PAST

PAleontological STatistics

Version 5.4

Reference manual

Øyvind Hammer

Natural History Museum

University of Oslo

ohammer@nhm.uio.no

1999-**2026**

Contents

Welcome to the PAST! 8
Installation 9
Windows 9
Mac 9
Quick start 10
How do I export graphics? 10
How do I organize data into groups? 10
The spreadsheet and the Edit menu 11
Entering data 11
Selecting areas 11
Moving a row or a column 12
Renaming rows and columns 12
Increasing the size of the array 12
Cut, copy, paste 12
Remove 12
Row colors and symbols 12
Selecting datatypes for columns, and specifying groups 12
Remove uninformative rows/columns 13
Transpose 13
Grouped columns to multivar 13
Grouped rows to multivar 14
Stack grouped rows into columns 14
Rows, cols, values to table 14
Value pairs to matrix 14
Samples to events (UA to RASC) 14
Events to samples (RASC to UA) 15
Loading and saving data 15
Importing data from Excel 16
Reading and writing Nexus files 16
Counter 16
Transform menu 17
Logarithm 17
Subtract mean 17
Remove trend 17
Row percentage 17
Row normalize length 17
Box-Cox 17
Compositional data transforms 18
Remove size from distances 20
Landmarks, Procrustes fitting 20
Landmarks, Bookstein fitting 21
Project to tangent space (not yet in Past 4) 21
Remove size from landmarks (not yet in Past 4) 21
Transform landmarks 22
Regular interpolation 22

Evaluate expression 22
Plot menu 24
Graph 24
XY graph 25
XY graph with error bars 26
Histogram 27
Bar chart/box plot 28
Pie chart 30
Stacked chart 31
Percentiles 32
Normal probability plot 33
Ternary 34
Bubble plot 35
Matrix plot 36
Mosaic plot 37
Venn diagram 38
Radar chart 40
Polar plot 41
Vector plot 42
Network plot 43
3D scatter/bubble/line plot 44
3D surface plot 45
3D parametric surface plot 46
Univariate menu 47
Summary statistics 47
One-sample tests 50
Two-sample tests 52
t test and related tests for equal means 52
F test for equal variances 55
Reference 55
Mann-Whitney test for statistic equality 55
Brunner-Munzel test for stochastic equality 57
Mood's median test for equal medians 57
Kolmogorov-Smirnov test for equal distributions 58
Anderson-Darling test for equal distributions 59
Epps-Singleton test for equal distributions 60
Coefficient of variation (Fligner-Kileen test) 61
F and *t* tests from parameters 62
Two-sample paired tests (*t*, sign, Wilcoxon) 63
Several-sample tests 65
Several-samples repeated measures tests 72

Two-way ANOVA 76
Two-way ANOVA without replication 79
Two-way repeated measures ANOVA 80
One-way ANCOVA 81
Correlation table 82
Intraclass correlation 87
Normality tests 89
Outlier tests 92
Contingency table (chi^2 etc.) 94
Cochran-Mantel-Haenszel test 96
Risk/odds 97
Single proportion 99
Multiple proportion confidence intervals 100
Ratio of counts confidence interval 101
Survival analysis (Kaplan-Meier curves, log-rank test etc.) 102
Combine errors 104
Propagate errors 104
Multivariate menu 106
Principal components 106
Principal coordinates 111
Non-metric MDS 112
Correspondence analysis 114
Detrended correspondence analysis 116
Canonical correspondence 117
Seriation 118
CABFAC factor analysis 119
Discriminant analysis 120
Two-block PLS 122
Redundancy analysis (RDA) 123
Nonlinear ordination (UMAP) 124
Nonlinear ordination (ISOMAP) 126
Classical cluster analysis 127
Neighbour joining 129
K-means clustering 130
K-medoids clustering 131
DBSCAN clustering 131
K-nearest neighbors classifier 132
Naïve Bayes classifier 133
Decision tree classification 134
Random forest classifier 135
Parsimony (cladistic) 136
Phylogenetically independent contrasts 139
Phylogenetic generalized least squares (PGLS) 140
Multivariate normality 141
Box's *M* 142
MANOVA 143
One-way ANOSIM 144
One-way PERMANOVA 145
Two-way ANOSIM 147
Two-way ANOSIM without replication 147

Two-way PERMANOVA 147
Test for multivariate dispersion (PERMDISP) 148
Mantel test and partial Mantel test 149
SIMPER 151
Indicator species analysis (IndVal) 152
Paired Hotelling 153
Modern Analog Technique 154
Weighted averaging calibration (WA) 156
Weighted averaging partial least squares (WA-PLS) 157
Similarity and distance indices 160
Genetic sequence stats 166
Model menu 167
Linear, bivariate 167
Linear, multivariate (one independent, n dependent) 172
Linear, multiple (one dependent, n independent) 173
Linear, multivariate multiple (m independent, n dependent) 174
Generalized Linear Model 175
Polynomial regression 177
Nonlinear 178
Sinusoidal regression 181
Smoothing spline 183
LOESS smoothing 184
Mixture analysis 186
Abundance models 188
Species packing (Gaussian) 190
Fit spiral 191
Changepoint modeling 192
Diversity menu 194
Alpha diversity indices 194
Quadrat richness 199
Beta diversity 202
Taxonomic 204
Individual rarefaction 206
Shareholder Quorum Subsampling (SQS) 209
Sample rarefaction (Mao's tau) 210
SHE analysis 212
Diversity permutation test 213
Diversity *t* test 214
Diversity profiles 215
Time series menu 216
Simple periodogram 216
REDFIT spectral analysis 217
Multitaper spectral analysis 219
Walsh transform 220
Evolutive Fourier transform 221
Wavelet transform 222
Wavelets for unequal spacing 224
Point events spectrum 225
Autocorrelation 227
Autoassociation 228

Cross-correlation 230
Mantel correlogram (and periodogram) 231
Runs test 233
Mann-Kendall trend test 234
Point events 235
Markov chain 237
ARMA (and intervention analysis) 238
Simple smoothers 240
FIR filter 242
Insolation (solar forcing) model 244
Date/time conversion 245
Geometrical menu 246
Circular (one sample) 246
Circular (two samples) 249
Circular correlation 251
Spherical (one sample) 252
Point pattern analysis - nearest neighbours 254
Ripley's *K* point pattern analysis 256
Correlation length analysis 258
Minimal spanning tree analysis 259
Kernel density 260
Point alignments 261
Quadrat counts 262
Spatial autocorrelation (Moran's *I*) 264
Gridding (spatial interpolation) 266
Multivariate allometry 268
PCA of 2D landmarks (relative warps) 269
Thin-plate splines for 2D landmarks 270
Linear regression of 2D landmarks 270
Common Allometric Component analysis for 2D landmarks 271
PCA of 3D landmarks 272
Linear regression of 3D landmarks 272
Common Allometric Component analysis for 3D landmarks 272
Size from landmarks (2D or 3D) NOT YET IN PAST 4 273
All distances from landmarks (EDMA) NOT YET IN PAST 3 274
Edit landmark lines/polygons 275
Elliptic Fourier shape analysis 276
Hangle Fourier shape analysis 277
Eigenshapes 278
Coordinate transformation 279
Open Street Map 280
Measure on image 281
Stratigraphy menu 282
Unitary Associations 282
Ranking-Scaling 287
Range confidence intervals 290
Distribution-free range confidence intervals 291
Sequence slotting 292
Diversity and rates 293
Stratigraphic chart 295

Radiocarbon calibration 297
Grain size analysis 300
Scripting 302
Language structure 302
The output window 309
Accessing the main Past spreadsheet and menus 311
Array and vector operations 312
Scalar math functions 312
File I/O 313
String operations 314
Other functions 316
Calling dll functions (Windows only) 317
Libraries and classes 318
Forms and components 319

Welcome to the PAST!

This program was originally designed as a follow-up to PALSTAT, a software package for paleontological data analysis written by P.D. Ryan, D.A.T. Harper and J.S. Whalley (Ryan et al. 1995). Through continuous development for more than twenty years, PAST has grown into a comprehensive statistics package used not only by paleontologists, but in many fields of life science, earth science, engineering and economics.

Further explanations of many of the techniques implemented together with case histories are found in the book "Paleontological data analysis" (Hammer & Harper 2024).

If you have questions, bug reports, suggestions for improvements or other comments, we would be happy to hear from you. Contact us at ohammer@nhm.uio.no. For bug reports, remember to send us the data used, as saved from PAST, together with a complete description of the actions that lead to the problem.

The latest version of Past, together with documentation and a link to the Past mailing list, is found at **https://www.nhm.uio.no/english/research/resources/past**

We are grateful if you cite PAST in scientific publications. The official reference is Hammer et al. (2001).

References

Hammer, Ø. & Harper, D.A.T. 2024. Paleontological Data Analysis, 2nd ed. Elsevier.

Hammer, Ø., Harper, D.A.T., and P. D. Ryan, 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica* 4(1): 9pp.

Harper, D.A.T. (ed.). 1999. Numerical Palaeobiology. John Wiley & Sons.

Installation

Windows

Download the past5setup5.zip (or with other version number) file, uncompress and run. Windows will consider this a breach of security and will ask if you trust the software provider. If you want to use the program, you will have to answer yes. This procedure requires administrator rights.

If you do not have admin rights, just download the file 'Past5.1.zip' (zipped) and put it anywhere on your hard disk. Double-clicking the file will start the program. However, with this installation mode, the map module (OpenStreet) will not work, and also there will be an error message on exiting Past, which you can disregard.

Mac

Past for Mac must be downloaded from the Apple Store. Search for "Past5" there or use the link on the Past home page. Only newer Macs with ARM processors (M1, M2, M3 etc.) are supported.

Command-line parameters

If you start Past from the command line, you can specify a data file name as the first parameter, and optionally a scripting file name as a second parameter (this script will be loaded and run automatically on startup).

Quick start

Past is to some extent self-explanatory, but a couple of important functions are a bit difficult to find:

How do I export graphics?

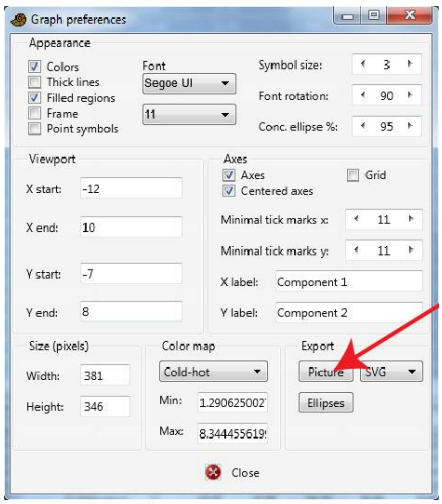
For publication quality, save the graphic in the SVG or PDF vector format. Click the “Graph settings” button next to the graphic. In the graph preferences window, click the “Export picture” button (arrow on the right). You can open SVG files in Adobe Illustrator, Corel Draw or the free program Inkscape. SVG files are supported by most web browsers, and can be placed directly on a web page.

