the samples. DECORANA removes this arch by a technique termed detrending. The tendency for points at each end of the first axis to be closer together than those in the middle is removed by segmenting the axis and expanding the terminal segments and compressing those towards the centre. Whereas RA scales the axes between 0 and 100 in relation to the magnitude of the eigenvalue, DECORANA scales in units of average standard deviation of species turnover. Therefore a change of 50% in species composition occurs in about 1 standard deviation.

Samples or species which are very different in their composition or distribution from others in the data set present problems for DECORANA and in general such outliers should be removed from the data before the method is applied.

When DECORANA is selected from the ordination drop-down menu, the Setup for DECORANA window is displayed. In many cases, the default settings can be used by simply clicking OK.



The options are described in turn below:

**Downweight rare** - Select this option if the influence of rare species is to be reduced. If selected, the abundances of species rarer than the frequency of the commonest species divided by 5 are down-weighted in proportion to their frequency.

**Number of axis rescalings** - Input an integer number; the default of 4 should generally be used.

**Rescaling threshold** - Axes shorter than this value will not be re-scaled. The default is zero.

**Number of segments** - This is the number of segments the axis is divided into for re-scaling. Input an integer number, the default of 26 should generally be used.

Output for DECORANA is presented under a number of tabbed components that can each be viewed by clicking on the tab. These are described in turn below:

[Computations - DECORANA](#)<sup>[71]</sup>

[Species Scores - DECORANA](#)<sup>[72]</sup>

[Sample Scores - DECORANA](#)<sup>[73]</sup>

[DECORANA plot](#)<sup>[74]</sup>

### 8.2.1 Computations - DECORANA

This text window gives the values of the eigenvalues calculated for each axis. These are estimated by iteration and this window also gives the residual term for each iteration. The eigenvalues are calculated to an accuracy of 0.000001. If after 999 iterations this residual value is not achieved, computation is halted and a warning message produced.

The length of the gradient is proportional to the rate of species' appearance and disappearance (species turnover) along that gradient.

To print this output use **File: Print**. Press Ctrl-Alt-C, or Edit: Copy All, to copy the entire text of this output. To copy a selected portion, select the text you require, and press Ctrl-C on your keyboard, or Edit: Copy.

## 8.2.2 Species Scores - DECORANA

This window presents in a grid the eigenvalues for each axis and the variable (species) scores for the first 4 axes. These scores are the coordinates of each variable used in the ordination plot (see [DECORANA plot](#)<sup>[74]</sup>).

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

The table can be sorted by ascending or descending order of any of the columns, simply by clicking on the top cell of the required column. Click the cell again to switch between ascending and descending order.

Species Scores					
N	Name ↴	Axis 1	Axis 2	Axis 3	Axis 4
	Eigenvalues =			621914	0.0227853
37	Anchovy			27	331
89	Angler fish	219	220	-13	-99
28	Bass	-33	136	198	-22
69	Blenny, Common (Shanny)	264	192	-2	-224
76	Blenny, Tompot	141	280	-8	-135
74	Bream	-182	139	-9	-31
67	Brill	10	245	82	-17
40	Bull rout (Sh. Sp. Sea scorpion)	283	-2	5	96
16	Cod	20	24	-28	35
52	Conger	101	271	95	84
9	Dab	176	128	-35	31
72	Dab, Long rough	347	-55	314	585
32	Dogfish, Lesser spotted	239	171	17	52
111	Dory (John dory)	22	281	124	275
23	Dragonet	205	141	-4	-62
18	Eel, Common	54	76	84	20
8	Flounder	79	99	21	101
68	Garfish	246	191	12	-239
107	Gilthead	201	243	721	-373
71	Goby, Black	82	233	109	-127
99	Goby, Common	178	150	195	225

Rearrange the sort order of the table by clicking on the title cell of any of the columns.

Raw Data Working Data Grouping Computations Species Scores Sample S

Species Scores					
N	Name ▾	Axis 1	Axis 2	Axis 3	Axis 4
	Eigenvalues =			621914	0.0227853
37	Anchovy			27	331
89	Angler fish	219	220	-13	-99
28	Bass	-33	136	198	-22
69	Blenny, Common (Shanny)	264	192	-2	-224
76	Blenny, Tompot	141	280	-8	-135
74	Bream	-182	139	-9	-31
67	Brill	10	245	82	-17
40	Bull rout (Sh.Sp.Sea scorpion)	283	-2	5	96
16	Cod	20	24	-28	35
52	Conger	101	271	95	84
9	Dab	176	128	-35	31
72	Dab, Long rough	347	-55	314	585
32	Dogfish, Lesser spotted	239	171	17	52
111	Dory (John dory)	22	281	124	275
23	Dragonet	205	141	-4	-62
18	Eel, Common	54	76	84	20
8	Flounder	79	99	21	101
68	Garfish	246	191	12	-239
107	Gilthead	201	243	721	-373
71	Goby, Black	82	233	109	-127
99	Goby, Common	178	150	195	225

Rearrange the sort order of the table by clicking on the title cell of any of the columns.

Raw Data Working Data Grouping Computations Species Scores Sample S

### 8.2.3 Sample Scores - DECORANA

This window presents in a grid the eigenvalues for each axis and the sample (quadrat) scores for the first 4 axes. These scores are the coordinates of each sample used in the ordination plot (see [DECORANA plot](#)<sup>[74]</sup>).

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

The table can be sorted by ascending or descending order of any of the columns, simply by clicking on the top cell of the required column. Click the cell again to switch between ascending and descending order.



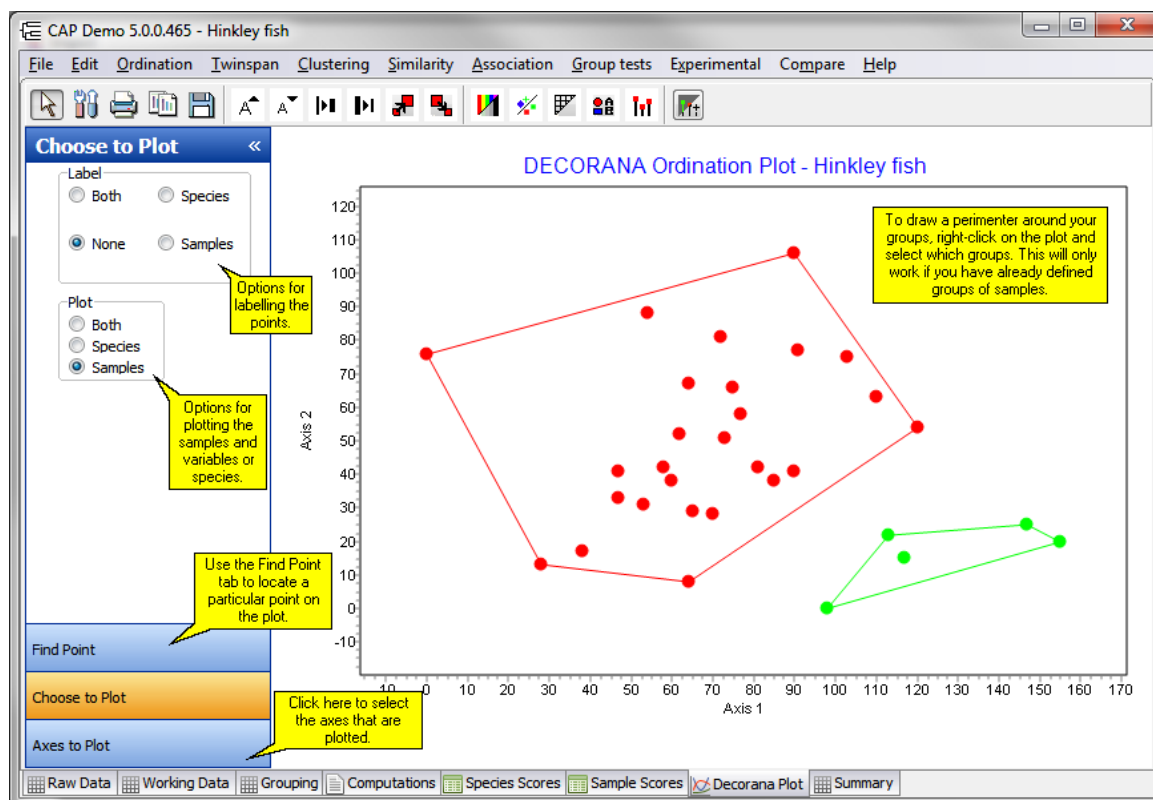
Site Scores (weighted mean Variables scores)					
N	Name	Axis 1	Axis 2	Axis 3	Axis 4
1	Hink	0.2020477	0.1189526	0.0621914	0.0227853
2	W t			50	35
3	Sizewell	105	86	27	44
4	Wylfa	150	110	45	22
5	Fawley	118	93	94	8
6	Oldbury	0	89	65	32
7	Heysham	110	79	50	39
8	Dung B	138	108	41	8
9	Hartlepool	121	3	0	51
10	Kingsnorth	91	60	44	21
11	Torness	182	83	8	0
12	Coolkeeragh	139	57	105	114
13	Ballylumford	202	72	62	70
14	Kilroot	177	53	75	92
15	Belfast West	149	21	57	40
16	graveline	99	33	81	64
17	paluel	97	0	136	53
18	Stallingborough	63	3	29	27

Rearrange the sort order of the table by clicking on the title cell of any of the columns.

## 8.2.4 DECORANA plot

Select this tab to display the ordination plots of the data. This window shows plots of both the sample and species scores. The display options are selected from the **Choose to Plot** panel to the left of the graph. This panel can be shrunk to the left to make the plot as large as possible. Above the plot there is a toolbar of commonly-used graphics editing tools.

You can also [draw a perimeter](#) <sup>[16]</sup> round a group, as shown in the plot below:



The label radio box is used to select labeling for the points. Select **Both** to display on the plot the names of the samples and species. Select **None** if no labels are required. Select **Species** (variables) or **Samples** to display variable or sample names respectively.

The Plot radio box allows sample, species or both ordinations to be plotted.

The axes displayed in the plot are selected from the **Axis to Plot** tab using the Plot x Axis, y Axis z Axis drop-down boxes. The default is Axis 1 and Axis 2, which will display the relative positions of the samples with respect to the two largest components. A 3D plot is produced if a z variable is selected.

### 8.3 Non-metric Multi-Dimensional Scaling

Multi-Dimensional Scaling (MDS) is a technique for expressing the similarities between different objects in a small number of dimensions. Hopefully, this allows a complex set of inter-relationships to be summarised in a simple figure. The method attempts to place the most similar objects (samples) closest together. The starting point for the calculations is a similarity or dissimilarity matrix between all the sites or quadrats. These can be non-metric distance measures for which the relationships between the sites/objects/samples (columns) cannot be plotted in a Euclidean space. The aim of Non-metric MDS is to find a set of metric coordinates for the sites which most closely approximates their non-metric distances.

The basic MDS algorithm is as follows:

1. Calculate the similarity or dissimilarity between sites.
2. Assign to each site a set of coordinates in p-dimensional space. These coordinates can be either chosen at random or chosen using Principal Coordinates Analysis (note, this is **not** the same as a [Principal Component Analysis](#)<sup>[63]</sup>). The value of p is chosen by the user.
3. Compute the Euclidean distance between these sites using the starting coordinates.
4. Compare the original dissimilarity between the sites with these Euclidean distances by calculating a stress function. The smaller the stress function, the closer the correspondence.
5. Adjust the positions so as to reduce the stress.



6. Repeat 2 to 4 until the stress is minimised or the maximum number of iterations is reached.

CAP uses Kruskal's least squares monotonic transformation to minimise the stress (see [Kruskal, 1964](#)<sup>[172]</sup>; [Kruskal & Wish](#)<sup>[172]</sup>, 1977). The program is designed to find an optimal two-dimensional representation of the data. It can happen that no useful two-dimensional representation can be produced. While it is possible to produce anything up to a six-dimensional solution, in practice this is of little use, as it cannot be displayed. When requested, CAP lists solutions for 3 or more dimensions, but does not plot them.

From the ordination drop-down menu select Non-metric MDS. The setup for Non-metric MDS is displayed. This offers a series of options that are described in [NMDS Starting Configuration](#)<sup>[76]</sup>. The program can usually be run with the default values.

Output from Non-metric MDS is presented under a number of tabbed components that can each be viewed by clicking on the tab. These are described in turn below:

[MDS setup](#)<sup>[76]</sup>

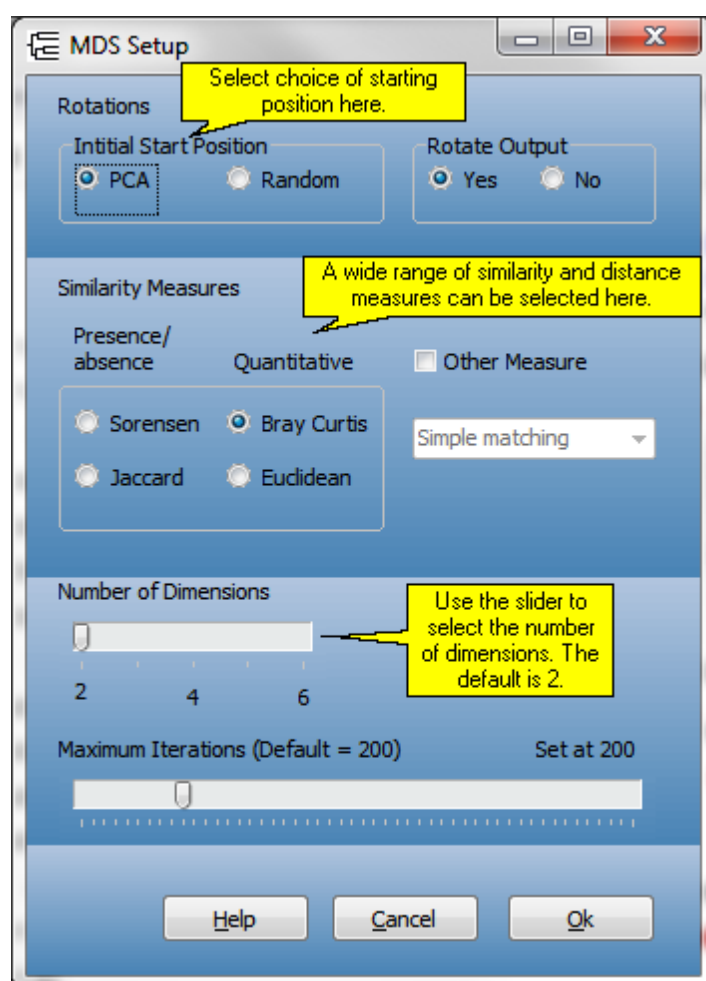
[MDS plots](#)<sup>[80]</sup>

[Sample coordinates](#)<sup>[77]</sup>

[Stress](#)<sup>[78]</sup>

### 8.3.1 Starting Configuration - MDS

When Non-metric MDS is selected, the MDS Setup dialog is activated.



#### Initial Start Position

This is a radio button selection. Select PCA to use a Principal Coordinates Analysis (note, this is **not** the same as a [Principal Component Analysis](#)<sup>[63]</sup>) to calculate the starting coordinates of the sites (columns). Select Random to use a random number generator to assign initial coordinates. The default is to use a PCA. The algorithm used by CAP to find the final coordinates will produce different results depending upon the starting configuration. Indeed, a poor initial choice may result in the algorithm becoming trapped before a low stress value is reached. Generally, PCA will produce a satisfactory choice. It is often useful to undertake a few runs with random starting points to satisfy yourself that the algorithm is finding a minimum-stress solution.

#### **Rotate output.**

Select Yes or No by clicking on the radio buttons. If Yes is selected a PCA is performed on the final site coordinates. The default is Yes, as this will usually produce the most suitable plot.

#### **Similarity measures.**

The measure to use is selected by clicking on the radio button. A number of different measures can be used to calculate the distances between sites. The regular options available are Euclidean distance, Sørensen, Jaccard and Bray Curtis. The default is Bray Curtis. If you prefer, tick the 'Other measure' box and select any of the other similarity measures offered by CAP.

#### **Number of dimensions.**

This is a slider bar activated by the mouse. The default value is 2. If a higher number is chosen, the program will calculate the configuration of the points from this dimension down to 1 dimension, list the final stress for each number of dimensions under the Stress tab, and display the change in the stress value with number of dimensions graphically. The coordinates of the points for the maximum number of dimensions are displayed under the Sample coordinates tab. However, the plot of the positions of the site is that found for a two-dimensional space.

#### **Maximum iterations.**

You can change the number of iterations used by the stress minimisation algorithm by using the mouse to move the slider bar. The default is 200. While there is a relationship between the number of iterations and the magnitude of the stress level achieved, in practice there is often little advantage in selecting a higher iteration number. You may like to vary this number to become satisfied that the minimum stress level possible has been achieved.

[Main MDS Page](#)<sup>[75]</sup>

### **8.3.2 Site coordinates**

This grid presents the final coordinates for each site or sample (column). The site identifier is given in the first column; in the second and third columns are listed the coordinates of each site for the two-dimensional solution. These are the coordinates plotted in the [MDS Plot](#)<sup>[80]</sup>.

Columns to the right of column 3 will also be filled if you specified a maximum number of dimensions above 2 (the default). In these columns are the coordinates of each site calculated for the model with the maximum dimension size selected. These are not available for plotting within CAP.

Results - Coordinates MDS		
Name	Axis 1 (2D)	Axis 2 (2D)
1981	-1.21462	-0.281509
1982	-1.00969	-0.808727
1983	-1.54757	-0.143054
1984	-0.499005	-0.230403
1985	-1.54571	0.573973
1987	1.20731	-0.0917531
1988	-0.628002	-0.172999
1989	-0.529236	0.321535
1990	-0.17165	0.706175
1991	-0.124583	-0.172774
1992	0.699875	-0.518585
1993	0.970876	-0.139061
1994	0.444878	0.174477
1995	-0.842218	0.083216
1996	0.331115	-0.278685
1997	0.923586	-0.490612
1998	2.19015	0.0215943
1999	0.364542	0.165458
2000	0.737097	0.268788
2001	-0.326702	-0.0617328
2002	0.544244	1.11484

Raw Data Working Data Grouping Similarity Sam

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

[Main MDS Page](#)<sup>[75]</sup>

### 8.3.3 Stress

This grid shows the iteration history for a two-dimensional non-metric multidimensional scaling. The first column gives the iteration number, the second the stress value at this iteration, and the third, the step size used to move the site positions.

The change in stress can also be plotted; see [MDS plots](#)<sup>[80]</sup>.

CAP Demo 5.0.0.465 - Hinkley fish

File Edit Ordination Twinspan Clustering Similarity

**Results - Stress and Final Stress**

Iteration	Stress	Step Size	Dimension No.	Final Stress
0	0.4494	0.1105	1	0.2669
1	0.3740	0.2962	2	0.1367
2	0.1974	0.8943		
3	0.1636	0.3644		
4	0.1542	0.1586		
5	0.1470	0.0949		
6	0.1406	0.0377		
7	0.1382	0.0238		

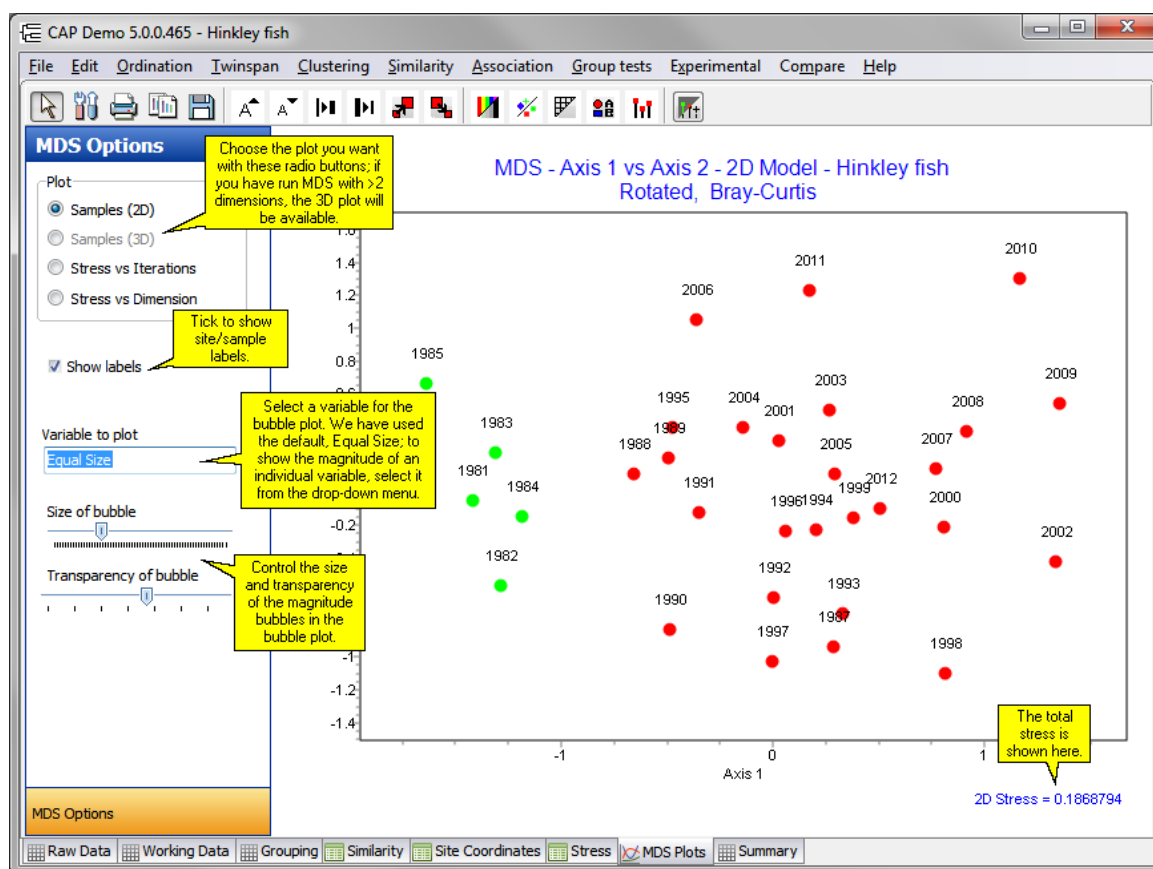
If a maximum number of dimensions above 2 is selected, this grid also tabulates the final stress value for models ranging from 1-dimensional to the maximum dimension selected by the user (the default is 2). These data are also available as a plot. These data can give an idea of the true dimensionality of the data set, as stress will reach a minimum value once the necessary number of dimensions required to fully express the distances between sites is reached. It gives an impression of the suitability of a two-dimensional plot to show between-site relationships. If a two-dimensional representation is satisfactory, then an increase in dimension number will not greatly decrease stress.

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

[Main MDS Page](#)<sup>[75]</sup>

### 8.3.4 MDS plots

This window can show four different plots, which are each selected using the radio buttons to the left of the graph in the MDS Options panel.



**Samples (2D)** - This shows the plot of the final positions for the samples or objects (columns).

**Samples (3D)** - If you selected more than 2 dimensions for the MDS, this will plot axes 1 to 3.

**Stress vs iterations** - This shows the plot of the stress against iteration number for a two-dimensional non-metric multidimensional scaling model. A successful analysis should show an approximately asymptotic decline in stress with iteration number.

**Stress vs Dimension** - This shows the plot of the final stress value plotted against the dimension of the model. The maximum dimension on the plot is the maximum number of dimensions selected during setup. This plot can give an idea of the true dimensionality of the data set as stress will reach a minimum value once the necessary number of dimensions to fully express the distances between sites is reached. It gives an impression of the suitability of a two-dimensional plot to show between-site relationships. If a two-dimensional representation is satisfactory, then an increase in dimension number will not greatly decrease stress.

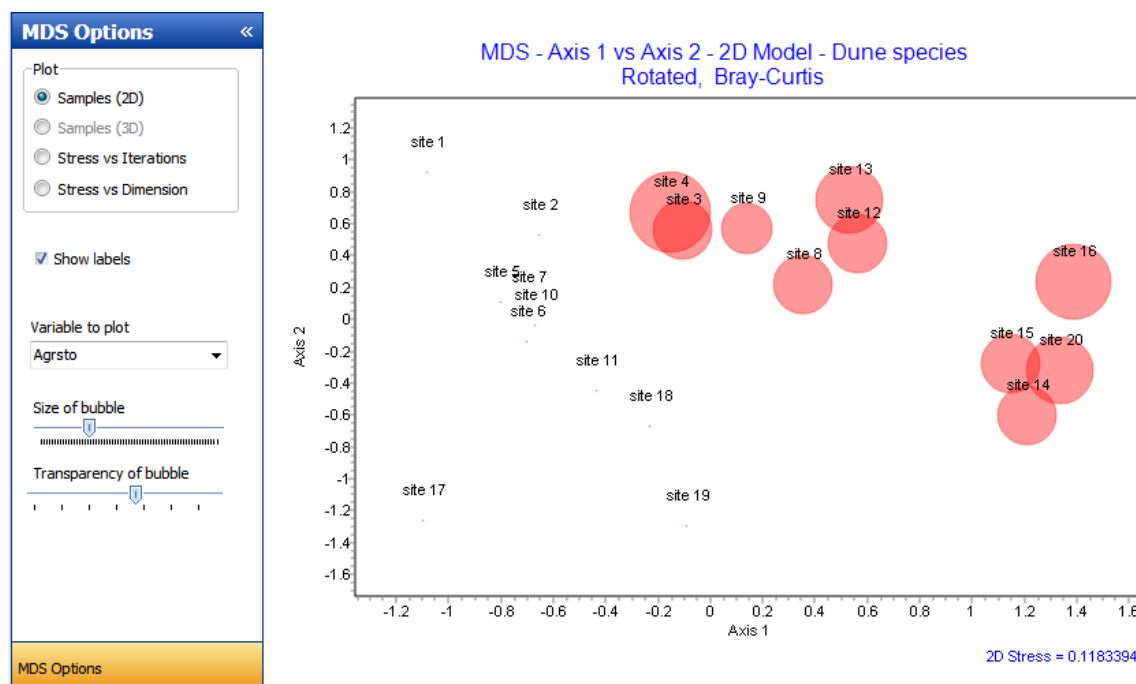
**Show labels** - This tick box switches object names on and off.

**Variable to plot** - In its default setting, the plot will show all sites/samples at an equal size. If you select one of the variables from the drop-down, then the size of the marker representing each site/sample will be proportional to the occurrence of that variable in the site/sample. This can be useful if a distinct group of sites/samples appears in the MDS plot, to detect which variables account for the grouping on the plot. Select variables from the drop-down menu; if all the sites/samples in the group are shown with larger bubbles, then it is likely that the magnitude of that variable's presence in those sites/samples is responsible for the grouping together of those sites/samples (see the example plot below). To scroll rapidly through the variables in the drop-down menu, click into the menu, then use the Up and Down arrows on your keyboard to scroll through the list of variables.

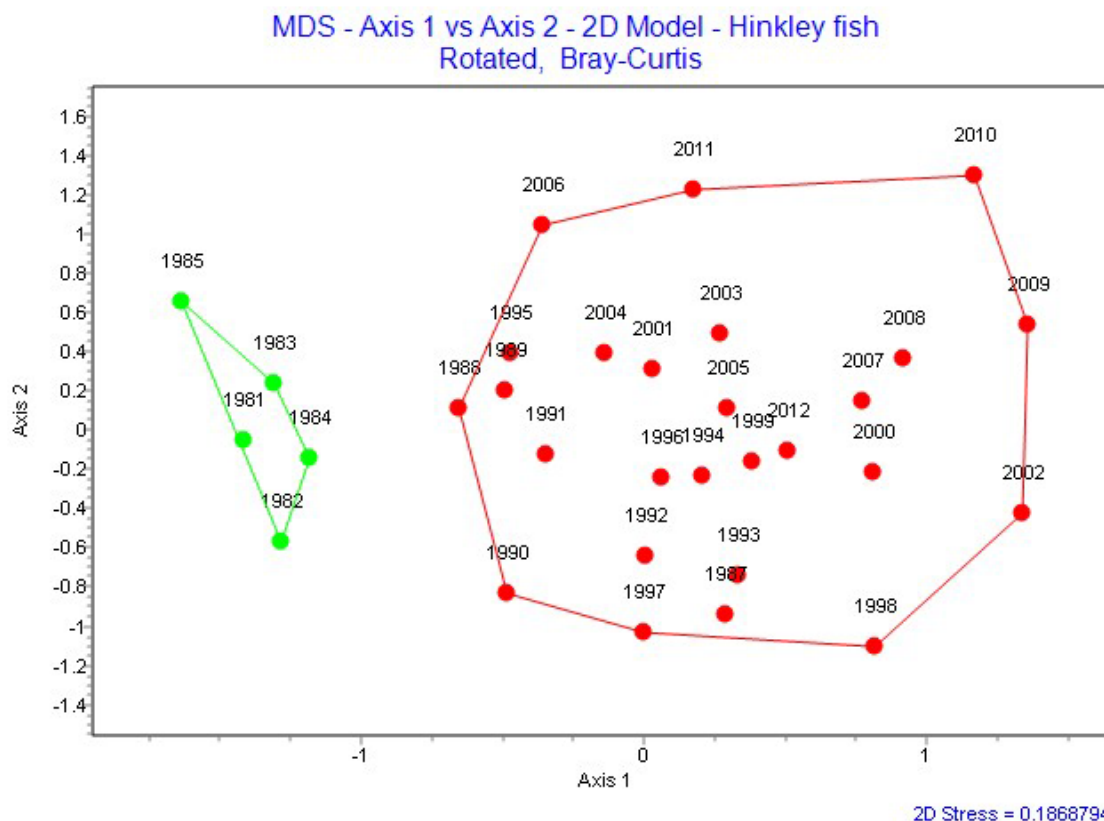
**Size of bubble** - Controls the maximum size of the bubble markers in the plot.

**Transparency of bubble** - sets the level of transparency of overlapping bubbles in the plot.

The plot below shows an MDS analysis of the *Dune species* demo data set, and demonstrates the use of the bubble plot to highlight the variables responsible for sites' grouping on the plot. Sites 3, 4, 8, 9, 12 and 13, and sites 14, 15, 16 and 20, have formed two distinct groups on the MDS plot. Using the drop-down menu to select individual variables, we can show that the magnitude of the occurrence of species 'Agrsto' (the grass, *Agrostis stolonifera*) is in part responsible for the positioning of the two groups of sites. Scrolling through the other variables, you may determine which other species contribute to the position of the two groups, and which species distinguish **between** the two groups.



Y



ou can also [draw a perimeter around groups](#)<sup>[161]</sup>:

See also:

[Editing charts](#)<sup>[157]</sup>

[Preparing charts for output](#)<sup>[162]</sup>

Printing charts

[Exporting charts](#)<sup>[153]</sup>

[Zooming on charts](#)<sup>[158]</sup>

[Themes for graphs](#)<sup>[167]</sup>

[Main MDS Page](#)<sup>[75]</sup>

## 8.4 Reciprocal Averaging - RA

This method, also called Correspondence Analysis, is a method of showing the relationship between both species and samples (quadrats) in a reduced space. Originally proposed by Hirschfeld (1935) and Fischer (1940) it was first used by ecologists in the 1960s (Roux & Roux, 1967; Benzécri, 1967) - see Kent & Coker (1992) for more details. The method is described by [Hill \(1973\)](#)<sup>[172]</sup> and a non-mathematical introduction to the technique is given in [Kent & Coker \(1992\)](#)<sup>[172]</sup>. RA uses Chi-squared distance values; this results in low abundance species (variables) having a possibly disproportionately large effect on the ordination produced, and can over-emphasise the difference in samples containing several infrequently-recorded species. RA performs best for analysing samples that were collected along an environmental gradient. If there are no clear environmental gradients in the habitat under study, or the gradients are short, then [PCA](#)<sup>[63]</sup> may give better results. RA can be applied to both presence/absence and quantitative data.

When Reciprocal Averaging is started, the **Setup for RA** dialog is displayed. There is a single option:

**Downweight rare variables.** Select this option if the influence of rare species or other variables is to be reduced. If selected, the abundances of variables rarer than the frequency of the commonest divided by 5 are down-weighted in proportion to their frequency. The default is no down-weighting. Click on OK to run the program.

Output for Reciprocal Averaging is presented under a number of tabbed components that can each be viewed by clicking on the tab. These are described in turn below.

[Computations](#)<sup>[83]</sup>

[Species Scores](#)<sup>[83]</sup>

[Sample Scores](#)<sup>[84]</sup>

[RA plot](#)<sup>[85]</sup>

### 8.4.1 Computations - Reciprocal Averaging

This text window gives the values of the eigenvalues calculated for each axis. These are estimated by iteration, and this window also gives the residual term for each iteration. The eigenvalues are calculated to an accuracy of 0.000001. If after 999 iterations, this residual value is not achieved, computation is halted and a warning message produced.

To print this output use **File: Print**. Press Ctrl-Alt-C, or Edit: Copy All, to copy the entire text of this output. To copy a selected portion, select the text you require, and press Ctrl-C on your keyboard, or Edit: Copy.

### 8.4.2 Species Scores - Reciprocal Averaging

This window presents in a grid the eigenvalues for each axis and the variable (species) scores for the first 4 axes. These scores are the coordinates of each variable used in the ordination plot (see [RA plot](#)<sup>[85]</sup>).

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

The table can be sorted by ascending or descending order of any of the columns, simply by clicking on the top cell of the required column. Click the cell again to switch between ascending and descending order.



Species Scores					
N	Name Δ	Axis 1	Axis 2	Axis 3	Axis 4
	Eigenvalues =			784337	0.0653890
1	Agonus cataphractus (L.)			25	48
2	Alosa fallax (Lacepede)	131	24	-50	226
3	Ammodytes tobianus L.	88	24	81	-28
4	Anguilla anguilla (L.)	293	-103	178	-126
5	Aphia minuta (Risso)	249	6	120	0
6	Atherina boyeri Risso	-54	-19	34	-45
7	Balistes carolinensis (Gmelin)	57	347	-183	-584
8	Belone bellone (L.)	-325	-416	145	24
9	Blennius gattorugine L.	75	159	-10	411
10	Buglossidium luteum (Risso)	-58	267	-285	-433
11	Callionymus lyra L.	10	57	11	-37
12	Centrolabrus exoletus (L.)	170	61	52	-707
13	Ciliata mustela (L.)	-3	145	5	22
14	Ciliata septentrionalis (Collet)	0	85	-88	-126
15	Clupea harengus L.	-193	416	416	91
16	Conger conger L.	99	47	-74	79
17	Crenilabrus melops (L.)	-159	-348	144	-48
18	Crenimugil labrosus (Risso)	-101	-256	83	51
19	Crystallogobius linearis (von Duben)	191	-113	27	-127
20	Ctenolabrus rupestris (L.)	359	-33	89	-46
21	Cydopterus lumpus L.	175	-147	67	-332

### 8.4.3 Sample Scores - Reciprocal Averaging

This window presents in a grid the eigenvalues for each axis and the sample (quadrat) scores for the first 4 axes. These scores are the coordinates of each sample or site used in the ordination plot (see [RA plot](#)<sup>[85]</sup>).

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

The table can be sorted by ascending or descending order of any of the columns, simply by clicking on the top cell of the required column. Click the cell again to switch between ascending and descending order.


### Site Scores (weighted mean Variables scores)

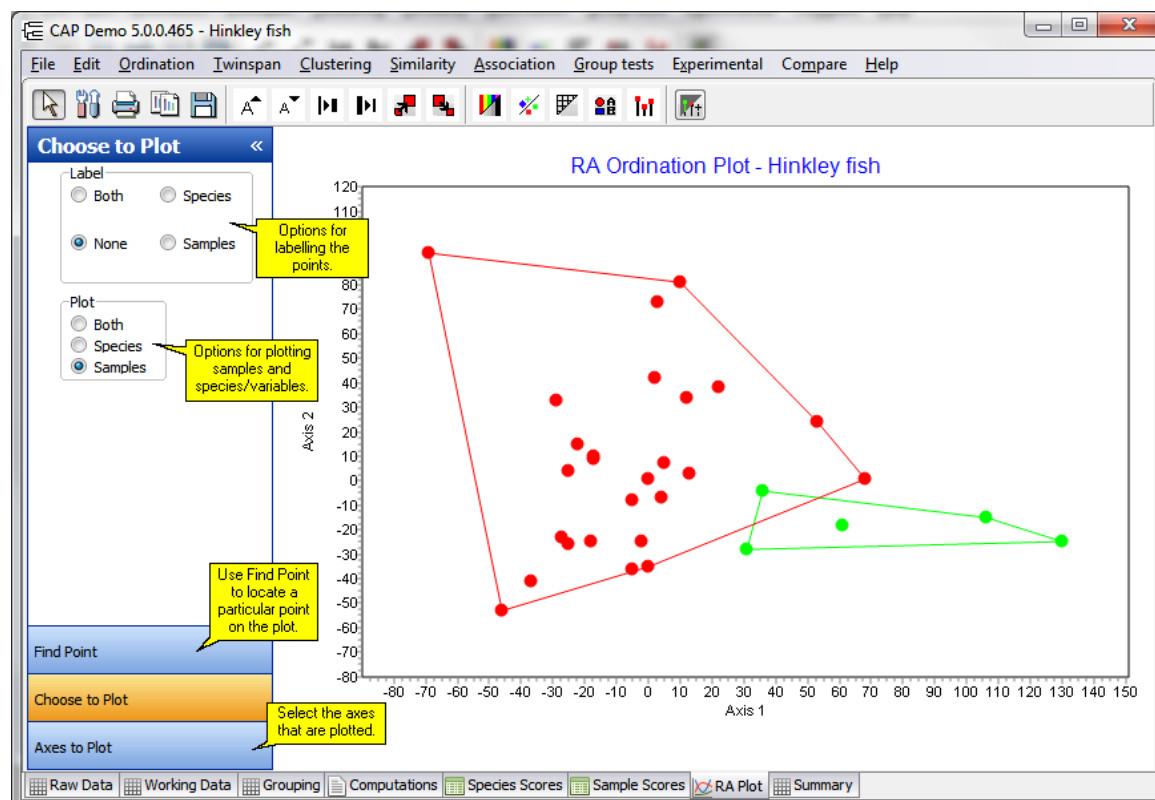
N	Name	Axis 1	Axis 2	Axis 3	Axis 4
	Eigenvalues =		0.5545	0.0784337	0.0653890
1	1981	18	24	-24	
2	1982	25	65	-17	
3	1983	106	-15	33	-25
4	1984	31	-28	13	-35
5	1985	36	-4	-17	-25
6	1987	-18	-25	-17	-7
7	1988	4	-7	-8	-12
8	1989	13	3	-26	22
9	1990	68	1	-25	71
10	1991	-2	-25	-6	5
11	1992	-37	-41	3	0
12	1993	0	-35	11	-7
13	1994	-5	-8	-24	-8
14	1995	5	7	-13	-7
15	1996	-25	-26	-4	-3
16	1997	-5	-36	9	16
17	1998	-46	-53	11	1
18	1999	0	1	-21	-4

Change the sort order of the table by clicking on the title cell of any of the columns.

#### 8.4.4 RA plot

Select this tab to display the ordination plots of the data. This window shows plots of both the sample and species scores. The display options are selected from the panel to the left of the graph. This panel can be shrunk to the left to make the plot as large as possible. Above the plot there is a toolbar of commonly-used graphics editing tools.

If you have assigned your sites/samples to groups, you can also [draw a perimeter](#)  round a group, as shown in the plot below.



The **Label** radio box is used to select labeling for the points. Select **Both** to display on the plot the names of the samples and species. Select **None** if no labels are required. Select **Species** (variables) or **Samples** to display variable or sample names respectively.

The **Plot** radio box allows sample, species or both ordinations to be plotted.

### Axes to Plot

The axes displayed in the plot are selected using the Plot x Axis, y Axis, z Axis drop-down boxes. The default is Axis 1 and Axis 2, which will display the relative positions of the samples with respect to the two largest components. A 3D plot is produced if a z variable is selected.

# Part

---



IX

## 9 TWINSpan

TWINSpan is somewhat complex divisive clustering method originally devised by [Hill](#)<sup>[172]</sup> for vegetation analysis, but quite suitable for animal communities as well. An interesting feature of TWINSpan is that it forms what are termed pseudospecies. These are separate variables for the different levels of abundance of a species. Samples are ordinated using Reciprocal Averaging (RA). A dichotomy is then made using the RA centroid line to divide the samples into two groups (negative and positive). This dichotomy is then refined using an iterative procedure. The clusters of samples obtained are then ordered so that similar clusters are near each other. This procedure continues in a hierarchical fashion to subdivide the groups until the minimum group size initially selected by the user is obtained. Species are then classified using the sample (quadrat) classification. In the original output a table is then produced showing species-by-site (quadrat or sample) relationships.