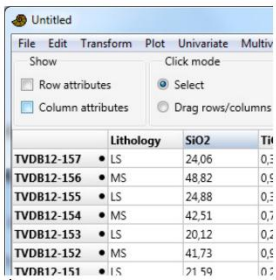
You can also export the picture in bitmap formats (JPG, TIF etc.), but the quality is lower, and you cannot easily edit the graphic. Or you can copy-paste the image as a bitmap by clicking the “Copy” button under the graphic.

Note: On Mac, you must *manually* provide the correct file extension (e.g. .svg or .jpg) when typing the file name, or the file type will not be recognized by other programs.

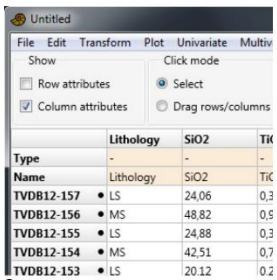
How do I organize data into groups?

This requires a separate group column, with a group identifier for each row. In the example (1) there is a group column “Lithology”, with two groups LS and MS. To specify that “Lithology” is a group column, first select the “Column attributes” box above the spreadsheet, or double-click the top cell (name cell) of any column. This will show two extra rows at the top of the spreadsheet (2). Then click a few times on the “Type” cell of the group column, to bring up a menu where you select “Group” (3). Then click elsewhere to update, and you can deselect “Column attributes” if you wish. The group column should now be marked with a G (4).

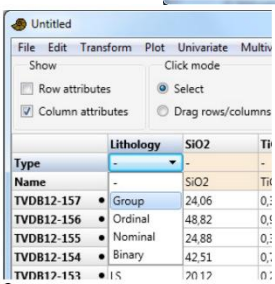




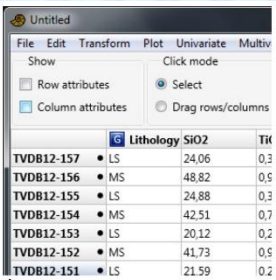
1



2



3



4

The spreadsheet and the Edit menu

PAST has a spreadsheet-like user interface. Data are entered as an array of cells, organized in rows (horizontally) and columns (vertically).

Entering data

To input data in a cell, click on the cell with the mouse and type in the data. The cells can also be navigated using the arrow keys. Any text can be entered in the cells, but most functions will expect numbers. Both comma (,) and decimal point (.) are accepted as decimal separators.

Absence/presence data are coded as 0 or 1, respectively. Any other positive number will be interpreted as presence. Absence/presence-matrices can be shown with black squares for presences by ticking the 'Square mode' box above the array.

Genetic sequence data are coded using C, A, G, T and U (lowercase also accepted).

Missing data are coded with question marks ('?'). Unless support for missing data is specifically stated in the documentation for a function, the function *may not handle missing data correctly*, so be careful.

The convention in PAST is that items occupy rows, and variables columns. Three brachiopod individuals might therefore occupy rows 1, 2 and 3, with their lengths and widths in columns A and B. Cluster analysis will always cluster *items*, that is rows. For Q-mode analysis of associations, samples (sites) should therefore be entered in rows, while taxa (species) are in columns. For switching between Q-mode and R-mode, rows and columns can easily be interchanged using the Transpose operation.

Selecting areas

Most operations in PAST are only carried out on the area of the array which you have *selected* (marked). If you try to run a function which expects data, and no area has been selected, you will get an error message.

- A row is selected by clicking on the row label (leftmost column).
- A column is selected by clicking on the column label (top row).
- Multiple rows are selected by selecting the first row label, then shift-clicking (clicking with the Shift key down) on the additional row labels.
- Multiple columns are similarly marked by shift-clicking the additional column labels.
- You can also select disjunct rows or columns by ctrl-clicking.
- The whole array can be selected by clicking the upper left corner of the array (the empty grey cell) or by choosing 'Select all' in the Edit menu.
- Smaller areas within the array can be selected by clicking and shift-clicking.

Moving a row or a column

Select the ‘Drag rows/columns’ button in the ‘Click mode’ box. A row or a column can now be moved simply by clicking on the label and dragging to the new position.

Renaming rows and columns

When PAST starts, rows are numbered from 1 to 99 and columns are labelled A to Z. For your own reference, and for proper labelling of graphs, you should give the rows and columns more descriptive but short names.

Select the 'Row attributes' option above the spreadsheet to see an editable column of the row names. Select the ‘Column attributes’ option to see an editable row of the column names.

Increasing the size of the array

By default, PAST has 99 rows and 26 columns. If you should need more, you can add rows or columns by choosing 'Insert more rows' or 'Insert more columns' in the Edit menu. Rows/columns will be inserted after the marked area, or at the bottom/right if no area is selected. When loading large data files, rows and/or columns are added automatically as needed.

Cut, copy, paste

The cut, copy and paste functions are found in the Edit menu. You can cut/copy data from the PAST spreadsheet and paste into other programs, for example Word and Excel. Likewise, data from other programs can be pasted into PAST – these need to be in a tab-separated text format.

Before pasting, select the top left cell of the spreadsheet area in Past you want to paste into. Take care not to paste into the possibly hidden column and row attribute fields, unless you mean to.

Remove

The remove function (Edit menu) allows you to remove selected row(s) or column(s) from the spreadsheet. The removed area is not copied to the paste buffer.

Row colors and symbols

Each row can be given a color and a symbol (dot, cross, square etc., or user-defined image). These will be used in scatter plots and other plots. Select the ‘Row attributes’ option (or double-click the name cell of any row) to edit the rows and colors individually, or use the ‘Row colors/symbols’ function to set all selected rows simultaneously (optionally based on the group, see below).

Selecting datatypes for columns, and specifying groups

Each column can be given a datatype using the ‘Column attributes’ mode. Select the ‘Column attributes’ box above the spreadsheet, or double-click the name cell of any column. Then click on the ‘Type’ cell of the column a few times to bring up a small menu where the data type can be selected.

The data types are as follows:

Unspecified (-)

This is the default datatype.

Ordinal, nominal or binary

Specifying one of these types is only required if you wish to use mixed similarity/distance measures.

Group

In a group column, you can enter identifiers for groups of data. You can use integers, or strings such as 'males' and 'females' (without the apostrophes). This will allow group-based polygons or ellipses in scatter plots. A group column is also required for many analyses, such as MANOVA. It is recommended to have rows in the same group as consecutive. Some analyses (e.g. two-way ANOVA) require two or even more group columns.

Note that unlike the first versions of Past, there are no automatic links between colors, symbols and groups. If you wish to use different colors and/or symbols for different groups, you can set up the group column first and then use the 'Row colors/symbols' function in the Edit menu to assign colors/symbols accordingly.

String

So far, this type is only used in the Stratigraphic chart module, for specifying names of periods, zones etc.

Remove uninformative rows/columns

Rows or columns can be uninformative especially with respect to multivariate analyses. Such rows and columns should be considered for removal. Several types can be searched for and removed: Rows or columns with only zeroes, rows or columns with only missing data ('?'), rows or columns with only one non-zero cell (singletons), rows or columns with constant values (zero variance).

Transpose

The Transpose function, in the Edit menu, will interchange rows and columns. This is used e.g. for switching between R mode and Q mode in cluster analysis.

Grouped columns to multivar

Converts from a format with multivariate items presented in consecutive groups of *N* columns to the Past format with one item per row and all variates along the columns. For *N*=2, two specimens and four variables a-d, the conversion is from

a_1 b_1 a_2 b_2

c_1 d_1 c_2 d_2

to

a_1 b_1 c_1 d_1
a_2 b_2 c_2 d_2

Grouped rows to multivar

Converts from a format with multivariate items presented in consecutive groups of *N* rows to the Past format with one item per row and all variates along the columns. For *N*=2, two specimens and four variables a-d, the conversion is from

a_1 b_1
c_1 d_1
a_2 b_2
c_2 d_2

to

a_1 b_1 c_1 d_1
a_2 b_2 c_2 d_2

Stack grouped rows into columns

Stacks groups horizontally along columns. This can be useful e.g. for performing univariate statistics on pairs of columns across groups.

Rows, cols, values to table

Expects three columns of data. The first column contains categories coded as numbers or strings, identifying rows in the output table. The second column contains categories identifying columns. The third column contains the values to be filled into the output table.

Value pairs to matrix

Similar to "Rows, cols, values to table" but expects only two columns of numbers or strings. Each row is one observation. Each column contains categories, e.g. Europe, Africa, Asia in the first column; Dogs, Cats, Foxes in the second column. The occurrences of different combinations are counted, giving a full data matrix, in this case with localities in columns and taxa in rows.

Samples to events (UA to RASC)

Given a data matrix of occurrences of taxa in a number of samples in a number of sections, as used by the Unitary Associations module, this function will convert each section to a single row with orders of events (FADs, LADs or both) as expected by the Ranking-Scaling module. Tied events (in the same sample) will be given equal ranking.

Events to samples (RASC to UA)

Expects a data matrix with sections/wells in rows, and taxa in columns, with FAD and LAD values in alternating columns (i.e. two columns per taxon). Converts to the Unitary Associations presence/absence format with sections in groups of rows, samples in rows and taxa in columns.

Loading and saving data

The 'Open' function is in the File menu. You can also drag a file from the desktop onto the PAST window. PAST uses a text file format for easy importing from other programs (e.g. Word), as follows:

The top left cell must contain a colon (:). Cells are tab-separated. There are two top rows with data types and column names, and three left columns with colors, symbols and row names. Here is an example:

:			-	-	-	Group
			Slow	Med	Fast	Species
Black	Dot	North	4	2	3	0
Black	Dot	South	4	3	7	0
Red	Dot	West	18	24	33	1
Red	Dot	East	10	6	7	1

Optional additional fields can be added to the end of the file, e.g.:

<image n>Filename Specifies an image file name to be used for the 'Image n' symbol, n=1 to 8.

In addition to this format, Past can also detect and open files in the following formats:

- Excel; only the first worksheet, and only .XLS format, not .XLSX.
- Nexus (see below), popular in systematics.
- TPS format developed by Rohlf. The landmark, outlines, curves, id, scale and comment fields are supported, other fields are ignored.
- NTSYS. Multiple tables and trees are not supported. The file must have the extension '.nts'.
- FASTA molecular sequence format, simplified specification according to NCBI.
- PHYLIP molecular sequence format. The file must have the extension '.phy'.
- Arlequin molecular sequence format. For genotype data the two haplotypes are concatenated into one row. Not all options are supported.
- TNT character matrix format, with some restrictions.
- BioGraph format for biostratigraphy (SAMPLES or DATUM format). If a second file with the same name but extension '.dct' is found, it will be included as a BioGraph dictionary.
- RASC format for biostratigraphy. You must open the .DAT file, and the program expects corresponding .DIC and .DEP files in the same directory.
- CONOP format for biostratigraphy. You must open the .DAT file (log file), and the program expects corresponding .EVT (event) and .SCT (section) files in the same directory.

If the file is not recognized, it is assumed to be a general text file with values separated by white space, tabs or commas. The program will then ask about the format of the file.

Importing data from Excel

There are several ways to get data from Excel to Past.

- Copy from Excel and paste into PAST. Make sure you click (select) the top left cell where the data should be placed in Past before pasting. This will depend on whether row or column attributes are included in the data.
- Open the Excel file from PAST (only .XLS, not .XLSX).
- Save as tab-separated text in Excel. The resulting text file can be opened in PAST.