Once the TWINSpan option is selected from the drop-down menu you will be presented with the [Setup for TWINSpan](#)<sup>[88]</sup> window in which options can be selected. If no changes are made prior to clicking OK, then the default settings will be used. For many vegetational studies the defaults are appropriate.

[Setup Window - TWINSpan](#)<sup>[88]</sup>

[TWINSpan Text](#)<sup>[89]</sup>

[Site summary](#)<sup>[90]</sup>

[Species summary](#)<sup>[91]</sup>

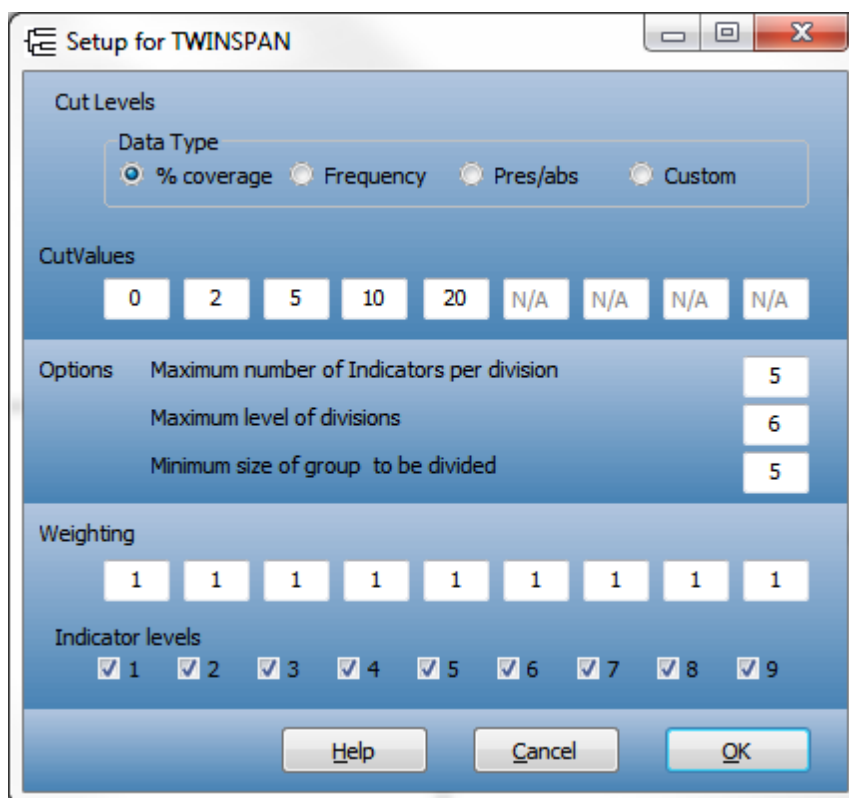
[Dendrogram samples](#)<sup>[92]</sup>

[Dendrogram species](#)<sup>[93]</sup>

See also [Maximum size of the data set](#)<sup>[94]</sup>.

### 9.1 Setup window - TWINSpan

The setup window is divided into four panels, which will be described in turn below.



**Cut levels.** This determines the abundance levels in the working data at which a species is divided into pseudospecies. A maximum of 9 cut levels can be defined. The default state is for 5 cut levels set at 0, 2, 5, 10 and 20. The radio box is set to % coverage as these are suitable values if the original data comprises percentage vegetation cover in quadrats. Select frequency if the original data are expressed in terms of relative frequency (maximum value 1). Select Pres/Abs if the working data are simply presence - absence (comprises 1s and 0s). To select your own cut levels select Custom and type in values in the line of text boxes arranged along the panel. The cut levels must be arranged from smallest to largest moving from left to right. The program ignores blank boxes.

**Options.** Three options are available. Maximum number of indicators per division defines the maximum number of indicator species that can be found per division. Maximum level of divisions defines the number of subdivisions of the groups that can be undertaken. The maximum is 9. Minimum group size per division gives the minimum number of samples (quadrats) in each group. Once a group size has reached this minimum level, no further subdivisions are undertaken.

**Weighting.** Each box in this panel gives the relative weighting given to each pseudospecies cut level. The default is all 1s, indicating that all pseudospecies are given equal weighting. If you wish to adjust the weighting given to a particular pseudospecies, for instance to give it double the weighting, simply enter '2' into the relevant box, leaving the other boxes at '1'.

**Indicator levels.** Defines which pseudospecies cut levels can act as indicators. The default is all; to deselect a level, uncheck a box.

Output from TWINSpan is presented on a series of tabbed pages. These are described in turn:

[TWINSpan Text](#)<sup>[89]</sup>

[Site Summary](#)<sup>[90]</sup>

[Species \(or Variables\) Summary](#)<sup>[91]</sup>

[Dendrogram Species](#)<sup>[93]</sup>

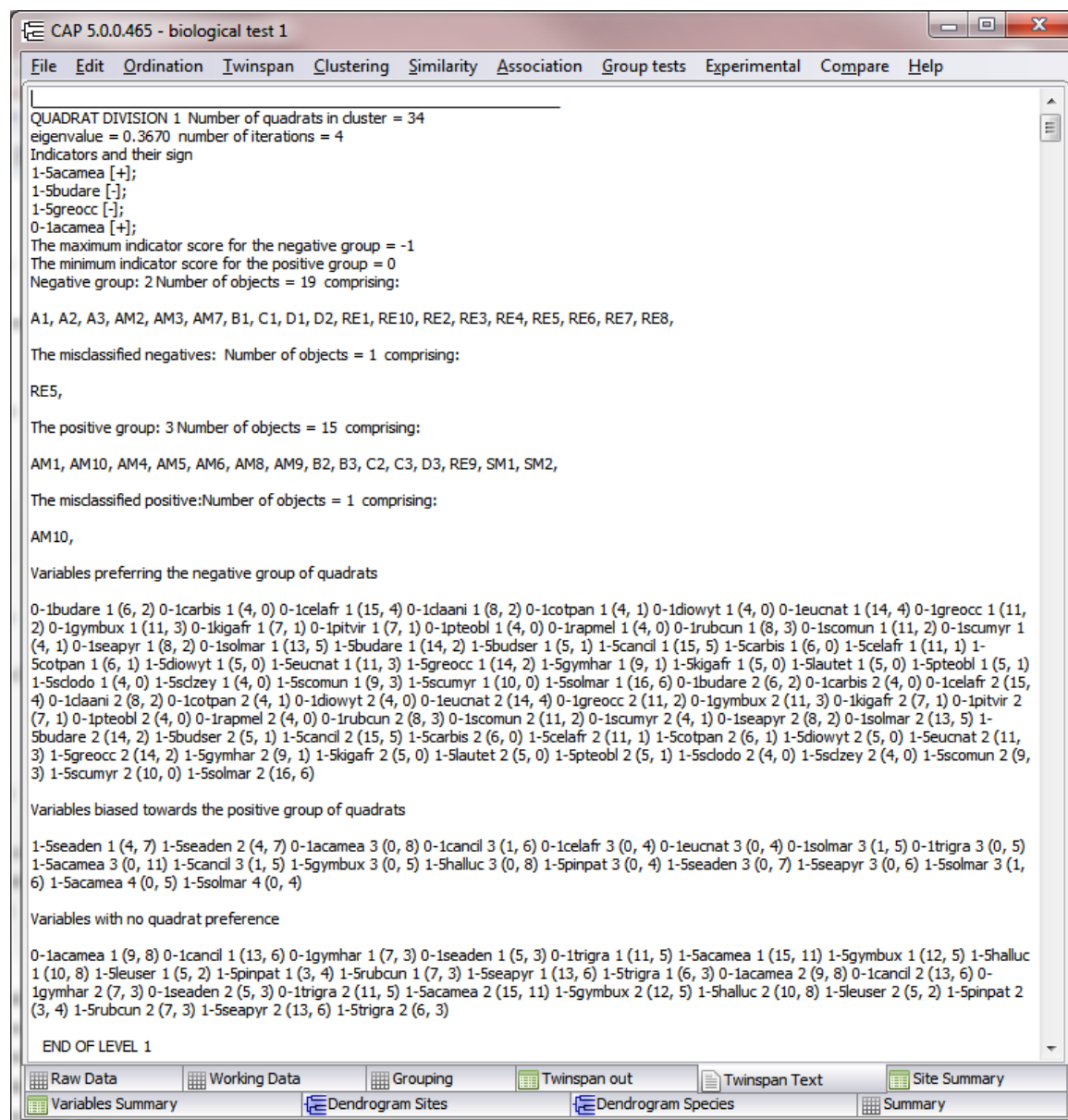
[Dendrogram Samples](#)<sup>[92]</sup>

[TWINSpan Out](#)<sup>[96]</sup>

## 9.2 TWINSpan Text

This window gives output similar in form to that produced by the original Fortran TWINSpan program for mainframe computers and DOS PCs. It first gives information on the clustering of the sites (quadrats), followed by the species, then finishes with the Classification Table. For large data sets it can be almost uninterpretable.

Press Ctrl-Alt-C, or Edit: Copy All, to copy the entire text of this output. To copy a selected portion, select the text you require, and press Ctrl-C on your keyboard, or Edit: Copy.



### 9.3 Site summary

This table gives a summary of the subdivisions undertaken on the sites (quadrats). In addition to the eigenvalue obtained from the RA, it gives the number of the sites classified into the negative, borderline and misclassified groups. See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

### Results - Site Summary

Division	Eigenvalue	No. Iterations	Group Size	No. -ve	No. Borderline -ve	No. Misclassified -ve
1	0.0722944	2	31	17	1	1
2	0.0638494	8	17	3		
3	0.0872626	3	14	11	1	
4	No Division		3			
5	0.0740461	4	14	8		
6	0.0846421	4	11	8		
7	No Division		3			
10	0.0930660	5	8	3		
11	0.1151659	2	6	4		
12	0.0910793	3	8	4		
13	No Division		3			
20	No Division		3			
21	0.1156901	1	5	2		
22	No Division		4			
23	No Division		2			
24	No Division		4			
25	No Division		4			
42	No Division		2			
43	No Division		3			

## 9.4 Variables summary

This table gives a summary of the subdivisions undertaken on the species or variables. In addition to the eigenvalue obtained from the RA, it gives the number of species (variables) classified into the positive, negative, borderline and misclassified groups. See [Printing and exporting text](#) <sup>[168]</sup> to save or print this table.

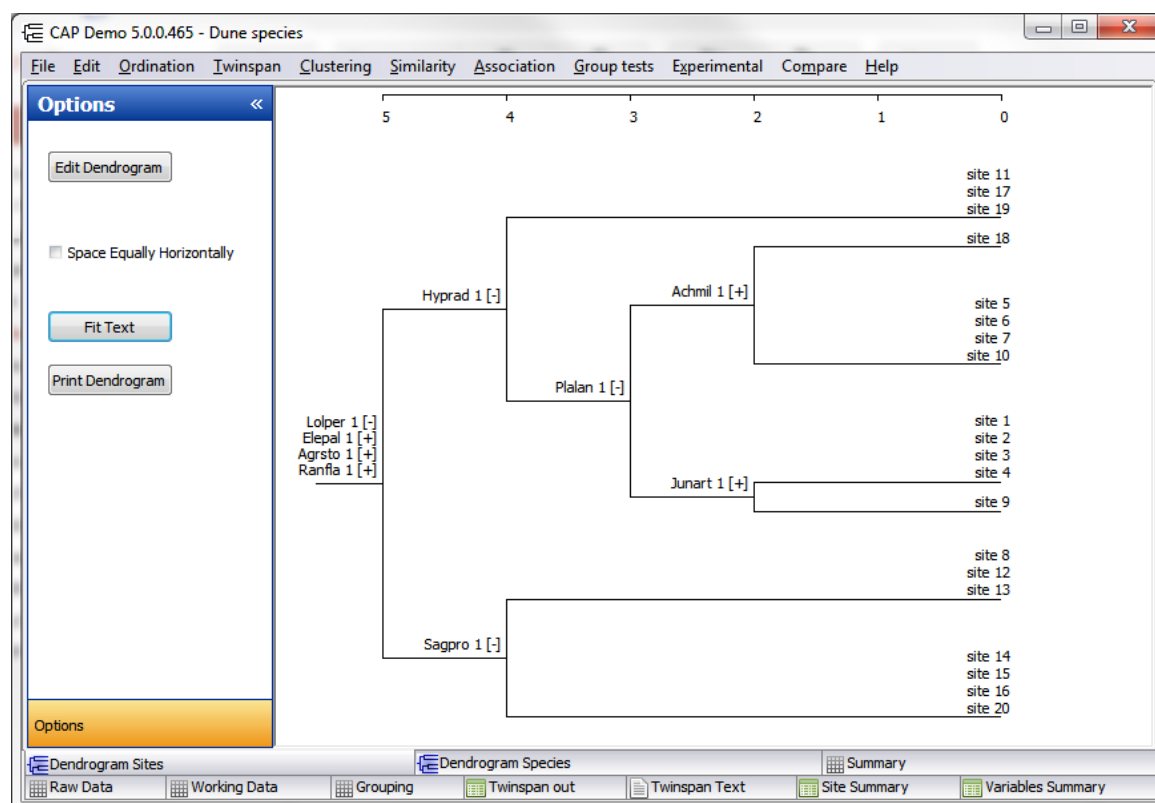


## Results - Variables Summary

Division	Eigenvalue	No. Iterations	Group Size	No. -ve	No. Borderline -ve	No. Misclassified -ve	No. +ve	No. Borderline +ve	No. Misclassified +ve
1	0.4927841	3	80	53			27		
2	0.4783570	2	53	21			32		
3	0.4588687	2	27	9			18		
4	0.2457848	4	21	11			10		
5	0.0530170	2	32	6			26		
6	0.2583189	4	9	8			1		
7	0.2423337	3	18	5			13		
8	0.1863913	3	11	7			4		
9	0.1799126	3	10	5			5		
10	0.1590113	2	6	4			2		
11	0.0340656	3	26	24			2		
12	0.1788557	3	8	7			1		
13	No Division		1						
14	0.2314341	1	5	2			3		
15	0.1426488	3	13	3			10		
16	0.1762352	2	7	6			1		
17	No Division		4						
18	0.1400564	1	5	2			3		
19	0.0819803	1	5	4			1		
20	No Division		4						

## 9.5 Dendrogram sites

This page displays the dendrogram of the TWINSpan analysis for the sites or samples; here we have used the *Dune species* demo data set. At each division the indicator species are shown. The look of the dendrogram can be edited with [Editing TWINSpan Dendrograms](#)<sup>[94]</sup>.



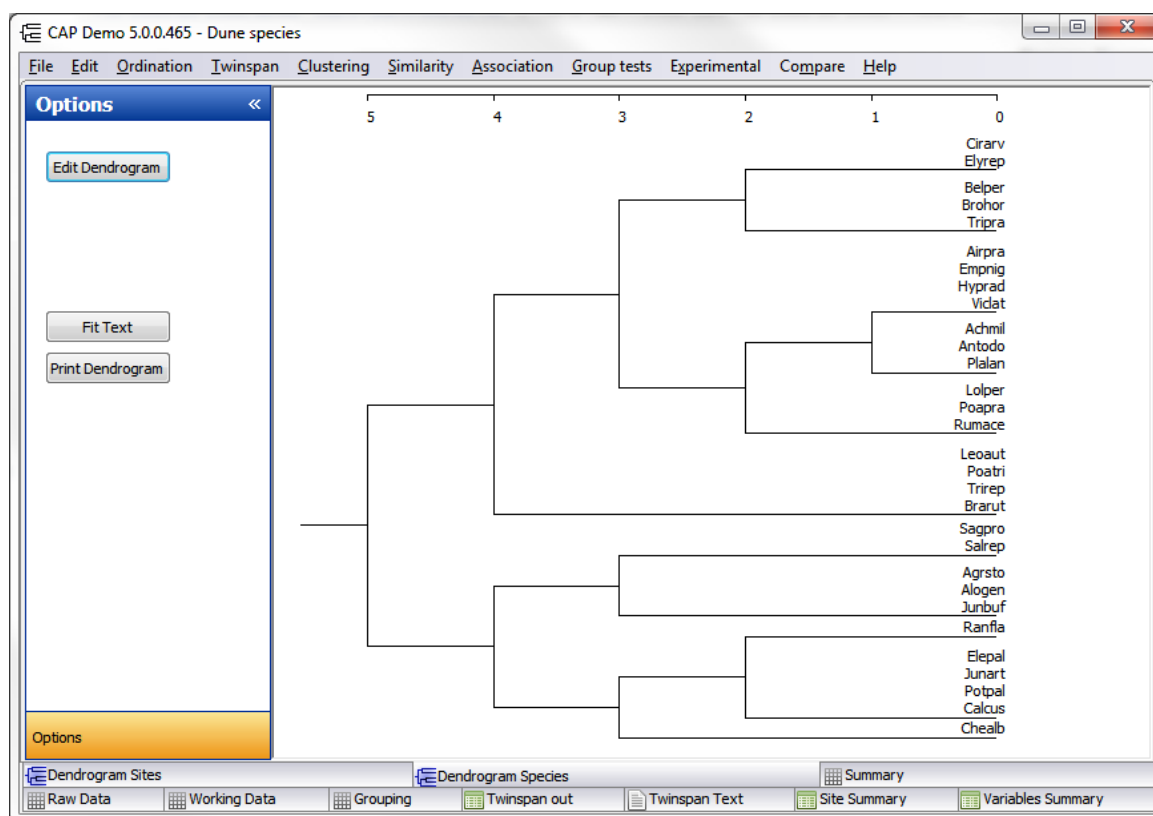
**Sizing the dendrogram to either fit the screen or make the sample titles readable.** Use the **Fit Text / Fit Screen** button to toggle between the two options. When Fit Screen is active the whole dendrogram is displayed in the window; for larger dendrograms the labels will overlap. To see the labels clearly, select Fit Text and use the scrollbar to move down the dendrogram.

To edit all other aspects of the dendrogram, use [Editing TWINSpan Dendrograms](#)<sup>[94]</sup> in the Options panel on the left-hand pane.

**Identifying the divisions.** To obtain details about the divisions produced, note the indicator species for the division in question on the dendrogram and then click on the [Twinspan Text](#)<sup>[89]</sup> tab. Scrolling down this output, each division is identified by its respective indicator species. Information about the division is then given in full.

## 9.6 Dendrogram species

This page displays the dendrogram of the TWINSpan analysis for the species or other variables; here we have used the *Dune species* demo data set. The species grouped together within each right-hand node are displayed at the right of the plot. The look of the dendrogram can be edited with [Editing TWINSpan Dendrograms](#)<sup>[94]</sup>.

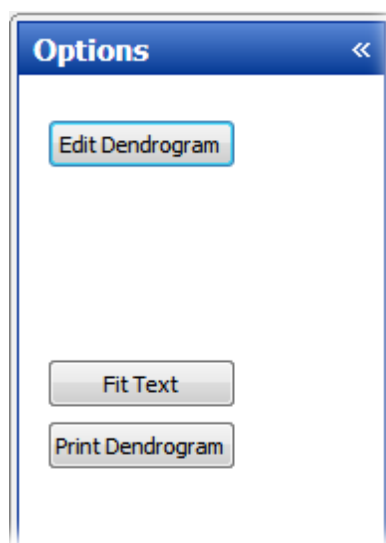


**Sizing the dendrogram to either fit the screen or make the species titles readable.** Use the **Fit Text / Fit Screen** button to toggle between the two options. When Fit Screen is active the whole dendrogram is displayed in the window; for larger dendrograms the labels will overlap. To see the labels clearly, select Fit Text and use the scrollbar to move down the dendrogram.

To edit all other aspects of the dendrogram, use [Editing TWINSpan Dendrograms](#)<sup>[94]</sup> in the Options panel on the left-hand pane.

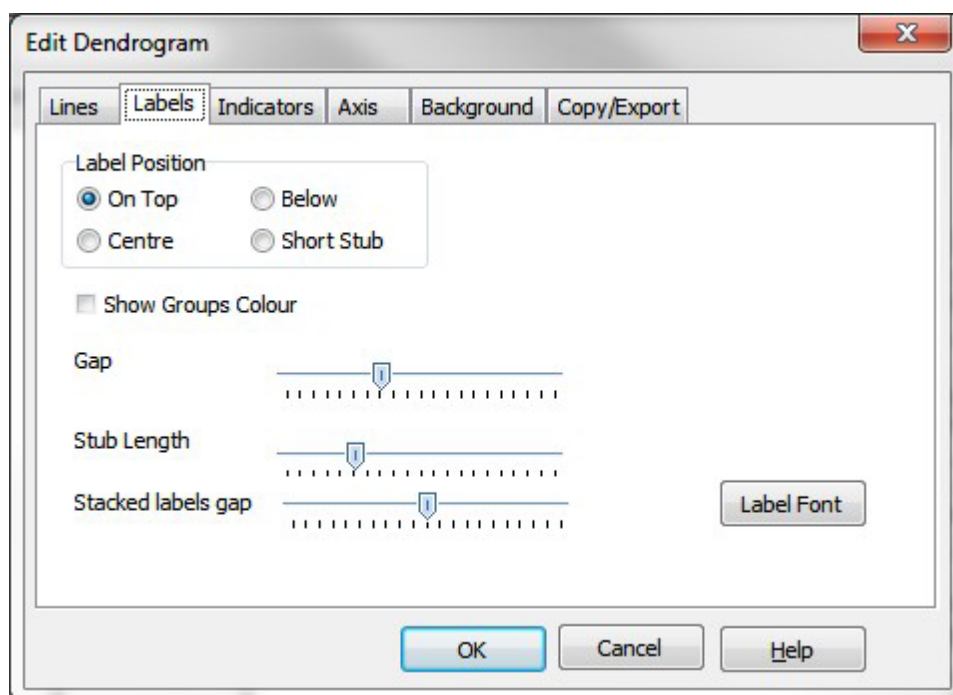
## 9.7 Editing TWINSpan Dendrograms

The Options pane of a TWINSpan dendrogram is shown below:



**Sizing the dendrogram to either fit the screen or make the species titles readable.** Use the **Fit Text / Fit Screen** button to toggle between the two options. When Fit Screen is active, the whole dendrogram is displayed in the window; for larger dendrograms the labels will overlap. To see the labels clearly, select Fit Text and use the scrollbar to move down the dendrogram.

You can change all other aspects of your dendrogram, and also the information displayed, using the **Edit Dendrogram** dialog:



The **Lines** tab offers options to change many aspects of the branches of the dendrogram.

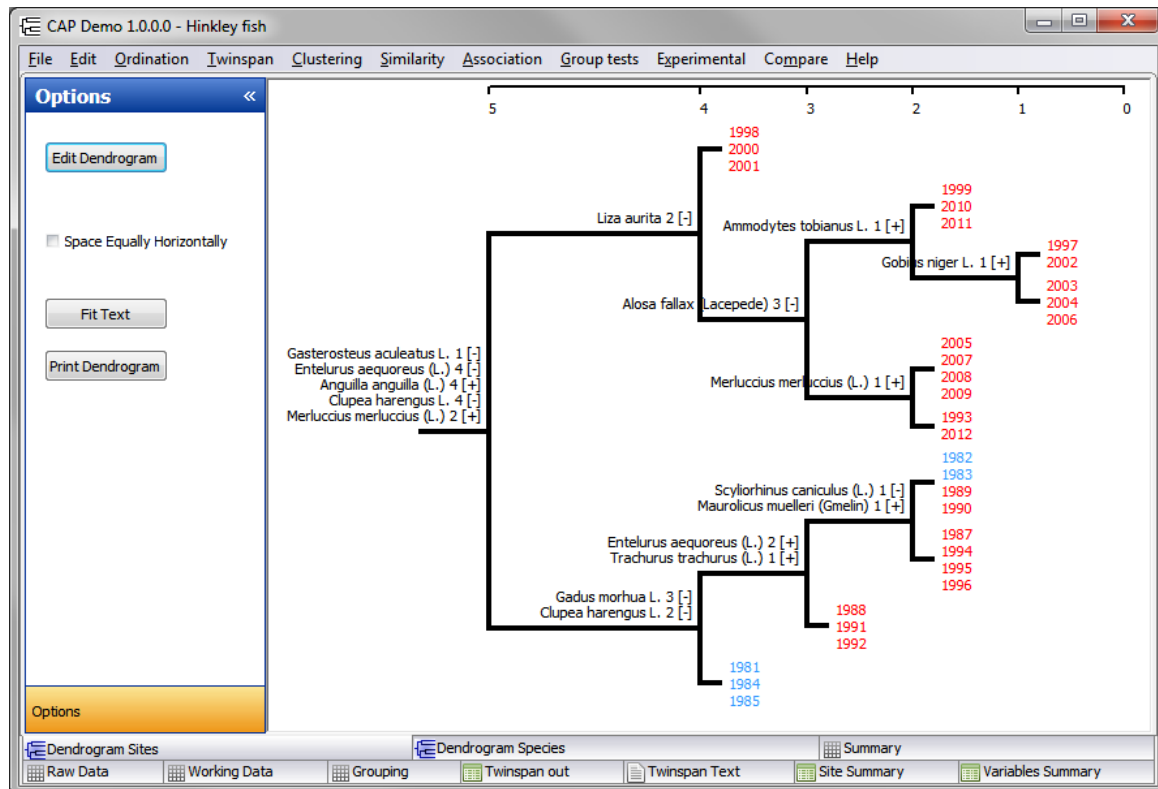
The **Labels** tab allows the font and position of the labels to be edited, and to show the colours assigned to groups in your data set. Shown below is an example dendrogram produced if you select Short stub.

The **Indicators** tab allows you to edit the font used for the indicator species labels, and rotate the label text.

The **Axis** tab gives options for the form of the axis.

The **Background** tab allows the background colour and the width of the left and right margins to be altered.

The **Copy/Export** tab offers output options, to copy the dendrogram as an image to paste into another program, or to save it.



## 9.8 TWINSpan Out

This grid shows the classification table used to produced the dendrograms. See [Printing and exporting text](#) to save or print this table.

CAP Demo 5.0.0.465 - Hinkley fish

File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare Help

**Results - Plot Table**

	17 - 1998	19 - 2000	20 - 2001	18 - 1999	29 - 2010	30 - 2011	16 - 1997	21 - 2002	22 - 2003	23 - 2004
8 - <i>Belone bellone</i> (L.)	1	0	0	0	0	0	0	0	0	0
10 - <i>Buglossidium luteum</i> (Risso)	0	2	0	0	0	0	0	0	0	0
17 - <i>Crenilabrus melops</i> (L.)	1	0	0	0	0	0	0	0	0	0
38 - <i>Liza aurita</i>	9	11	3	1	0	0	0	0	0	0
79 - <i>Ammodytes marinus</i>	0	1	0	0	0	0	0	0	0	0
30 - <i>Gobius paganellus</i>	0	2	0	0	3	0	0	0	0	0
18 - <i>Crenimugil labrosus</i> (Risso)	4	0	0	1	0	0	0	0	1	0
44 - <i>Micromesistius poutassou</i>	0	0	1	0	5	0	0	0	53	2
48 - <i>Nerophis lumbriciformis</i>	0	0	0	0	0	0	1	0	0	0
49 - <i>Petromyzon marinus</i>	0	0	0	1	0	0	0	0	0	0
71 - <i>Trachinus vipera</i> Cuvier	0	0	0	0	0	0	0	0	0	0
41 - <i>Mauroliscus muelleri</i> (Gmelin)	1	0	1	1	0	1	10	4	1	2
45 - <i>Microstomus kitt</i>	0	0	0	1	0	0	0	0	0	0
32 - <i>Hyperoplus lanceolatus</i> (Lesauvage)	0	1	0	1	0	0	0	1	0	0
54 - <i>Pomatoschistus microps</i> (Kroyer)	2	2	0	2	3	22	0	0	5	0
64 - <i>Scyliorhinus caniculus</i> (L.)	0	1	3	3	6	6	1	2	2	3
6 - <i>Atherina boyeri</i> Risso	2	2	5	0	0	3	3	0	1	1
28 - <i>Gasterosteus aculeatus</i> L.	2	2	12	1	2	1	2	1	1	2
56 - <i>Psetta maxima</i> (L.)	6	0	5	3	1	2	0	1	0	0
69 - <i>Syngnathus acus</i> (L.)	9	6	4	2	0	0	1	5	0	0
31 - <i>Gobius niger</i> L.	2	0	5	0	0	0	0	0	2	3
3 - <i>Ammodytes tobianus</i> L.	2	0	3	0	0	0	2	1	2	1
15 - <i>Clupea harengus</i> L.	10	163	125	3	1091	43	18	4	70	254
24 - <i>Entelurus aequoreus</i> (L.)	5	3	1	12	21	4	0	1	1	13
35 - <i>Lampetra fluviatilis</i> (L.)	2	0	0	1	1	0	1	0	0	0
25 - <i>Eutrigla gurnardus</i> (L.)	4	12	15	23	0	2	1	8	2	5
27 - <i>Gaidropsaurus vulgaris</i> (Cloquet)	0	2	0	0	0	4	0	0	0	0
26 - <i>Gadus morhua</i> L.	14	214	31	91	111	13	28	22	12	9
1 - <i>Agonus cataphractus</i> (L.)	11	5	5	4	11	16	9	10	5	2
5 - <i>Aphia minuta</i> (Risso)	2	20	2	18	6	11	6	14	31	2

Raw Data Working Data Grouping Twinspan out Twinspan Text Site Summary  
Variables Summary Dendrogram Sites Dendrogram Species Summary

# Part

---



## 10 Clustering

When numbers of sites or habitats are to be compared, the [similarity measures](#)<sup>[110]</sup> offered by CAP can form the basis of cluster analysis, which seeks to identify groups of sites, or stations that are similar in their species composition.

Classification methods comprise two principal types, hierarchical, where objects are assigned to groups that are themselves arranged into groups, as in a dendrogram, and non-hierarchical, where the objects are simply assigned to groups. The methods are further classified as either [agglomerative](#)<sup>[99]</sup>, where the analysis proceeds from the objects by sequentially uniting them, or [divisive](#)<sup>[106]</sup>, where all the objects start as members of a single group which is repeatedly divided. For computational and presentational reasons hierarchical-agglomerative methods are the most popular.

CAP includes the following methods of cluster analysis:

Agglomerative clustering

[Ward's](#)<sup>[100]</sup>

[Single linkage](#)<sup>[100]</sup>

[Complete linkage](#)<sup>[100]</sup>

[Average linkage](#)<sup>[100]</sup>

[McQuitty's](#)<sup>[101]</sup>

[Gower's](#)<sup>[101]</sup>

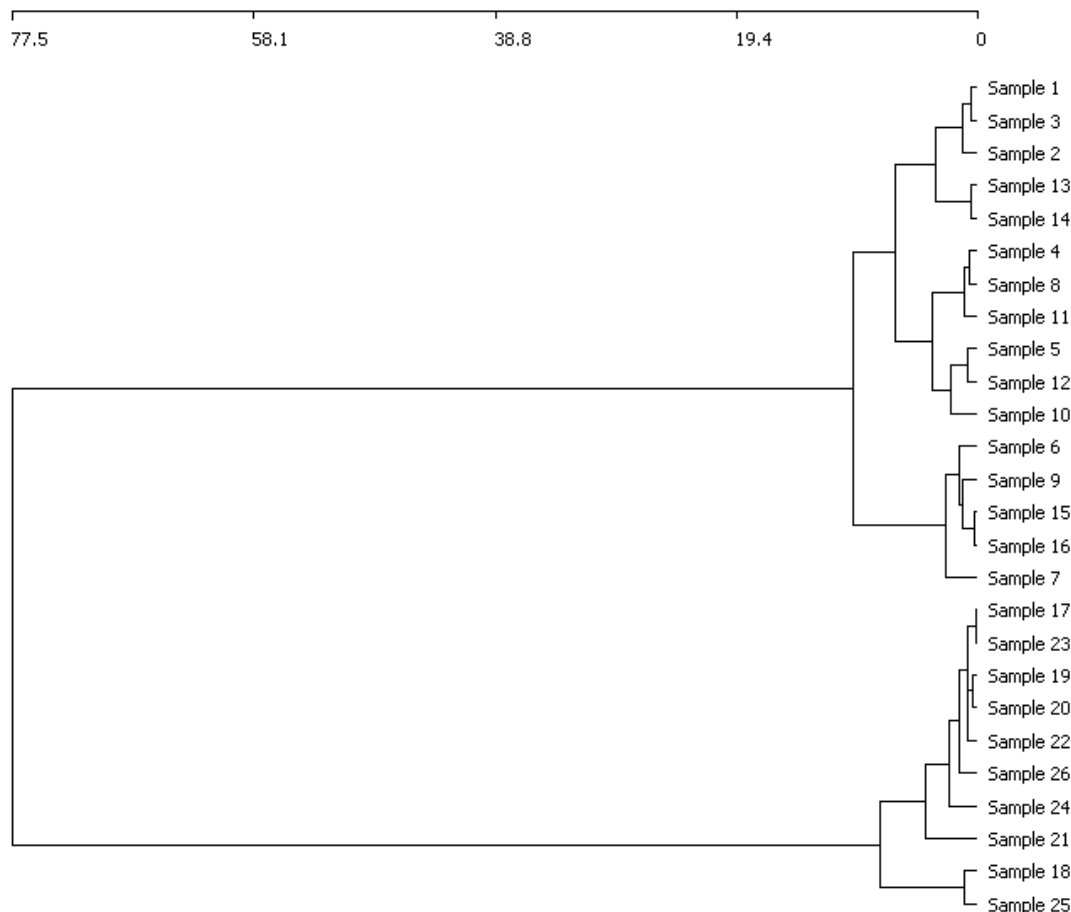
[Centroid](#)<sup>[101]</sup>

[Divisive clustering](#)<sup>[106]</sup>

The basic computational scheme used in cluster analysis can be illustrated using single linkage cluster analysis as an example. This is the simplest procedure and consists of the following steps.

1. Start with n groups each containing a single object (sites or variables).
2. Calculate, using the similarity measure of choice, the array of between-object similarities.
3. Find the two objects with the greatest similarity, and group them into a single object.
4. Assign similarities between this group and each of the other objects using the rule that the new similarity will be the greater of the two similarities prior to the join.
5. Continue steps 3 and 4 until only one object is left.

The results from a cluster analysis are usually presented in the form of a dendrogram:



The problem with all classification methods is that there can be no objective criteria of the best classification; indeed even randomly-generated data can produce a pleasing and convincing dendrogram. Always consider carefully whether the groupings identified seem to make sense and reflect some feature of the natural world.

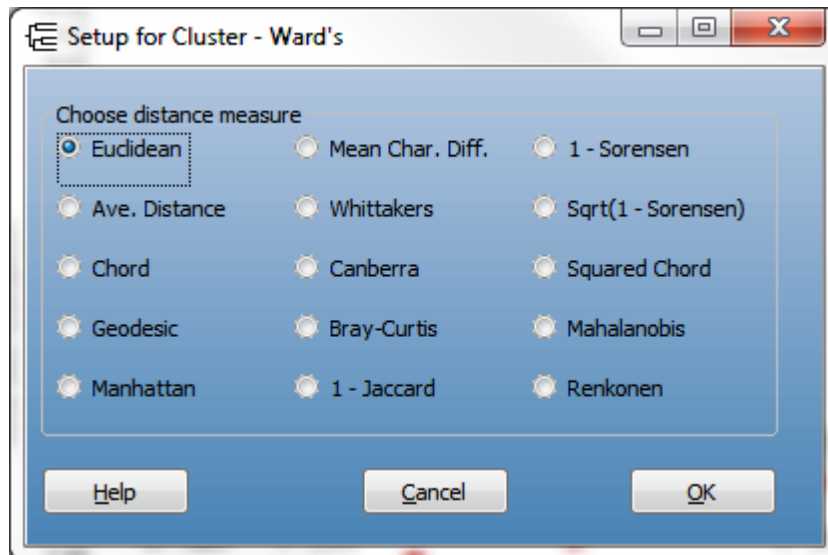
## 10.1 Agglomerative cluster analysis

Hierarchical agglomerative cluster analysis is selected by choosing **Clustering: Agglomerative** from the drop-down menu. The following methods for linking groups are available:

[Ward's](#)<sup>[100]</sup>  
[Single linkage](#)<sup>[100]</sup>  
[Complete linkage](#)<sup>[100]</sup>  
[Average linkage](#)<sup>[100]</sup>  
[McQuitty's](#)<sup>[101]</sup>  
[Gower's](#)<sup>[101]</sup>  
[Centroid](#)<sup>[101]</sup>

Whichever method you select, in the Setup for Cluster box, CAP then offers a choice of 15 distance measures: [Euclidean](#)<sup>[119]</sup>, [Geodesic](#)<sup>[120]</sup>, [Whittakers](#)<sup>[121]</sup>, [Ave. Distance](#)<sup>[119]</sup>, [Manhattan](#)<sup>[120]</sup>, [Canberra](#)<sup>[121]</sup>, [Chord](#)<sup>[119]</sup>, [Mean Character Difference \(Czekanowski\)](#)<sup>[119]</sup>, [Bray-Curtis](#)<sup>[121]</sup>, [Squared Chord](#)<sup>[121]</sup>, [Mahalanobis](#)<sup>[122]</sup>, 1-[Jaccard](#)<sup>[114]</sup>, 1-[Sørensen](#)<sup>[114]</sup> and Square root 1-[Sørensen](#)<sup>[114]</sup> and [Renkonen](#)<sup>[122]</sup> distance. Each of these distance measures is different, and will influence the outcome of cluster analysis. If you have presence-absence data, it is worthwhile choosing a distance measure specifically designed for it, such as Jaccard's or Sorensen's.





Once a measure has been selected the cluster analysis will immediately be run. Output is presented on a series of tabbed pages. These are described in turn below.

[Cluster summary](#)<sup>[105]</sup>

[Cluster groups](#)<sup>[101]</sup>

[Dendrogram - CLUSTER ANALYSIS](#)<sup>[102]</sup>

### 10.1.1 Ward's

Also termed minimum variance or error sums of squares clustering. At each iteration, all possible pairs of groups are compared and the two groups chosen for fusion are those which will produce a group with the lowest variance.

### 10.1.2 Single linkage

Also termed minimum or nearest neighbour method. At each iteration, the clusters are compared in terms of the similarity of their most similar samples (columns) and the two clusters that hold the most similar samples are fused.

### 10.1.3 Complete linkage

This is also called furthest neighbour sorting. At each step, the clusters are compared in terms of the similarity of their least similar members and the two clusters that are most similar are fused.

### 10.1.4 Average linkage

Also known as group-average sorting ([Lance & Williams, 1966](#)<sup>[172]</sup>). At each step, the clusters are compared in terms of the average similarity of their members and the two clusters that are most similar are fused.

### 10.1.5 McQuitty's

Let  $D_{kl}$  = any distance or dissimilarity measure between clusters  $C_k$  and  $C_l$ . If clusters  $C_k$  and  $C_l$  are merged to form  $C_m$ , the formula giving the distance between the new cluster  $C_m$  and any other cluster  $C_j$  is:

$$D_{jm} = (D_{jk} + D_{jl})/2 .$$

The method was independently developed by Sokal and Michener (1958) and McQuitty (1966).

### 10.1.6 Gower's

At each step, the centroids of the clusters are compared, and the two clusters that are most similar are fused. The position of the centroid is calculated using the formula of Gower (1967).

### 10.1.7 Centroid

At each iteration, the clusters are compared in terms of the similarity of their most similar samples (columns) and the two clusters that hold the most similar samples are fused. The average of the attributes of the fused group is calculated and the similarity between average properties are used in subsequent iterations.

### 10.1.8 Cluster groups

This grid gives the membership of each cluster arranged by rows. The first column gives the cluster formation number - this is the cycle during which the cluster was formed. The second column, termed the group label, gives an identifier for the cluster based on the membership. The group label is the lowest number of the column numbers of the samples included within the group. Thus if samples in columns 23 and 45 are joined to form a cluster the group label will be 23. If at some later point these samples are joined with the sample in column 1 then the group label for this new cluster will be 1. From column 3 onwards the members of each cluster are given.

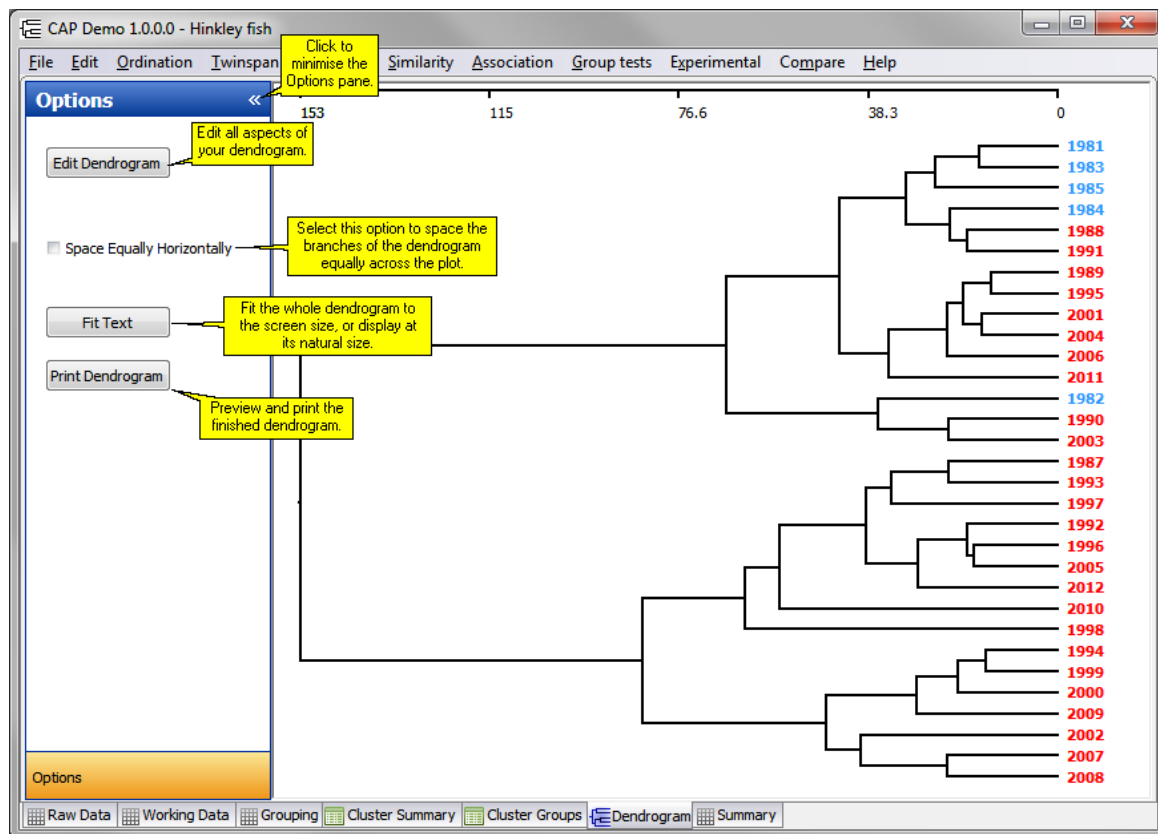
To print this output use **File: Print**.

Results - Groups										
Cluster	Group Label	Sites								
1	7	1988	2001							
2	8	1989	2004							
3	8	1989	2004	1995						
4	1	1981	1983							
5	15	1996	2005							
6	4	1984	1988	2001						
7	13	1994	1999							
8	9	1990	2003							
9	5	1985	2006							
10	4	1984	1988	2001	1991					
11	26	2007	2008							
12	1	1981	1983	1985	2006					
13	11	1992	1997							
14	19	2000	2009							
15	15	1996	2005	2012						
16	13	1994	1999	2000	2009					
17	4	1984	1988	2001	1991	2011				
18	11	1992	1997	1993						
19	21	2002	2007	2008						
20	1	1981	1983	1985	2006	1989	2004	1995		

The first cluster is formed by the samples from 1988 and 2001; it is assigned Group Label 7 because the first sample in the group is from column 7 of the original data set.

### 10.1.9 Dendrogram - Cluster analysis

The dendrogram of the cluster relationships is shown by clicking on the Dendrogram tab. Many aspects of the dendrogram can be edited by clicking the [Edit Dendrogram](#) button in the Options pane.

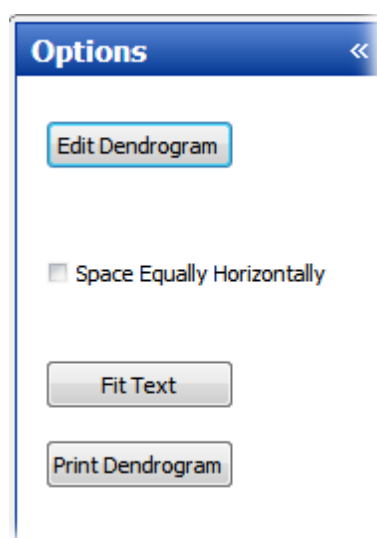


**Space Equally Horizontally.** Selecting this option will spread the branches of the dendrogram equally across the plot, otherwise the divisions are shown at the point they occur on the axis scale.

**Sizing the dendrogram to either fit the screen or make the species titles readable.** Use the **Fit Text / Fit Screen** button to toggle between the two options. When Fit Screen is active, the whole dendrogram is displayed in the window; for larger dendrograms the labels will overlap. To see the labels clearly, select Fit Text and use the scrollbar to move down the dendrogram.

#### 10.1.9.1 Edit Dendrogram

The Options pane of a cluster dendrogram is shown below:

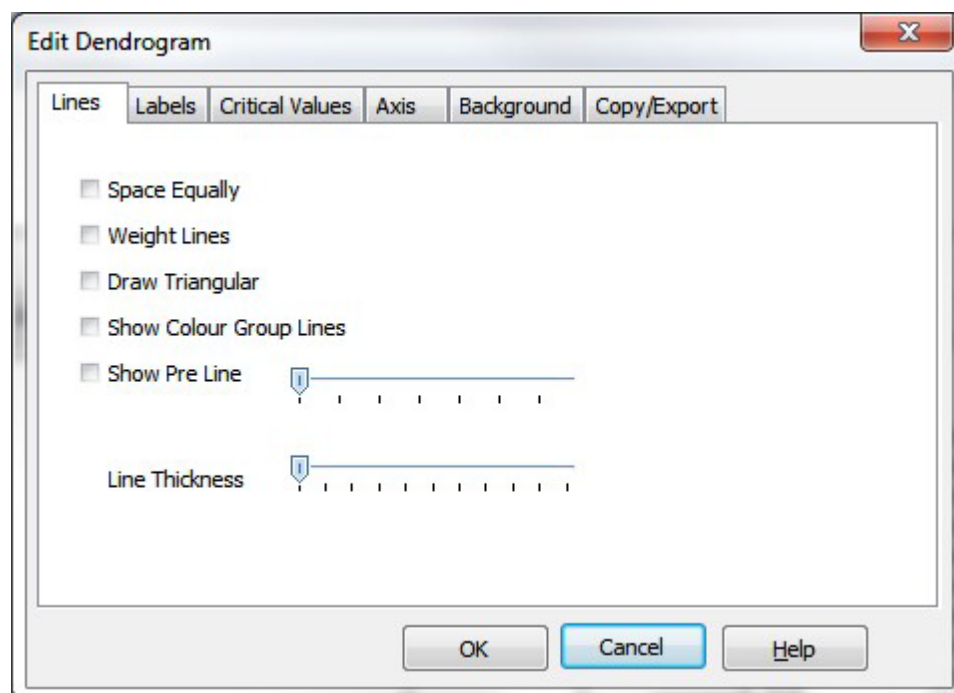


**Space Equally Horizontally.** Selecting this option will spread the branches of the dendrogram

equally across the plot, otherwise the divisions are shown at the point they occur on the axis scale.

**Sizing the dendrogram to either fit the screen or make the species titles readable.** Use the **Fit Text / Fit Screen** button to toggle between the two options. When Fit Screen is active, the whole dendrogram is displayed in the window; for larger dendrograms the labels will overlap. To see the labels clearly, select Fit Text and use the scrollbar to move down the dendrogram.

You can change all other aspects of your dendrogram, and also the information displayed, using the **Edit Dendrogram** dialog:



The **Lines** tab offers options to change many aspects of the branches of the dendrogram.

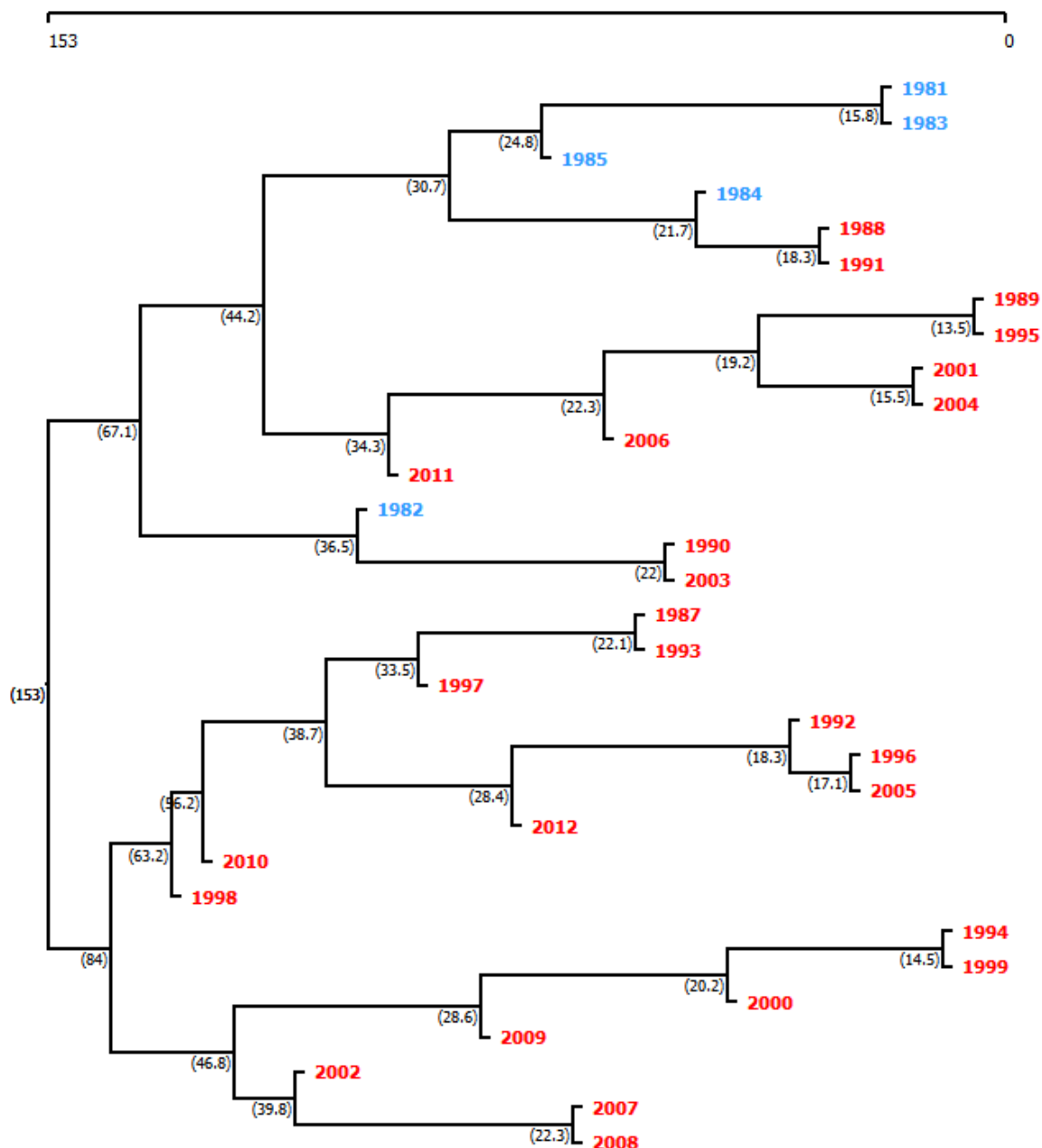
The **Labels** tab allows the font and position of the labels to be edited, and to show the colours assigned to groups in your data set. Shown below is an example dendrogram produced if you select Short stub.

The **Critical Values** tab allows you to put on the dendrogram the distance values which cause the samples to be combined. These are shown on the plot below.

The **Axis** tab gives options for the form of the axis.

The **Background** tab allows the background colour and the width of the left and right margins to be altered.

The **Copy/Export** tab offers output options, to copy the dendrogram as an image to paste into another program, or to save it.



### 10.1.10 Cluster summary

This window presents a table of the sequence of cluster formation. The first column, entitled Cluster, gives a sequential number to each cluster formed. The group 1 and group 2 columns identify the groups which are united to form the cluster. Initially every sample (column) is a separate group numbered sequentially. Thus the first row of the table gives the column numbers of the first two samples to be joined to form a cluster. The column headed Dissimilarity gives the dissimilarity measure between the two groups which are united. Group label gives the number by which this new group will be referenced further down the table. Group size gives the number of items (samples or columns) in the group.

Results - Clustering					
Cluster	Group 1	Group 2	Dissimilarity	Group Label	Group Size
1	15	20	6.7823300	15	2
2	5	18	7.3801000	5	2
3	11	18	7.3801000	11	2
4	2	10	7.3801000	2	2
5	12	13	7.3801000	12	2
6	17	19	8.3666000	17	2
7	15	16	8.4629831	15	3
8	5	6	8.4722757	5	3
9	3	8	8.6023254	3	2
10	3	9	9.4965467	3	3
11	3	4	10.3260145	3	4
12	14	15	10.7176962	14	4
13	1	11	11.7137060	1	3
14	3	12	13.3561411	3	6
15	2	5	13.6868610	2	5
16	1	17	14.3670368	1	5
17	1	2	21.0936699	1	10
18	3	14	28.8158321	3	10
19	1	3	36.7146873	1	20
20	0	0	0.0000000	0	

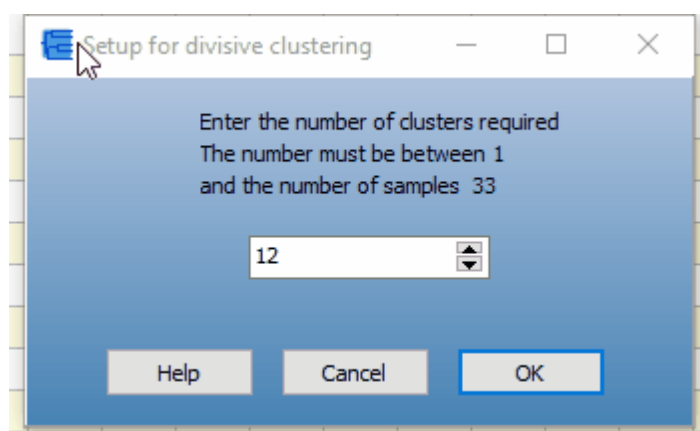
Dissimilarity between the 2 sites

Cluster 1 is formed from sites 15 and 20 - the two most similar in the data set.

This cluster is assigned Group Label 15 because the first site comes from column 15 of the original data.

## 10.2 Divisive cluster analysis

In divisive clustering methods, the samples (columns) all start as members of a single group which is subsequently subdivided. [TWINSPAN](#)<sup>[88]</sup> is an example of such an approach. The method available under **Clustering: Divisive** minimises the sums of squares of the dissimilarity of each cluster formed. To carry out the analysis, you must specify the number of clusters required in the Setup dialog window. The number of clusters must lie between 1 and the number of samples present in your data set; the setup dialog will inform you how many samples there are (see image below).



To gain an idea for a suitable number of clusters you might find it useful to examine a [PCA](#)<sup>[63]</sup> or [MDS](#)<sup>[75]</sup> plot. Although the method will always group the samples into the number of clusters you

request, the results will not necessarily be meaningful for every data set, and so the method is best used when some other analysis such as [TWINSPAN](#)<sup>[88]</sup> or PCA has suggested the presence of a number of groups, and you wish to see if an independent method can produce the same group membership.

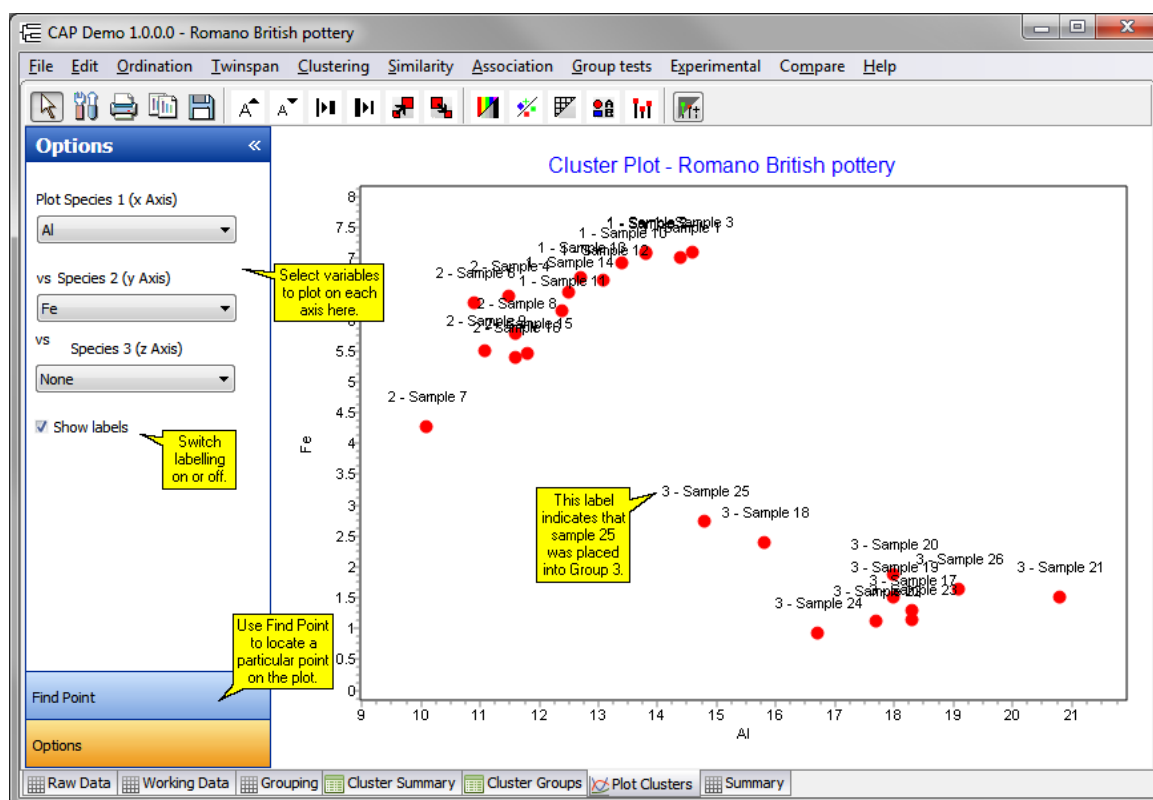
[Cluster Summary](#)<sup>[108]</sup>

[Cluster Groups](#)<sup>[108]</sup>

[Plot Clusters](#)<sup>[107]</sup>

## 10.2.1 Plot Clusters

The Plot Clusters tab presents a plot of the position of the samples (objects) in space, as defined by two variables. Beside each sample is a number that identifies the cluster into which the sample has been assigned. The particular variables used to define the space within which the samples are plotted are selected from drop-down menus in the Options pane.



See also

[Preparing charts for output](#)<sup>[162]</sup>

Printing charts

[Exporting charts](#)<sup>[153]</sup>

[Zooming on charts](#)<sup>[158]</sup>

[Themes for graphs](#)<sup>[167]</sup>



### 10.2.2 Cluster Groups

This grid itemises the cluster to which each sample has been assigned.

In this example the samples are labeled 1981, 1982 etc... Both 1981 and 1983 have been assigned to cluster 3.

Results Group	
Sample	Cluster Number
1981	3
1982	2
1983	3
1984	4
1985	5
1987	6
1988	4
1989	1
1990	9
1991	4
1992	10

To print this output use **File: Print**.

### 10.2.3 Cluster Summary

This grid gives the Sums of Squares and the number of samples or objects allocated to each cluster.

The example below is from an analysis of the *Romano British pottery* data set, where 3 clusters were specified.

Results - Clustering		
Cluster	No of Members	Sums of Squares
1	3	133963
2	1	0
3	2	96395
4	5	719166
5	2	148815
6	3	1.408E006
7	1	0
8	6	2.4912E006
9	2	139100
10	4	936204
11	3	1.55488E006
12	1	0

To print this output use **File: Print**.

**Part**

---

**XI**

## 11 Similarity and Distance Measures

Choose the similarity measure you wish to calculate from the Similarity drop-down menu. The calculations will appear on the Similarity tab at the bottom of the program window. For ease of use, the program will highlight sites with similarity above a certain level. You can set this level by entering the number in the **Set threshold level** box at the bottom of the page.

### The similarity measures used.

These are simple measures of either the extent to which two habitats have species in common (Q analysis) or which variables (species) have habitats in common (R analysis). Binary similarity coefficients use presence-absence data; following the introduction of computers, more complex quantitative coefficients became practicable. Analysis of quantitative, rather than presence-absence, data with a binary method may report a perfect similarity between every sample/site in data sets (such as the *Romano British pottery* demo data set) in which each variable is present in every sample.

Both groups of indices can be further divided between those which take account of the absence from both communities (double zero methods) and those which do not. In most ecological applications it is unwise to use double-zero methods as they assign a high level of similarity to localities which both lack many species; a problem which becomes particularly acute in habitats which have a potentially extremely large species list, such as the marine benthos.

A good account of similarity and distance measures is given in [Legendre & Legendre \(1983\)](#)<sup>[172]</sup>. Because of division by zero problems for some data sets not all measures can be calculated. When a division by zero error would occur CAP gives an index of -99.

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		Sample 1	
		Species Present	Species Absent
Sample 2	Species Present	<b>a</b>	<b>b</b>
	Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

### Binary - double zeros

[Simple matching](#)<sup>[111]</sup>

[Rogers Tanimoto](#)<sup>[112]</sup>

[S3](#)<sup>[111]</sup>

[S4](#)<sup>[112]</sup>

[S5](#)<sup>[113]</sup>

[S6](#)<sup>[113]</sup>

### Binary - no double zeros

[Jaccards](#)<sup>[114]</sup>

[Sørensen](#)<sup>[114]</sup>

[S9](#)<sup>[115]</sup>

[S10](#)<sup>[115]</sup>

[Russel & Rao](#)<sup>[116]</sup>

[Kulczynski](#)<sup>[116]</sup>

[S13](#)<sup>[116]</sup>

[Ochiai](#)<sup>[117]</sup>

**Quantitative**[Q1](#)<sup>[117]</sup>[Q2](#)<sup>[118]</sup>[Steinhaus](#)<sup>[119]</sup>[Kulczynski-Quantitative](#)<sup>[118]</sup>**Distance measures**[Euclidean](#)<sup>[119]</sup>[Mahalanobis](#)<sup>[122]</sup>[Average](#)<sup>[119]</sup>[Chord](#)<sup>[119]</sup>[Geodesic](#)<sup>[120]</sup>[Manhattan](#)<sup>[120]</sup>[Mean character difference](#)<sup>[120]</sup>[Whittaker](#)<sup>[121]</sup>[Canberra](#)<sup>[121]</sup>[Bray-Curtis](#)<sup>[121]</sup>[Squared chord](#)<sup>[121]</sup>[Renkonen](#)<sup>[122]</sup>

## 11.1 Simple matching

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		<b>Sample 1</b>	
<b>Sample 2</b>	Species Present	Species Present	Species Absent
	Species Absent	<b>a</b>	<b>b</b>
		<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

Simple matching similarity is  $(a+d) / N$ .

Note that this measure gives equal weighting to double zeros and species which are present in both samples. This is rarely useful in ecological studies.

## 11.2 S3

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		<b>Sample 1</b>	
<b>Sample 2</b>	Species Present	Species Present	Species Absent
	Species Absent	<b>a</b>	<b>b</b>
		<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$\frac{(2a + 2d)}{(2a + b + c + 2d)}$$

### 11.3 Rogers\_Tanimoto

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		<b>Sample 1</b>	
<b>Sample 2</b>	Species Present	Species Present	Species Absent
	Species Absent	<b>a</b>	<b>b</b>
		<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$\frac{(a + d)}{(a + 2b + 2c + d)}$$

### 11.4 S4

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		<b>Sample 1</b>	
<b>Sample 2</b>	Species Present	Species Present	Species Absent
	Species Absent	<b>a</b>	<b>b</b>
		<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$\frac{(a+d)}{(b+c)}$$

## 11.5 S5

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

Sample 2	Sample 1	
	Species Present	Species Absent
Species Present	<b>a</b>	<b>b</b>
Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$\frac{1}{4} \left[ \left( \frac{a}{a+b} \right) + \left( \frac{a}{a+c} \right) + \left( \frac{d}{b+d} \right) + \left( \frac{d}{c+d} \right) \right]$$

## 11.6 S6

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

Sample 2	Sample 1	
	Species Present	Species Absent
Species Present	<b>a</b>	<b>b</b>
Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$\left( a / \sqrt{(a+b)(a+c)} \right) \left( d / \sqrt{(b+d)(c+d)} \right)$$

## 11.7 Jaccards

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		Sample 1	
		Species Present	Species Absent
Sample 2	Species Present	<b>a</b>	<b>b</b>
	Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$a / (a + b + c)$$

## 11.8 Sørensen

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		Sample 1	
		Species Present	Species Absent
Sample 2	Species Present	<b>a</b>	<b>b</b>
	Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$2a / (2a + b + c)$$

## 11.9 S9

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

<b>Sample 2</b>	<b>Sample 1</b>	
	Species Present	Species Absent
Species Present	<b>a</b>	<b>b</b>
Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$\frac{3a}{3a + b + c}$$

## 11.10 S10

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

<b>Sample 2</b>	<b>Sample 1</b>	
	Species Present	Species Absent
Species Present	<b>a</b>	<b>b</b>
Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$\frac{a}{a + 2b + 2c}$$



### 11.11 Russell & Rao

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		<b>Sample 1</b>	
		Species Present	Species Absent
<b>Sample 2</b>	Species Present	<b>a</b>	<b>b</b>
	Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore  $a+b+c+d$ .

This measure is calculated as:  $a/N$

### 11.12 Kulczynski

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		<b>Sample 1</b>	
		Species Present	Species Absent
<b>Sample 2</b>	Species Present	<b>a</b>	<b>b</b>
	Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore  $a+b+c+d$ .

This measure is calculated using:

$$\frac{a}{b+c}$$

### 11.13 S13

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		<b>Sample 1</b>	
		Species Present	Species Absent
<b>Sample 2</b>	Species Present	<b>a</b>	<b>b</b>
	Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample

1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$1/2 \left[ \left( \frac{a}{a+b} \right) + \left( \frac{a}{a+c} \right) \right]$$

## 11.14 Ochiai

For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

Sample 2	Sample 1	
	Species Present	Species Absent
	<b>a</b>	<b>b</b>
Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore a+b+c+d.

This measure is calculated using:

$$a / \sqrt{(a+b)(a+c)}$$

## 11.15 Q1

This is simply the agreement (number of variables or species which have the same abundance in the two samples) divided by the total number of variables or species.

Example: Ten species at two sites:

	Site 1	Site 2	Agreement
Sp. 1	1	3	0
Sp. 2	4	4	1
Sp. 3	6	9	0
Sp. 4	22	23	0
Sp. 5	7	7	1
Sp. 6	5	5	1
Sp. 7	5	2	0
Sp. 8	8	4	0
Sp. 9	9	9	1
Sp. 10	10	1	0

4

$Q1 = 4 \text{ agreements} / 10 \text{ species} = 0.4$

### 11.16 Q2

This is similar to [Q1](#)<sup>[117]</sup> but takes account of double zeros. It is calculated as agreement divided by the total number of species minus the number of double zeros.

Example: Ten species at two sites:

	Site 1	Site 2	Agreement
Sp. 1	1	3	0
Sp. 2	4	4	1
Sp. 3	6	9	0
Sp. 4	22	23	0
Sp. 5	7	7	1
Sp. 6	5	5	1
Sp. 7	0	0	1
Sp. 8	8	4	0
Sp. 9	2	2	1
Sp. 10	4	1	0
			5

$Q2 = 5 \text{ agreements} / (10 \text{ species} - 1 \text{ double zero}) = 0.555$

### 11.17 Kulczynski-Quantitative

Using the same nomenclature as for the [Steinhaus coefficient](#)<sup>[119]</sup>, this measure is given by:

$$\frac{1}{2} \left[ \left( \frac{W}{A} \right) + \left( \frac{W}{B} \right) \right]$$

## 11.18 Steinhaus

There is some confusion in the literature as to the correct name for this distance measure. It is attributed to Steinhaus by Motyka (See Legendre & Legendre, 1983) and is also called the Mean Character Difference (the name used by CAP).

For two samples the distance is given by:

$$\frac{2W}{(A+B)}$$

where  $w$  is the sum of the minimum abundances of the species in the two samples,  $A$  is the sum of species abundance in sample 1, and  $B$  is the sum of species abundance in sample 2.

## 11.19 Euclidean

This is the most commonly-used metric distance measure. If  $s_{i1}$  and  $s_{i2}$  are the abundances of species  $i$  in samples 1 and 2 respectively, then the Euclidean distance is:

$$\sqrt{\sum_{i=1}^{i=n} (s_{i1} - s_{i2})^2}$$

where  $n$  is the total number of species (variables).

## 11.20 Average

This is a variant on the [Euclidean distance](#)<sup>[119]</sup> given by:

$$\sqrt{\frac{1}{n} \sum_{i=1}^{i=n} (s_{i1} - s_{i2})^2}$$

Notation as for the [Euclidean measure](#)<sup>[119]</sup>.

## 11.21 Chord

This is a [Euclidean distance](#)<sup>[119]</sup> measure after normalisation of each sample vector to unit length. Thus, differences caused by the relative magnitudes of the number of individuals (counts) in each sample are removed from the comparison.

$$\sqrt{2 \left( 1 - \frac{\sum_{i=1}^{i=n} s_{i1} s_{i2}}{\sqrt{\sum_{i=1}^{i=n} s_{i1}^2} \sqrt{\sum_{i=1}^{i=n} s_{i2}^2}} \right)}$$

## 11.22 Geodesic

This is related to the [Chord distance](#)<sup>[119]</sup>, as follows:

$$D_4 = \arccos \left[ 1 - \frac{D_3^2}{2} \right]$$

where D3 is the Chord distance.

## 11.23 Manhattan

This measure is also known as the taxicab or city-block metric and is calculated as follows:

$$\sum_{i=1}^{i=n} |s_{i1} - s_{i2}|$$

where si1 and si2 are the abundances of species i in samples 1 and 2 respectively.

## 11.24 Mean Character Difference (Czekanowski)

This is the mean species difference:

$$\frac{1}{n} \sum_{i=1}^{i=n} |s_{i1} - s_{i2}|$$

where si1 and si2 are the abundances of species i in samples 1 and 2 respectively.

## 11.25 Whittaker

Whittaker's index of association is designed for species abundance data, as the abundance of each species in a sample is expressed as the fraction of the total number of individuals in the sample. The distance measure calculated by CAP is the complement of Whittaker's association index and is given by:

$$\frac{1}{2} \left| \frac{s_{i1}}{\sum_{i=1}^{i=n} s_{i1}} - \frac{s_{i2}}{\sum_{i=1}^{i=n} s_{i2}} \right|$$

where  $s_{i1}$  and  $s_{i2}$  are the abundances of species  $i$  in samples 1 and 2 respectively

## 11.26 Canberra

Lance & Williams (1967) introduced this variant of the Manhattan metric calculated as follows:

$$\sum_{i=1}^{i=n} \frac{|s_{i1} - s_{i2}|}{(s_{i1} + s_{i2})}$$

where  $s_{i1}$  and  $s_{i2}$  are the abundances of species  $i$  in samples 1 and 2 respectively.

## 11.27 Bray-Curtis

This measure is also termed the percentage difference, and is related to the complement of the [Steinhaus](#)<sup>[119]</sup> similarity measure. It is calculated as follows:

$$\frac{\sum_{i=1}^{i=n} |s_{i1} - s_{i2}|}{\sum_{i=1}^{i=n} (s_{i1} + s_{i2})}$$

Using the same notation as for the Steinhaus method this equates to  $1 - 2W/(A + B)$

## 11.28 Squared Chord Distance

This distance measure is popular with paleontologists and in studies on pollen. It is defined as:

$$D = \sum_{i=1}^{i=n} (\sqrt{x_i} - \sqrt{y_i})^2$$

where  $x_i$  is the number of observations for species  $i$  in sample  $x$  and  $y_i$  is the number of observations of species  $i$  in sample  $y$ .

## 11.29 Mahalanobis distance

This distance measure was designed by [P. C. Mahalanobis in 1936](#)<sup>[172]</sup>. It differs from Euclidean distance in that it takes into account the correlations between variables in the data set and is scale-invariant.

It is defined by the equation:

$\mu = (\mu_1, \mu_2, \mu_3, \dots, \mu_p)^T$  and [covariance matrix](#) for a multivariate vector  $x = (x_1, x_2, x_3, \dots, x_p)^T$  is defined as:

$$D_{x,\mu} = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

where  $x$  and  $\mu$  are the two vectors of variables between which the distance is measured and  $S^{-1}$  is the inverse of the covariance matrix between the variables.

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance.

See [PCA -Cor -Outlier R](#)<sup>[188]</sup> and [PCA - Covar - Outlier R](#)<sup>[190]</sup> if you wish to use the Mahalanobis distance to test for outliers following a PCA

## 11.30 Renkonen

The percent similarity or Renkonen index is defined as:

$$S_r = \sum_1^i \min(p_{1,i}, p_{2,i})$$

where  $p_{1,i}$  is the frequency of species  $i$  in collection 1.

This measure of sample similarity is considered one of the best quantitative similarity coefficients because it not heavily influenced by sample size and species number.

**Part**

---

**XII**



## 12 Association analysis

The association between species (rows) in the working data is shown by selecting **Association: Chi squared** from the drop-down menu. For each species pair, the contingency table of presence/absence is used to calculate a Chi-squared value. For measures of similarity between samples based on species presence-absence, the observations can be summarised in a simple frequency table:

		Sample 1	
		Species Present	Species Absent
Sample 2	Species Present	<b>a</b>	<b>b</b>
	Species Absent	<b>c</b>	<b>d</b>

where the number of species present in both samples is a, the number of species present in sample 1 but missing from sample 2 is b, the number of species missing in sample 1 but present in sample 2 is c and the number of species missing from both samples is d. The total number of species, N, is therefore  $a+b+c+d$ .

The Chi-squared value for each contingency table is calculated in the normal fashion using the Yate's correction for small numbers of observations.

The Chi-squared values are presented in a grid displayed by clicking on the Chi-squared tab, with the following display conventions:

1. Chi-squared values which indicate a significant degree of positive or negative association between the species at the 5% level are shown in bold.
2. Positive associations are displayed in blue, and negative associations as negative numbers in red.
3. If either of the two species is present in every sample the contingency table is meaningless; N/A is displayed.

Please note that association analysis will give misleading results if the data comprise many zeros caused by low sampling effort, rather than true species presence/absence.

## CAP Demo 1.0.0.0 - Powerstation fish

File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare

## Results - Association

	Sprat	Whiting	Goby,Sand	Herring	Pipefish, Greater	Pipefish, Nilsson's	Sole (Dover sole)
Sprat			Significant negative association at the 5% level between sand goby and whiting.				
Whiting	N/A						
Goby,Sand	N/A	<b>-3.9861592</b>					
Herring	N/A	N/A	N/A				
Pipefish, Greater	N/A	0.1323529	0.1323529	N/A			
Pipefish, Nilsson's	N/A	0.0550038	Significant positive association.		0.0292208		
Sole (Dover sole)	N/A	<b>-1.6213235</b>			1.7578125	0.1826299	
Flounder	N/A	<b>3.9861592</b>	<b>-3.9861592</b>	N/A	0.1323529	0.0550038	<b>-1.6213235</b>
Dab	N/A	1.6213235	<b>-1.6213235</b>	N/A	1.7578125		<b>0.4394531</b>
Pout	N/A	<b>-3.9861592</b>	<b>-3.9861592</b>	N/A	0.1323529		<b>-1.6213235</b>
Hooknose (Pogge)	N/A	0.4726891	<b>-0.4726891</b>	N/A	1.9687500		<b>0.0100446</b>
Plaice	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Stickleback, 3-Spined	N/A	0.8470588	0.8470588	N/A	0.4500000	0.1870130	<b>-0.1125000</b>
Smelt	N/A	0.0000000	0.0000000	N/A	2.2500000	0.0000000	0.5625000
Sea snail, Common	N/A	0.0550038	0.0550038	N/A	1.4318182	0.5950413	1.2345779
Cod	N/A	0.8470588	<b>-0.8470588</b>	N/A	0.4500000	0.1870130	<b>-0.1125000</b>
Goby, Transparent	N/A	<b>-0.0550038</b>	0.0550038	N/A	<b>-0.0292208</b>	3.1087873	0.1826299

## CAP Demo 1.0.0.0 - Powerstation fish

File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare

## Results - Association

	Sprat	Whiting	Goby,Sand	Herring	Pipefish, Greater	Pipefish, Nilsson's	Sole (Dover sole)
Sprat			Significant negative association at the 5% level between sand goby and whiting.				
Whiting	N/A						
Goby,Sand	N/A	<b>-3.9861592</b>					
Herring	N/A	N/A	N/A				
Pipefish, Greater	N/A	0.1323529	0.1323529	N/A			
Pipefish, Nilsson's	N/A	0.0550038	Significant positive association.		0.0292208		
Sole (Dover sole)	N/A	<b>-1.6213235</b>			1.7578125	0.1826299	
Flounder	N/A	<b>3.9861592</b>	<b>-3.9861592</b>	N/A	0.1323529	0.0550038	<b>-1.6213235</b>
Dab	N/A	1.6213235	<b>-1.6213235</b>	N/A	1.7578125		<b>0.4394531</b>
Pout	N/A	<b>-3.9861592</b>	<b>-3.9861592</b>	N/A	0.1323529		<b>-1.6213235</b>
Hooknose (Pogge)	N/A	0.4726891	<b>-0.4726891</b>	N/A	1.9687500		<b>0.0100446</b>
Plaice	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Stickleback, 3-Spined	N/A	0.8470588	0.8470588	N/A	0.4500000	0.1870130	<b>-0.1125000</b>
Smelt	N/A	0.0000000	0.0000000	N/A	2.2500000	0.0000000	0.5625000
Sea snail, Common	N/A	0.0550038	0.0550038	N/A	1.4318182	0.5950413	1.2345779
Cod	N/A	0.8470588	<b>-0.8470588</b>	N/A	0.4500000	0.1870130	<b>-0.1125000</b>
Goby, Transparent	N/A	<b>-0.0550038</b>	0.0550038	N/A	<b>-0.0292208</b>	3.1087873	0.1826299

**Part**

---

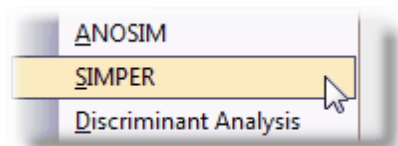


## 13 Group Tests

Samples must be [allocated to groups](#)<sup>[53]</sup> prior to running a [Discriminant Analysis](#)<sup>[131]</sup>, [Analysis of Similarity](#)<sup>[128]</sup> or [SIMPER analysis](#)<sup>[129]</sup>.

This menu allows you to:

- (1) undertake an Analysis of Similarity ([ANOSIM](#)<sup>[128]</sup>),
- (2) identify those species which most contribute to the similarity (or dissimilarity) between samples ([SIMPER](#)<sup>[129]</sup>), and
- (3) undertake a [Discriminant or Canonical Variates Analysis](#)<sup>[131]</sup>.

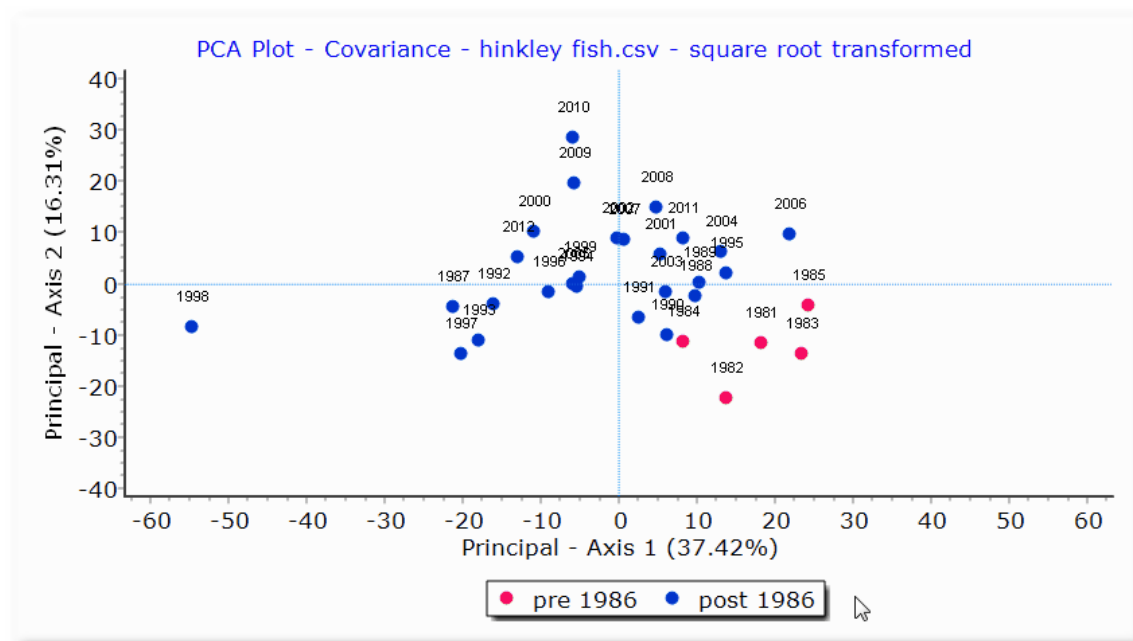


For video demonstrations see the **Help: Guides: Grouping from Form**.