Reading and writing Nexus files

The Nexus file format is used by many systematics programs. PAST can read and write the Data (character matrix) block of the Nexus format. Interleaved data are supported. Also, if you have performed a parsimony analysis and the 'Parsimony analysis' window is open, all shortest trees will be written to the Nexus file for further processing in other programs (e.g. MacClade or Paup). Note that not all Nexus options are presently supported.

Counter

A counter function is available in the Edit menu for use e.g. at the microscope when counting microfossils of different taxa. A single row (sample) must be selected. The counter window will open with a number of counters, one for each selected column (taxon). The counters will be initialized with the column labels and any counts already present in the spreadsheet. When closing the counter window, the spreadsheet values will be updated.

Count up (+) or down (-) with the mouse, or up with the keys 0-9 and a-z (only the first 36 counters). The bars represent relative abundance. A log of events is given at the far right - scroll up and down with mouse or arrow keys. An optional auditive feedback has a specific pitch for each counter.

Transform menu

These routines subject your data to mathematical operations. This can be useful for bringing out features in your data, or as a necessary preprocessing step for some types of analysis.

Logarithm

The Log function in the Transform menu log-transforms your data using the base-10 logarithm. If the data contain zero or negative values, it may be necessary to add a constant (e.g. 1) before log-transforming (use Evaluate Expression x+1).

This is useful, for example, to compare your sample to a log-normal distribution or for fitting to an exponential model. Also, abundance data with a few very dominant taxa may be log-transformed in order to downweight those taxa.

Missing data supported.

Subtract mean

This function subtracts the column mean from each of the selected columns. The means cannot be computed row-wise.

Missing values supported.

Remove trend

This function removes any linear trend from a data set (two columns with X-Y pairs, or one column with Y values). This is done by subtraction of a linear regression line from the Y values. Removing the trend can be a useful operation prior to time series analyses such as spectral analysis, auto- and cross-correlation and ARMA.

Missing data supported.

Row percentage

All values converted to the percentage of the row sum. Missing values supported.

Row normalize length

All values divided by the Euclidean length of the row vector. This is sometimes called chord transformation. Missing values supported.

Box-Cox

The Box-Cox transformation is a family of power transformations with the purpose of making data x more normally distributed. The transformation has a parameter λ:

$$y = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases}$$

If the smallest input value is zero or negative (which would invalidate the transform), a constant is added to all data such that the minimum input value becomes 1.

The default value of the parameter is calculated by maximizing the log likelihood function:

$$L(\lambda) = -\frac{n}{2} \ln \hat{\sigma}_{\lambda}^{2} + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \ ,$$

where σ^2_λ is the variance of the transformed data. Alternatively, λ can be found by maximizing the Shapiro-Wilk test statistic *W* for the transformed data.

This optimal value can be changed by the user, limited to the range $-4 \leq \lambda \leq 4$.

Missing values supported.

Compositional data transforms

Multivariate data which sum to a constant by design, such as percentages summing to 100, are called compositional data (Aitchison 1986). Such data contain "spurious" correlations because as one value increases, the others will have to decrease. Some multivariate analyses and tests such as PCA can be negatively affected by this. Past includes three commonly used transforms that can be applied to compositional data before further analysis.

The data should have the usual multivariate format, with variables in columns and items in rows. The values in each row should sum to a constant, e.g. 1 or 100. Negative values are not allowed, and missing data are not supported.

Additive logratio (ALR)

The input data (one row) is a vector ***x*** with *N* dimensions

$$\mathrm{alr}(\boldsymbol{x}) = \left[ln \frac{x_1}{x_N}, \quad \cdots, \quad ln \frac{x_N}{x_N} \right]$$

That last element is equal to zero, showing that the transformed data have dimension *N*-1. As the ALR is computed with respect to the last element x_N, it may be a good idea to place a low-noise variable with high values in the last column.

Center logratio (CLR)

$$\mathrm{clr}(\boldsymbol{x}) = \left[\ln \frac{x_1}{g(\boldsymbol{x})}, \quad \ldots, \quad \ln \frac{x_N}{g(\boldsymbol{x})} \right]$$

where $g(\boldsymbol{x})$ is the geometric mean of the data vector. This is equivalent to a simple log transform followed by subtraction of the mean.

Isometric log ratio (ILR)

The isometric logratio transform was introduced by Egozcue et al. (2003). It has some good theoretical properties, but the results are difficult to interpret because the transformed variables are complicated combinations of the original variables.

Define a matrix $\mathbf{H_0}$ with dimensions NxN, with ones in the first row, and each subsequent row j containing j-1 ones followed by the value –(j-1) on the diagonal followed by zeros:

$$\mathbf{H_0} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & -2 & 0 & \cdots & 0 \\ 1 & 1 & 1 & -3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & -(N-1) \end{bmatrix}$$

Then, normalize each row to unit length. This gives the so-called *Helmert matrix* **H** of order N. Remove the first row of ones, giving an (N-1) x N matrix **V** (we had that first row there just to make the connection to the Helmert matrix). Finally, apply the center logratio to the data and pre-multiply with –**V** to obtain the ilr:

$$ilr(\boldsymbol{x}) = -\boldsymbol{V}\, clr(\boldsymbol{x})$$

The negative sign is included only to reproduce the results of the compositional package in R. Also note that R defines the Helmert matrix as the transpose of **H** above.

The transformed data vector has dimensions N-1. The last column N is filled with zeros for reference, and should not be included in further analysis.

Treatment of zero values

All three transforms involve log-transforming the original data, so zero values cannot be transformed directly. Past treats zero values following equation (6) in Martin-Fernandez et al. (2003):

$$r_j = \begin{cases} \delta & \text{if } x_j = 0 \\ \left(1 - \frac{z}{c}\right) x_j & \text{if } x_j > 0 \end{cases}$$

Here, δ is a small value, approximating to the lower detection limit of the measurements. This value can be set by the user in the "Zero threshold" box (default 0.01). The z is the number of zeroes in the original data vector, while c is the total sum, computed from the data (e.g. 100 for percentages).

References

Aitchison J. 1986. The Statistical Analysis of Compositional Data. Chapman & Hall.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35:279-300.

Martin-Fernandez, J.A., Barcelo-Vidal, C., Pawlowsky-Glahn, V. 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* 35:253-278

Remove size from distances

Attempts to remove the size component from a multivariate data set of measured distances (specimens in rows, variables in columns). Three methods are available.

- *Isometric Burnaby's method* projects the set of measured distances onto a space orthogonal to the first principal component. Burnaby's method may (or may not!) remove isometric size from the data, for further "size-free" data analysis. Note that the implementation in PAST does not center the data within groups - it assumes that all specimens (rows) belong to one group.
- *Allometric Burnaby's method* will log-transform the data prior to projection, thus conceivably removing also allometric size-dependent shape variation from the data.
- *Allometric vs. standard* estimates allometric coefficients with respect to a standard (reference) measurement *L* such as overall length (Elliott et al. 1995). This standard variable should be placed in the first column. Each additional column is regressed onto the first column after log-transformation, giving a slope (allometric coefficient) *b* for that variable. An adjusted measurement is then computed from the original value *M* as

$$M_{adj} = M\left(\frac{\bar{L}}{L}\right)^b$$

Reference

Elliott, N.G., K. Haskard & J.A. Koslow 1995. Morphometric analysis of orange roughy (*Hoplostethus atlanticus*) off the continental slope of southern Australia. *Journal of Fish Biology* 46:202-220.

Landmarks, Procrustes fitting

Transforms your measured point coordinates to Procrustes coordinates. There is also a menu choice for Bookstein coordinates. Specimens go in different rows and landmarks along each row. If you have three specimens with four landmarks in 2D, your data should look as follows:

x1 y1 x2 y2 x3 y3 x4 y4
x1 y1 x2 y2 x3 y3 x4 y4
x1 y1 x2 y2 x3 y3 x4 y4

For 3D the data will be similar, but with additional columns for *z*.

Landmark data in this format could be analyzed directly with the multivariate methods in PAST, but it is recommended to standardize to Procrustes coordinates by removing position, size and rotation. A further transformation to Procrustes residuals (approximate tangent space coordinates) is achieved

by selecting 'Subtract mean' in the Edit menu. You must convert to Procrustes coordinates first, then to Procrustes residuals.

The “Rotate to major axis” option places the result into a standard orientation for convenience.

The “Keep size” option adds a final step where the shapes are scaled back to their original centroid sizes.

The “Allow mirroring” (reflection) selection is only available for 2D. For 3D, this option is always on. Note that the default in the procGPA function in R is “off”, while in MorphoJ it is “on”. It will usually not make any difference, as reflection will only be optimal for data sets with very large variance.

A thorough description of Procrustes and tangent space coordinates is given by Dryden & Mardia (1998). The algorithms for Procrustes fitting are from Rohlf & Slice (1990) (2D) and Dryden & Mardia (1998) (3D). It should be noted that for 2D, the iterative algorithm of Rohlf & Slice (1990) often gives slightly different results from the direct algorithm of Dryden & Mardia (1998). Past uses the former in order to follow the “industry standard”.

Missing data is supported but only by column average substitution, which is perhaps not very meaningful.

References

Dryden, I.L. & K.V. Mardia 1998. Statistical Shape Analysis. Wiley.

Rohlf, F.J. & Slice, D. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology* 39:40-59.

Landmarks, Bookstein fitting

Bookstein fitting has a similar function as Procrustes fitting, but simply standardizes size, rotation and scale by forcing the two first landmarks onto the coordinates (0,0) and (1,0). It is not in common use today. Bookstein fitting is only implemented for 2D.

Project to tangent space (not yet in Past 4)

After Procrustes or Bookstein fitting, some statistical procedures are ideally carried out on tangent space projected coordinates (usually it doesn’t make any difference, but don’t quote us on that!). With *d* the number of dimensions and *p* the number of landmarks, the projection is

$$\mathbf{X}' = \mathbf{X}\left(\mathbf{I}_{dp} - \mathbf{X}_c^{\mathrm{t}}\mathbf{X}_c\right).$$

Here, **X** is the *nxdp* matrix of *n* specimens, **X’** is the transformed matrix, **I** the *dpxdp* identity matrix, and $\mathbf{X}_c$ the mean (consensus) configuration as a *dp*-element row vector.

Remove size from landmarks (not yet in Past 4)

The 'Remove size from landmarks' option in the Transform menu allows you to remove size by dividing all coordinate values by the centroid size for each specimen (Procrustes coordinates are also normalized with respect to size).

See Dryden & Mardia (1998), p. 23-26.

Reference

Dryden, I.L. & K.V. Mardia 1998. Statistical Shape Analysis. Wiley.

Transform landmarks

Allows rotation of the point cloud in steps of 90 degrees, and top-bottom or left-right flipping (mirroring), mainly for plotting convenience. The mirror operation may be useful for reducing a bilaterally symmetric landmark data, by Procrustes fitting the left half to a mirrored version of the right half (and optionally averaging the two).

Only for 2D coordinates.