Why does CAP allow you to assign samples to groups? You might for example allocate a series of samples to polluted and control groups, or in a benthic study, to different physical conditions such as mud, sand and gravel substrates. In another type of study you might want to define as a group all samples that held a species of particular interest.

By [allocating samples to groups](#)<sup>[53]</sup> you can show the different groups in ordination plots and other graphs.

For example, in the *hinkley fish.csv* example dataset supplied with CAP, the species of fish caught in different years are grouped as pre- and post-1986. If a PCA is run on these data the two groups can be clearly seen to cluster around different positions within the ordination space. Note that the data were square-root transformed to get the clear result shown.



Once samples have been placed into groups you can test if the similarity of samples within each group is greater than the similarity that would occur by random chance, using [ANOSIM](#)<sup>[128]</sup>, or you

can undertake a [Discriminant Analysis](#)<sup>[131]</sup>.

## 13.1 Analysis of Similarity (ANOSIM)

To undertake this test you must first have defined the group membership of the individual samples (see [Allocating samples to Groups](#)<sup>[531]</sup>).

This test was developed by [Clark \(1988, 1993\)](#)<sup>[172]</sup> as a test of the significance of the groups that had been defined *a priori*. The idea is simple; if the assigned groups are meaningful, samples within groups should be more similar in composition than samples from different groups. The method uses the [Bray-Curtis](#)<sup>[121]</sup> measure of similarity. The null hypothesis is therefore that there are no differences between the members of the various groups.

[Clark \(1988, 1993\)](#)<sup>[172]</sup> proposed the following statistic to measure the differences between the groups:

$$R = \frac{\bar{r}_B - \bar{r}_W}{n(n-1)/4}$$

where

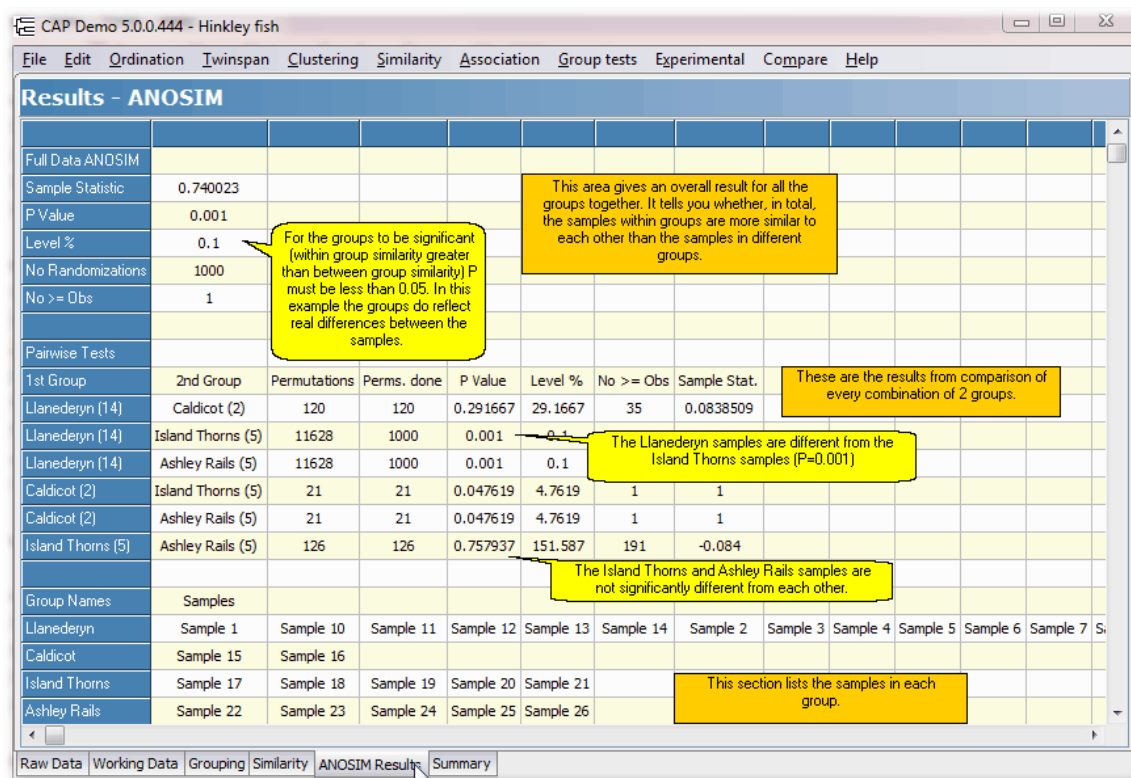
$\bar{r}_B, \bar{r}_W$  are the mean of the ranked similarity BETWEEN groups and WITHIN groups respectively and n is the total number of samples (objects).

R scales from +1 to -1. +1 indicates that all the most similar samples are within the same groups. R = 0 occurs if the high and low similarities are perfectly mixed and bear no relationship to the group. A value of -1 indicates that the most similar samples are all outside of the groups. While negative values might seem to be a most unlikely eventuality it has been found to occur with surprising frequency.

To test for significance, the ranked similarity within and between groups is compared with the similarity that would be generated by random chance. Essentially the samples are randomly assigned to groups 1000 times and R calculated for each permutation. The observed value of R is then compared against the random distribution to determine if it is significantly different from that which could occur at random.

If the value of R is significant, you can conclude that there is evidence that the samples within groups are more similar than would be expected by random chance.

The layout of the ANOSIM results is explained below:



## 13.2 Similarity Percentages (SIMPER)

To undertake this test, you must first have defined the group membership of the individual samples (see [Allocating samples to groups](#)<sup>[53]</sup>).

This analysis breaks down the contribution of each species (or other variable) to the observed similarity (or dissimilarity) between samples. It will allow you to identify the species that are most important in creating the observed pattern of similarity. The method uses the [Bray-Curtis](#)<sup>[12]</sup> measure of similarity, comparing in turn, each sample in Group 1 with each sample in Group 2. The Bray-Curtis method operates at the species level, and therefore the mean similarity between Groups 1 & 2 can be obtained for each species.

In the following example, using the *Romano British pottery.csv* data file, the data have been divided into 3 location groups.

The SIMPER results for the SIMPER Within tab shown below indicate that Al (aluminum) is the variable that contributes the most to the within-group similarities at every site.

CAP Demo 1.0.0.0 - Romano British pottery

File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare Help

### Results - SIMPER similarity

The average similarity between group members, based on the Bray-Curtis similarity measure, is 91.83%.

Each group is considered in turn. For the Llanderyn group it is Al (aluminium) which contributes most highly to the similarity between group members.

Aluminium contributed about 53% to the total similarity, followed by iron (Fe) at about 27%.

Name	Ave. Abund	Ave. Simil	% Contribution	Cumulative %
Al	12.5643	48.4685	52.7804	52.7804
Fe	6.3721	24.5102	26.6907	79.4711
Mg	4.8264	17.3618	18.9063	98.3774

Average Sim 91.8305

Name	Ave. Abund	Ave. Simil	% Contribution	Cumulative %
Al	11.7000	54.4218	55.0024	55.0024
Fe	5.4150	25.2874	25.5571	80.5595
Mg	3.8550	17.6871	17.8758	98.4353

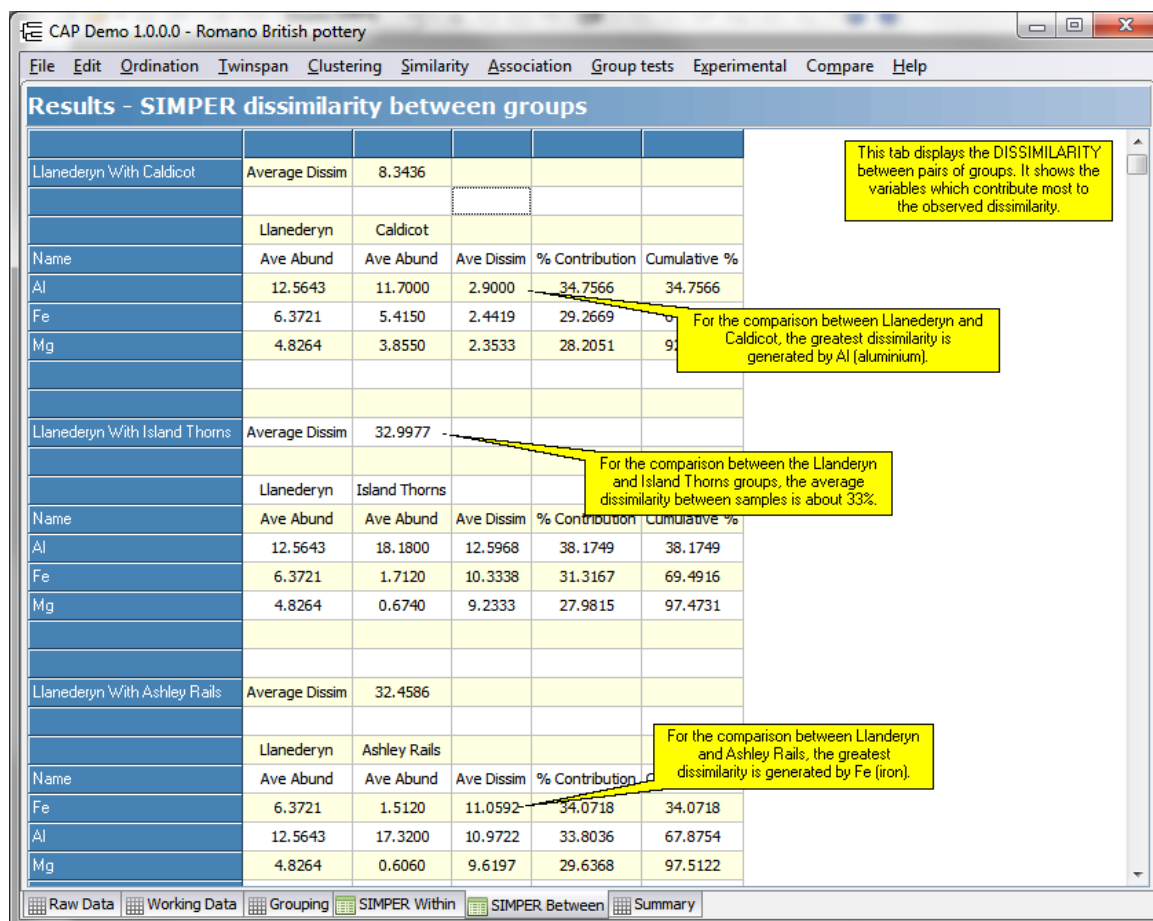
Average Sim 98.9444

Name	Ave. Abund	Ave. Simil	% Contribution	Cumulative %
Al	18.1800	83.0652	88.8091	88.8091
Fe	1.7120	7.0503	7.5379	96.3470

Average Sim 93.5323

Raw Data Working Data Grouping SIMPER Within SIMPER Between Summary

Below, the results for the between-groups analysis is shown (SIMPER Between tab). There is one of these results panels for each pair-wise combination of groups. Note that on this panel it is actually the **dissimilarity** that is displayed.



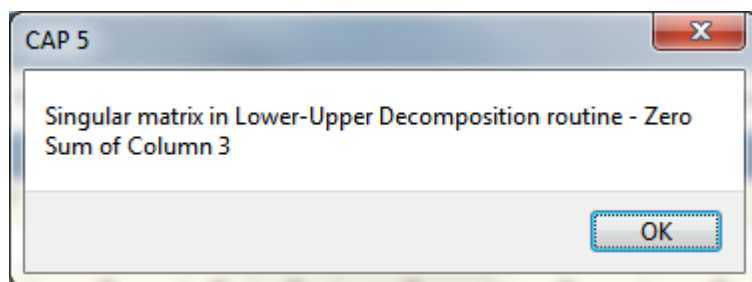
### 13.3 Discriminant Analysis

Discriminant analysis, or canonical variates analysis, is a standard method for testing the significance of previously-defined groups, identifying and describing which variables distinguish between groups, and producing a model to allocate new samples to a group. DA allows the relationship between groups of samples to be displayed graphically. A key goal of a discriminant analysis is to produce a simple function of the variables to classify the objects to their correct group.

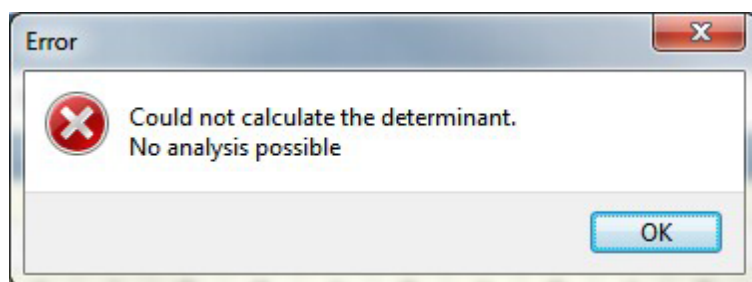
DA is closely related to multivariate analysis of variance (MANOVA). It provides information on the relative importance or contribution of each variable to the group structure and produces a method to allocate new observations to a group. To use DA you must [allocate samples to groups](#) <sup>[53]</sup>. For video demonstrations of how to group samples see the **Help: Guides: Grouping from Form** and **Help: Guides: Grouping from Plots**.

Some data sets simply will not work with a Discriminant Analysis, if there is low (or zero) variability in the samples/sites, or if there are close correlations between sites or variables. The program will show two error messages; the first shows where in the data set the problem lies:





while the second states that no analysis is possible:



See also:

[Eigenvalues - DA](#)<sup>[132]</sup>

[Discriminant Function Coefficients](#)<sup>[133]</sup>

[Fisher's Discriminant Functions](#)<sup>[134]</sup>

[Group Centroids -DA](#)<sup>[135]</sup>

[Significant tests - DA](#)<sup>[136]</sup>

[Plot- DA](#)<sup>[137]</sup>

[Dispersion Matrices](#)<sup>[137]</sup>

[Coordinates- DA](#)<sup>[138]</sup>

[Predictive Validation](#)<sup>[138]</sup>

### 13.3.1 Eigenvalues - DA

The Eigenvalues tab screen displays the eigenvalues for the discriminant functions and the correlation between the discriminant scores and the group membership variable. See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

CAP Demo 1.0.0.0 - Irises							
File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare H							
Results - Eigenvalues							
Number of roots extracted	2.0000						
Percent of trace extracted	100.0000						
Roots of the W inverse time B Matrix							
	No.	Eigenvalue	Proportion	Canonical Correlation	Chi-Squared	D.F.	Prob.
	1.0000	32.1919	0.9912	0.9848	544.2386	8.0000	0.0000
	2.0000	0.2854	0.0088	0.4712	36.4041	3.0000	0.0000

**Eigenvalues** measure the amount of variance in the grouping variable explained by the predictors in the discriminant function. There is one eigenvalue for each discriminant function, so if you only have two groups there will only be 1 eigenvalue, which will account for 100% of the explained variation.

The discriminant functions are arranged in terms of their discriminatory power so that the first has the largest eigenvalue. The ratio of the eigenvalues for each discriminant function measures the relative importance of each function in discriminating between the groups. In the example above, for example, the largest function (eigenvalue 32.1919, proportion 0.9912) has by far the greatest power to discriminate between the groups.

**Proportion** is the fraction of the total sum of the eigenvalues represented by any one eigenvalue.

**Canonical Correlation** is a measure of the association between the groups and those defined by the discriminant function. If there are two groups, then the canonical correlation is the Pearson correlation between discriminant scores and the group membership variable (eg group 1 = 1, group 2 = 2 etc.). It measures the usefulness of the function in discriminating between the groups. A value of zero indicates no relationship and no discriminatory ability. A value of 1 indicates that all the variability in the discriminant scores for objects is generated by a single discriminant function.

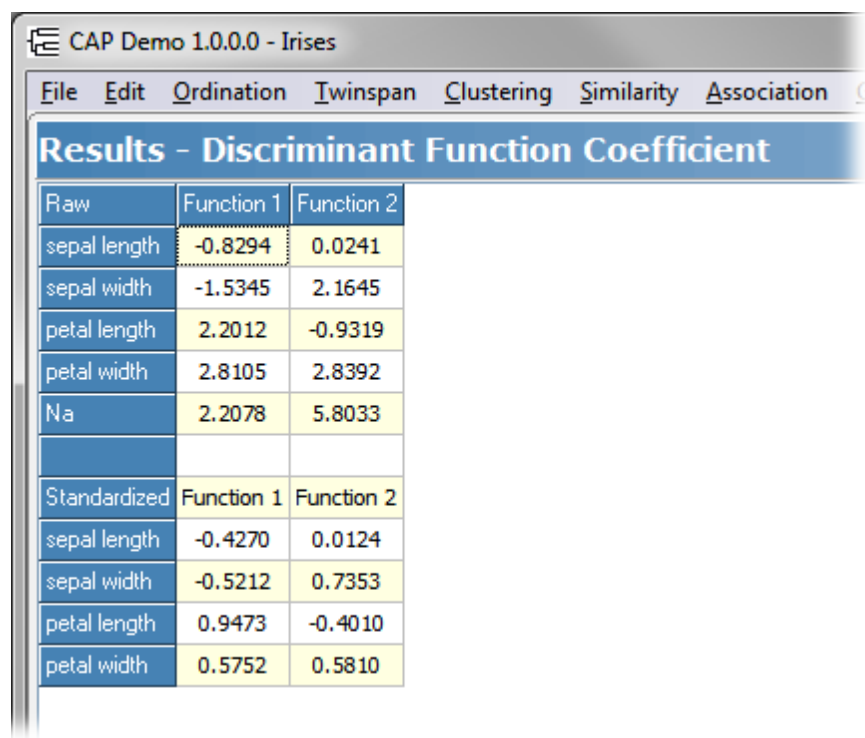
**Chi-squared** is the test statistic for the significance of the observed canonical correlation.

**D.F.** is the degrees of freedom for the Chi-squared test statistic.

**Prob.** is the probability that a correlation of the size observed could be generated by random chance.

### 13.3.2 Discriminant Function Coefficients

The Disc. Func. Coef. tab presents tables for the raw and standardised discriminant function coefficients. The raw coefficients are used in the linear discriminant functions with the observations to generate the score used to allocate to group. The standardised discriminant function coefficients are used to compare the relative importance of the variables to the generation of the discriminant score. The greater the magnitude of a standardised coefficient, the more important the role in the discriminant function.



CAP Demo 1.0.0.0 - Irises		
File Edit Ordination Twinspan Clustering Similarity Association		
Results - Discriminant Function Coefficient		
Raw	Function 1	Function 2
sepal length	-0.8294	0.0241
sepal width	-1.5345	2.1645
petal length	2.2012	-0.9319
petal width	2.8105	2.8392
Na	2.2078	5.8033
Standardized	Function 1	Function 2
sepal length	-0.4270	0.0124
sepal width	-0.5212	0.7353
petal length	0.9473	-0.4010
petal width	0.5752	0.5810

See [Printing and exporting text](#) <sup>168</sup> to save or print this table.

### 13.3.3 Fisher's Discriminant Functions

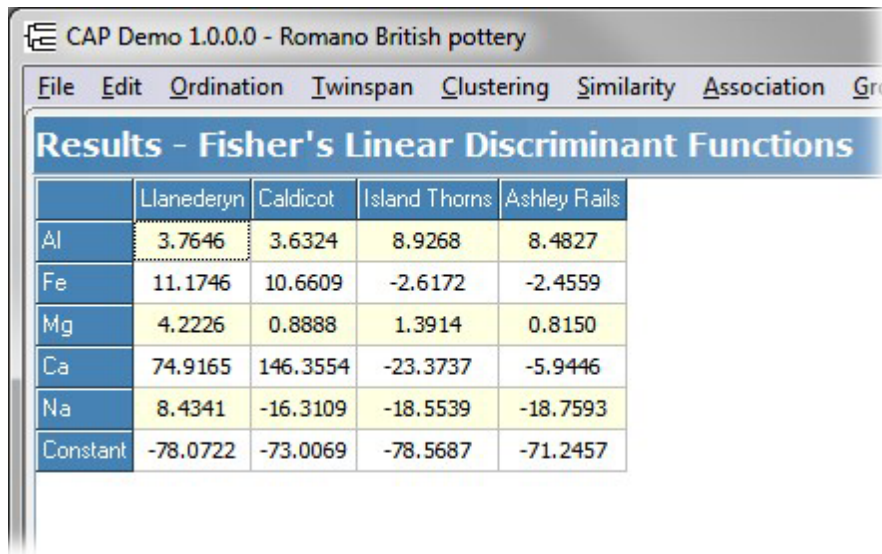
There are many ways in which the classification of the samples can be accomplished. Fisher's discriminant functions is the simplest method.

To assign an object to one of the pre-defined groups, we use the classification equations generated by DA to give a score for a sample. These equations have the form:

$$C_j = c_{j0} + c_{j1}X_1 + c_{j2}X_2 + \dots c_{jp}X_p$$

where  $c_j$  is a classification function coefficient,  $j$  is the group and  $p$  the number of variables.

Fisher's discriminant functions give the coefficients,  $c$ , in the above equation.



The screenshot shows a software window titled 'CAP Demo 1.0.0.0 - Romano British pottery'. The menu bar includes File, Edit, Ordination, Twinspan, Clustering, Similarity, Association, and Groups. The main window displays 'Results - Fisher's Linear Discriminant Functions' with a table of coefficients for four groups: Llanederyn, Caldicot, Island Thorns, and Ashley Rails. The rows represent elements: Al, Fe, Mg, Ca, Na, and a Constant term.

	Llanederyn	Caldicot	Island Thorns	Ashley Rails
Al	3.7646	3.6324	8.9268	8.4827
Fe	11.1746	10.6609	-2.6172	-2.4559
Mg	4.2226	0.8888	1.3914	0.8150
Ca	74.9165	146.3554	-23.3737	-5.9446
Na	8.4341	-16.3109	-18.5539	-18.7593
Constant	-78.0722	-73.0069	-78.5687	-71.2457

Using the example above, from the *Romano British pottery* demo data set, the functions for each group are:

#### Caldicot:

$$C_{cald} = -76.217 + 3.73X_{Al} + 11.17X_{Fe} + 0.84X_{Mg} + 155.68X_{Ca} + -17.22X_{Na}$$

#### Llanederyn

$$C_{lla} = -80.81 + 3.75X_{Al} + 11.75X_{Fe} + 4.17X_{Mg} + 85.62X_{Ca} + 8.73X_{Na}$$

#### New Forest

$$C_{NF} = -75.29 + 8.75X_{Al} - 2.52X_{Fe} + 0.7X_{Mg} - 1.49X_{Ca} - 19.67X_{Na}$$

A sample is allocated to the group for which it has the highest classification score.

For example, a sample with 14% Al, 7% Fe, 4% Mg, 0.1% Ca and 0.5% Na gives

$$C_{cald} = -76.217 + 3.73 \times 14 + 11.17 \times 7 + 0.84 \times 4 + 155.68 \times 0.1 - 17.22 \times 0.5$$

so

$$C_{cald} = 64.511$$

and in similar fashion

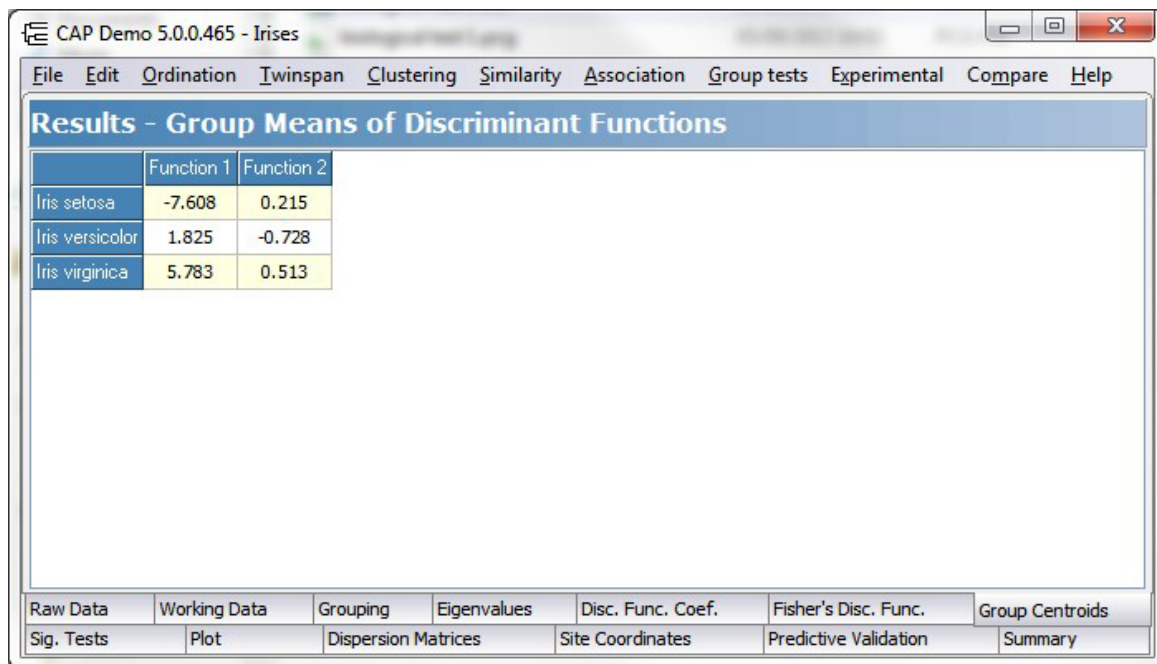
$$C_{lla} = 83.547$$

$$C_{NF} = 22.286$$

This sample is therefore allocated to the Llanederyn group, as it has the highest score.

### 13.3.4 Group Centroids - DA

The Group Centroids tab tabulates the mean values (centroids) of the discriminant functions for each group. To have groups that are clearly different, the means need to be spaced well apart.

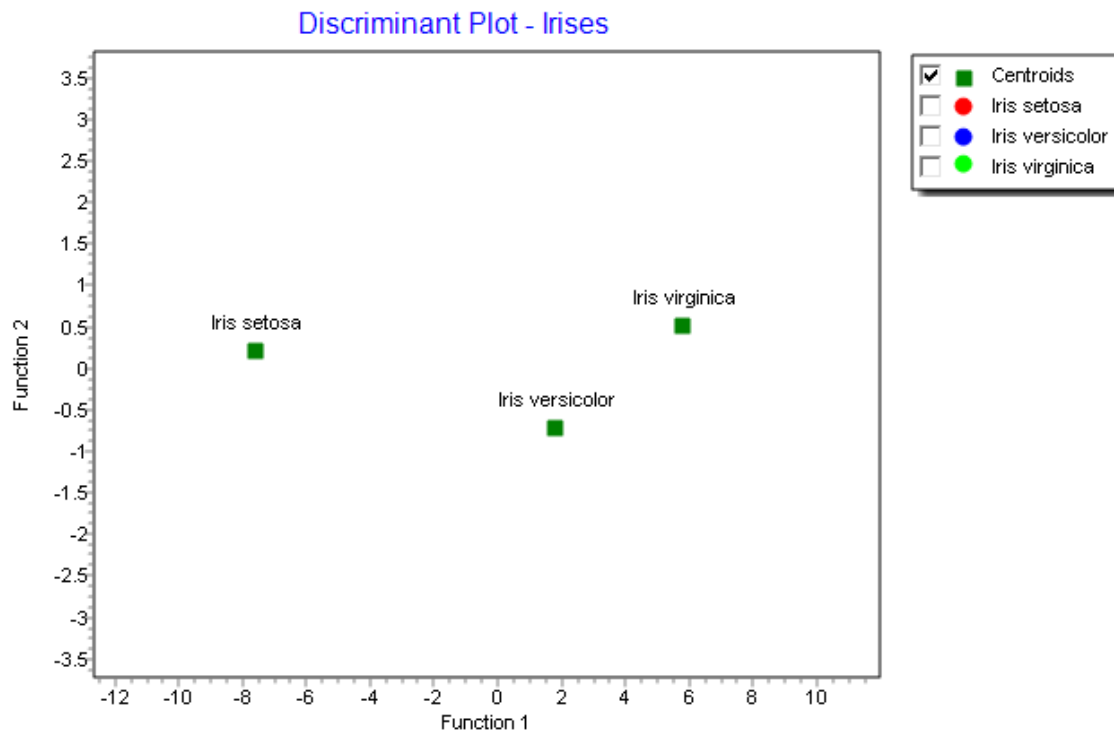


The screenshot shows a software window titled 'CAP Demo 5.0.0.465 - Irises'. The menu bar includes File, Edit, Ordination, Twinspan, Clustering, Similarity, Association, Group tests, Experimental, Compare, and Help. The main content area displays a table titled 'Results - Group Means of Discriminant Functions'. The table has three columns: the group name, Function 1, and Function 2. The data rows are for Iris setosa, Iris versicolor, and Iris virginica. At the bottom of the window, there is a tabbed interface with the following tabs: Raw Data, Working Data, Grouping, Eigenvalues, Disc. Func. Coef., Fisher's Disc. Func., Group Centroids, Sig. Tests, Plot, Dispersion Matrices, Site Coordinates, Predictive Validation, and Summary. The 'Group Centroids' tab is currently selected.

	Function 1	Function 2
Iris setosa	-7.608	0.215
Iris versicolor	1.825	-0.728
Iris virginica	5.783	0.513

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

CAP also plots these centroids; see the plot below:



### 13.3.5 Significance tests - DA

This tab presents the results of a number of standard tests in a Rich Text box.

**Wilks' lambda** is used to test the significance of the discriminant function as a whole. Generally, you will want a probability less than 0.05, as this indicates a rejection of the null hypothesis at the 5% level that the groups have the same mean discriminant function scores. If the groups have different scores, then the model is discriminating between the groups.

**DF** is the degrees of freedom for Wilks' lambda.

**Prob** is the probability that the groups have the same discriminant scores.

**Pillai's trace** is a second overall test for a significant ability to discriminate between the groups.

**Bartlett's Test of Sphericity** tests if the variables are uncorrelated. If this is the case then the correlation matrix will be an identity matrix. The null hypothesis is that the correlation matrix comes from a population in which the variables are noncollinear (i.e. an identity matrix) and that the non-zero correlations in the sample matrix are due to sampling error. For a successful discriminant analysis you need to be able to reject the null hypothesis.

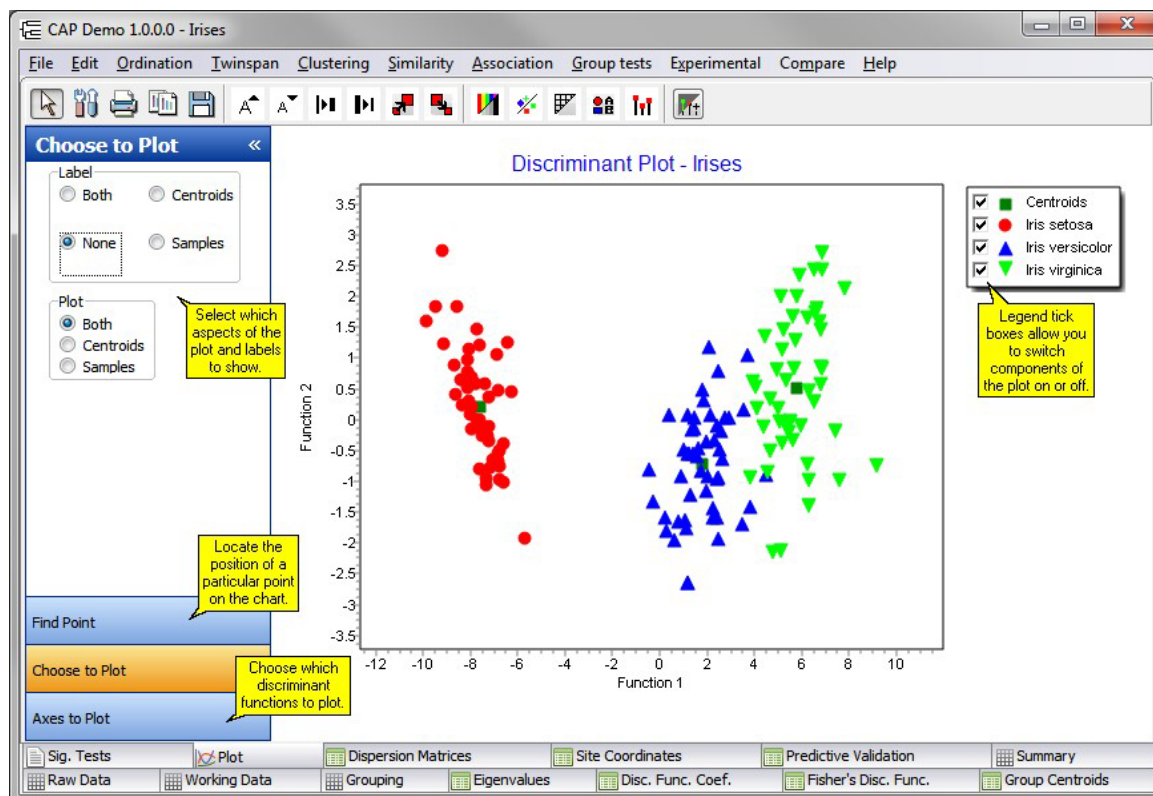
**DF** is the degrees of freedom for the Bartlett's test.

**Prob** is the probability that the variables are all uncorrelated.

Press Ctrl-Alt-C, or Edit: Copy All, to copy the entire text of this output. To copy a selected portion, select the text you require, and press Ctrl-C on your keyboard, or Edit: Copy.

### 13.3.6 Discriminant analysis plot

This window plots the discriminant scores for each sample, and also the centroid for each group.



The **Label** radio box is used to select labelling for the samples.

Select: **Both** to label centroids and samples.

**Centroids** to show the group name of the centroid.

**Samples** to show the sample names.

**None** to remove labels from the plot.

The **Plot** radio box allows the selection of centroids and samples for plotting.

The axes displayed in the plot are selected using the Plot x Axis, y Axis z Axis drop-down boxes on the **Axes to Plot** tab. The default is axis 1 and axis 2, which will display the relative positions of the samples using the scores for the two functions with the largest eigenvalues.

In two-dimensional plotting it is also possible to [draw a perimeter line](#)<sup>[161]</sup> around the members of each predefined group.

See also

[Preparing charts for output](#)<sup>[162]</sup>

Printing charts

[Exporting charts](#)<sup>[153]</sup>

[Zooming on charts](#)<sup>[158]</sup>

[Themes for graphs](#)<sup>[167]</sup>

### 13.3.7 Dispersion matrices - DA

Select the Dispersion Matrices tab to view the covariance matrices. The Total, within-group and between-group covariance matrices are all presented on the same grid. See the key at the bottom of the grid to distinguish between the 3 matrices. The between-group covariance is simply the difference between the total and within-group covariance.

CAP Demo 5.0.0.465 - Irises

File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare Help

### Results - Covariances

	sepal length	sepal width	petal length	petal width
sepal length	0.686 <b>0.265</b> 0.421	-0.042 <b>0.093</b> -0.135	1.274 <b>0.168</b> 1.107	0.516 <b>0.038</b> 0.478
sepal width	-0.042 <b>0.093</b> -0.135	0.190 <b>0.115</b> 0.075	-0.330 <b>0.055</b> -0.385	-0.122 <b>0.033</b> -0.154
petal length	1.274 <b>0.168</b> 1.107	-0.330 <b>0.055</b> -0.385	3.116 <b>0.185</b> 2.931	1.296 <b>0.043</b> 1.253
petal width	0.516 <b>0.038</b> 0.478	-0.122 <b>0.033</b> -0.154	1.296 <b>0.043</b> 1.253	0.581 <b>0.042</b> 0.539
Key	Total Covariance Matrix <b>Pooled within Group Covariance Matrix</b> Between Group Covariance Matrix			

Set threshold level 0.90

Sig. Tests	Plot	Dispersion Matrices	Site Coordinates	Predictive Validation	Summary
Raw Data	Working Data	Grouping	Eigenvalues	Disc. Func. Coef.	Fisher's Disc. Func.
					Group Centroids

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

### 13.3.8 Site Coordinates - DA

The Site Coordinates tab holds the score for each sample (object) for each discriminant function. These are the values that are [plotted](#)<sup>[137]</sup>.

### Results - Coordinates

	F1	F2	F3	F4	F5	F6	F7	F8	F9
Function 1	-8.0618	-7.1287	-7.4898	-6.8132	-8.1323	-7.7019	-7.2126	-7.6053	-6.56
Function 2	0.3004	-0.7867	-0.2654	-0.6706	0.5145	1.4617	0.3558	-0.0116	-1.01

See [Printing and exporting text](#)<sup>[168]</sup> to save or print this table.

### 13.3.9 Predictive Validation

The Predictive Validation tab compares the pre-assigned (original) group membership of a sample or object with the membership assigned by discriminant analysis. This will show you to what extent the discriminant functions are able to produce a classification corresponding to that originally given to the samples.

CAP Demo 1.0.0.0 - Irises

File Edit Ordination Twinspan Clustering Similarity Association

### Results - Predictive Validation

	Original Iris setosa	Original Iris versicolor	Original Iris virginica	No. Correct	% Correct
Predicted Iris setosa	50.0000	0.0000			0.0000
Predicted Iris versicolor	0.0000	48.0000			5.0000
Predicted Iris virginica	0.0000	2.0000	49.0000	49.0000	98.0000
Total	50.0000	50.0000	50.0000	147.0000	98.0000

This cell shows that 2 of the flowers originally classified as Iris versicolor were classified by DA as Iris virginica.

See [Printing and exporting text](#)<sup>168</sup> to save or print this table.



**Part**

---



## 14 Variable Filtering

Variable filtering is a novel ordination method developed by Peter Henderson and Richard Seaby of Pisces Conservation Ltd. In earlier versions of CAP it was called Species filtering, however since the variables are not inevitably species, its name has been changed to reflect this. The description below refers throughout to 'species', but plainly the variables could consist of any other suitable entity.

The utility of the method is yet to be assessed, but we believe that it has many useful features. In particular, it produces a 2-dimensional ordination of the sites that has a clear biological interpretation in terms of the species present in each sample. Our objective was a final result that would produce an ordination of sites with the greatest possible discrimination of sites along axes that allowed simple ecological interpretation. Ideally the method would work for both presence/absence and quantitative abundance data.

The method ordinated the sites in a two-dimensional space. The first dimension is an ordination in terms of species content and the second is simply a plot of the number of species present. The key feature is the unique way in which the sites are ordinated along axis 1. This will be described in detail below.

Imagine a sieve that allowed all the samples that did not hold a species to pass through it, but retained those samples that contained the species. We term such a sieve a negative filter. In contrast a filter that will only allow samples containing a particular species to pass through would be a positive filter. Now consider a series of species sieves placed in line so that progressively more and more of the samples are retained. To produce a useful ordination we need a set of rules that will determine which species should be used as sieves and in which order the sieves should be placed. The proposed rules are quite simple.

1. Species present at all sites are excluded. This is because they cannot be used to differentiate between sites.
2. When two or more species have the same pattern of occurrence at the sites a single sieve represents them all.
3. The order of sieves should be that which will produce the most even distribution of sites along the axis.

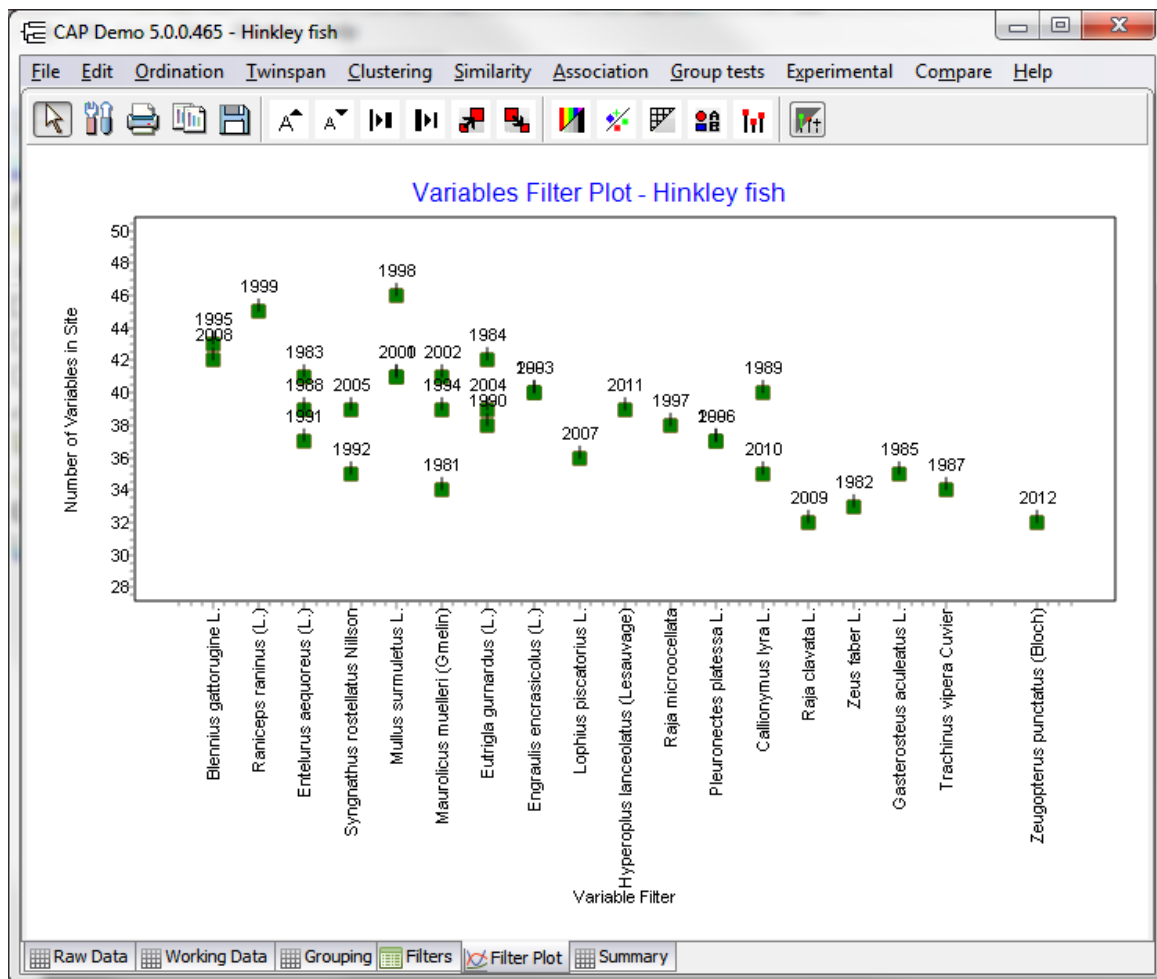
Rule 3 is the key feature that will lead to a useful ordination, and in practice is more difficult to discover than you might suppose. The possible arrangement of species filters will increase factorially with the number of species present so that even with quite a modest species number it is impossible to consider the merits of all possible combinations. The solution to this problem is to use a numerical method termed annealing to seek an optimal solution. While the search for a good solution requires considerable computation, our experience suggests that even with quite large data sets a useful ordination is created within 2 to 5 minutes.

When using a negative filter, which we recommend, the final ordination along the first axis will tend to arrange the sites (samples) in a clear and quite particular order. Starting from the left the sites will initially be characterised by their presence of rare or unusual species. At the opposite end of the axis will be sites that only hold the most frequently present species. Sites at the left end of the ordination can be classified into two groups, those that contain both infrequent and frequent species, and sites that hold only infrequent species. Sites that hold both will tend to have higher species richness than those that only hold rarer forms, and this immediately suggests that the second axis should simply be the total number of species in the sample.

Annealing works by taking an initially, possibly arbitrary, arrangement of the species and changing this arrangement in a manner that will make it more likely that any superior order will be accepted. There are two types of change allowed. (1) A randomly-chosen length of the species sequence is removed and replaced at a randomly-chosen site or (2) the order of a randomly-chosen length of the species sequence is reversed. When any such change is made a test is undertaken to see if the new arrangement produces a more even distribution of the sites. If it does, the new arrangement is

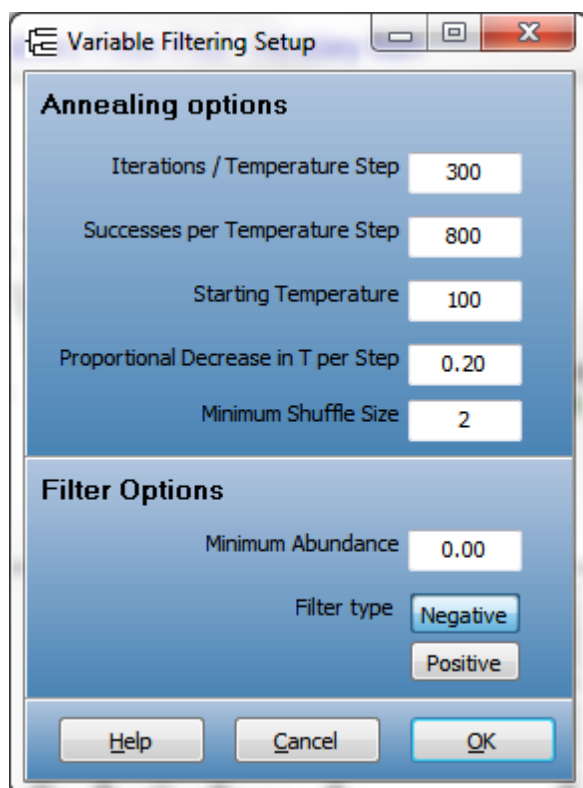
accepted. The temperature comes into play by allowing possibly inferior arrangements to be accepted. The higher the temperature the more likely this is to occur. The reason for this feature is that it tends to allow the program to find a global optimal solution and not become trapped by a local minimum. The higher the temperature the more likely an inferior arrangement is likely to be accepted. Annealing proceeds by a steady and gradual reduction in temperature.

The first axis therefore consists of a series of species, which are labelled on the output. These species are the key members of the community, whose presence or absence can be used to distinguish between sites. The plot below shows the result of a variable filter analysis of the *Hinkley fish* demo data set:



## 14.1 Variable Filtering - Setup

When variable filtering is selected, you are initially presented with an options screen which can be used to tune the annealing and also select key features of the filter options.



### Annealing options:

**Iterations / temperature set** – This option sets the maximum number of rearrangements of the species filter order that will be attempted at each temperature.

**Successes per Temperature Step** – This option determines how many superior arrangements in the species order will be found at each temperature before the temperature declines.

**Starting temperature** – The higher the initial temperature the more mixing of the species order will occur. In practice a temperature above 100 (boiling point) will be unlikely to be useful.

**Proportional decrease in T per Step** – This option should be less than 1. A value of 0.2 will result in a reduction to 20% of the initial value at each step.

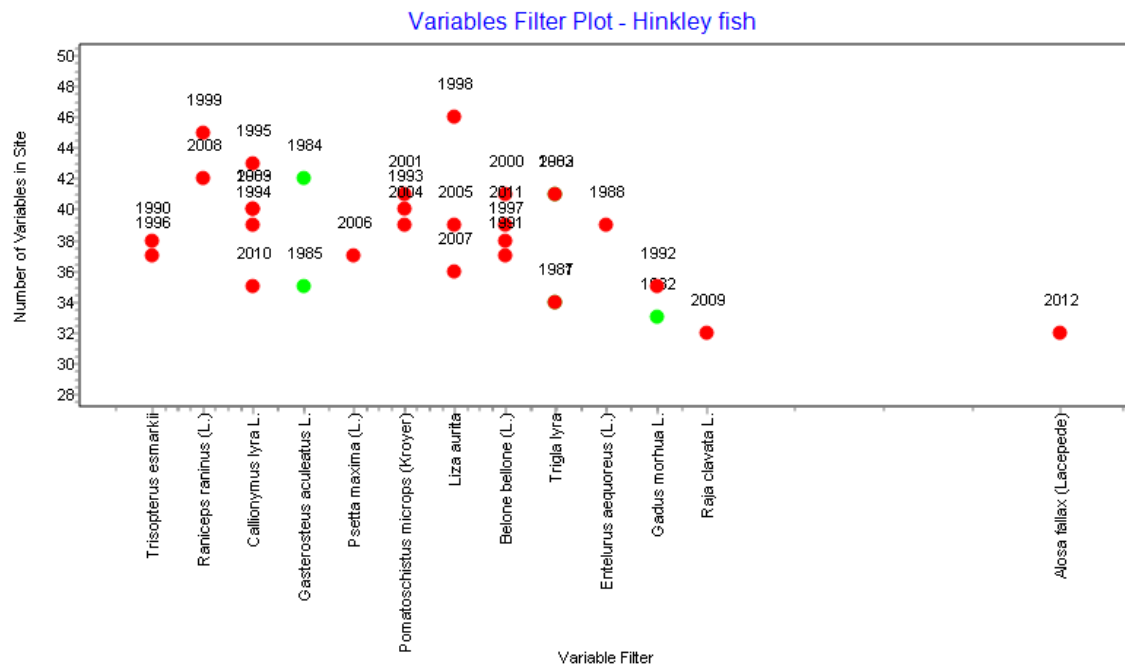
**Minimum shuffle size** – This sets the minimum length of the species sequence that will be rearranged. For example if it is set at 3, no sequence of fewer than 3 species will be moved or reversed.

### Filter Options:

**Minimum abundance** – For presence/absence data this should be set to 0. For quantitative or percentage composition data it can be set to a higher number. If it is set, for instance, to 4, then negative filters will only retain sites (samples) where the abundance of the species exceeds 4.

**Filter type** - This allows a switch from negative to positive filtration. Negative is normally recommended. In negative filtration a sample is retained if the species is present.

While the computation is running, the graph shows the progress of the calculation downwards towards the most stable solution. To show the chart of the whole series of iterations, untick the box marked 'Track 350'. Tick the box to chart just the last 350 iterations. When the computation has finished, the progress chart disappears, and the results are shown on the Filters and Filter Plot pages. You can halt the computation at any time by pressing the 'End' button; the most recent iteration will be shown on the Filter Plot page:



# Part

---



**XV**

## 15 Compare

This menu introduces four methods which can be used to compare samples within your data set, and explore the distribution of species or other variables within the data set:

[Compare Samples](#)<sup>[146]</sup> - examine pairs of samples or groups, and see which variables they have in common, and which are exclusive to one group or another.

[Profile plot](#)<sup>[147]</sup> - see which samples certain species or variables occur in.

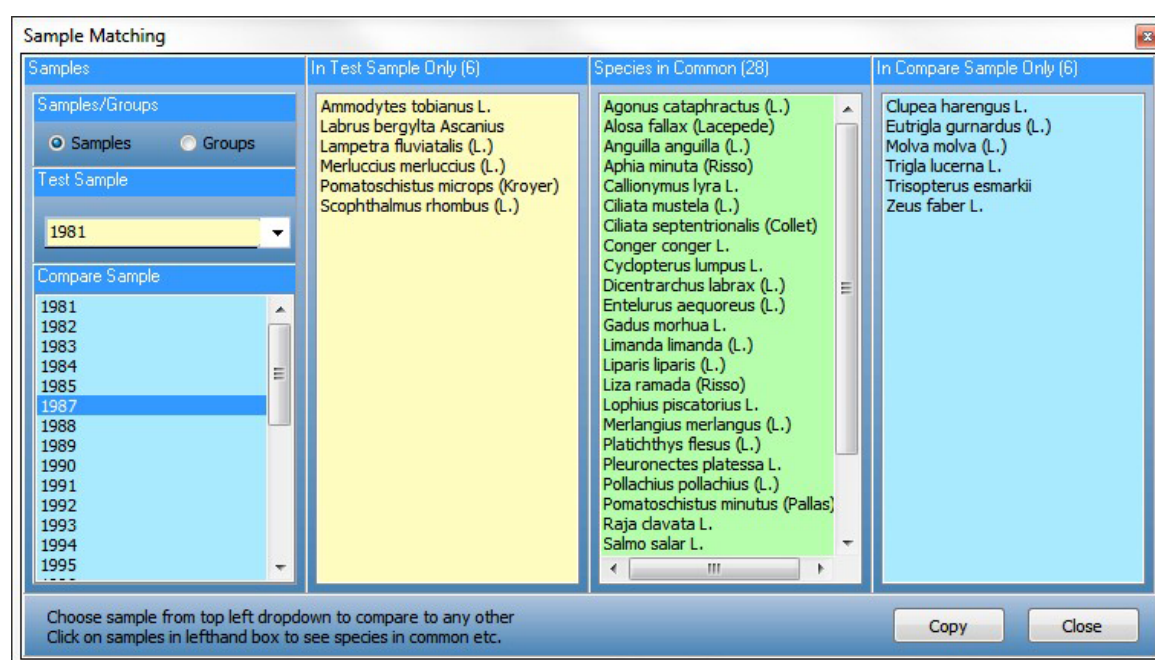
[Scatter plot](#)<sup>[148]</sup> - compare pairs of sites, or pairs of variables.

[Matrix plot](#)<sup>[149]</sup> - look at the relationship between a number of variables (species) or samples simultaneously

### 15.1 Compare samples

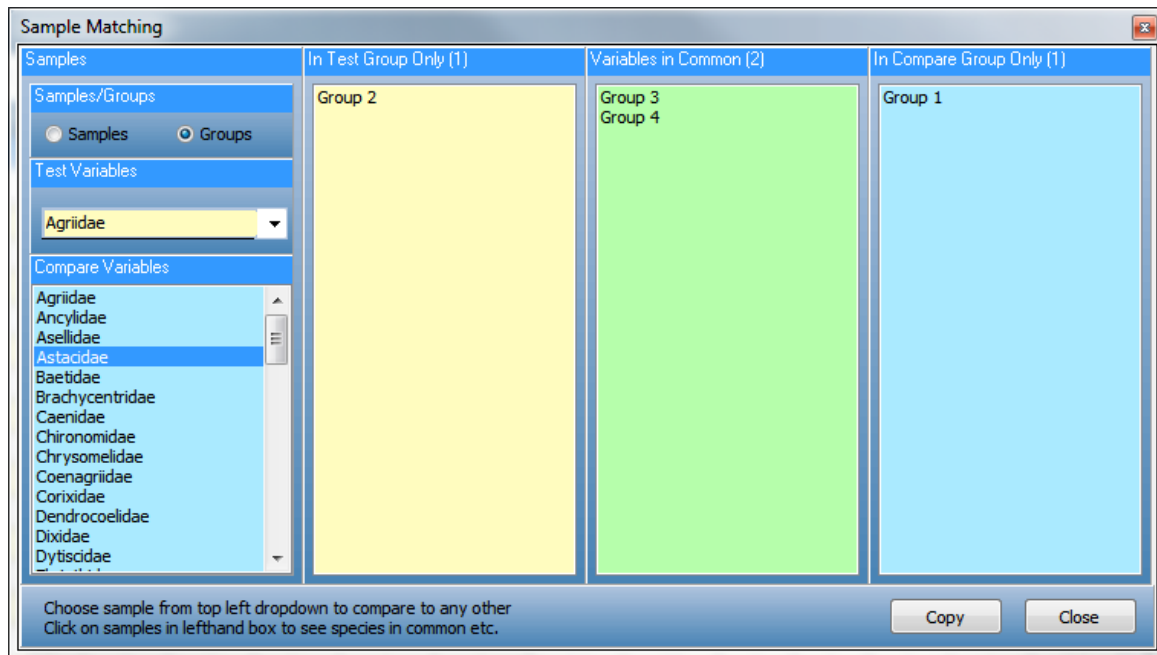
This window allows you to compare the variables present in two samples, or the group membership of two species. It is particularly suitable for comparing the species present in different samples, but can also be used to compare types of pottery, fossils etc. When used to compare group membership, it can be used to help determine which species characterise a particular group.

Comparing two samples is a straightforward operation; each sample will contain a subset of the full list of species present in all the samples, and in most cases, the two subsets of species will overlap; it is a Venn diagram in list form. There is a fourth category which is not shown by this function - species which do not occur in either of the two samples being compared.



In the case of the Groups comparison, it is the *variables' membership of groups*, rather than the groups themselves, which is being compared. In the following example, using the *River inverts* demo data set, we are comparing the occurrence of the Agriidae (the demoiselles) with the Astacidae (freshwater crayfish); the Agriidae occur in a Group 2 sample, Agriidae and Astacidae occur together in one of the samples of Groups 3 and 4, and Astacidae occur without the Agriidae in a sample from Group 1.

To continue our Venn diagram analogy, this function of CAP categorises a group in one of 3 ways: "Agriidae only", "Agriidae+Astacidae", and "Astacidae only":

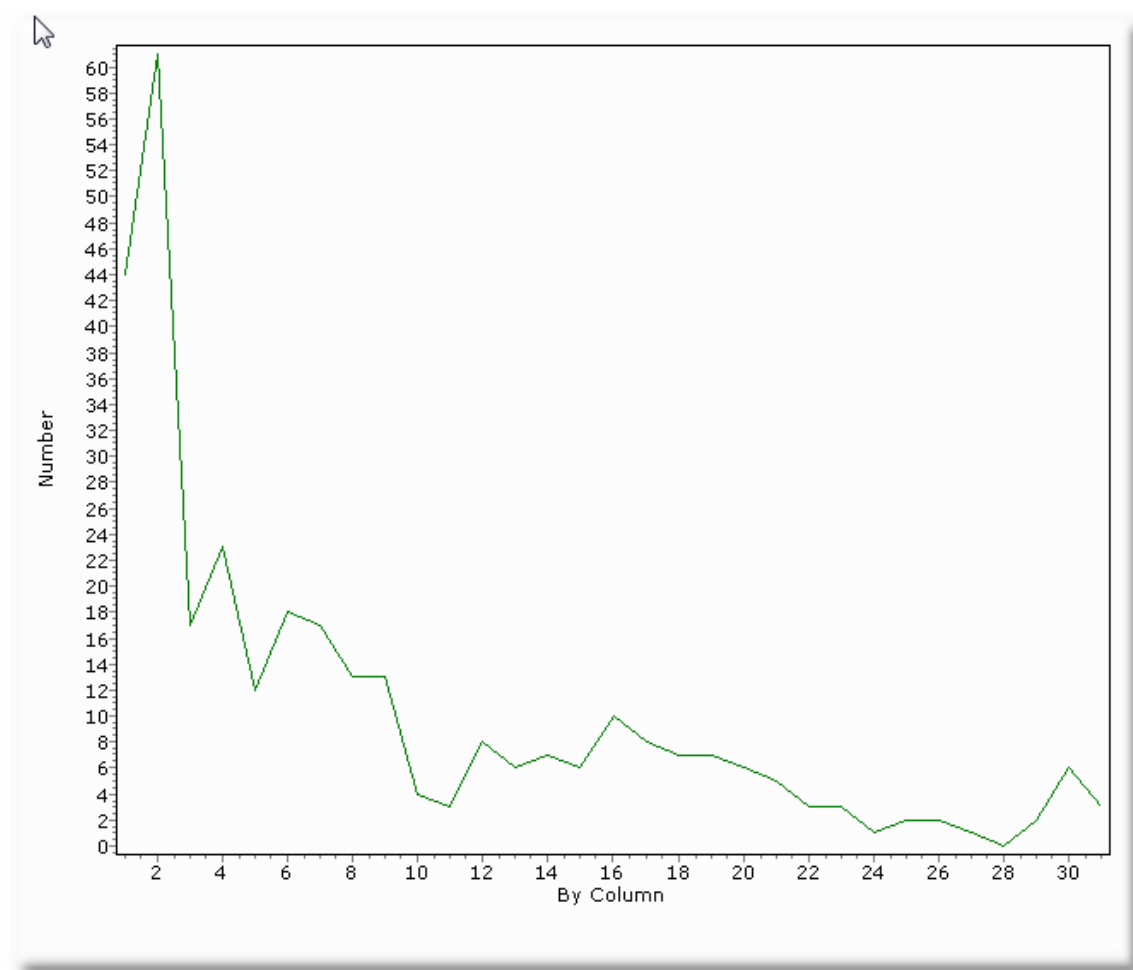


As in the previous example, there remains a fourth classification, which is not shown: "Not-Agnriidae+Not-Astacidae", groups where neither variable occurs.

## 15.2 Profile Plot

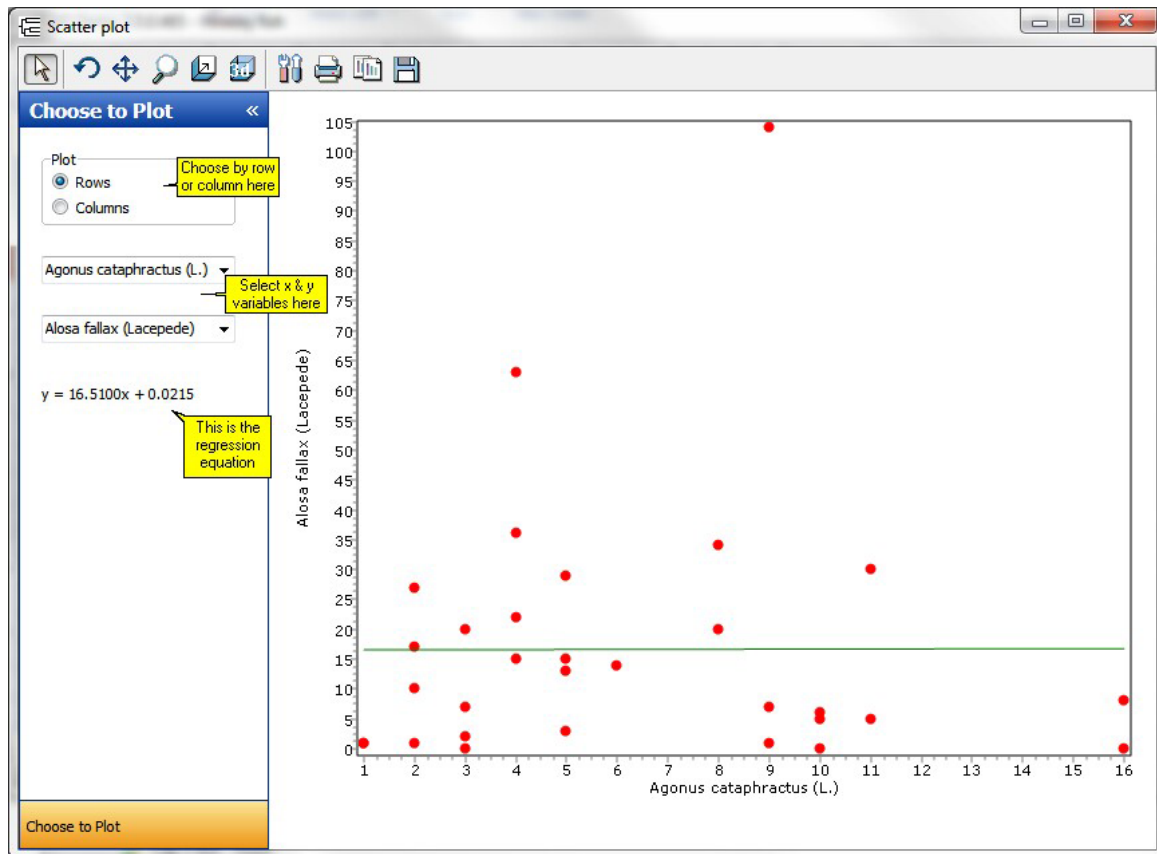
A profile plot simply plots all the values along a row or column of your data array. If your columns were samples arranged by year then you will see a plot of the selected species through time. The plot below shows the plot of a single fish species, the eel, *Anguilla anguilla*, in the *hinkley fish* data set.





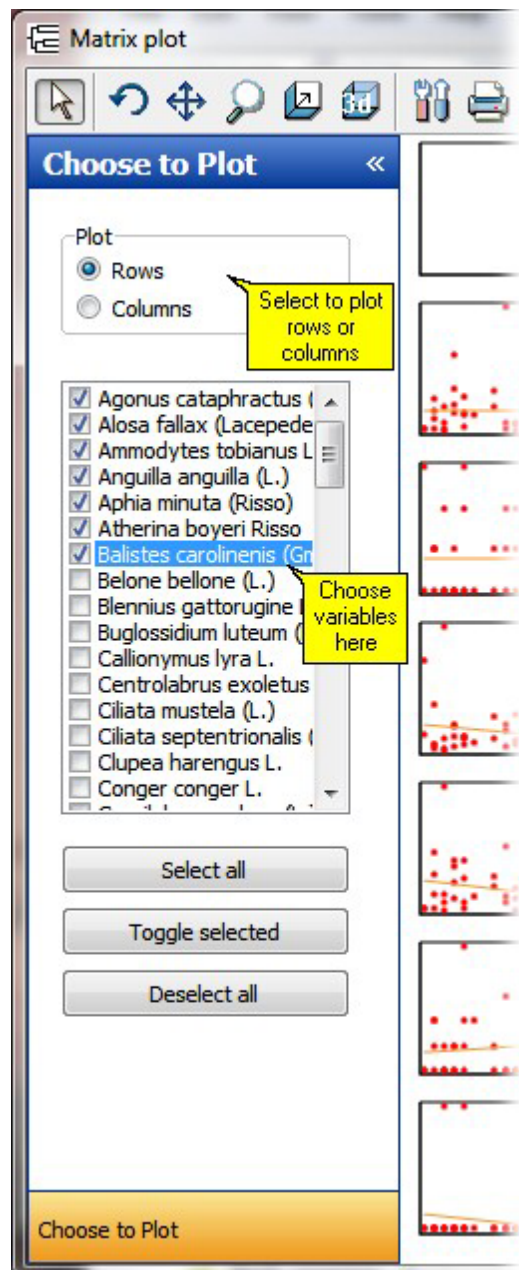
## 15.3 Scatter Plot

This is a simple scatter plot of one selected variable against another. The variables to be plotted against each other are chosen in the panel on the left hand side. Also shown on the left is the linear regression equation for the plot.

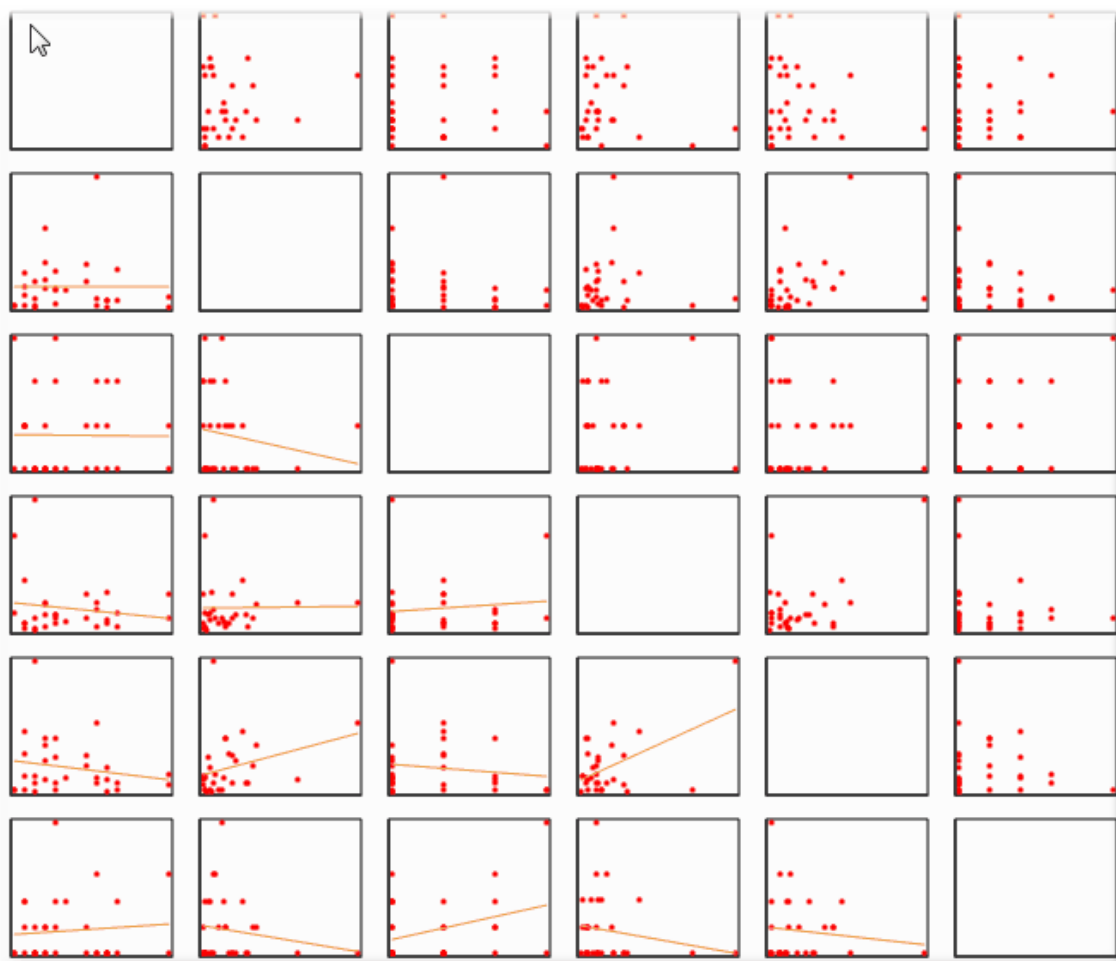


## 15.4 Matrix Plot

Use this option to look at the relationship between a number of variables (species) or samples simultaneously. The program will not allow you to choose more than 15 variables at once, as the resultant plot would take a very long time to produce, and would not be viewable.



The following demonstration output is for the first 6 fish (alphabetically by Latin name) in the *hinkley fish* data set. The yellow lines on the plots are linear regression plots. These will help to point out the variables that are positively or negatively correlated.



**Part**

---

**XVI**

## 16 Printing, editing and saving results

Output, both graphical and text, can be saved as a file, copied to the clipboard or printed. CAP also offers a wide variety of options for editing and designing your graphs, including pre-defined [themes](#) <sup>[167]</sup>. See topics below for further details:

[Exporting and copying charts](#) <sup>[153]</sup>

[Exporting dendrograms](#) <sup>[154]</sup>

[Printing charts and dendrograms](#) <sup>[155]</sup>

[Editing charts](#) <sup>[157]</sup>

[Zooming on charts](#) <sup>[158]</sup>

[Preparing charts for output](#) <sup>[162]</sup>

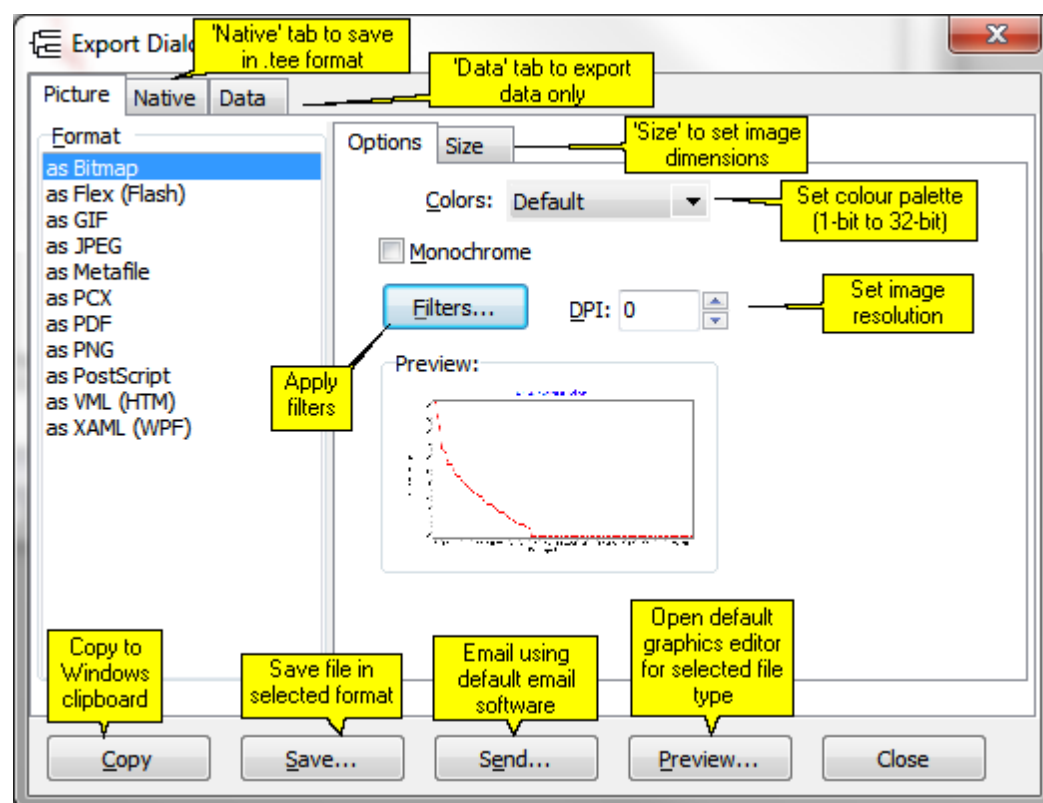
[Themes for graphs](#) <sup>[167]</sup>

[Printing text and grid output](#) <sup>[168]</sup>

For fuller information and a video demonstration see the **Help: Guides: Output**.

### 16.1 Exporting and copying charts

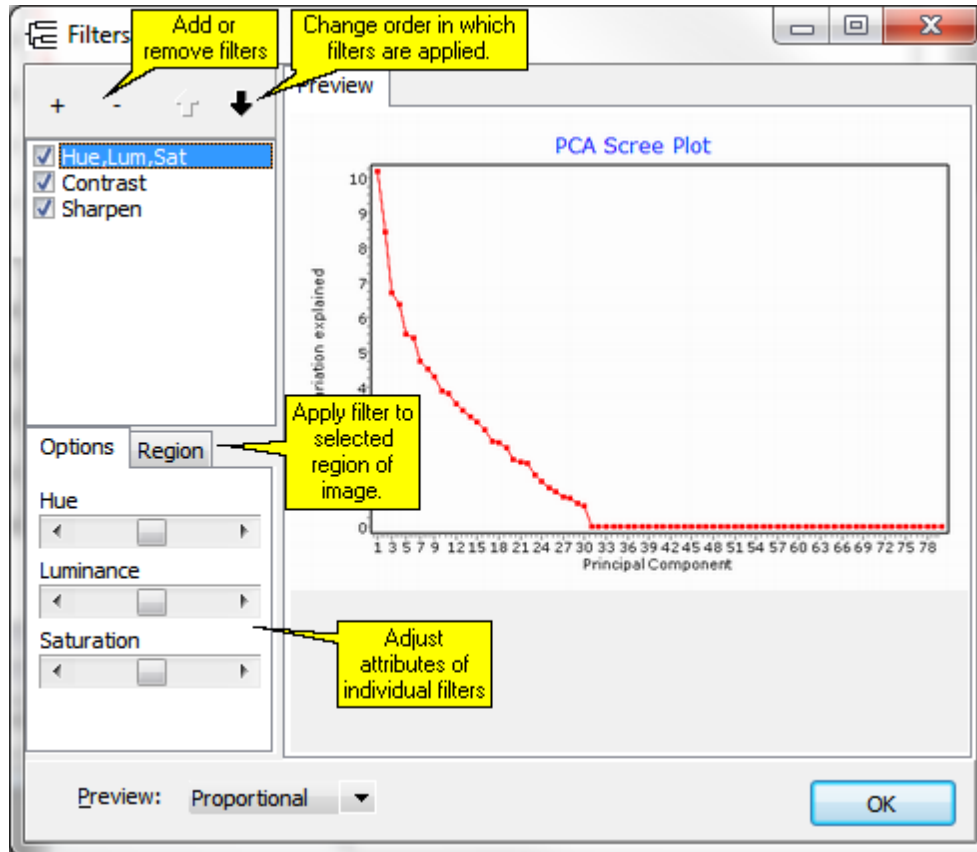
To export the image of a chart select **File: Export**. This will open the Export Dialog.



The chart can be saved in a wide range of different file formats, including Enhanced Metafile (\*.emf), Bitmap (\*.bmp), JPEG (\*.jpg), PCX, PNG, GIF or Native (\*.tee). Each file format has advantages and disadvantages. Enhanced Metafiles, when pasted into a Word document, can be resized by dragging, without losing resolution. Bitmaps are a lossless method of saving; the stored file will not lose any of the original's detail; however, the file size will be much larger than compressed files such as Enhanced Metafiles or JPEGs. JPEGs are file formats which can be compressed to take up less space - useful if you wish to send one by email, put it on a website, or paste it in to a document. If they are compressed too heavily, they can lose resolution and detail, and spoil colours. GIFs are also compressed files useful for web sites; they have a considerably smaller colour palette than JPEGs or BMPs; 256 colours, as opposed to many millions. This means that while a JPEG or BMP

image can show a smooth gradation of colour (for instance in a graded background), saved as a GIF image, it will appear broken up into jagged zones of colour. GIF images are therefore better suited to images showing large discrete blocks of single colours. The Native (\*.tee) format saves all the chart attributes, and the data series, rather than the image itself. This means that you can save the chart, and open it again at a later date to edit it, using the free TeeReader software supplied on the installation CD. Tee files tend to be very small indeed; often less than 1KB.

If you select the Bitmap (.bmp) format, you can apply a range of filters to the image:



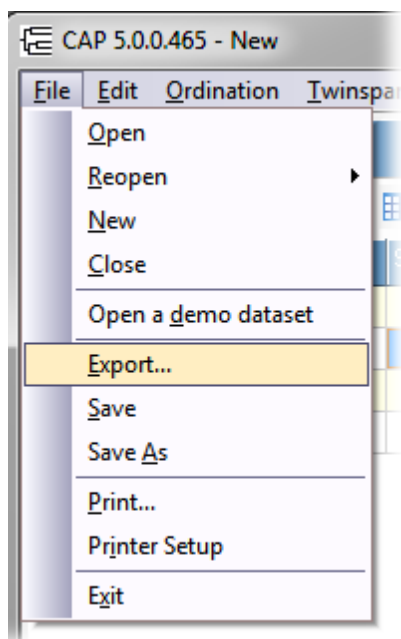
This export dialog can also be used to export the data shown on the chart, using the 'Data' tab.

You can also copy a chart to the Windows clipboard using the Copy button on the graphics toolbar (4th button from left), or by pressing **Ctrl-C** on your keyboard.

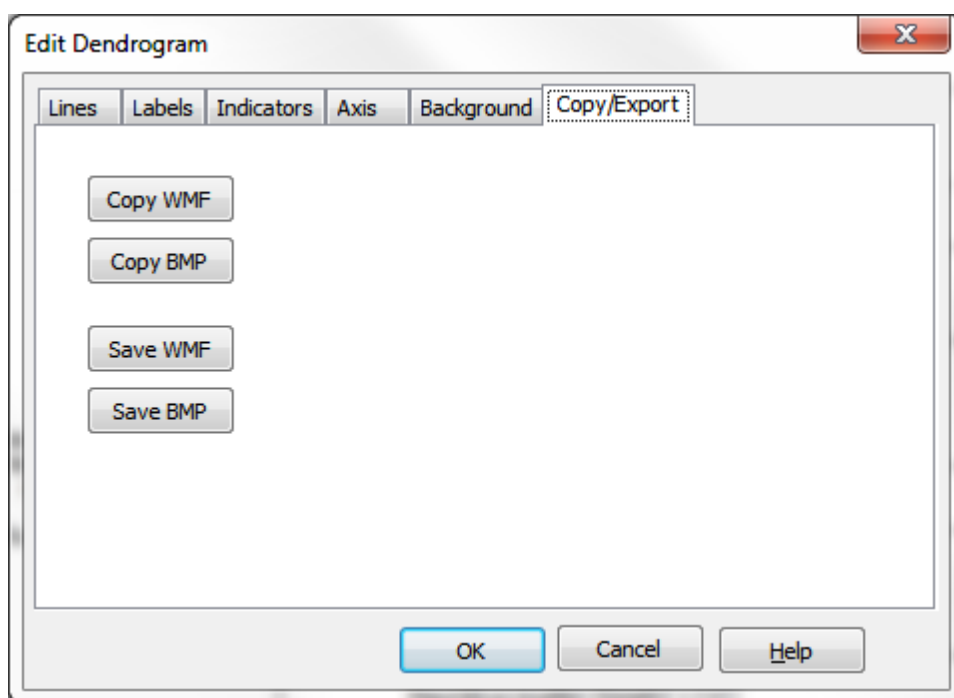


## 16.2 Exporting dendrograms

Use **File: Export**, **Edit: Copy** or **Ctrl-C** on your keyboard to copy the dendrogram to the Windows clipboard, from where it can be pasted into other Windows applications. The chart will be exported in Metafile format; when pasted into a Microsoft Word document, it can be resized by dragging the corner or edge of the image, without distorting or losing resolution.