Regular interpolation

Interpolates an irregularly sampled time series or transect (possibly multivariate) into a regular spacing, as required by many methods for time series analysis. The *x* values should be in the first selected column. These will be replaced by a regularly increasing series. All additional selected columns will be interpolated correspondingly. The perils of interpolation should be kept in mind.

You can either specify the total number of interpolated points, or the new point spacing. Four interpolation methods are available. The antialiasing interpolation uses a 50-point sinc (FIR) filter with a Hamming window, low-pass filtering at half the new sampling frequency (averaged for uneven sampling) to reduce aliasing when downsampling.

Evaluate expression

This powerful feature allows flexible mathematical operations on the selected array of data. Each selected cell is evaluated, and the result replaces the previous contents. A mathematical expression must be entered, which can include any of the operators +, -, *, /, ^ (power), and mod (modulo). Also supported are brackets (), and the functions abs, atan, asin, cos, gtzer (greater than zero; 1 if x>0, 0 if x<=0), sin, exp, ln, sqrt, sqr, round, tan, and trunc.

The following values are also defined:

- x (the contents of the current cell)
- l (the cell to the left if it exists, otherwise 0)
- r (the cell to the right)
- u (the cell above, or up)
- d (the cell below, or down)
- mean (the mean value of the current column)
- median (the median of the current column)
- min (the minimum value)
- max (the maximum value)
- n (the number of cells in the column)
- i (the row index)
- j (the column index)
- random (uniform random number from 0 to 1)

- normal (Gaussian random number with mean 0 and variance 1).
- integral (running sum of the current column)
- stdev (standard deviation of the current column)
- iqr (interquartile range of the column)
- sum (total sum of the current column)
- pi

In addition, other columns can be referred to using the column name preceded by '%', for example %A.

Examples:

sqrt(x)	Replaces all numbers with their square roots
(x-mean)/stdev	Mean and standard deviation normalization, column-wise
x-0.5*(max+min)	Centers the values around zero
(u+x+d)/3	Three-point moving average smoothing
x-u	First-order difference
I	Fills the column with the row numbers (requires non-empty cells, such as all zeros)
sin(2*3.14159*i/n)	Generates one period of a sine function down a column (requires non-empty cells)
5*normal+10	Random number from a normal distribution, with mean of 10 and standard deviation of 5.

Missing values supported.

Plot menu

Graph

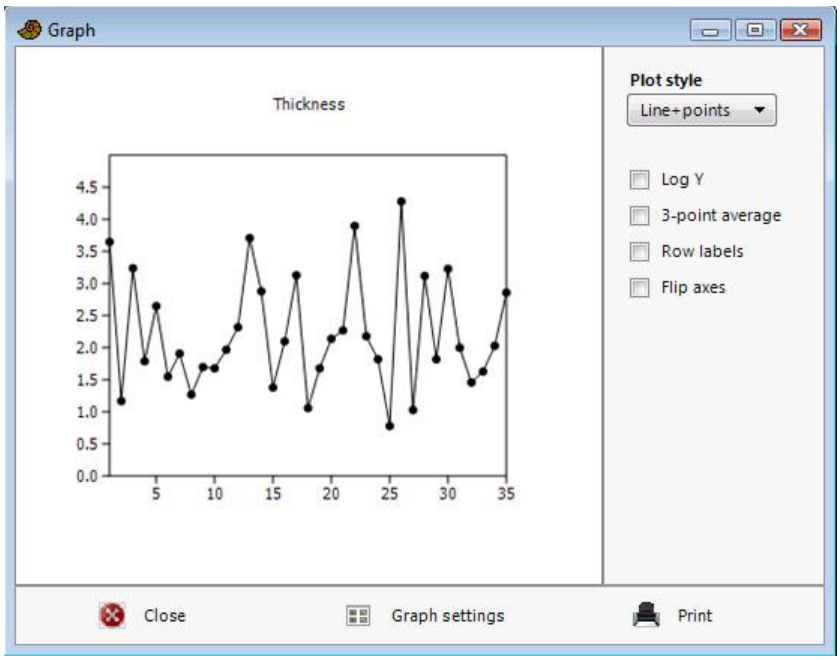
Plots one or more columns as separate graphs. You can also use a group column. It is also possible to show each row, instead of each column, as a separate graph, with the “Plot rows” option. The *x* coordinates are set automatically to 1,2,3,...

There are six plot styles available: Line, points, line with points, bars, steps and stems (vertical lines). The 'Row labels' option sets the *x* axis labels to the appropriate row names.

The “Log Y” option log-transforms the values to base 10. For values <=0, the log value is set to 0.

The sequence can be smoothed with a 3-point moving average.

Missing values are disregarded.



XY graph

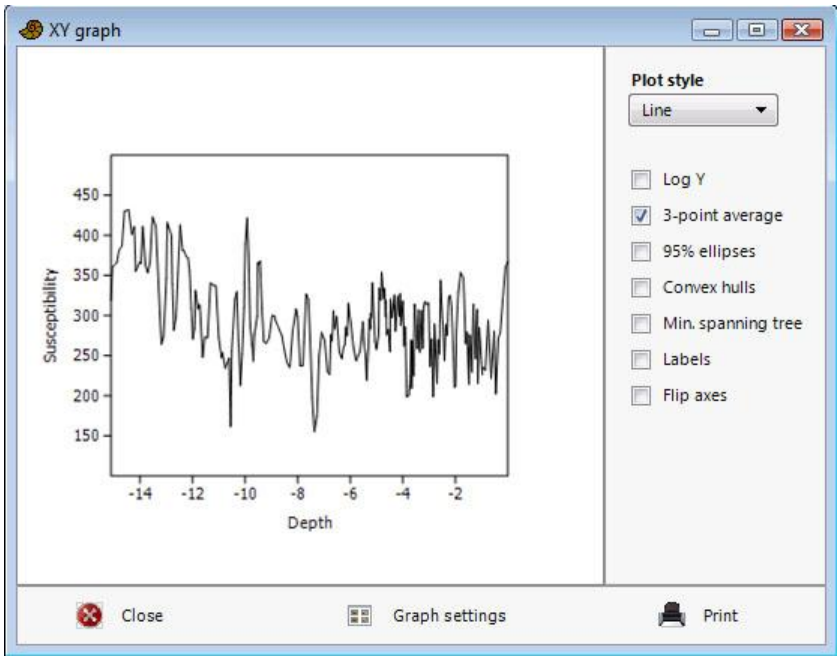
Plots one or more pairs of columns containing x/y coordinate pairs. The 'log Y' option log-transforms your Y values (zero or negative values are set to 0). The curve can also be smoothed using 3-point moving average.

95% concentration ellipses can be plotted in most scatter plots in PAST, such as scores for PCA, CA, DCA, PCO and NMDS. The calculation of these ellipses assumes bivariate normal distribution. They estimate a region where 95% of population points are expected to fall, i.e. they are not confidence regions for the mean.

Convex hulls can also be drawn in the scatter plots, in order to show the areas occupied by points of different groups. The convex hull is the smallest convex polygon containing all points.

The minimal spanning tree is the set of lines with minimal total length, connecting all points. In the XY graph module, Euclidean lengths in 2D are used.

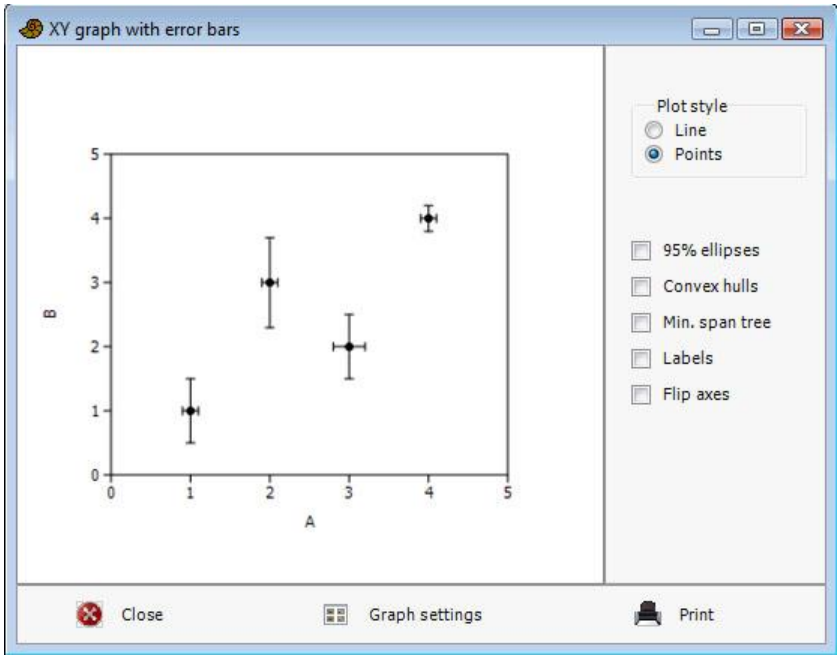
Points with missing values in X and/or Y are disregarded.



XY graph with error bars

As XY graph, but expects four columns (or a multiple), with x, y, x error and y error values. Symmetric error bars are drawn around each point, with half-width as specified. If an error value is set to zero or missing, the corresponding error bar is not drawn.

Points with missing values in X and/or Y are disregarded.



Histogram

Plots histograms (frequency distributions) for one or more columns. You can also use a group column. The number of bins is by default set to an "optimal" number (the zero-stage rule of Wand 1997), with bin width

$$h = 3.49 \min(\mathrm{s}, IQ/1.349) n^{-1/3}$$

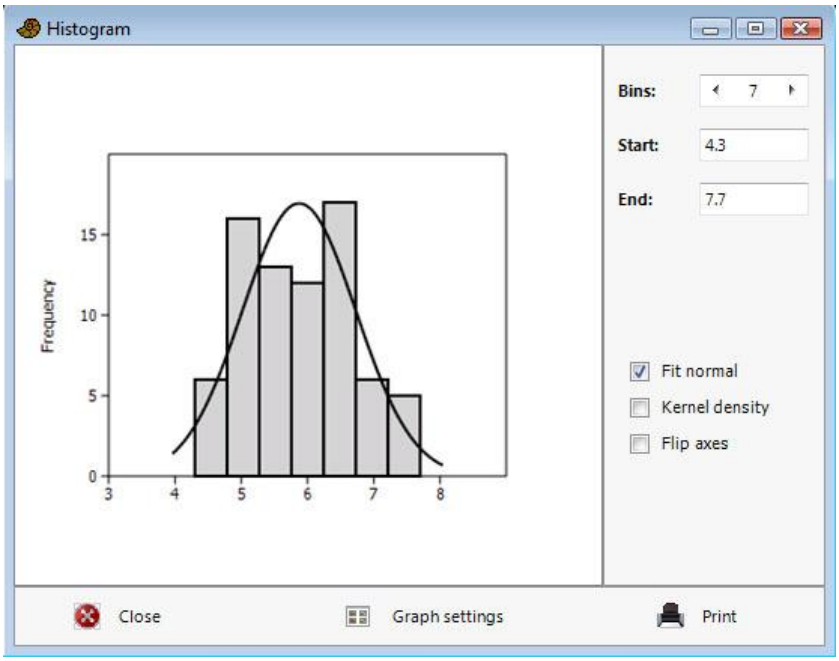
where *s* is the sample standard deviation and *IQ* the interquartile range. The number of bins can be changed by the user. When two columns are selected, they can be plotted as a bihistogram, i.e. “mirror image” histograms which are easily compared.

The "Fit normal" option draws a graph with a fitted normal distribution (Parametric estimation, not Least Squares).

Kernel Density Estimation is a smooth estimator of the histogram. PAST uses a Gaussian kernel with range according to the rule given by Silverman (1986):

$$h = 0.9 \min(s, IQ/1.34) n^{-1/5}.$$

Missing values are deleted.



References

Silverman, B.W. 1986. Density estimation for statistics and data analysis. Chapman & Hall.

Wand, M.P. 1997. Data-based choice of histogram bin width. *American Statistician* 51:59-64.

Bar chart/box plot

Bar plot, box plot, mean-and-whisker plot, jitter plot or violon plot for one or several columns (samples) of univariate data. Alternatively, you can use a group column. If you include *two* group columns, you get a grouped bar/box plot with the first group column specifying the group of bars, and the second the bar within the group. There is also an option for using the first value in each column for setting the x position of the column in the plot. Missing values are disregarded.

Bar chart

For each sample, the mean value is shown by a bar. In addition, “whiskers” can optionally be shown. The whisker interval can represent a one-sigma or a 95% confidence interval (1.96 sigma) for the estimate of the mean (based on the standard error), or a one-sigma or 95% concentration interval (based on the standard deviation).

Box plot

For each sample, the 25-75 percent quartiles are drawn using a box. The median is shown with a horizontal line inside the box. The minimal and maximal values are shown with short horizontal lines ("whiskers").

If the "Outliers" box is ticked, another box plot convention is used. The whiskers are drawn from the top of the box up to the largest data point less than 1.5 times the box height from the box (the "upper inner fence"), and similarly below the box. Values outside the inner fences are shown as circles, values further than 3 times the box height from the box (the "outer fences") are shown as stars.

The “Notches” option visualizes an approximate 95% confidence interval for the median.

The quartile methods (rounding or interpolation) are described under “Percentiles” below.

Mean and whisker plot

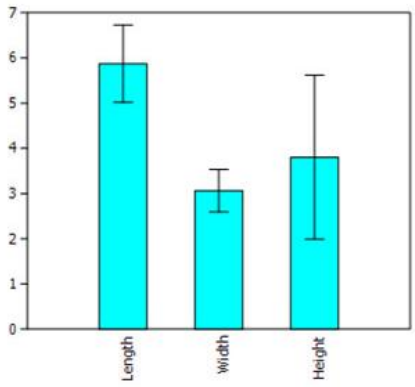
Similar to bar chart, but without the bar, showing the mean as a point with whiskers for the standard error, standard deviation or 95% intervals.

Jitter plot

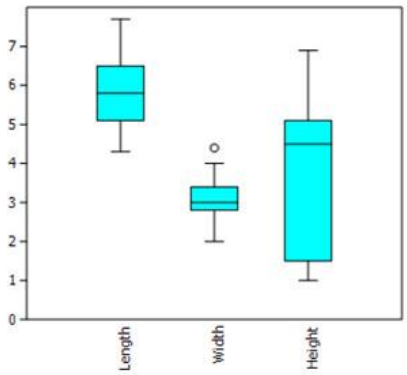
Each value is plotted as a dot. To show overlapping points more clearly, they can be displaced using a random “jitter” value controlled by a slider.

Violin plot

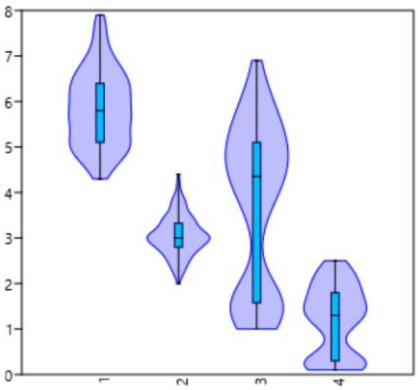
The violin plot shows a kernel density plot (“continuous histogram”) for each sample. The plot ranges from the minimum to the maximum value. A box plot (described above) can optionally be shown on top of the violin.



Bar chart



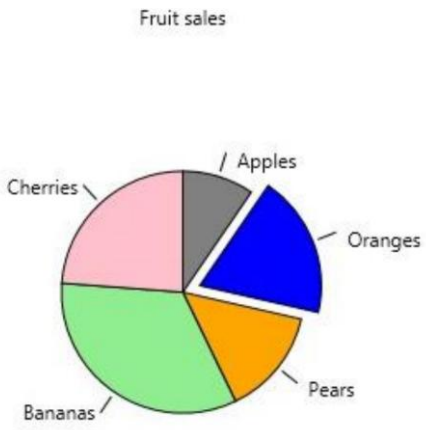
Box plot



Violin plot with box plot

Pie chart

Plots a pie chart or doughnut chart from a single column of data, or several columns for multiple charts. A sector can be emphasized by “explosion”:



With the “Coords in first rows” option, the program expects x-y coordinates in the first two rows of each column, followed by the pie chart data. This allows a combination scatterplot/pie chart, with multiple pies in a coordinate system.

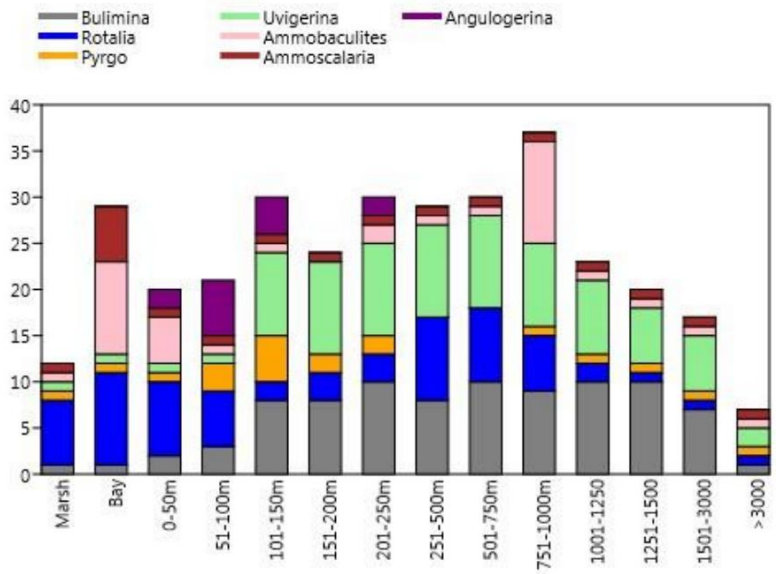
The “Sort sizes” option will sort the slices from largest to smallest, in clockwise order.

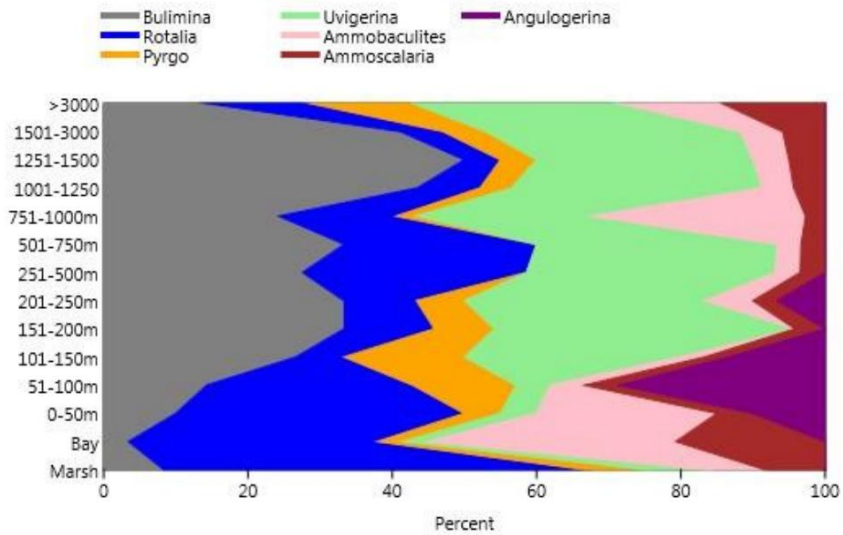
Missing values are disregarded.

Stacked chart

One or more rows of data can be plotted as stacked bar chart or stacked area chart. Each bar represents one row, and the data along columns are plotted cumulatively. The 'Percentage' option converts to percentages of the row total, so that all bars will be of equal height (100%).

Missing data are treated as zero.



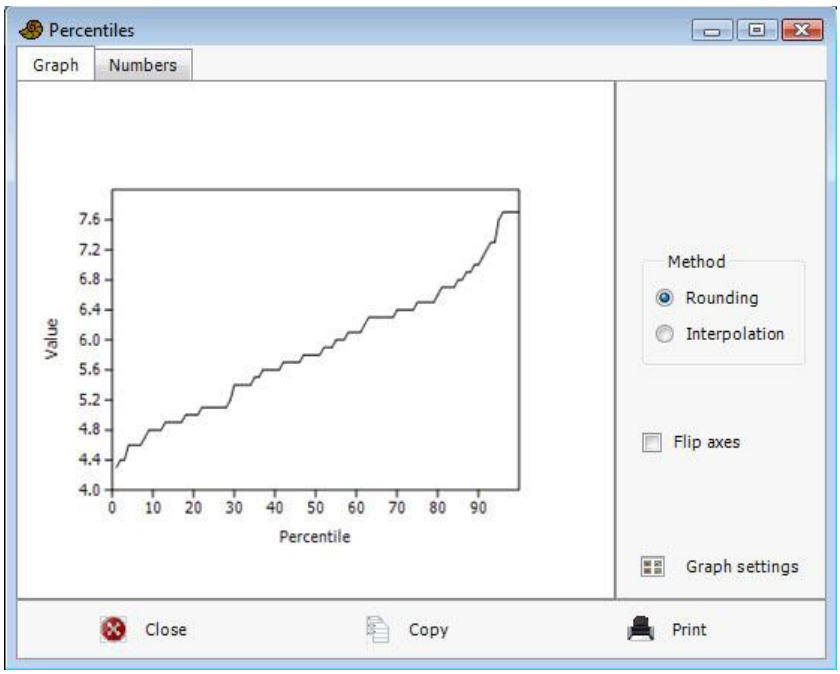


Stacked area chart with percentage option, axes flipped

Percentiles

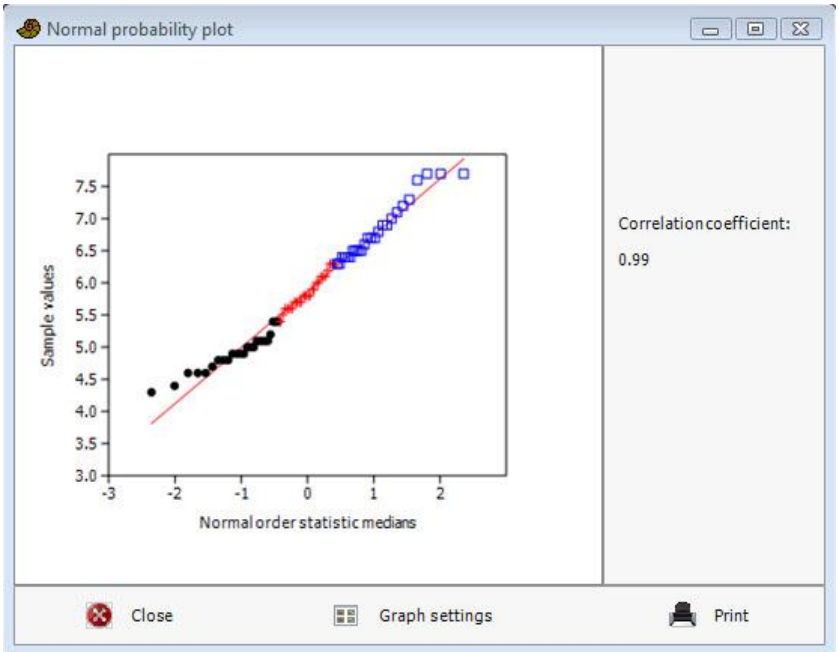
For each percentile *p*, plots the value *y* such that *p* percent of the points are smaller than *y*. Multiple plots can be produced using several data columns or a group column. Two popular methods are included. For a percentile *p*, the rank is computed according to $k=p(n+1)/100$, and the value that corresponds to that rank taken. In the rounding method, *k* is rounded to the nearest integer, while in the interpolation method, non-integer ranks are handled by linear interpolation between the two nearest ranks.