Dendrograms can also be exported as a Metafile or Bitmap from the Edit Dendrogram button:



## 16.3 Printing charts and dendrograms

Dendrograms can be printed from the **Print dendrogram** button, or by clicking **File: Print**. Charts can be printed by using the **Print** button on the graphics toolbar:

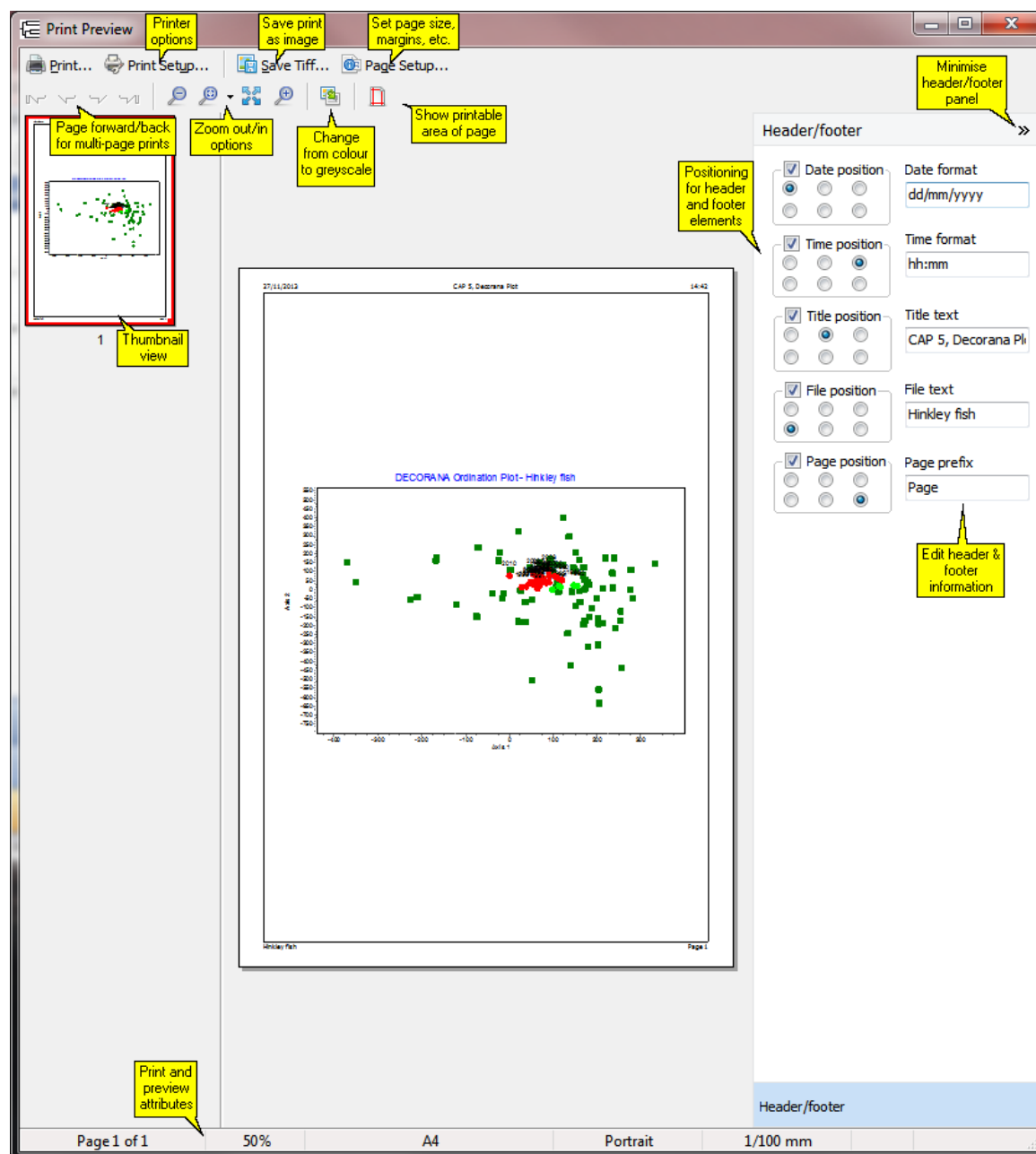


or by clicking **File: Print**. In both cases, this will bring up the Print Preview dialog, where you can alter many attributes of the printed page.



## Print preview

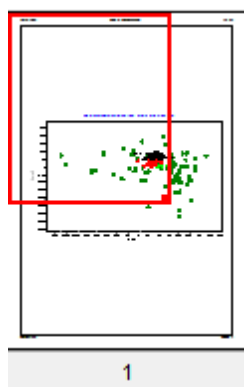
The Print Preview dialog holds many options for previewing and completing the job of printing your document:



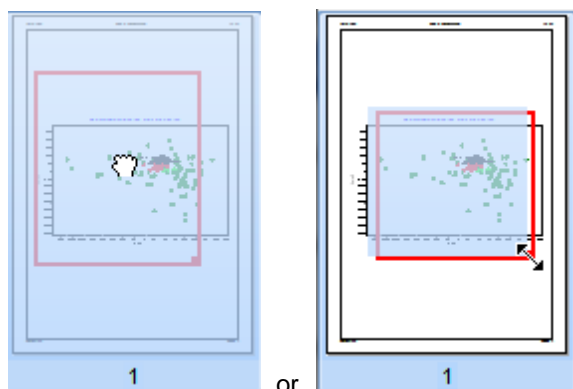
In the case of very large dendrograms, it is possible that the default page setup of fitting the plot onto a single page, will render the plot so small as to be unreadable. In this case, it may be better to save the plot as a [metafile](#)<sup>[154]</sup>, and print that instead.

## Thumbnail view

The Thumbnail view is useful when you have zoomed in to view a portion of the print; it shows which portion of the page is displayed:



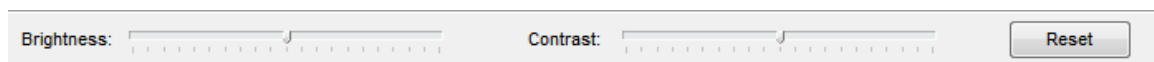
Click and drag on the red box to show a different portion of the page, or click/drag the square at the bottom right hand corner of the red box to zoom in or out further:



(You can also click and drag the main preview page to move it around).

### Colour/greyscale printing

When you switch from Colour to Greyscale view using the Colour/greyscale button, the panel at the bottom of the page allows you to alter the brightness and contrast of the greyscale print:



This panel disappears again when you switch back to a colour print.

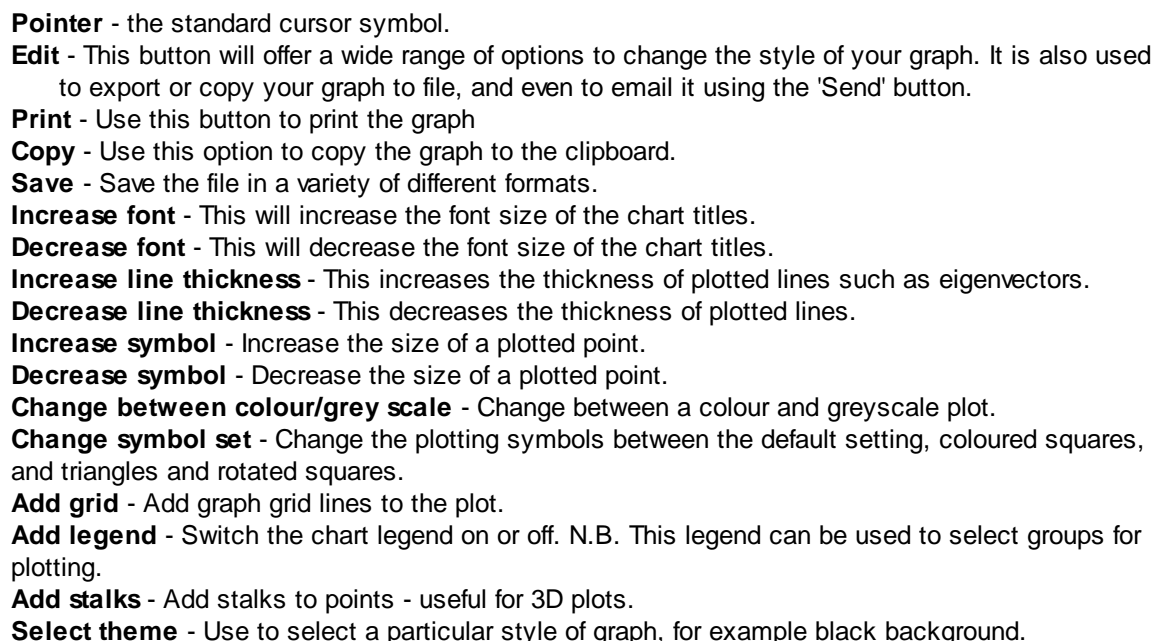
Charts and dendrograms can also be [copied to the clipboard](#)<sup>[154]</sup> using **File: Export, Edit: Copy** or **Ctrl-C** on your keyboard, and then pasted in the normal Windows fashion into a Word document or other suitable document for subsequent printing.

To save the chart as a PDF document, use the [File: Export](#)<sup>[153]</sup> facility and select PDF as the required file type. Alternatively, if you have the full version of Adobe Acrobat (**not** the free Acrobat Reader) installed on your computer, you will be able to convert the chart directly to a .pdf file by Adobe PDF from the list of available printers in the Print dialog box.

## 16.4 Editing charts

Almost every aspect of your graphs can be edited.

The graph option buttons on the Chart Toolbar are described in order from left to right below. A hint will pop up if you hover over a button.

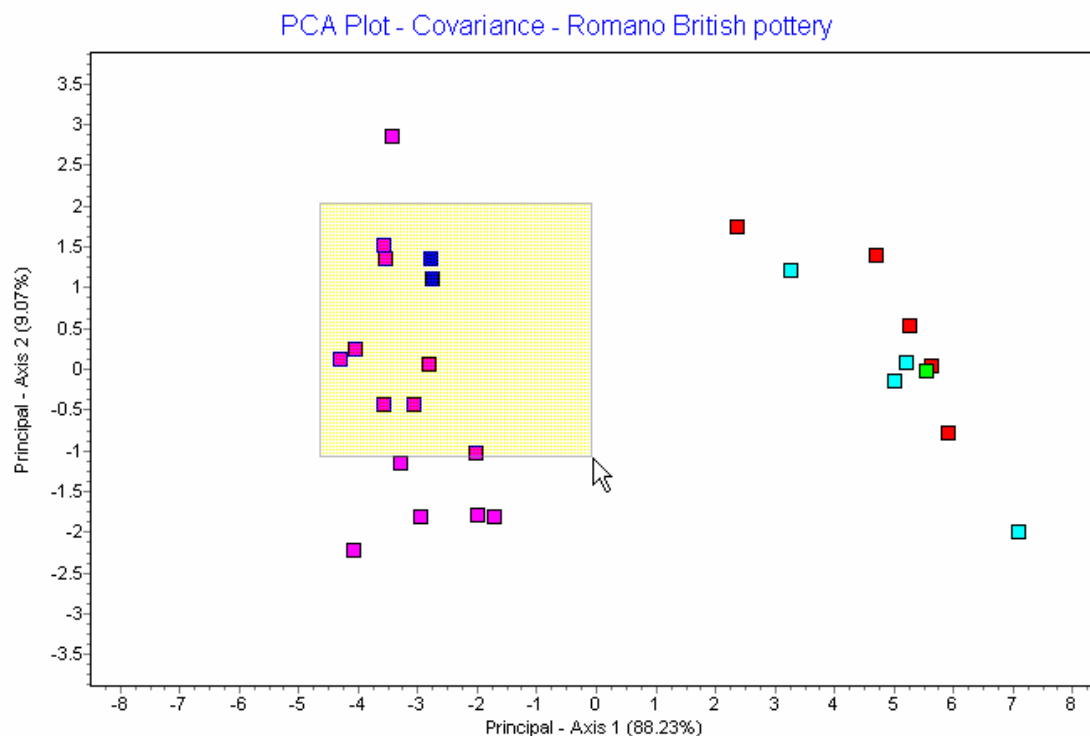


See also:

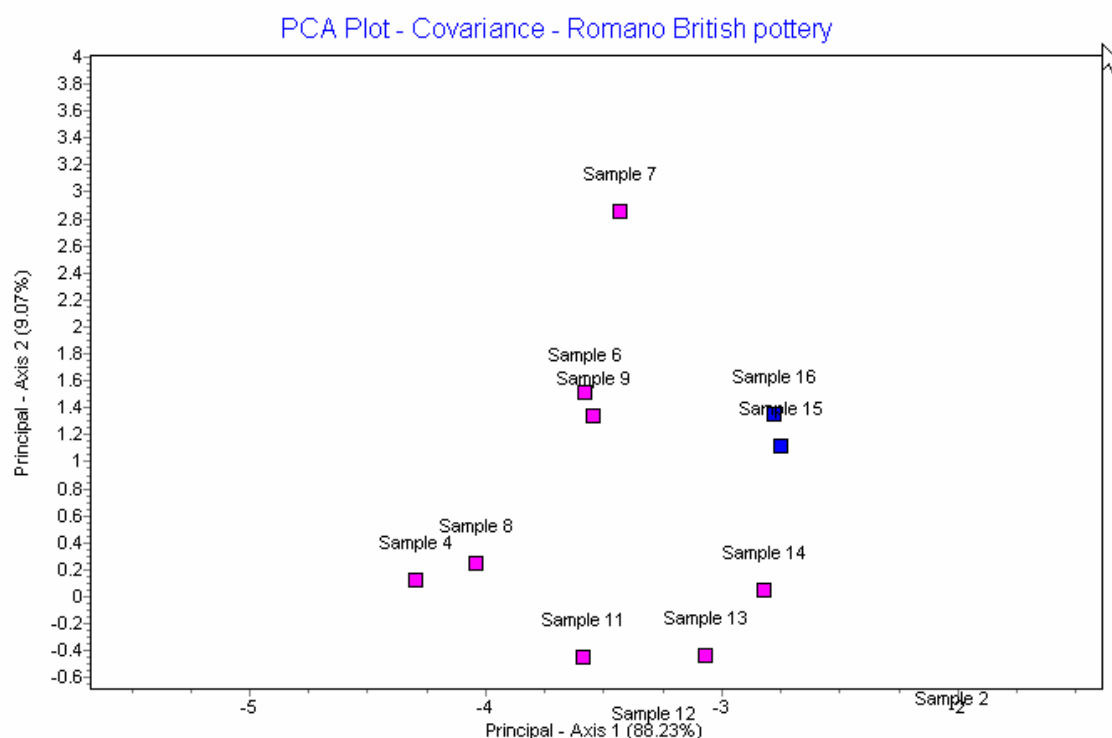
- [The Editing Chart dialog](#)<sup>[160]</sup>
- [Drawing a perimeter around a group of points](#)<sup>[161]</sup>
- [Preparing charts for output](#)<sup>[162]</sup>
- [Themes for charts](#)<sup>[167]</sup>
- [Zooming on charts](#)<sup>[158]</sup>
- [Printing charts and dendrograms](#)<sup>[155]</sup>

Tight clusters of points which cannot be differentiated can occur in, for example, the PCA plot. To zoom in on an area, move to the top left corner of the area to be enlarged, then hold down the left hand mouse button and drag to the lower right hand corner of the area you require, and release the button. An enlarged view of the selected area will be displayed.

As you drag down from upper left to lower right corner of the area to be enlarged the area will appear as a coloured panel on the plot.



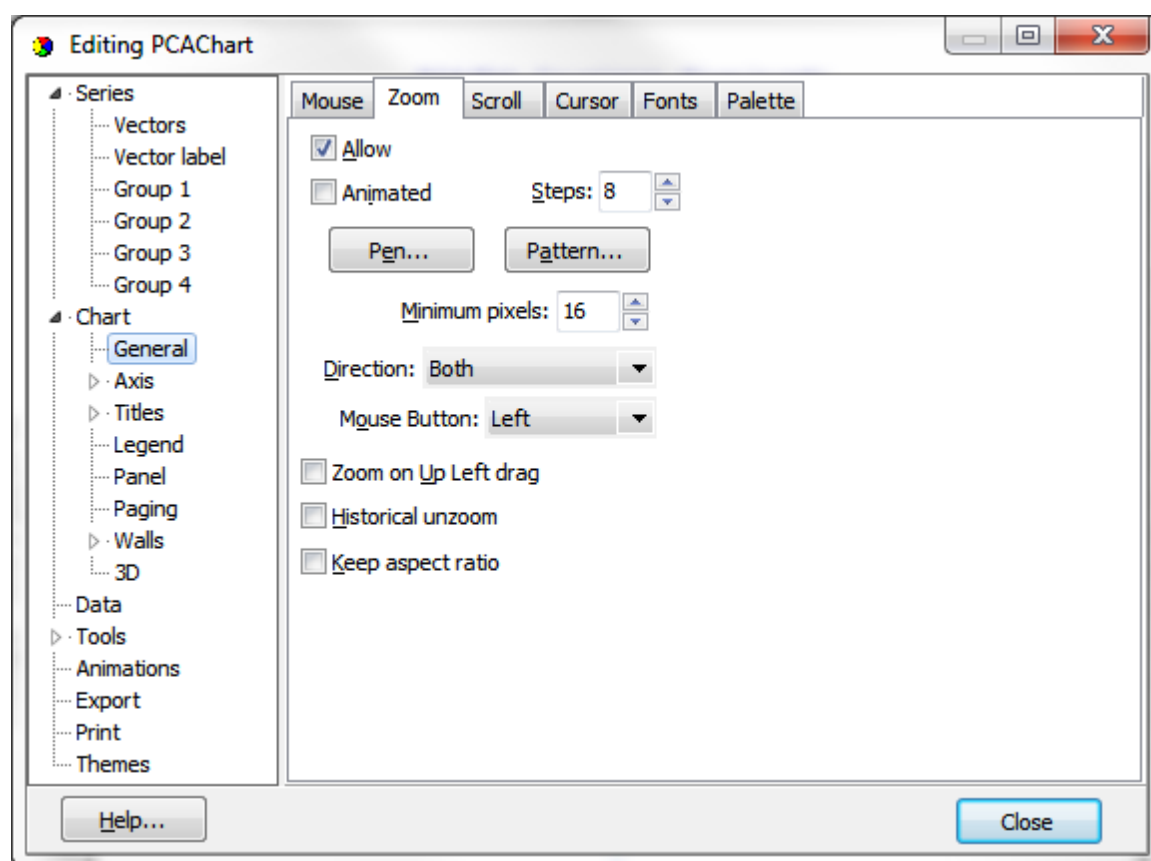
When you release the button this area will fill the plot. Using this zoom facility can allow you to read labels and distinguish points that are very close together.



To return to the original view, hold down the left hand mouse button and move upwards from the lower right to the upper left corner and release.

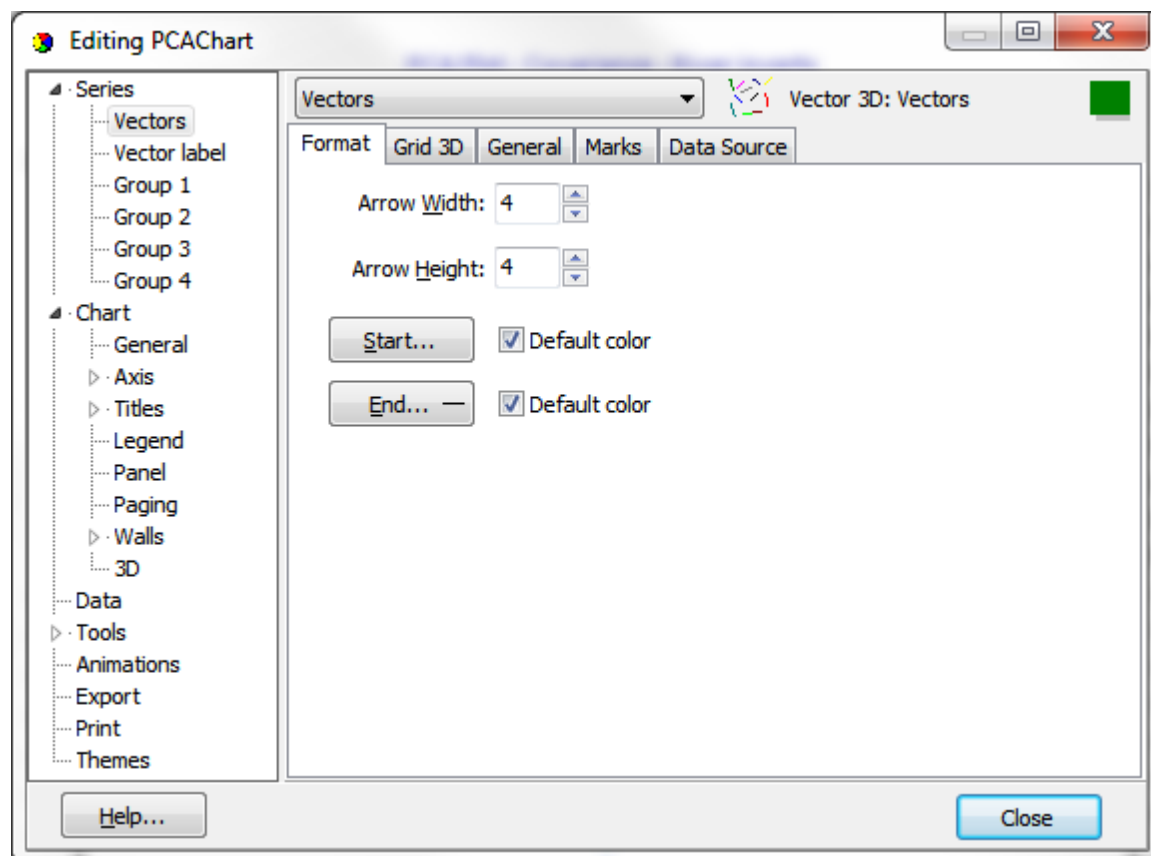
All aspects of zooming behaviour can be edited; click the Edit Chart button, and on the Editing Chart

dialog choose **Chart: General: Zoom:**



### 16.4.2 The Editing Chart dialog

The Editing Chart dialog is an extremely versatile tool which allows you to alter every single aspect of your chart, as well as exporting it, saving it and printing. The dialog has an expanding menu in the left-hand pane with the options that can be changed arranged into groups:



**Series** - These are the actual plotted components of the chart; the data points and (in charts that include them) centroids, and the gradients or vectors along which the data points are arranged.

**Chart** - All the other components of your graph; the axes, title, legend and captions, fonts, frames and borders, as well as the colour scheme, 3D plotting, and additional items like mouse and cursor behaviour, [zooming](#)<sup>[158]</sup> and scrolling.

**Data** - all of the data on which the plot is based.

**Tools** - a wide range of tools that you can use to further manipulate the plot. For more details, see [Preparing charts for output](#)<sup>[162]</sup>.

**Export** - allows you to [export the chart](#)<sup>[153]</sup> in a wide range of different formats, or send it by email.

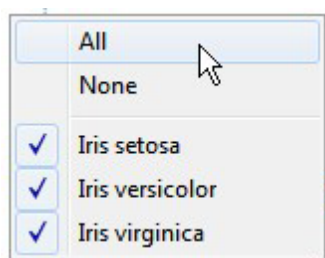
**Print** - Send the chart to a printer.

**Themes** - apply a range of [pre-defined themes](#)<sup>[167]</sup> to your chart.

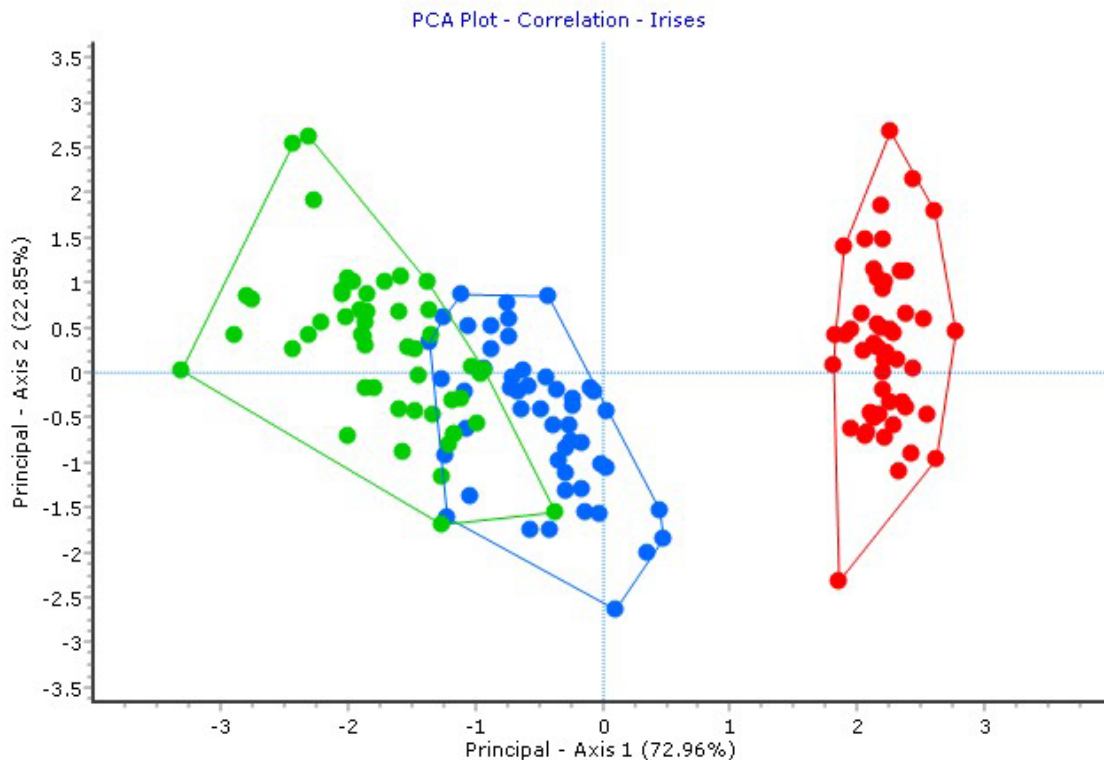
For further help on the Editing Chart dialog, click the Help button to launch the dedicated TeeChart help system.

### 16.4.3 Drawing a perimeter

It is possible to display the perimeter of each predefined group. Right-click on the plot to display a menu that enables you to choose which group(s) you wish to outline. Remember you must have already [defined group membership](#)<sup>[53]</sup>.



In this example, we have carried out a PCA-Correlation on the *Iris* demo data set. Select from the menu which groups you wish to add a perimeter to. Here is the result:



Note this line is dynamically created and will need to be refreshed if any other changes are made to the plot, such as changing the colour of the data points. This option is also available in 3D plotting, but may not be helpful, depending on the nature of the plot.

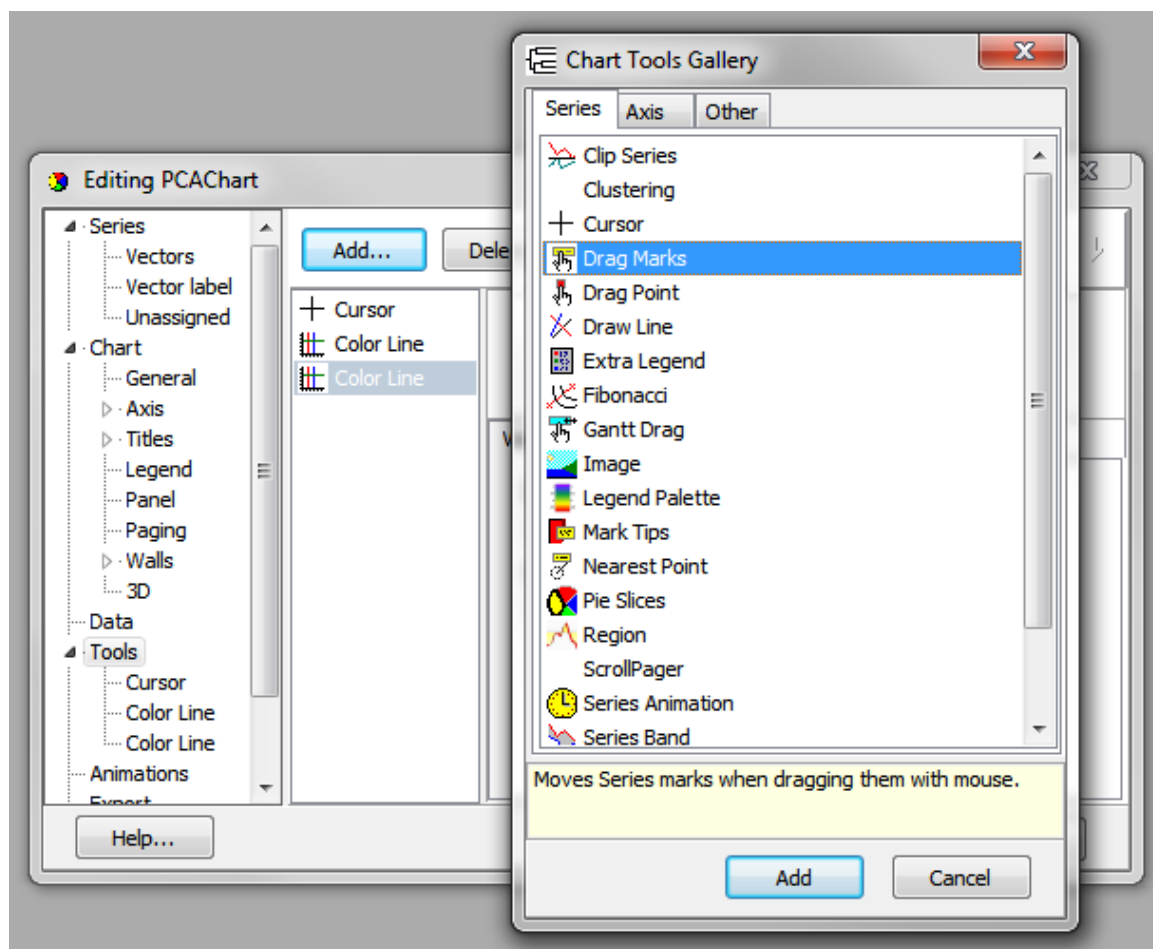
#### 16.4.4 Preparing charts for output - Chart Tools

As well as editing the standard chart options, there are many tools you can use to improve your charts for printing or publication. These appear in the Chart Tools Gallery of the Edit Chart dialog. We feature a few below; there are many more to explore.

##### Moving Chart Labels

Click the Edit button to change a wide range of aspects of the graph, add titles, and use various tools to customise the chart. One of the most useful tools allows you to drag the chart labels - very useful if you have a cluster of points with overlapping labels. Click the Edit button, then the Tools tab of the Edit dialog, the Add... button, and choose 'Drag Marks'. Click the "Add" button, then on the next dialog box, click "Close". When the chart is displayed, you will find that the cursor displays as a hand symbol on hovering over a data label, allowing you to drag the label to the desired position.

**NOTE:** any alteration of other aspects of the graph will cause the labels to return to their original position. Therefore, moving the labels should be the **last** alteration you make before printing or exporting the graph.



### Drawing Lines

If the new [group perimeter](#)<sup>[16]</sup> method is not appropriate to your needs, you may find it useful to draw lines on a chart, to separate groups or clusters of points, for instance on an MDS plot (below).

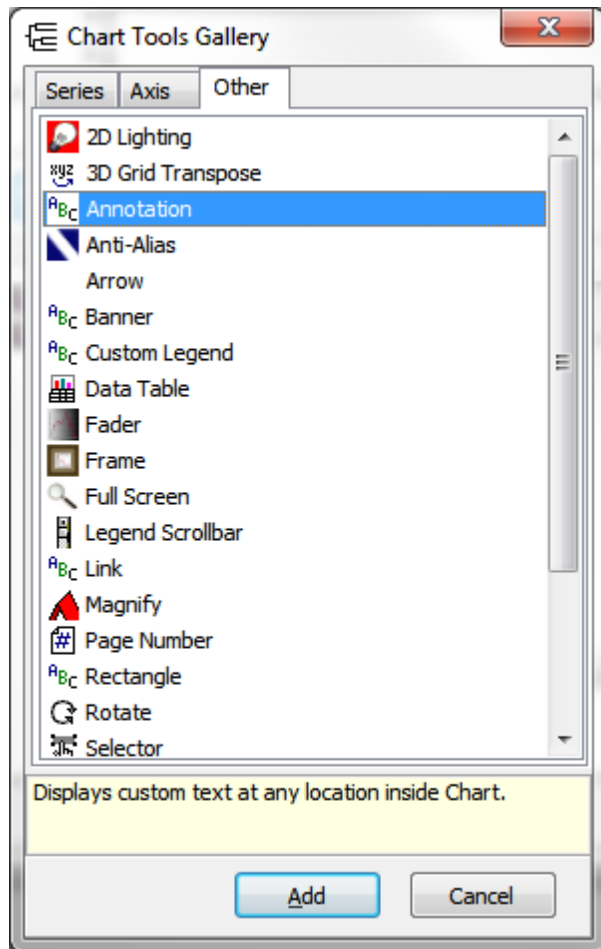
- Click the Chart Edit button, choose the Tools tab - Add... button, and from the list of tools, choose Draw Line. If you click on the Pen button, you can specify the type and colour of the line.
- When you have made your choice, click 'Close' to return to the program, where you can simply draw lines with the mouse. Lines, once drawn, can be dragged to the position required.
- To alter the line type or colour, open the Tools tab again, click on Draw Line, and make the changes you require.
- To remove lines from a plot, go to the Tools tab, click on Draw Line, and press the Delete button.





Annotations can be very useful if, for instance, there are a large number of points on your chart and you wish only to label a few of them. To add annotations:

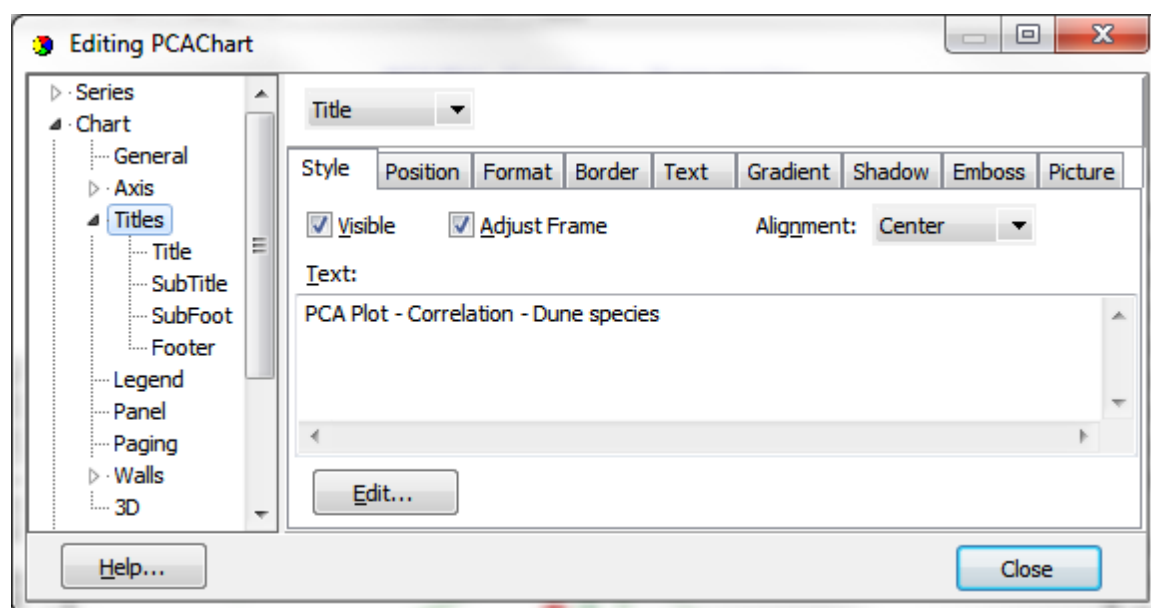
- With the chart showing, click on the Edit button to show the editing screen, select the Tools tab; click Add, and from the 'Other' tab on the Chart Tools Gallery screen, select Annotation from the list of tools.



- Enter the text you want on your annotation into the text box on the Options tab, and then use Position: Custom to manoeuvre the label into the desired position. Use the options on the other tabs to change font, shading, outline and other details.
- To add another annotation, click Add, and select Annotation again.

### **Altering chart titles**

- With the chart showing, click on the set-square icon to show the editing screen. Click on the Titles tab (on the Chart tab) to show the following screen:

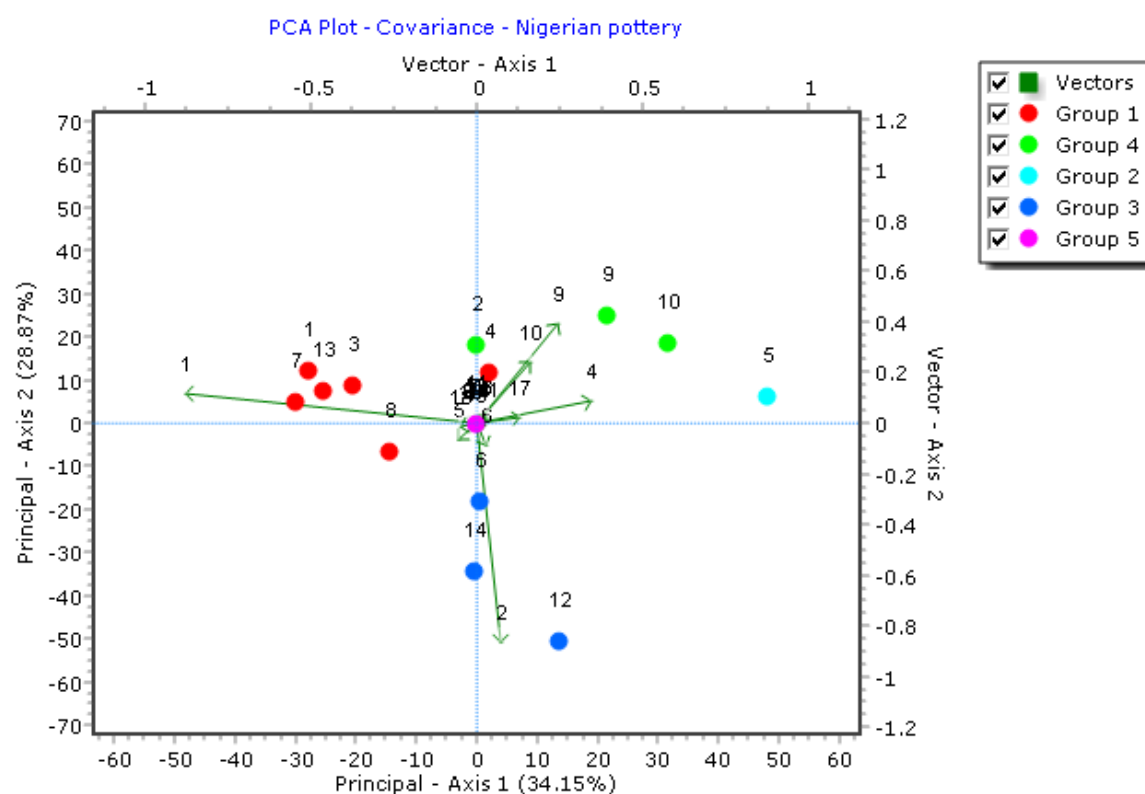


Enter the new title text into the box and click 'Close' to save.