Missing values are disregarded.



Normal probability plot

Plots a normal probability (normal QQ) plot for one or more columns of data. A normal distribution will plot on a straight line. For comparison, an RMA regression line is given, together with the Probability Plot Correlation Coefficient.



(three groups were given in this example)

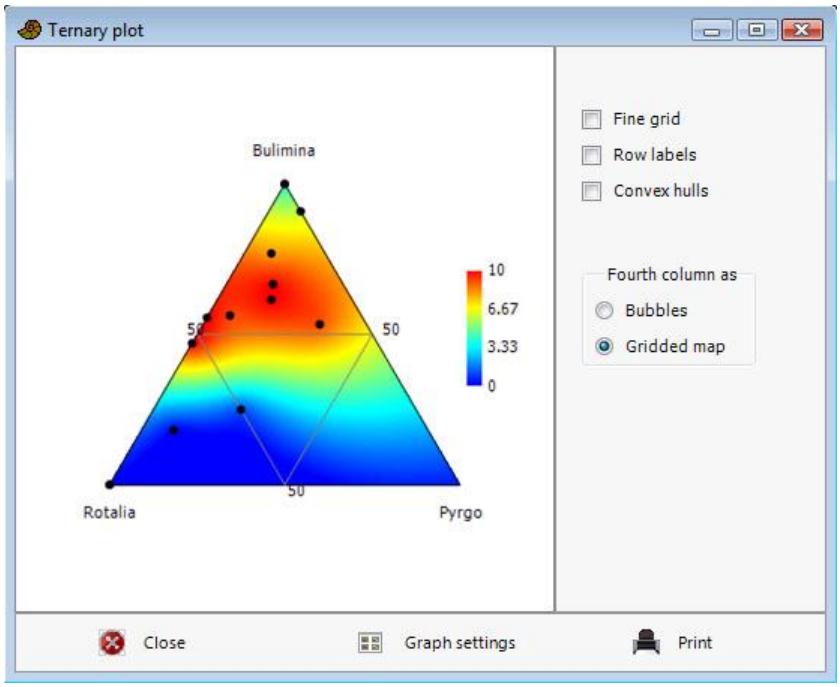
Missing values are disregarded.

The normal order statistic medians are computed as $N(i) = G(U(i))$, where G is the inverse of the cumulative normal distribution function and U are the uniform order statistic medians:

$$U(i) = \begin{cases} 1 - U(n), & i = 1 \\ i - 0.3175/(n + 0.365) & i = 2,3,\ldots n-1 \\ 0.5^{1/n} & i = n \end{cases}$$

Ternary

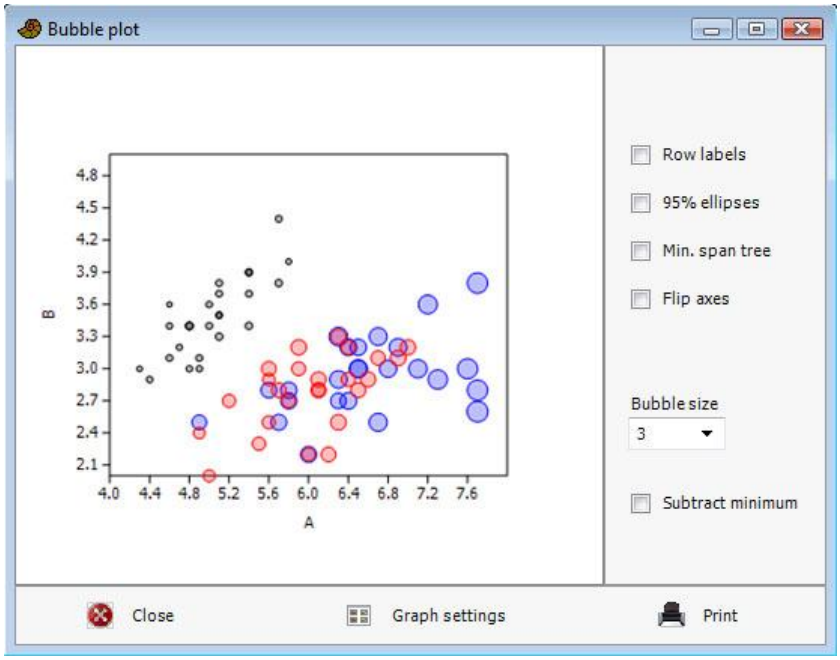
Ternary plot for three columns of data, normally containing proportions of compositions. A color map of point density (computed with a kernel density method) can also be shown. If a fourth column is included, it will be shown using either a bubble representation or as a color/grayscale map.



Rows with missing value(s) in any column are disregarded. When using the color map option, rows with only the fourth variable missing are included in the plot but do not contribute to the map.

Bubble plot

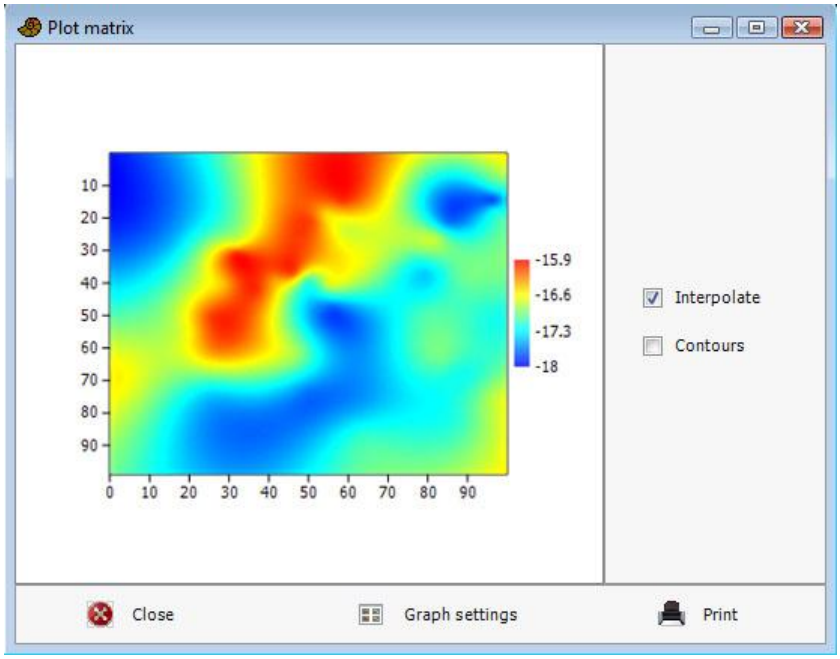
Plotting 3D data (three columns) by showing the third axis as size of disks. Negative values are not shown. Select "Subtract min" to subtract the smallest third axis value from all values - this will force the data to be positive. The "Size" slider scales the bubbles relative to unit radius on the x axis scale.



Rows with missing value(s) in any column are disregarded.

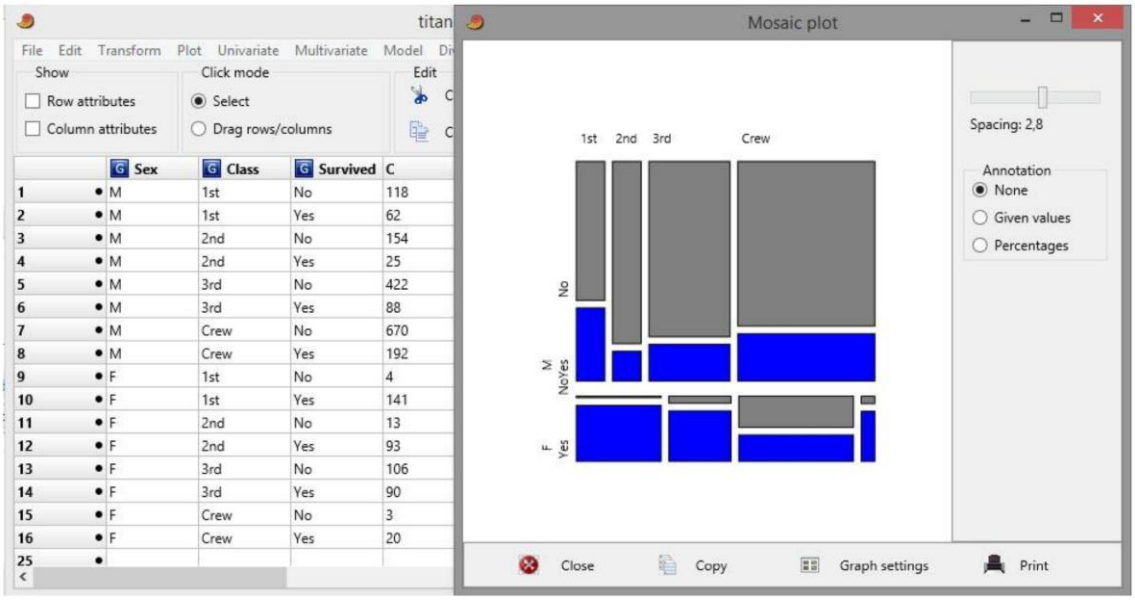
Matrix plot

Two-dimensional plot of the data matrix, using a grayscale with white for lowest value, black for highest, or a colour scale. Includes contouring. Use to get an overview over a large data matrix. Missing values are plotted as blanks (allowing holes and non-square boundaries).



Mosaic plot

Shows proportions in a two-way or three-way contingency table as areas of rectangles. A two-way table can be given as a simple data matrix or with two group columns and a single data column (there must be exactly one row for each combination of group levels). Column widths will reflect column totals, and row heights will reflect row totals. The interpretation of a three-way contingency table (specified with three group columns) is a little more complicated – see below for the standard "Titanic" example, described on e.g. the Wikipedia page about mosaic plots.



Venn diagram

Plotting of Venn diagrams for two or three sets, with a number of options. Circles can be plotted with equal sizes, or with sizes and overlaps proportional to the number of members.

Two sets

The input counts for two sets A and B can be given as a 2x2 contingency table on this form:

A and B	A, not B
B, not A	Not A, not B

The "not A, not B" value is only used when the "Show none" box is ticked.

Alternatively, the values can be given in a single column with 3 or 4 rows as follows:

A, not B
B, not A
A and B
Not A, not B (this value is optional)

Finally, you can supply a binary (0/1) matrix with any number of items in rows, and two columns A and B corresponding to the two sets. Each item (row) can be a member of set A (value 1 in column A) or set B (value 1 in column B), or neither (0 in both columns) or both (1 in both columns).

Three sets

The input counts for three sets A, B and C are given in a single column with 7 or 8 rows:

A, not B, not C (ABC=100)
B, not A, not C (ABC=010)
A and B, not C (ABC=110)
C, not A, not B (ABC=001)
A and C, not B (ABC=101)
B and C, not A (ABC=011)
A and B and C (ABC=111)
not A, not B, not C (ABC=000) (this value is optional)

Example data:

20
10
12
8
9
4
3

80

Alternatively, you can supply a binary matrix with 3 columns and any number of rows, similarly as described above for two sets.

Plotting a fully size-proportional 3-set Venn diagram with circles is not generally possible. Past prioritizes the total sizes of circles and the pairwise overlapping regions. The algorithm is inspired by the “matplotlib-venn” Python code by Konstantin Tretyakov.

Four sets

Only the binary input format is available, with four columns corresponding to the four sets, and items in rows. A full, four-set Venn diagram cannot be constructed with circles, so ellipses are used. Proportional sizes are not attempted.

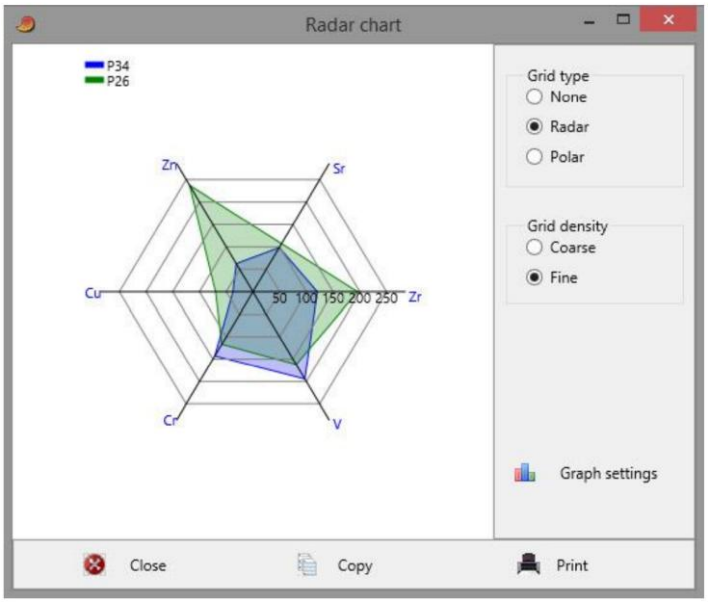
Radar chart

For visualizing multivariate data. One radar chart (polygon) will be plotted for each row in the data. In the example below, the data consisted of two rows (P34 and P26) and six columns (variables). The grid lines can be polygons (radar chart) or circles (polar chart).

The module will also accept a single column of data.

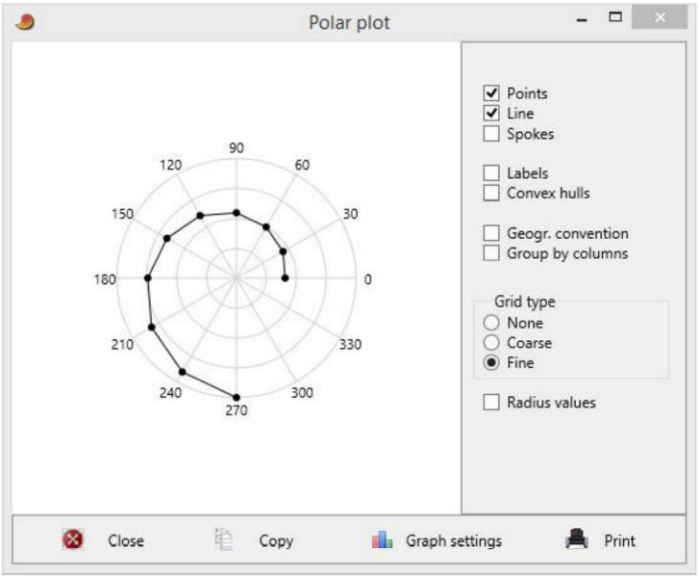
Another use of this module is for visualizing circular or cyclic data such as animal activity through 24 hours or sunlight through 12 months.

Missing values are treated as zero.



Polar plot

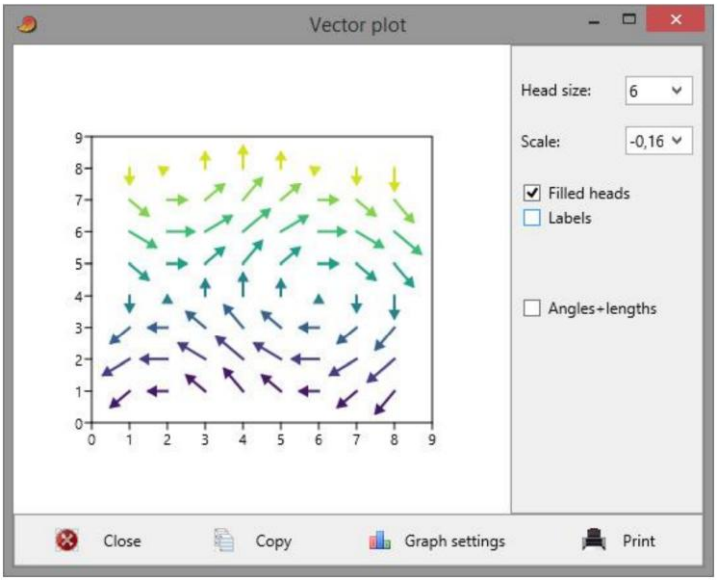
This plot accepts polar coordinates, with angles in degrees in the first column and radius values in the second column. By default, the angles are assumed to go counterclockwise from east (see figure). By ticking "Geographical convention" the angles are assumed to go clockwise from north.



Vector plot

Accepts four columns with vector start points (x and y) and vector x and y components. An optional fifth column can specify line thickness.

There is also an option to allow the vector components (columns 3 and 4) to be specified in polar coordinates, as angles in degrees (counter-clockwise from East) and lengths.



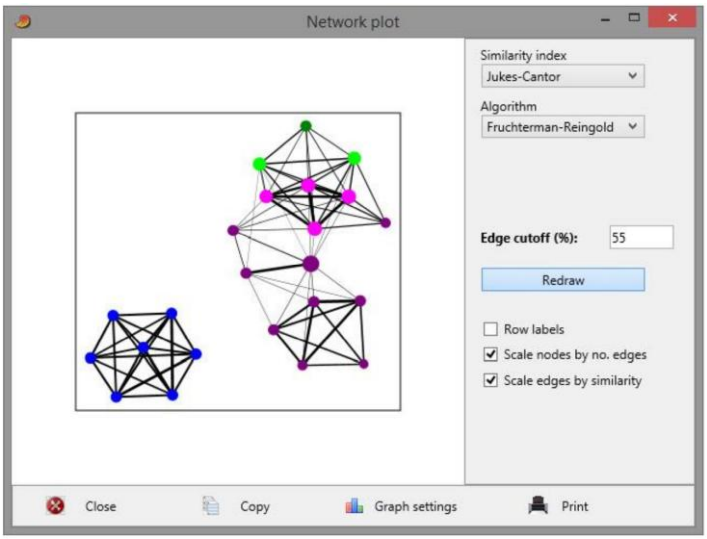
(In this example, the colors were specified by the row colors in the spreadsheet).

Network plot

This module plots networks (graphs), with nodes (rows in the spreadsheet) connected by edges. You can specify the network with an adjacency matrix in the spreadsheet (only the lower triangle needs to be given). In this matrix, a 1 in row *i*, column *j* implies an edge from node *i* to node *j*. All other cells should be zero. For this type of input data, you must select "User-defined similarity" as the similarity index. You can also specify similarities other than 1 for variable edge thickness (see below).

Alternatively, you can plot a network showing similarities between rows, using your raw data matrix and any similarity or distance measure you choose. You can then choose a similarity cutoff (in percent) to control the number of edges included. Zero percent cutoff will give a fully connected (complete) network with edges between all pairs of nodes; fifty percent will show only edges between nodes that are more than 50% similar.

The "Scale nodes by no. edges" option will set the diameter of nodes proportional to the number of edges connected to it. The "Scale edges by similarity" will set the thickness of edges proportional to the similarity.



The nodes can be arranged in a circle, or they can be positioned with the Fruchterman-Reingold algorithm (1991). Using a random starting position, this algorithm will produce a new layout each time, so click "Redraw" a few times until you get a pleasing result.

The "Bipartite linear" option plots a bipartite graph with the data rows shown as graph nodes in one vertical column and the data columns as nodes in another column. A non-zero value in a data cell will produce an edge from the corresponding row node to the column node. This format is typically used for ecological data matrices with samples in rows and taxa in columns.

"Bipartite Fruchterman-Reingold" produces a similar bipartite graph but using the F-R algorithm.

Network (graph) statistics

The following statistics are calculated for the unweighted graph:

- Number of nodes *M*

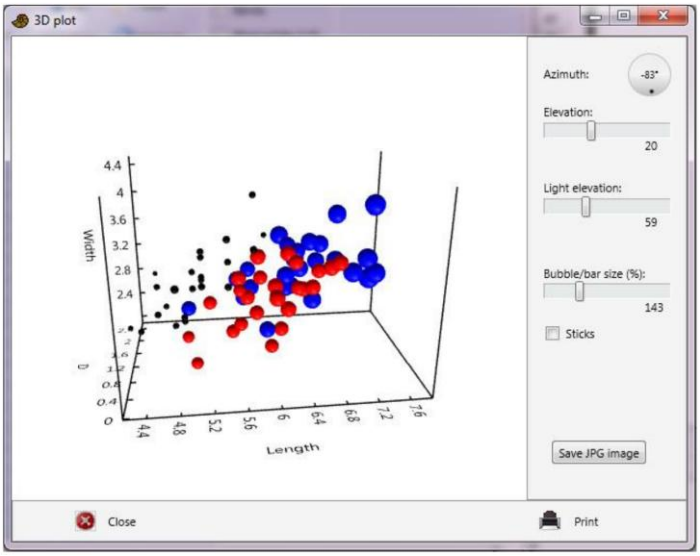
- Number of edges *N*
- Graph density: Number of edges divided by the maximal possible number of edges, *N* / (*M*(*M*-1)/2)
- Average number of neighbors
- Graph diameter (largest eccentricity, which is the max distance from each node)
- Graph radius (smallest eccentricity)
- Characteristic path length (average shortest distance between any pair of nodes)
- Local clustering coefficient

Reference

Fruchterman, T. M. J. & Reingold, E. M. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience* 21:1129–1164.