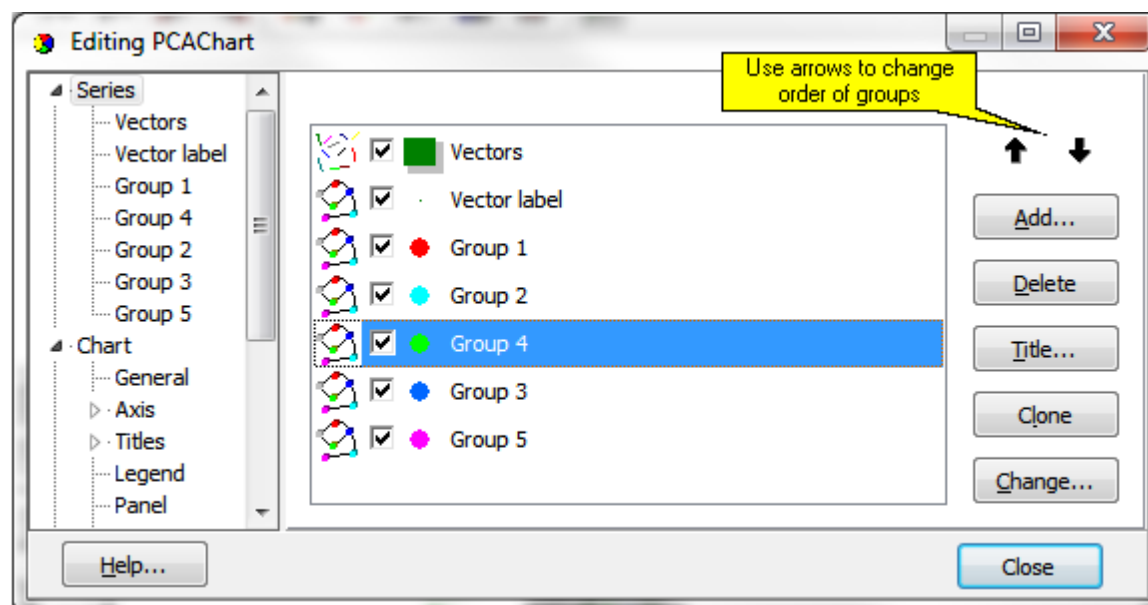
Note that there are very many options on the Edit screen and it is worth a few minutes' experimentation to see what sort of effects are available to enhance your charts for publication or display.

### Changing the order of items in the chart legend

When you have performed an analysis, if you enable the plot legend, the groups will be displayed on the legend:



If you wish, you can change the order in which the Groups are displayed in the legend; click the Edit Chart button, and on the Series tab, select a Group, and use the Up or Down arrow to move it into the order you want. When an item is shifted up or down the list, its position on the chart legend will change accordingly.



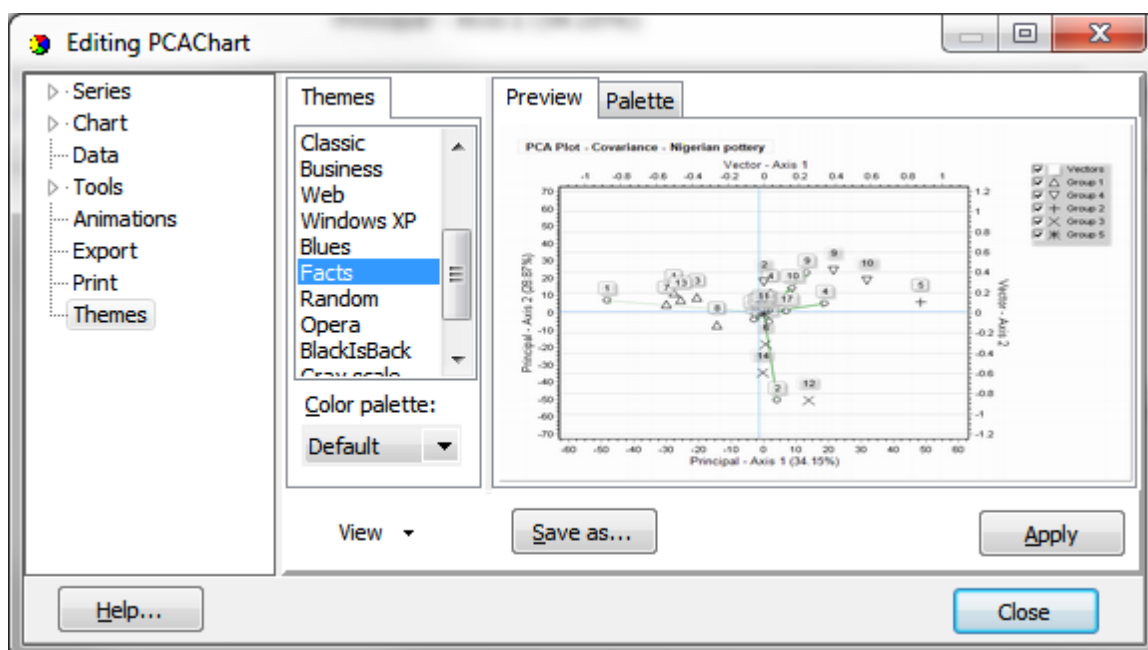
### 16.4.5 Themes for charts

Themes are general graph plotting styles that the user can select or make.

To choose a theme, click on the Themes button at the end of the graphics tool bar.



This will display the themes dialog with a number of preset themes ready for selection:



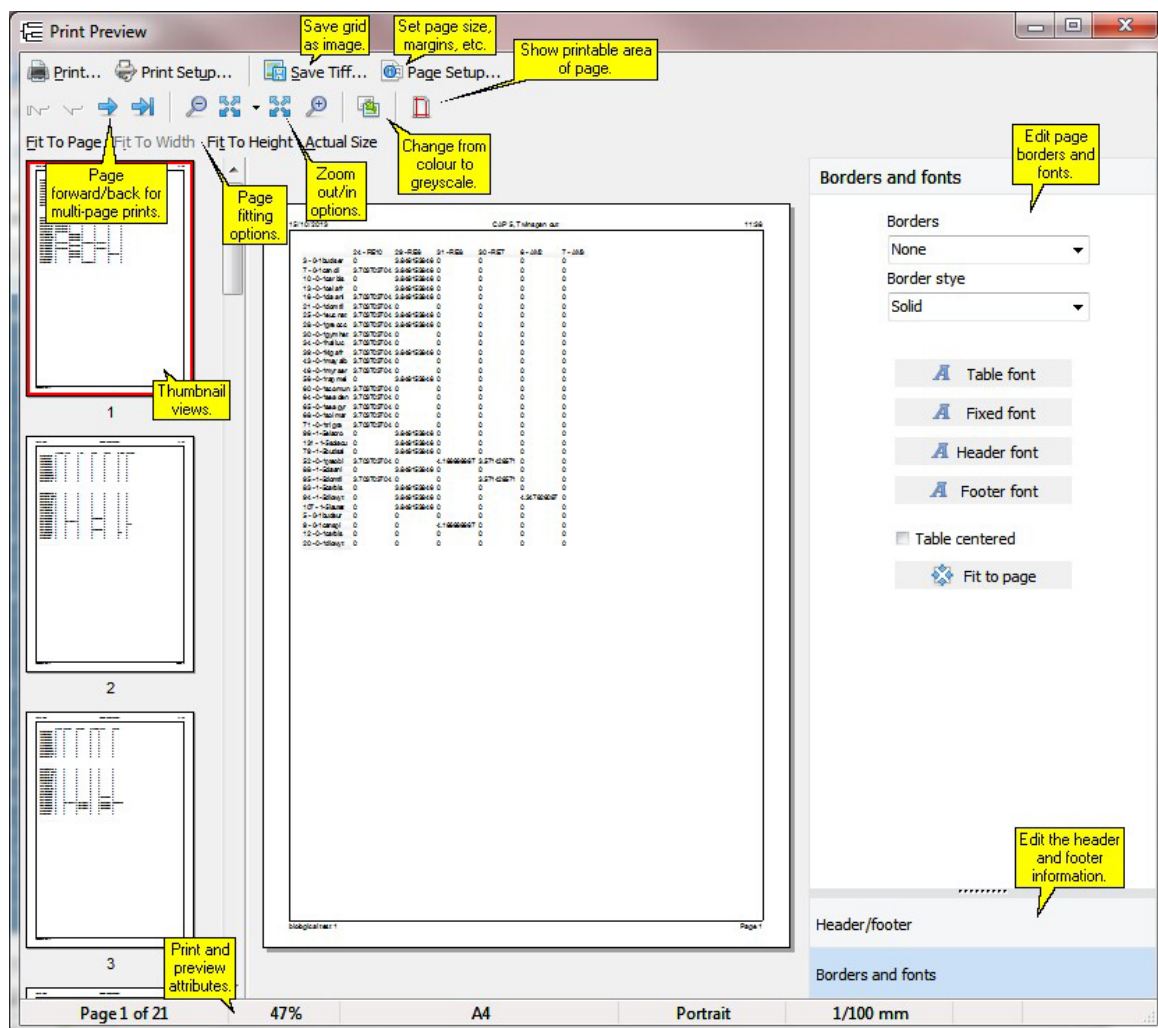
Once a theme has been applied to a chart, it can be difficult to return to the plain default style used by the program, particularly if you have made modifications of your own to labels and legends, etc. Try selecting the theme titled "Pisces General"; failing that, close the program and re-perform the analysis to regain the default style of chart.

## 16.5 Printing and exporting grid and text output

### Printing, saving and exporting grid output.

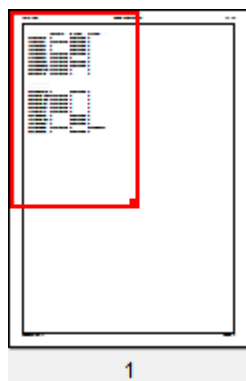
#### 1. Printing

With an output grid showing simply choose **File: Print**, and the Print Preview dialog will be activated:

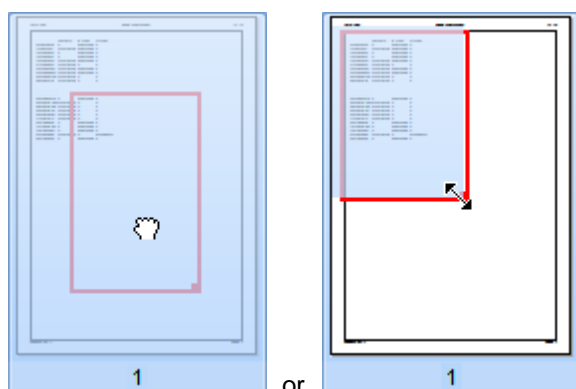


#### Thumbnail view

The Thumbnail view is useful when you have zoomed in to view a portion of the print; it shows which portion of the page is displayed:



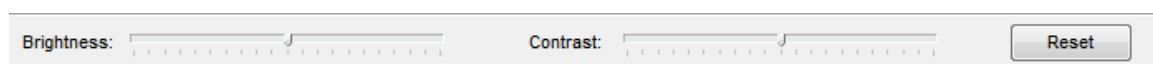
Click and drag on the red box to show a different portion of the page, or click/drag the square at the bottom right hand corner of the red box to zoom in or out further:



(You can also click and drag the main preview page to move it around).

### Colour/greyscale printing

When you switch from Colour to Greyscale view using the Colour/greyscale button, the panel at the bottom of the page allows you to alter the brightness and contrast of the greyscale print:



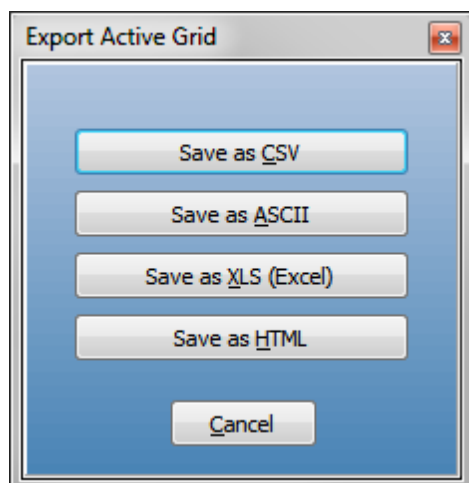
This panel disappears again when you switch back to a colour print.

### 2. Saving a grid as an image.

To save the grid as an image, click 'Save Tiff' on the Print Preview dialog; it will be saved as a .tif image, with multiple pages if necessary.

### 3. Exporting a grid.

You can save any of the grids in a variety of file formats. To save the results to a file choose **File: Export** and select the format you want from the dialogue box. You will be able to save the text as a comma delimited file (\*.csv), ASCII text file (\*.txt), Excel file (\*.xls) or HTML:



Output grids can also be copied to the clipboard using **Edit: Copy** or **Ctrl-C** on your keyboard.

### **Printing, saving and exporting text output.**

#### **1. Printing text output.**

Text output can be printed in exactly the same way as the plots and grids; click **File: Print** to open up the Print Preview dialog, which is explained above.

#### **2. Saving text output.**

Text output such as [Twinspan Text](#)<sup>[89]</sup>, [DECORANA Computations](#)<sup>[71]</sup>, [Reciprocal Averaging Computations](#)<sup>[83]</sup> or the [Significance Tests](#)<sup>[136]</sup> in Discriminant Analysis can be saved as a Rich Text (rtf) file, which is openable in most word processors. Click **File: Export**, and choose the file name and location to save the .rtf.

#### **3. Saving text output as an image.**

To save the text output as an image, click 'Save Tiff' on the Print Preview dialog; it will be saved as a .tif image, with multiple pages if necessary.

# Part

---





## 17 Obtaining help

For most active windows context sensitive help can be obtained by pressing F1, clicking on the Help button or selecting the Help drop-down menu. or clicking on the right-hand mouse button and choosing help from the pop-up menu. If pressing F1, make sure that the window that you are seeking help for is the active one.

CAP also includes 'Instant Assist' - a continuously visible advisory panel that tells you about the methods available on the main menu bar. This can be placed anywhere on your screen while you work and switched on or off under the Help menu.

When learning to use CAP you will find the video guides invaluable.



If you have problems using the program or entering data which you cannot solve then contact Pisces Conservation Ltd by e-mailing [pisces@pisces-conservation.com](mailto:pisces@pisces-conservation.com) or by phone to +44 (0)1590 674000 during office hours (09.00 to 17.00 GMT/BST).

PISCES Conservation Ltd,  
IRC House, The Square  
Pennington, Lymington  
Hants, SO41 8GN  
UK

Telephone      44 (0) 1590 674000  
Fax              44 (0) 1590 675599

For more information, details of our other software, and answers to a range of technical queries, visit our web site at <http://www.pisces-conservation.com>

For details about our consultancy and other work, visit <http://consult.pisces-conservation.com>

### 17.1 References

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

Benzécri, J.P. 1967. Statistical analysis as a tool to make patterns emerge from data. In: Watanabe, S. (ed) *Methodologies of Pattern Recognition*. Academic Press, New York, 35-60.

Clarke, K. R. 1988. Detecting change in benthic community structure. 131-142 in R. Oger [ed.} *Proceedings of invited papers, 14th international biometric conference, Namour, Belgium*.

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. *Aust.*

J. Ecol. 18, 117-143.

Digby, P. G. N. & Kempton, R. A. 1987. Multivariate analysis of ecological communities. Chapman & Hall, London.

Fischer, R. A. 1940. The precision of discriminant functions. Annals of Eugenics, 10, 422-429.

Hirschfeld, H. O. 1935. A connection between correlation and contingency. Proceedings of the Cambridge Philosophical Society, 31, 520-527.

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

Hill, M. O. 1979 DECORANA--a FORTRAN program for detrended correspondence analysis and reciprocal averaging. Ithaca, NY. Ecology and Systematics, Cornell University.

Kent, M. and Coker, P. (1992). Vegetation description and analysis: A practical approach. John Wiley and Sons, Chichester, 363 pp.

Legendre, L. and Legendre, P. (1983). Numerical Ecology. Elsevier Scientific Publishing Company, Amsterdam, 419 pp.

Mahalanobis, P.C. (1936). On the generalised distance in statistics, Proceedings of the National Institute of Science of India 12, 49-55.

McCune, B. 1994. Improving community analysis with the Beals smoothing function. Ecoscience 1: 82-86.

Kruskal, J. B. (1964) Nonmetric multidimensional scaling: a numerical method. Psychometrika 29, 115-129.

Kruskal, J. B. & Wish, M. (1977) Multidimensional Scaling. Sage Publications. Beverly Hills. CA. USA.

Roux, G. & Roux, M. 1967. A propos de quelques methodes de classification en phytosociologie. Revue de statistique Appliquee. 15, 59-72.

Sokal, R. R. & Sneath, P. H. A. (1963). Principles of numerical taxonomy. Freeman and Co. San Francisco, 859 pp.

## 17.2 Citation

For publication purposes this product should be cited as follows:

Community Analysis Package Version 6, 2019, Pisces Conservation Ltd. Lymington, UK ([www.pisces-conservation.com](http://www.pisces-conservation.com))

or alternatively, if you prefer:

Henderson, P.A. & Seaby, R. M. H., 2019, Community Analysis Package Version 6, Pisces Conservation Ltd, Lymington, UK.

# Part

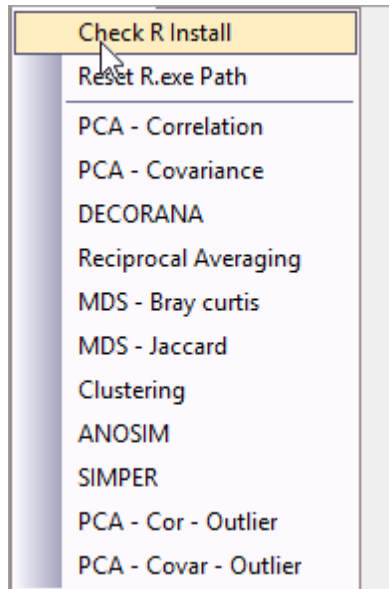
---



## 18 Run R code

Choose this drop-down menu to run an analysis using R. This will only work if you have down-loaded and installed R and the required libraries onto your computer.

[Setting up R](#)<sup>[175]</sup> for use with CAP



[Check R Install](#)<sup>[175]</sup> - Choose to check you have R installed

[Reset R.exe Path](#)<sup>[177]</sup> - Choose to tell CAP where on your computer the R executable is placed

[PCA - Correlation](#)<sup>[179]</sup>

[PCA - Covariance](#)<sup>[180]</sup>

[DECORANA](#)<sup>[181]</sup>

[Reciprocal Averaging](#)<sup>[182]</sup>

[MDS - Bray Curtis](#)<sup>[183]</sup>

[MDS - Jaccard](#)<sup>[184]</sup>

[Clustering](#)<sup>[185]</sup>

[ANOSIM](#)<sup>[186]</sup>

[SIMPER](#)<sup>[187]</sup>

[PCA - Cor - Outlier](#)<sup>[188]</sup> - This will calculate [Mahalanobis distances](#)<sup>[122]</sup>

[PCA - Covar - Outlier](#)<sup>[190]</sup> - This will calculate [Mahalanobis distances](#)<sup>[122]</sup>

### 18.1 Setting up R

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To use R for multivariate analyses in CAP you will need to undertake the following.

First, if you have not already got R you will need to install it on your computer and note the directory where it is installed. You can use the default or choose your own location.

You will find many useful YouTube videos taking you through installation, for example:

<https://www.youtube.com/watch?v=MFfRQuQKGyG> for Windows

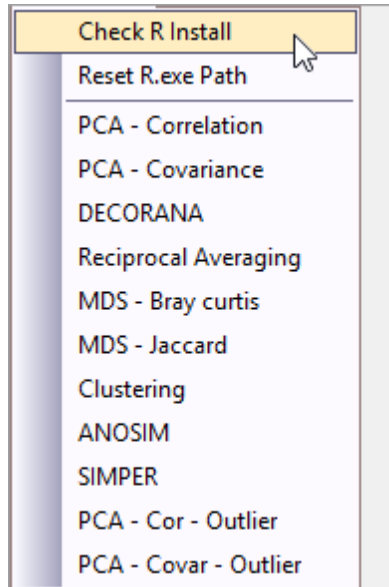
To download R, choose your preferred CRAN mirror at <https://cran.r-project.org/mirrors.html> (any will do).

For example, download and install the current Windows version of R from <https://cran.r-project.org/>

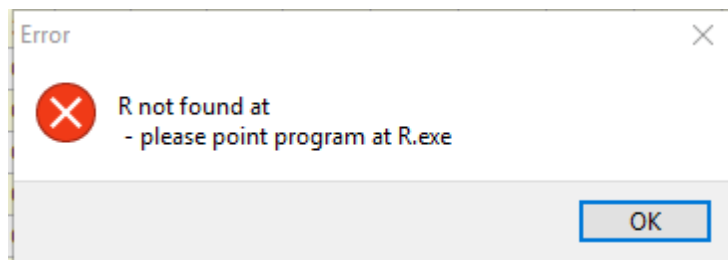
[bin/windows/base/](#)

When installing, you can usually just accept the default settings.

Within CAP run Check R Install from the drop-down menu. **NOTE:** a data set needs to be loaded in CAP before the Check R Install function will work.

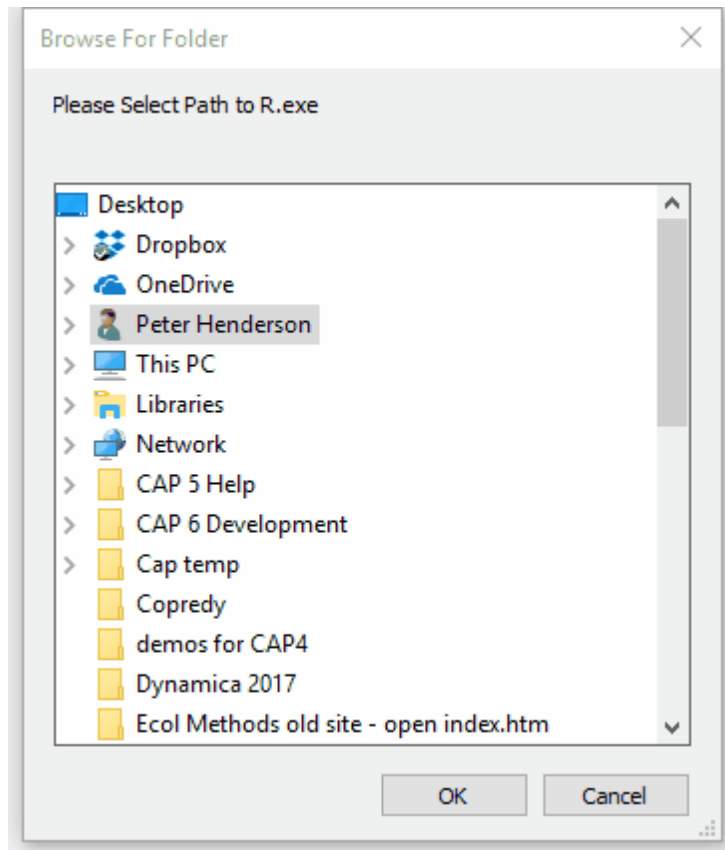


If it is not found you will get an *R not found* error and will be asked to tell CAP where R is on your machine.



If R is not found you will be asked to select the path to the file **R.exe**.

**For example my path is C:\Program Files\R\R-3.4.3\bin**



If you need to reinstall R or get CAP to work with a new R update select [Reset R exe Path](#)<sup>[177]</sup>

### 18.1.1 Upgrading R

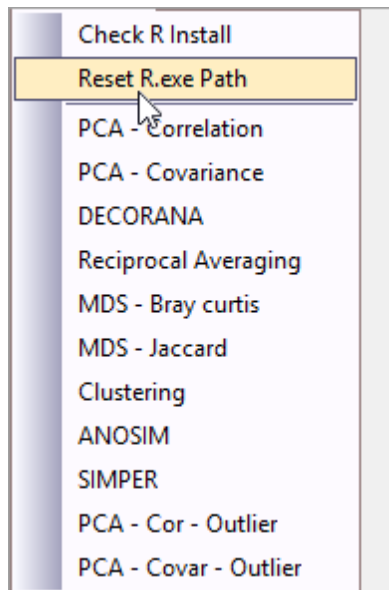
If you should upgrade R to a newer version after you have installed R you will need to do the following.

1. Reset the R exe path to the folder holding the new version of the R executable file. Select [Reset R exe Path](#)<sup>[177]</sup>
2. Now find the folder R library in your documents folder and delete it.
3. Now select [Check R install](#)<sup>[175]</sup> and run the installation to install the R packages used by CAP.

## 18.2 Reset R exe Path

Select this drop-down option if you wish to select a new directory where CAP will access the R executable code.

You will need this option if you install an R update to ensure CAP uses the new R version.



As an example a typical default location could be *C:\Program Files\R\R-3.6.0\bin*  
In this location there will be a file *R.exe*.

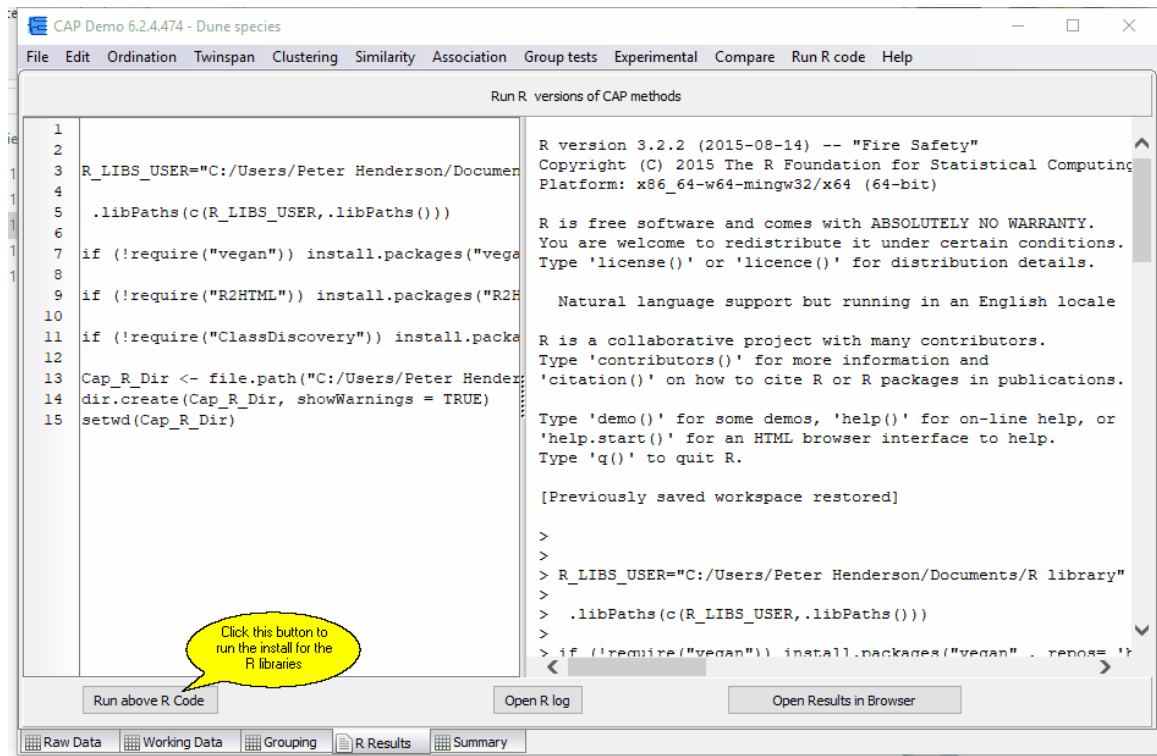
## 18.3 Installing R packages

To undertake the various analyses you will also need to have a number of packages installed.  
For example the **vegan** package.

CAP will do this for you. Once R is installed, selecting **Check R Install** will produce the following screen. Just click on the **Run above R Code** button and the required R code listing to install the packages will be run for you.

Remember, you can go online and get help for all R packages.

If the packages are already installed on your computer, CAP will use these, otherwise it will install the packages.



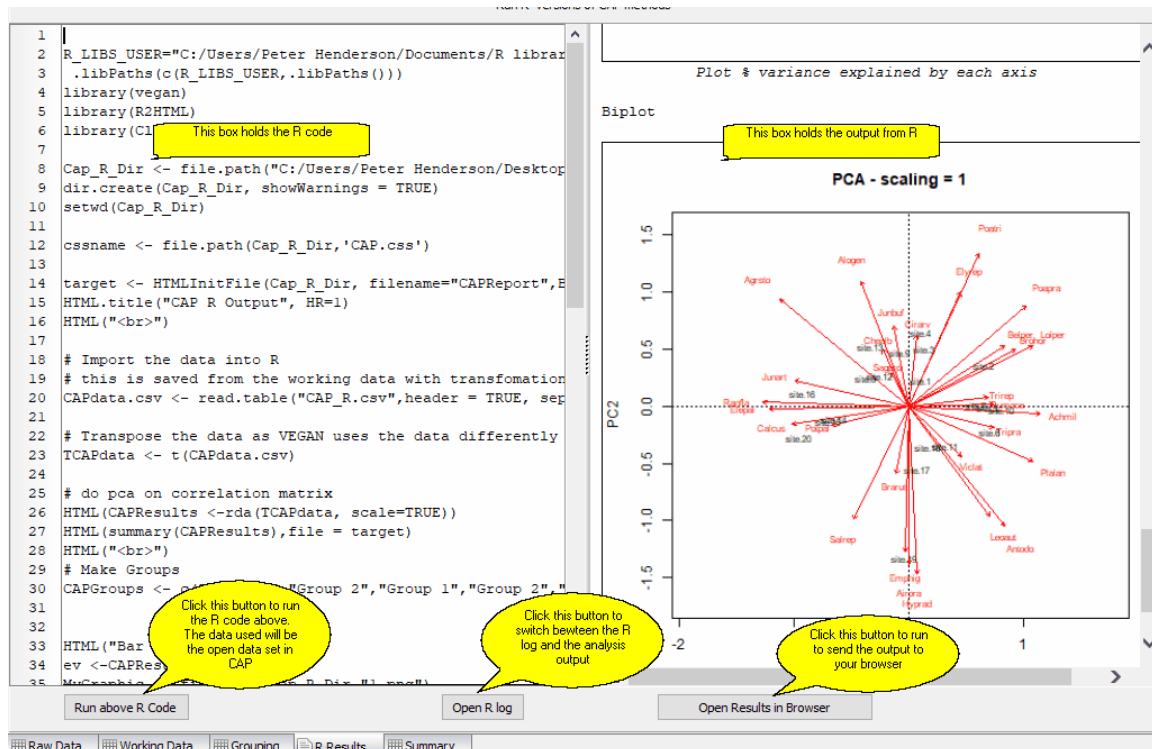
## 18.4 PCA - Correlation in R

Select **PCA - Correlation** to run a Principal Component Analysis using the correlation matrix. The *vegan* package in R is used. The data set used will be your working data.

The relationship between samples (columns) in terms of their variables cannot normally be visualised because this would require a plot with as many axes as there are variables (rows). If your study only includes 3 variables this is possible, but is quite impossible given 4 or more variables or species. PCA is a technique that may summarise the relationship between the samples in a small number of axes that can be plotted. For such a summarisation to work, there must be some degree of correlation between the descriptive variables so that the effect of a number of these variables can be combined into a single axis. For good general introductions to PCA for non-mathematicians see [Kent & Coker<sup>\[172\]</sup> \(1992\)](#) and [Legendre & Legendre<sup>\[172\]</sup> \(1983\)](#).





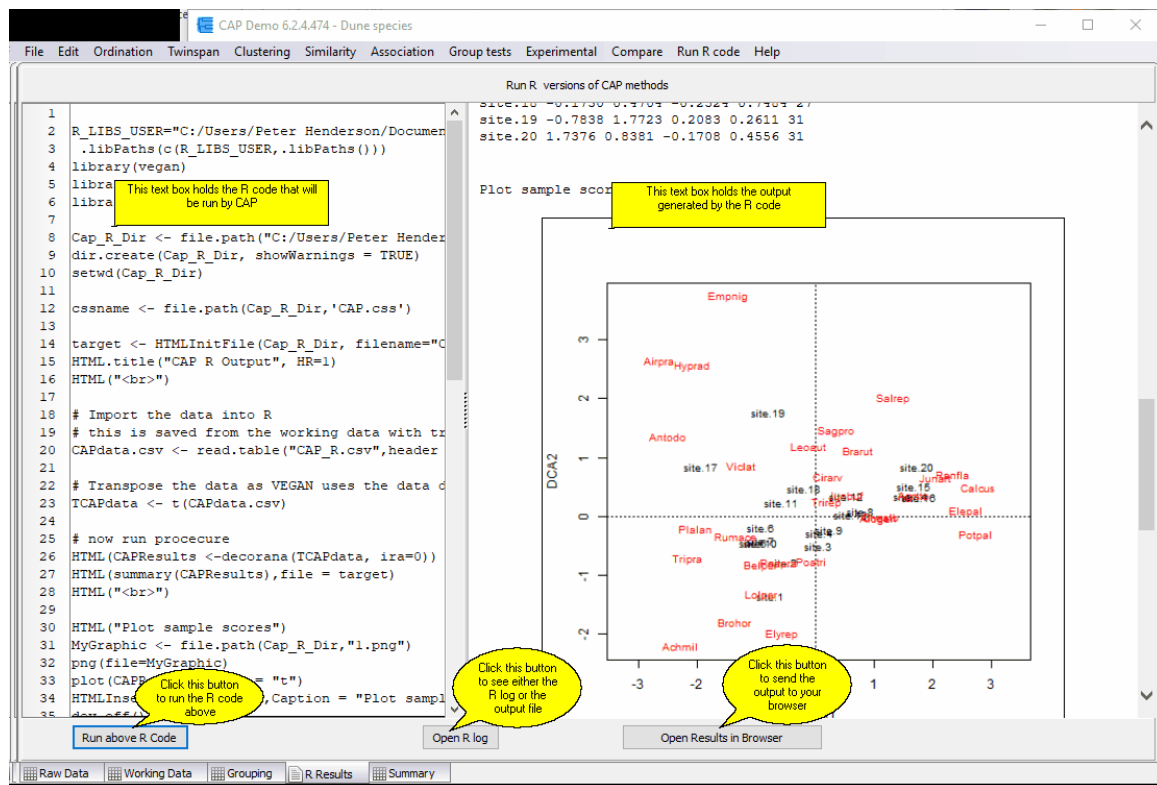


## 18.6 DECORANA R

Select this option to undertake a Detrended Correspondence Analysis using the **vegan** package in R.

The data set used will be your working data.

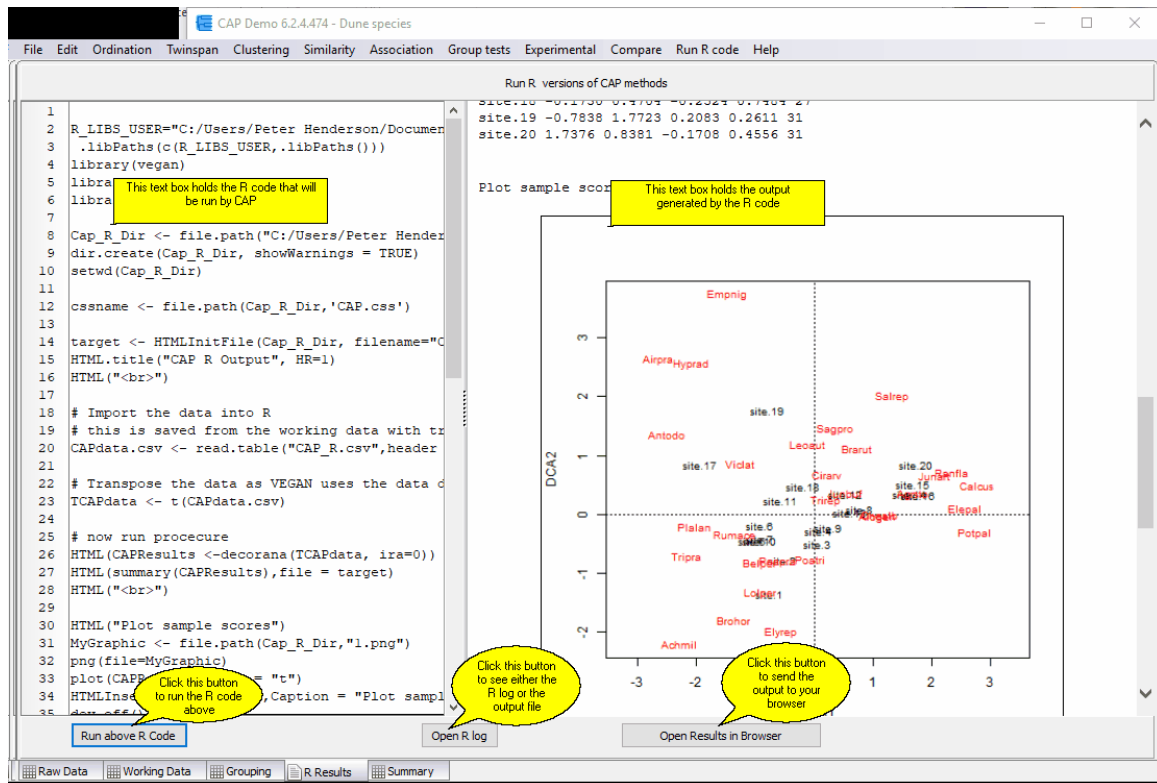
Detrended Correspondence Analysis was devised by Hill (1979) as an attempt to improve upon reciprocal averaging (RA). Two problems that occur with reciprocal averaging are termed the 'arch effect' and 'end point compression'. When the first and second axes produced by RA are plotted, it is often observed that the points are arranged in an arch, because of the quadratic relationship between the axes, rather than because of any relationship between the samples. DECORANA removes this arch by a technique termed detrending. The tendency for points at each end of the first axis to be closer together than those in the middle is removed by segmenting the axis and expanding the terminal segments and compressing those towards the centre. Whereas RA scales the axes between 0 and 100 in relation to the magnitude of the eigenvalue, DECORANA scales in units of average standard deviation of species turnover. Therefore a change of 50% in species composition occurs in about 1 standard deviation.



## 18.7 Reciprocal Averaging

Select this option to undertake Reciprocal Averaging using the **vegan** package in R. The data set used will be your working data.

This method, also called Correspondence Analysis, is a method of showing the relationship between both species and samples (quadrats) in a reduced space. Originally proposed by Hirschfeld (1935) and Fischer (1940) it was first used by ecologists in the 1960s (Roux & Roux, 1967; Benzécri, 1967) - see Kent & Coker (1992) for more details. The method is described by Hill (1973)<sup>[172]</sup> and a non-mathematical introduction to the technique is given in Kent & Coker (1992)<sup>[172]</sup>. RA uses Chi-squared distance values; this results in low abundance species (variables) having a possibly disproportionately large effect on the ordination produced, and can over-emphasise the difference in samples containing several infrequently-recorded species. RA performs best for analysing samples that were collected along an environmental gradient. If there are no clear environmental gradients in the habitat under study, or the gradients are short, then PCA<sup>[63]</sup> may give better results. RA can be applied to both presence/absence and quantitative data.



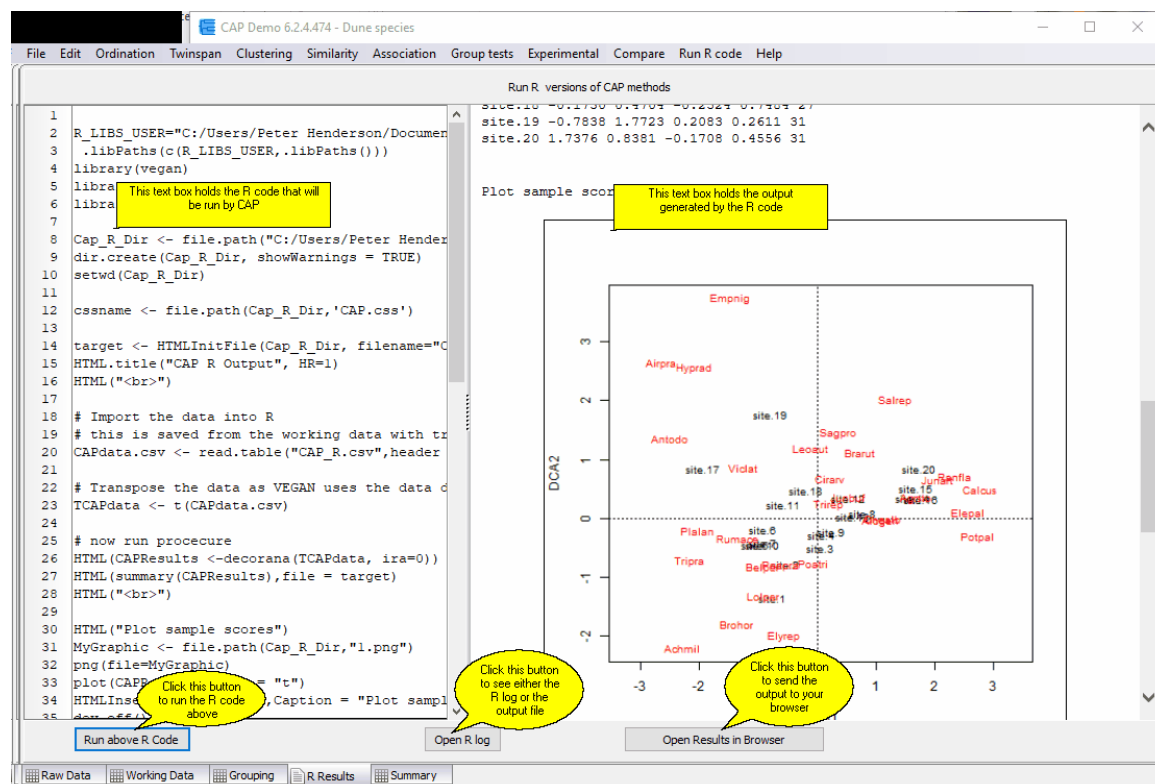
## 18.8 MDS - Bray Curtis R

Select this option to undertake Multi-Dimensional Scaling using the **vegan** package in R. The [Bray-Curtis](#) similarity measure will be used. This is considered a good measure for quantitative data. The data set used will be your working data.

Multi-Dimensional Scaling (MDS) is a technique for expressing the similarities between different objects in a small number of dimensions. Hopefully, this allows a complex set of inter-relationships to be summarised in a simple figure. The method attempts to place the most similar objects (samples) closest together. The starting point for the calculations is a similarity or dissimilarity matrix between all the sites or quadrats. These can be non-metric distance measures for which the relationships between the sites/objects/samples (columns) cannot be plotted in a Euclidean space. The aim of Non-metric MDS is to find a set of metric coordinates for the sites which most closely approximates their non-metric distances.

The basic MDS algorithm is as follows:

1. Calculate the similarity or dissimilarity between sites.
2. Assign to each site a set of coordinates in p-dimensional space. These coordinates can be either chosen at random or chosen using Principal Coordinates Analysis (note, this is **not** the same as a [Principal Component Analysis](#)). The value of p is chosen by the user.
3. Compute the Euclidean distance between these sites using the starting coordinates.
4. Compare the original dissimilarity between the sites with these Euclidean distances by calculating a stress function. The smaller the stress function, the closer the correspondence.
5. Adjust the positions so as to reduce the stress.
6. Repeat 2 to 4 until the stress is minimised or the maximum number of iterations is reached.



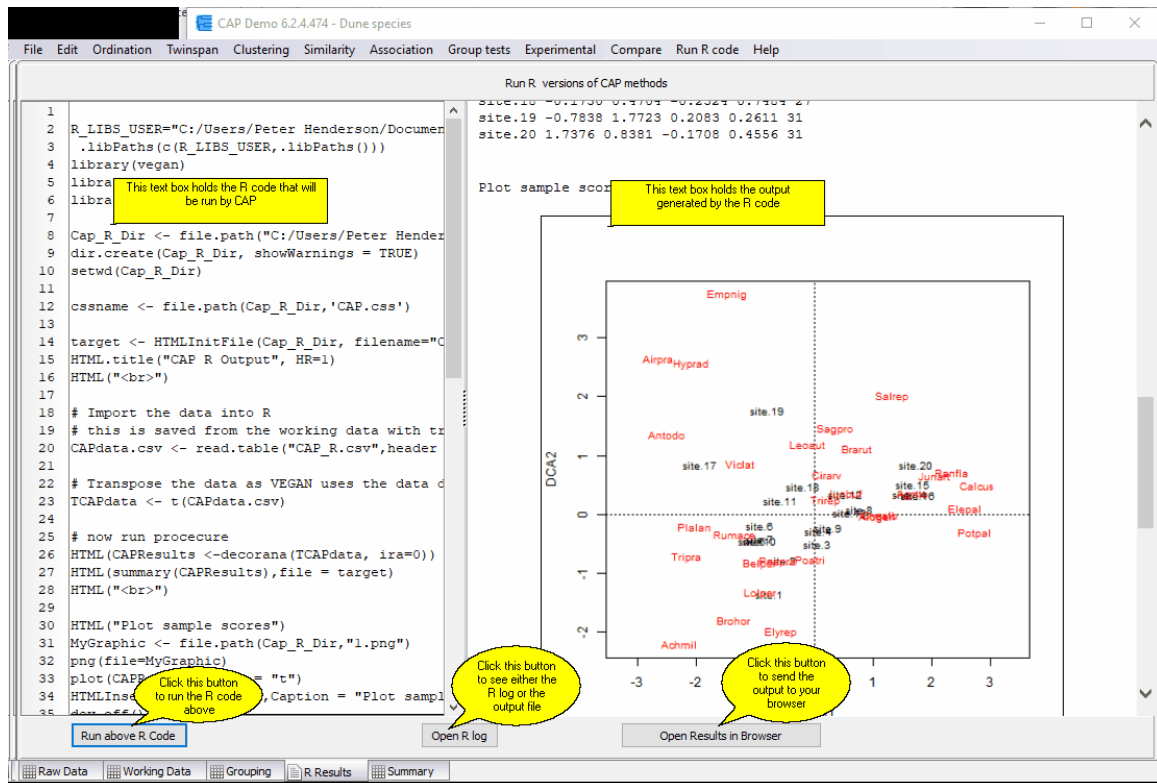
## 18.9 MDS - Jaccard R

Select this option to undertake Multi-Dimensional Scaling using R. The [Jaccard](#)<sup>[114]</sup> similarity measure will be used, this is considered a good measure for qualitative (presence/absence) data. The data set used will be your working data.

Multi-Dimensional Scaling (MDS) is a technique for expressing the similarities between different objects in a small number of dimensions. Hopefully, this allows a complex set of inter-relationships to be summarised in a simple figure. The method attempts to place the most similar objects (samples) closest together. The starting point for the calculations is a similarity or dissimilarity matrix between all the sites or quadrats. These can be non-metric distance measures for which the relationships between the sites/objects/samples (columns) cannot be plotted in a Euclidean space. The aim of Non-metric MDS is to find a set of metric coordinates for the sites which most closely approximates their non-metric distances.

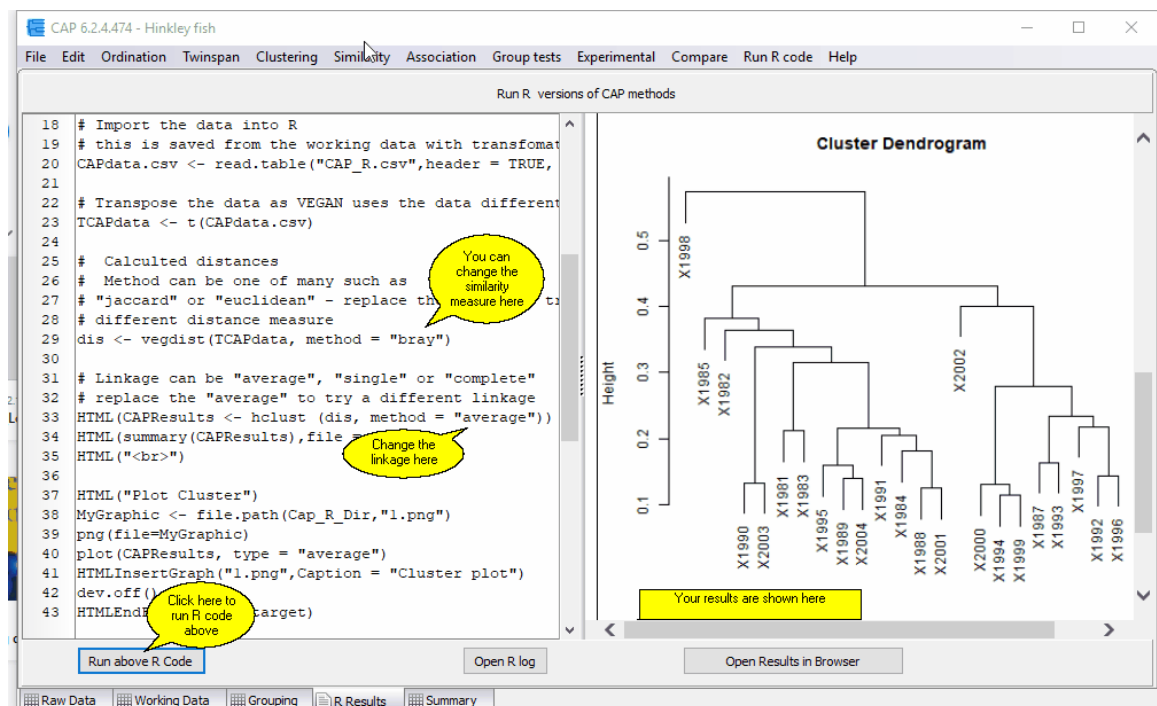
The basic MDS algorithm is as follows:

1. Calculate the similarity or dissimilarity between sites.
2. Assign to each site a set of coordinates in p-dimensional space. These coordinates can be either chosen at random or chosen using Principal Coordinates Analysis (note, this is **not** the same as a [Principal Component Analysis](#)<sup>[63]</sup>). The value of p is chosen by the user.
3. Compute the Euclidean distance between these sites using the starting coordinates.
4. Compare the original dissimilarity between the sites with these Euclidean distances by calculating a stress function. The smaller the stress function, the closer the correspondence.
5. Adjust the positions so as to reduce the stress.
6. Repeat 2 to 4 until the stress is minimised or the maximum number of iterations is reached.



## 18.10 Clustering R

Choose this menu to undertake an agglomerative cluster analysis and produce a dendrogram using the **vegan** package in R.



## 18.11 ANOSIM R

To undertake this test you must first have defined the group membership of the individual samples (see [Allocating samples to Groups](#)<sup>[53]</sup>).

This test was developed by [Clark \(1988, 1993\)](#)<sup>[172]</sup> as a test of the significance of the groups that had been defined *a priori*. The idea is simple; if the assigned groups are meaningful, samples within groups should be more similar in composition than samples from different groups. The method uses the [Bray-Curtis](#)<sup>[12]</sup> measure of similarity. The null hypothesis is therefore that there are no differences between the members of the various groups.

[Clark \(1988, 1993\)](#)<sup>[172]</sup> proposed the following statistic to measure the differences between the groups:

$$R = \frac{\bar{r}_B - \bar{r}_W}{n(n-1)/4}$$

where

$\bar{r}_B, \bar{r}_W$  are the mean of the ranked similarity BETWEEN groups and WITHIN groups respectively and n is the total number of samples (objects).

R scales from +1 to -1. +1 indicates that all the most similar samples are within the same groups. R = 0 occurs if the high and low similarities are perfectly mixed and bear no relationship to the group. A value of -1 indicates that the most similar samples are all outside of the groups. While negative values might seem to be a most unlikely eventuality it has been found to occur with surprising frequency.

To test for significance, the ranked similarity within and between groups is compared with the similarity that would be generated by random chance. Essentially the samples are randomly assigned to groups 1000 times and R calculated for each permutation. The observed value of R is then compared against the random distribution to determine if it is significantly different from that which could occur at random.

If the value of R is significant, you can conclude that there is evidence that the samples within groups are more similar than would be expected by random chance.

An example output is shown below

Call:

```
anosim(x = dis, grouping = CAPGroups)
```

Dissimilarity: bray

ANOSIM statistic R: 0.8731 - **This is the test statistic**

Significance: 0.001 - **This is the significance of the grouping, in this case it is highly significant**

Permutation: free

Number of permutations: 999

Upper quantiles of permutations (null model):

90% 95% 97.5% 99%

0.0846 0.1232 0.1697 0.2201

Dissimilarity ranks between and within classes:

0% 25% 50% 75% 100% N

Between 26 184.75 231.5 278.25 325 188

Ashley Rails 1 20.00 71.0 115.00 164 45

Caldicot 3 3.00 3.0 3.00 3 1

Llanederyn 7 48.00 85.0 129.50 166 91

## 18.12 SIMPER

To undertake this test, you must first have defined the group membership of the individual samples (see [Allocating samples to groups](#)<sup>[53]</sup>).

This analysis breaks down the contribution of each species (or other variable) to the observed similarity (or dissimilarity) between samples. It will allow you to identify the species that are most important in creating the observed pattern of similarity. The method uses the [Bray-Curtis](#)<sup>[12]</sup> measure of similarity, comparing in turn, each sample in Group 1 with each sample in Group 2. The Bray-Curtis method operates at the species level, and therefore the mean similarity between Groups 1 & 2 can be obtained for each species.

In the following example, using the *Romano British pottery.csv* data file, the data have been divided into 3 location groups.

cumulative contributions of most influential species:

\$Llanederyn\_Caldicot

Al Fe Mg

0.3475663 0.6402357 0.9222869 - **These values are the cumulative proportion of the dissimilarity from each element between the 2 groups**

\$`Llanederyn\_Ashley Rails`

Al Fe Mg

0.3600725 0.6869017 0.9749251

\$`Caldicot\_Ashley Rails`

Al Fe

0.4515666 0.7374816

Contrast: Llanederyn\_Caldicot



```

average sd ratio ava avb cumsum
Al 0.029000 0.018356 1.580 12.5643 11.700 0.3476 - Average abundance Al in
Llanederyn = 12.56, in Caldicot = 11.7
Fe 0.024419 0.010210 2.392 6.3721 5.415 0.6402 - The contribution of Al and Fe to
dissimilarity = 0.6402 = 64%
Mg 0.023533 0.019578 1.202 4.8264 3.855 0.9223
Na 0.004391 0.002558 1.716 0.2507 0.050 0.9749
Ca 0.002093 0.001169 1.790 0.2021 0.295 1.0000

```

Contrast: Llanederyn\_Ashley Rails

```

average sd ratio ava avb cumsum
Al 0.117845 0.049056 2.402 12.5643 17.750 0.3601
Fe 0.106965 0.018641 5.738 6.3721 1.612 0.6869
Mg 0.094265 0.021610 4.362 4.8264 0.640 0.9749
Na 0.004501 0.002631 1.711 0.2507 0.051 0.9887
Ca 0.003706 0.001470 2.521 0.2021 0.039 1.0000

```

Contrast: Caldicot\_Ashley Rails

```

average sd ratio ava avb cumsum
Al 0.1450682 0.0347999 4.169 11.700 17.750 0.4516
Fe 0.0918517 0.0137446 6.683 5.415 1.612 0.7375
Mg 0.0777523 0.0040412 19.240 3.855 0.640 0.9795
Ca 0.0062040 0.0008860 7.002 0.295 0.039 0.9988
Na 0.0003792 0.0003395 1.117 0.050 0.051 1.0000
Permutation: free
Number of permutations: 0

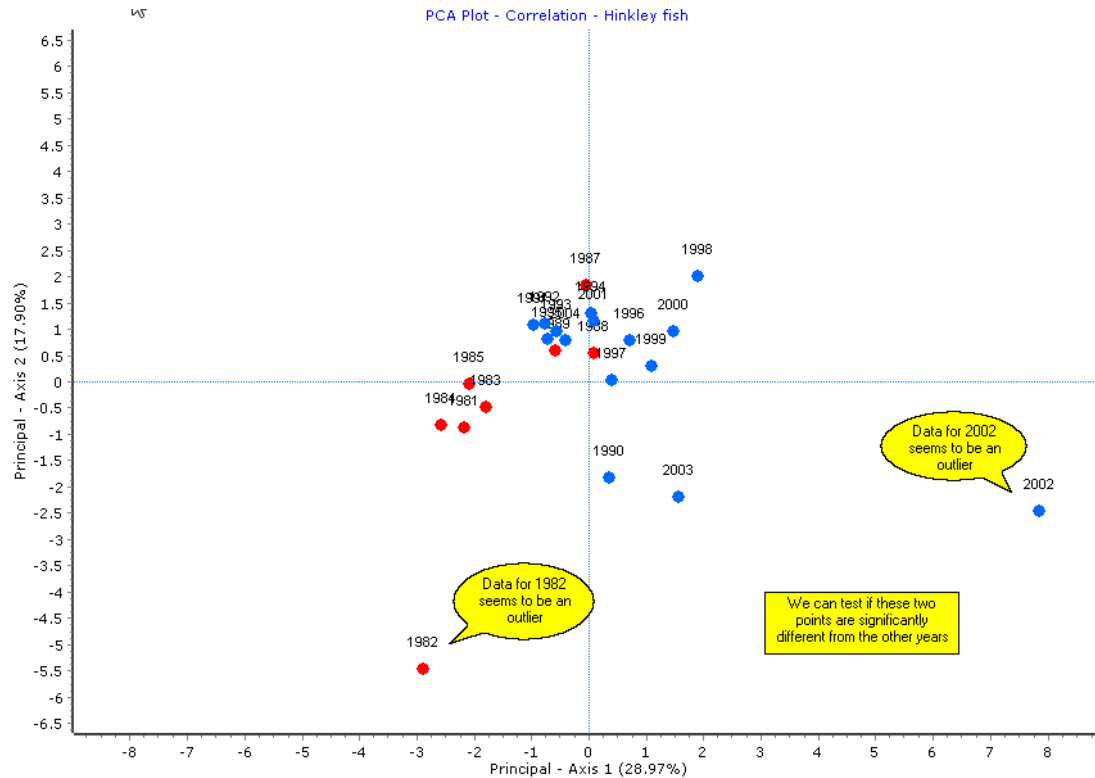
```

## 18.13 PCA - Cor - Outlier R

Having completed a Principal Component Analysis (PCA) it is sometimes useful to consider if any of the observations are significantly different from the others.

For example, in the PCA plot below Hinkley fish data for 1982 and 2002 seem to be notably different from other years.

This R code undertakes an outlier analysis using the correlation values between the variables. The [Mahalanobis distance](#)<sup>[122]</sup> is used. The default calculations are undertaken using the two largest principal components. This can be changed - see image below.



The following image shows the output - note that 1982 and 2002 are indeed significantly different from the other years.

change the  $k$  to a higher number.

Run R versions of CAP methods

```

1 R_LIBS_USER="C:/Users/Peter Henderson/Documents/R library"
2 .libPaths(c(R_LIBS_USER, libPaths()))
3 library(vegan)
4 library(R2HTML)
5 library(ClassDiscovery)
6
7 Cap_R_Dir <- file.path("C:/Users/Peter Henderson/Desktop/CAP 6 Development/RTemp")
8 dir.create(Cap_R_Dir, showWarnings = TRUE)
9 setwd(Cap_R_Dir)
10
11 casname <- file.path(Cap_R_Dir, 'CAP.cas')
12
13 target <- HTMLInitFile(Cap_R_Dir, filename="CAPReport", BackgroundColor="#FFFFFF")
14 HTML.title("CAP R Output", HR=1)
15 HTML("<br>")
16 HTML("<br>")
17
18 # Import the data into R
19 # this is saved from the working data with transformations etc
20 CAPdata <- read.table("CAP_R.csv", header = TRUE, sep = ",", row.names = 1)
21
22 # do poa outlier on correlation matrix
23 HTML("Mahalanobis Distances")
24 HTML("<br>")
25 HTML("<br>")
26 HTML("MahalanobisQC calculated using the first two principal components")
27 HTML("The calculation of the Mahalanobis distance for each sample and their significance")
28 HTML("Removed outliers from the sample and their significance")
29 HTML("<br>")
30 HTML("Be aware that if your data encompasses nonlinear relationships Mahalanobis distance can be misleading as it assumes linear relationships between variables.")
31 CAPResults <- SamplePCA(CAPdata, nrecor=TRUE)
32 HTML(mahalanobisQC(CAPResults, 2))
33 HTML("<br>")
34 HTMLEndFile(file = target)

```

Be aware that if your data encompasses nonlinear relationships Mahalanobis distance can be misleading as it assumes linear relationships between variables.

1982 is a significant outlier

2002 is a significant outlier

	statistic	p.value
X1981	1.286	5.3e-01
X1982	21.657	2.0e-05
X1983	0.767	6.8e-01
X1984	1.707	4.3e-01
X1985	0.936	6.3e-01
X1987	1.200	5.5e-01
X1988	0.102	9.5e-01
X1989	0.185	9.1e-01
X1990	1.208	5.5e-01
X1991	0.587	7.5e-01
X1992	0.539	7.6e-01
X1993	0.368	8.3e-01
X1994	0.573	7.5e-01
X1995	0.333	8.5e-01
X1996	0.307	8.6e-01
X1997	0.033	9.8e-01
X1998	2.213	3.3e-01
X1999	0.281	8.7e-01
X2000	0.760	6.8e-01
X2001	0.450	8.0e-01
X2002	36.757	1.0e-08
X2003	2.252	3.2e-01
X2004	0.236	8.9e-01

Run above R Code

Open R log

Open Results in Browser

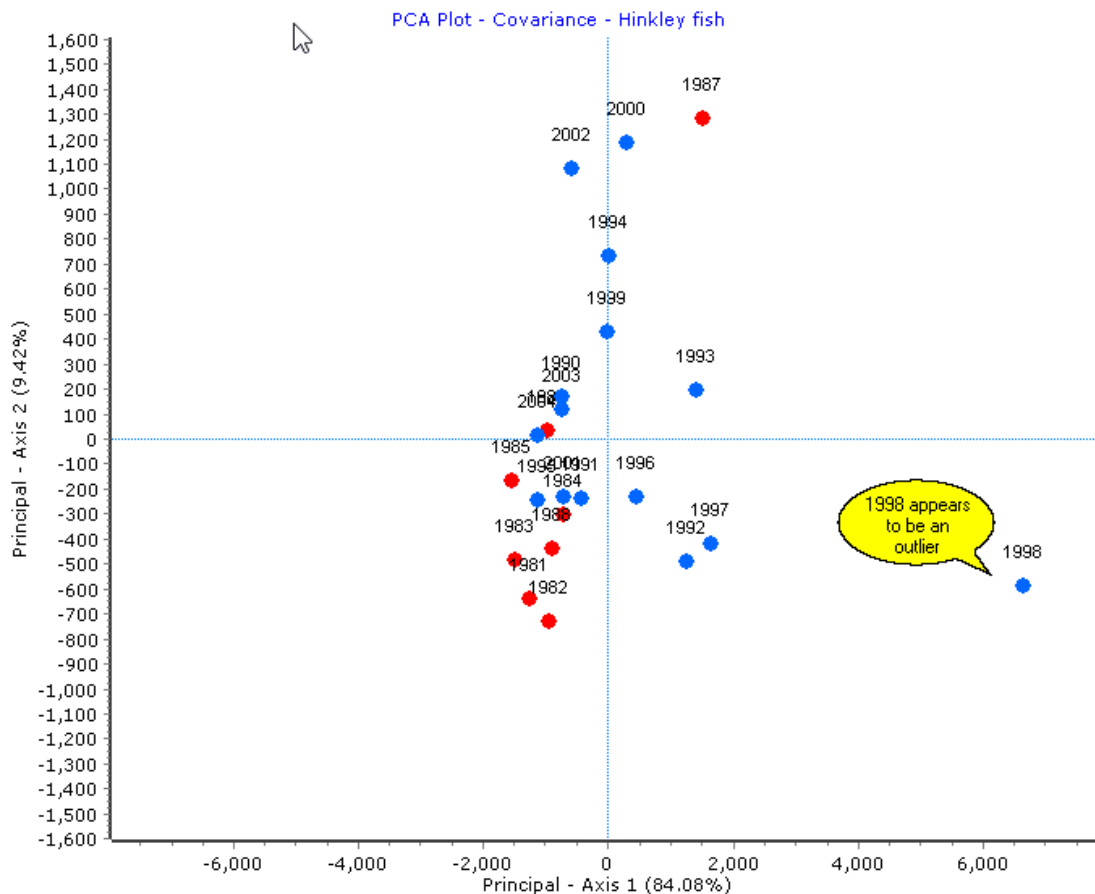
Raw Data Working Data Grouping R Results Summary

## 18.14 PCA - Covar - Outlier R

Having completed a Principal Component Analysis (PCA) it is sometimes useful to consider if any of the observations are significantly different from the others.

For example, in the PCA plot below Hinkley fish data for 1998 seems to be notably different from other years.

This R code undertakes an outlier analysis using the variance - covariance matrix values between the variables. The [Mahalanobis distance](#)<sup>[122]</sup> is used. The default calculations are undertaken using the two largest principal components. This can be changed - [see image](#)<sup>[188]</sup>.



change the  $\lambda$  to a higher number.

File Edit Ordination Twinspan Clustering Similarity Association Group tests Experimental Compare Run R code Help

Run R versions of CAP methods

1		X1985	0.84	6.6e-01
2	R_LIBS_USER="C:/Users/Peter H	X1987	6.72	3.5e-02
3	.libPaths(c(R_LIBS_USER,.lib	X1988	0.81	6.7e-01
4	library(vegan)	X1989	0.31	8.6e-01
5	library(R2HTML)	X1990	0.25	8.8e-01
6	library(ClassDiscovery)	X1991	0.22	9.0e-01
7		X1992	1.22	5.4e-01
8	Cap_R_Dir <- file.path("C:/Us	X1993	0.75	6.9e-01
9	dir.create(Cap_R_Dir, showWarn	X1994	1.62	4.4e-01
10	setwd(Cap_R_Dir)	X1995	0.58	7.5e-01
11		X1996	0.22	8.9e-01
12	cssname <- file.path(Cap_R_Di	X1997	1.41	4.9e-01
13		X1998	45.41	1.4e-10
14	target <- HTMLInitFile(Cap_R_	X1999	0.52	7.7e-01
15	HTML.title("CAP R Output", HR	X2000	4.91	8.6e-02
16	HTML(" ")	X2001	0.32	8.5e-01
17		X2002	3.98	1.4e-01
18	# Import the data into R	X2003	0.20	9.0e-01
19	# this is saved from the work	X2004	0.41	8.1e-01
20	CAPdata <- read.table("CAP_R.			
21				
22	# do pca outlier on covarianc			
23	HTML("Mahalanobis Distaces")			
24	HTML("covariance")			
25	HTML(" ")			
26	HTML("MahalanobisQC calculate			
27	HTML("The calculation of the			
28	HTML("Removed outliers from t			
29	HTML(" ")			
30	HTML("Be aware that if your d			
31	CAPResults <-SamplePCA(CAPdat			
32	HTML(mahalanobisQC(CAPResults			
33	HTML(" ")			

1998 is shown to be highly significantly different.

# Index

## - " -

" is not a valid floating point value 10

## - . -

.csv 6  
.xls 6  
.xlsx 6

## - A -

abundance data 17  
Access Violation error 10  
add annotation 162  
Add grid 157  
Add legend 157  
Add stalk 157  
agglomerative clustering 14, 98, 100, 101  
agglomerative clustering methods 99  
allocate 53  
allocating 53  
Analysis of Similarity 128  
annotations 162  
ANOSIM 127, 128, 186  
apply transformations 48, 49  
arch effect 70  
arcsine 49  
ascii 168  
assigning 53  
association analysis 124  
average 59, 119  
average linkage 100  
axes 15, 16  
axis 16, 63  
axis vs variable plot 69

## - B -

Bartlett's test 136  
base10 49  
Beals 49  
between-group covariance 137  
binary 49  
binary data 110, 124

biplot 66  
bitmap 153, 154  
bmp 153  
borderline groups 90, 91  
Bray Curtis 183  
bray-curtis 121  
Bray-Curtis similarity 128  
bubble plot 3, 80

## - C -

Canberra 121  
canonical correlation 132  
canonical variates analysis 17, 127, 131  
cell contents 44  
centroid 14, 101  
centroids 135  
Change between colour/grey scale 157  
change cell content 44  
change chart theme 167  
change data 48, 49  
change groups 53  
change order of chart groups 162  
Change symbol set 157  
characteristics 56  
Chart 160  
chart export 153  
chart format 153  
chart options 160  
chart themes 167  
chart titles 162  
chart tools 162  
charts 153, 162  
Chi-squared 82, 124  
choose data 8  
choose demo data set 8  
chrod 119  
cicada songs 30  
citation 173  
classification 14, 15  
classification table 96  
classifying 14  
clipboard 154  
cluster 185  
cluster analysis 14, 98, 99, 100, 101, 102, 105, 106  
cluster dendrogram 98, 102, 103  
cluster group membership 101  
cluster groups 101, 105, 108  
cluster plot 107  
cluster summary 105, 108

clustering 98  
 clustering results 101  
 clustering setup 99, 106  
 clusters of points 158  
 coefficients 133, 134  
 colour palette 153  
 colours 53  
 column statistics 59  
 column title 44  
 column titles 42  
 columns 35, 49, 50, 51  
 common errors 10  
 community structure 2, 15  
 compare groups 146  
 compare samples 146, 149  
 compare variables 148, 149  
 comparison 186  
 comparisons 14  
 complete linkage 100  
 computations 71, 83  
 constant 49  
 contact pisces 172  
 Copy 42, 153, 154, 157  
 copy dendrogram 94, 103  
 copying 153  
 correlation 16, 63, 65, 66  
 correlation between variables 65  
 correspondence analysis 70, 71, 72, 73, 74, 82  
 counting zeros 59  
 covariance 16, 63, 65, 66, 137  
 create data set 38, 41  
 create groups 53  
 creating a data set 34  
 creating new data file 45  
 critical values 94  
 cross products 66  
 csv 6, 36, 37, 168  
 cumulative 70  
 customise charts 167  
 cut levels 88  
 cut values 88  
 czekanowski 119

## - D -

Data 9, 19, 34, 36, 38, 45, 160  
 data file 45  
 data grid 34, 41  
 data input 42  
 data preparation 35

data problems 6  
 data set 9  
 data structure 35  
 data transformations 48, 49  
 decimal places 5  
 DECORANA 70, 71, 72, 73, 74, 82  
 Decrease font 157  
 Decrease line thickness 157  
 Decrease symbol 157  
 default chart theme 167  
 define perimeter 161  
 defined groups 131  
 delete 44, 50  
 delete column 42  
 delete row 42  
 demo data 8  
 demonstration data 19  
 demos 19  
 demote 42  
 Dendrogram 14, 16, 92, 93, 98, 102, 154, 155, 185  
 dendrogram edit 102, 103  
 dendrogram options 94, 103  
 dendrograms 154  
 detrended 70, 71  
 Detrended Correspondence Analysis 181  
 different methods 4  
 dimensions 76, 78  
 discriminant analysis 17, 127, 131, 132, 133, 134, 135, 136, 137, 138  
 discriminant analysis plot 137  
 discriminant function coefficients 133  
 discriminant scores 138  
 dispersion matrices 137  
 dispersion matrix 65, 66  
 dissimilarity 75, 105, 106, 129  
 distance measures 99, 110, 119, 120, 121, 122  
 diversity 2  
 divisions 92  
 divisive clustering 14, 88, 98, 106, 107, 108  
 divisive clustering setup 106  
 double-zero methods 110  
 download 4  
 downweight rare variables 70, 82  
 draw lines 162

## - E -

ecom 2  
 Edit 157  
 edit chart 157, 160

edit charts 167  
 edit data 44, 48  
 edit dendrogram 94, 102, 103  
 edit footer 155  
 edit graphs 167  
 edit header 155  
 edit titles 162  
 editing chart dialog 160  
 editing results 153  
 eigenvalues 63, 65, 66, 71, 72, 73, 83, 84, 90, 91, 132  
 eigenvectors 16, 65, 66  
 email 153  
 end point compression 70  
 enhanced metafile 153  
 enlarge 158  
 enter data 38  
 environmental gradient 17  
 errors 6, 10  
 Euclidean 119  
 Euclidean distance 75  
 examples 19  
 excel 6, 36, 37, 168  
 existing data 44  
 export 45, 154, 160, 168  
 export charts 153  
 export dendrogram 94, 103  
 export pdf 155  
 exporting 153, 154

## - F -

faq 10  
 file format 153  
 filter options 143  
 filtering 141  
 filters 153  
 finding errors 6  
 Fisher's discriminant coefficients 134  
 footer 155  
 fuzzy grouping 2  
 fuzzy logic 2

## - G -

general 5, 49  
 geodesic 120  
 gif 153  
 Gower's 101  
 gradient 17, 71

graph export 153  
 graph themes 167  
 graphics 66, 74, 153, 154  
 graphs 153  
 group 56  
 group centroids 135  
 group colours 53  
 group means 59  
 group membership 3  
 group names 53  
 group properties 56  
 group tests 127, 128, 131  
 grouping 53  
 groups 17, 53, 56, 98, 161  
 groups labels 105  
 guides 8

## - H -

header 155  
 help 10, 172  
 hierarchical classification 14  
 hierarchical clustering 98, 99  
 hill 16, 70  
 Hinkley 19  
 how to cite CAP 173  
 html 168

## - I -

I/O error 103 10  
 I/O error 32 - access denied 10  
 image format 153  
 import data 37  
 improvements 3  
 Increase font 157  
 Increase line thickness 157  
 Increase symbol 157  
 index 14  
 indicator 16, 89  
 indicator levels 88  
 indicator species 92  
 indices 14  
 initial start position 76  
 input 5, 6, 9, 34, 36, 38  
 insert 44  
 insert column 42  
 insert row 42  
 install 4  
 installation 4, 175

installing 175, 178  
 instant assist 5  
 instructions 4, 5  
 introduction 2  
 invalid floating point value 10  
 invertebrates 20  
 iterations 75, 76

## - J -

Jaccard 75, 184  
 Jaccards 114  
 japanese pottery 22  
 Jomon Hall 22  
 jpeg 153

## - K -

Kruskals's least squares monotonic transformation  
 75  
 Kulczynski 116  
 Kulczynski quantitative 118  
 kurtosis 59

## - L -

label columns 42  
 label rows 42  
 labels 53, 66, 162  
 large array 9  
 large data sets 9  
 large files 9  
 large matrix 9  
 last-used file 5  
 latest 3  
 lines 162  
 list combiner 9  
 loading 4  
 loading from Excel 37  
 log 49

## - M -

magnify 158  
 Mahalanobis 122  
 Mahalanobis distance 188, 190  
 manhattan 120  
 MANOVA 131  
 matrix 51, 65, 66  
 matrix plot 149

maximum 9, 59  
 McQuitty's 101  
 MDS 75, 76, 77, 78  
 MDS plot 80  
 mean 49, 59  
 mean character difference 119, 120  
 median 49, 59  
 memory 9  
 menus 5  
 metafile 153  
 metafile format 154  
 methods 4  
 methods of analysis 14  
 methods of cluster analysis 98  
 Microsoft Excel 37  
 minimum 59  
 misclassified groups 90, 91  
 move labels 162  
 move left 42  
 move right 42  
 MS Excel 37  
 multidimensional 75  
 multidimensional scaling 3  
 Multi-dimensional scaling 183, 184  
 multivariate 15  
 multivariate analysis 2  
 multivariate analysis of variance 131

## - N -

name groups 53  
 natural 49  
 negative groups 90, 91  
 nethods of ordination 63  
 new data set 34, 45  
 new methods 3  
 NMDS 75, 76, 77, 78  
 NMDS plot 80  
 NMDS setup 76  
 node 92  
 nodes 93  
 non-hierarchical classification 14  
 non-hierarchical clustering 98  
 non-metric 75  
 notation 5  
 number of clusters 106  
 number of columns 9  
 number of dimensions 76, 78  
 number of rows 9



## - O -

Ochiai 117  
 opening a data set 6  
 options 162  
 order 162  
 Ordination 16, 63, 75, 82, 179, 180, 181, 182  
 ordination methods 141  
 ordination space 15  
 organising 35  
 Outlier analysis 188, 190  
 outline 161  
 output 66, 74, 153, 155

## - P -

packages 178  
 paste 42, 154  
 pasting data 42  
 PCA 16, 63, 64, 65, 66, 69, 179, 180, 188, 190  
 pcg 53  
 pdf 153, 155  
 Pearson correlation 132  
 percent similarity 122  
 percentage cover 16  
 perimeter 161  
 petrology.csv 26  
 Pillai 136  
 plot clusters 107  
 plots 66, 74, 80  
 plotting groups 56  
 png 153  
 Pointer 157  
 positive groups 91  
 power 49  
 precision 5  
 predicted group 138  
 predictive validation 138  
 preferences 5  
 presence-absence data 14, 16, 110, 124  
 preview 168  
 principal axes 66  
 principal axis 69  
 principal axis vs variable plot 69  
 principal component analysis 63, 64, 65, 66  
 principal component scores 64  
 principal components 15, 16  
 Principal Components Analysis 179, 180, 188, 190  
 print 157, 160, 168

print as pdf 155  
 print dendrogram 155  
 print preview 168  
 printing 5, 153, 168  
 printing data errors 6  
 printing dendrogram 155  
 printing grids 168  
 printing output 168  
 problems 10  
 profile plot 147  
 promote 42  
 properties 56  
 pseudospecies 16, 88

## - Q -

Q analysis 14, 110  
 q1 117  
 q2 118  
 quadrat 16  
 quantitative data 17, 110  
 quick guide 6

## - R -

R 175, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187  
 R analysis 14, 110  
 RA 82, 83  
 RA computations 83  
 RA plot 85  
 RA sample scores 84  
 RA species scores 83  
 rare variables 82  
 raw data 34, 44  
 recent files 5  
 Reciprocal averaging 17, 70, 82, 182  
 reciprocal averaging computations 83  
 reciprocal averaging plot 85  
 reciprocal averaging sample scores 84  
 reciprocal averaging species scores 83  
 references 172  
 relative 49  
 relativisations 49  
 removing zeros 50  
 renkonen 122  
 resize grid 41  
 Rogers-Tanimoto 112  
 row statistics 59  
 row title 44

row titles 42  
 rows 35, 49, 50, 51  
 rtf 168  
 run demo data 8  
 running a data set 6  
 running R 175  
 Russell & Rao 116

## - S -

s10 115  
 s13 116  
 s3 111  
 s4 112  
 s5 113  
 s6 113  
 s9 115  
 sample differences 146  
 sample names 5  
 sample number 59  
 sample scores 73, 84  
 samples 2  
 Save 153, 157  
 save as 45  
 save text 168  
 save working data 45  
 saving 45, 153  
 saving grids 168  
 saving output 168  
 scaling 75  
 scatter plot 69, 148  
 scientific notation 5  
 scores 83, 84  
 scree 70  
 scree plot 3  
 select demo data 8  
 Select theme 157  
 selecting 53  
 send by email 153  
 Series 160  
 sets 53  
 Setting up R 175  
 settings 56  
 setup 5, 75  
 setup for agglomerative clustering 99  
 setup for divisive clustering 106  
 setup for reciprocal averaging 82  
 setup for TWINSpan 88  
 setup MDS 76  
 show me 8

significance 17  
 significance of defined groups 131  
 significance tests 136  
 similarity 3, 14, 75, 98, 129, 186, 187  
 similarity between groups 128, 129  
 similarity measures 76, 110, 111, 112, 113, 114, 115, 116, 117, 118  
 Similarity Percentages 129  
 similarity within groups 128, 129  
 SIMPER 127, 129, 187  
 simple matching 111  
 single linkage 100  
 Singular Matrix in Lower-Upper Decomposition routine 10  
 site coordinates 77, 138  
 site summary 90  
 sites dendrogram 92  
 size 36  
 size of data grid 41  
 size of data set 9  
 skewness 59  
 Sorensen 75, 114  
 sorting 53  
 space 63  
 sparse 50  
 species 2, 83  
 species dendrogram 93  
 species diversity & richness 2  
 species filtering 141  
 species filtering setup 143  
 species in common 146  
 species number 59  
 species occurrence in samples 147  
 species score 72  
 species scores 83  
 species summary 91  
 speed 9  
 sphericity 136  
 spreadsheet 36  
 square root 49  
 squared chord distance 121  
 standardised coefficients 133  
 starting 6  
 statistical methods 4  
 step size 78  
 stream invertebrates 20  
 stress 75, 78, 80  
 styles 167  
 subdivisions 90, 91  
 summary 59

summary statistics 59  
sums of squares 59, 106, 108  
system requirements 4

## - T -

techniques 14  
TeeReader 153  
text 168  
Themes 160, 167  
titles 42, 44, 162  
tools 160, 162  
total covariance 137  
total variance 63  
transfer data 42  
transformations 48, 49  
transposing 51  
TWINSpan 9, 16, 88  
TWINSpan dendrogram 94  
TWINSpan options 88  
TWINSpan results 89, 96  
TWINSpan setup 88  
TWINSpan site summary 90  
TWINSpan sites dendrogram 92  
TWINSpan species dendrogram 93  
TWINSpan species summary 91  
TWINSpan text 89  
TWINSpan variables summary 91

## - U -

user preferences 5  
using CAP 6

## - V -

valid floating point value 10  
variable filtering 141  
variable filtering setup 143  
variable names 5  
variance 59, 63, 65, 66, 70  
variance-covariance matrix 63  
vegan 178, 179, 180, 181, 182, 183, 185  
vegetation analysis 16  
version 3  
video guides 3, 8  
videos 8

## - W -

Ward's 100  
weighting 88  
Whittaker 121  
Wilks' lambda 136  
worked examples 19, 20, 22, 26, 30  
working data screen 48

## - X -

xls 6, 37  
xlsx 37

## - Y -

Yate's correction 124  
YouTube 3

## - Z -

zero 50, 59  
zoom 158

Back Cover