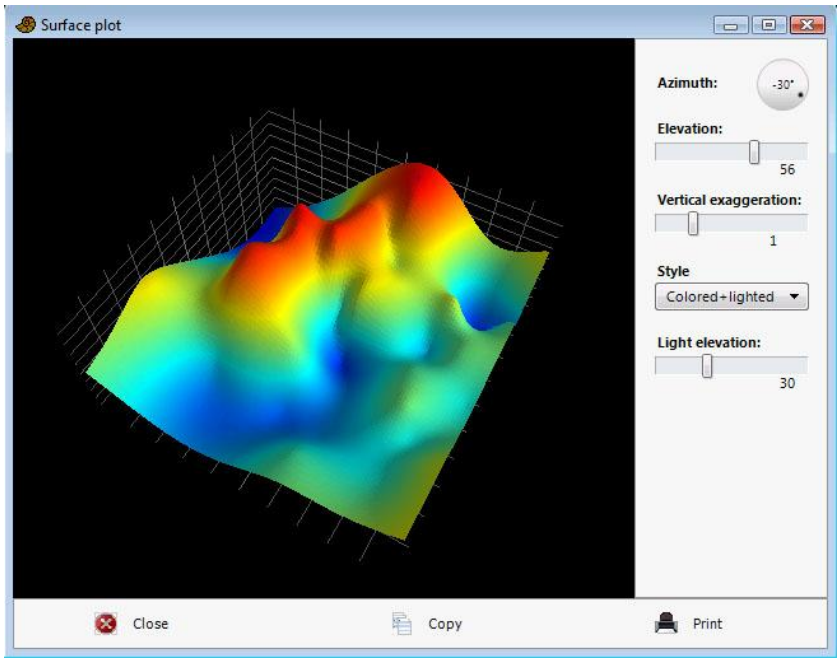
3D scatter/bubble/line plot

Requires three or four columns of data. For three columns, the data are plotted as fixed-size spheres (or other symbols as given in the spreadsheet) with the given xyz coordinates. An optional fourth column is shown as sizes of bubbles. The coordinate system is right-handed, with the z axis vertical (positive up). Sticks can be added to emphasize the positions in the xy plane. Select the 'Lines' box to draw lines between the points.



3D surface plot

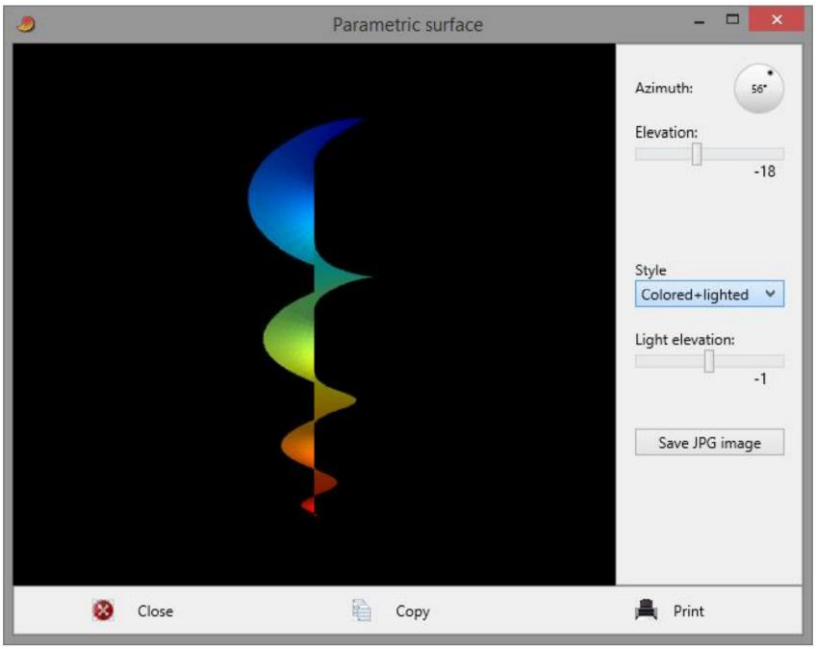
Three-dimensional landscape plot of a data matrix of elevation values. Colors are assigned according to height, and/or the surface can be gray-shaded using a lighting model. Vertical exaggeration is adjustable. Missing values are replaced with the average. The data in the example below are the same as for the matrix plot above.



3D parametric surface plot

Plots a 3D parametric surface given as a matrix of xyz triples. Such a matrix will usually be generated by a script. Consider for example the following Past script, writing a parabolic helicoid into the spreadsheet:

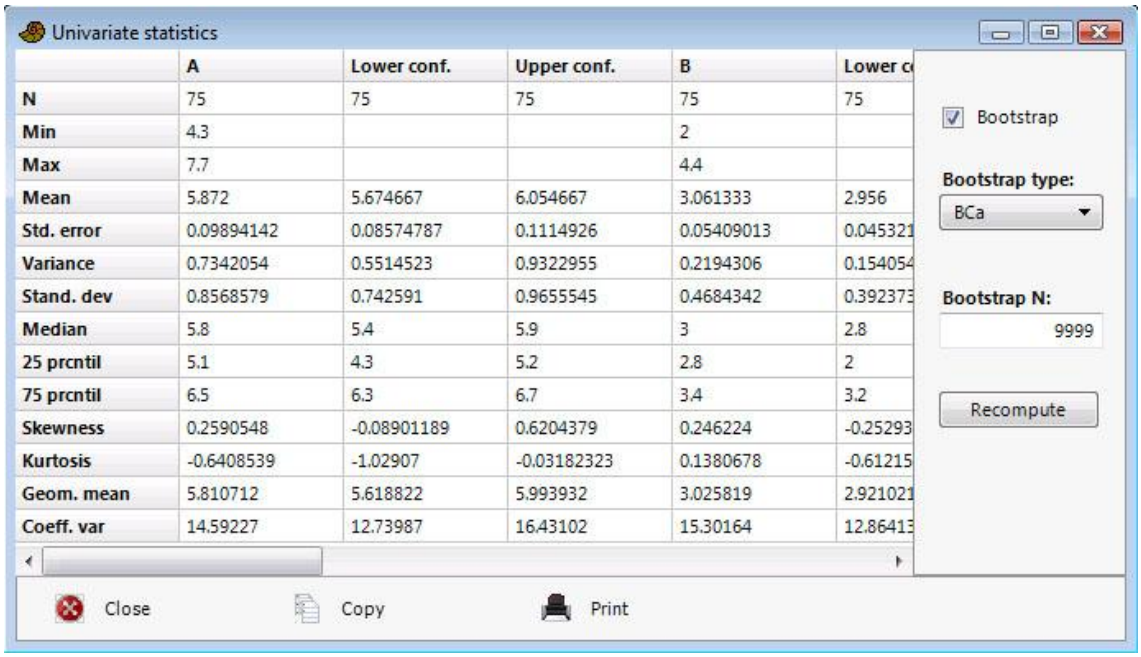
```
tablesize(10, 400);
for i:=1 to 100 do begin
      phi:=4*2*pi*(i-1)/100;
      z:=0.2*phi*phi;
      for j:=1 to 8 do begin
            r:=phi*(j-1)/7;
            x:=r*cos(phi);
            y:=r*sin(phi);
            tableout(j-1, (i-1)*3, x);
            tableout(j-1, (i-1)*3+1, y);
            tableout(j-1, (i-1)*3+2, z);
      end;
end;
```



Univariate menu

Summary statistics

This function computes descriptive statistics for one or more samples (columns) of univariate data. The samples can be given in one or more separate columns or with a single data column and a group column. Each sample must have at least 3 values. The columns can have different numbers of values.



	A	Lower conf.	Upper conf.	B	Lower c
N	75	75	75	75	75
Min	4.3			2	
Max	7.7			4.4	
Mean	5.872	5.674667	6.054667	3.061333	2.956
Std. error	0.09894142	0.08574787	0.1114926	0.05409013	0.045321
Variance	0.7342054	0.5514523	0.9322955	0.2194306	0.154054
Stand. dev	0.8568579	0.742591	0.9655545	0.4684342	0.392373
Median	5.8	5.4	5.9	3	2.8
25 prcntil	5.1	4.3	5.2	2.8	2
75 prcntil	6.5	6.3	6.7	3.4	3.2
Skewness	0.2590548	-0.08901189	0.6204379	0.246224	-0.25293
Kurtosis	-0.6408539	-1.02907	-0.03182323	0.1380678	-0.61215
Geom. mean	5.810712	5.618822	5.993932	3.025819	2.921021
Coeff. var	14.59227	12.73987	16.43102	15.30164	12.86413

The following numbers are shown for each sample:

N: The number of values *n* in the sample

Min: The minimum value

Max: The maximum value

Mean: The estimate of the mean, calculated as

$$\bar{x} = \frac{\sum x_i}{n}$$

Std. error: The standard error of the estimate of the mean, calculated as

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where *s* is the estimate of the standard deviation (see below).

Variance: The sample variance, calculated as

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 .$$

Stand. dev.: The sample standard deviation, calculated as

$$s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}\ .$$

Median: The median of the sample. For *n* odd, the given value such that there are equally many values above and below. For *n* even, the average of the two central values.

25 prcntil: The 25th percentile, i.e. the given value such that 25% of the sample is below, 75% above. The "interpolation" method is used (see Percentile plot above).

75 prcntil: The 75th percentile, i.e. the given value such that 75% of the sample is below, 25% above. The "interpolation" method is used (see Percentile plot above).

Mode: The most common value. If there is no single most common value (e.g. if all values are different), the mode is reported as "NA" (not available)

Skewness: The sample skewness, zero for a normal distribution, positive for a tail to the right. Calculated as

$$G_1 = \frac{n}{(n-1)(n-2)} \frac{\sum(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}\right)^3}\ .$$

Note there are several versions of this around – Past uses the same equation as SPSS and Excel. Slightly different results may occur using other programs, especially for small sample sizes.

Kurtosis: The sample kurtosis, zero for a normal distribution. Calculated as

$$G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum(x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}\right)^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}\ .$$

Again, Past uses the same equation as SPSS and Excel.

Geom. mean: The geometric mean, calculated as $(x_1 x_2 \cdots x_n)^{1/n}$. Logarithms are used internally.

Harm. mean: The harmonic mean, $n/\sum 1/x_i$.

Coeff.var: Coefficient of variation, or ratio of standard deviation to the mean, in percent:

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}}{\bar{x}} \cdot 100$$

Bootstrapping

Selecting bootstrapping will compute lower and upper limits for 95% confidence intervals, using the specified number of bootstrap replicates. Confidence intervals for the min and max values are not

given, because bootstrapping is known to not work well for these statistics. Three different bootstrap methods are available (cf. Davison and Hinkley 1997):

Simple (basic):

The statistic estimated from the original sample is *t*. The simulated estimates from *R* bootstrap replicates are t_1^*, t_2^*, …, t_R^*. For a 95% CI, we set the one-tailed error α=0.025. The simple (or basic) bootstrapped CI is then

$$\left[2t - t^*_{(R+1)(1-\alpha)},\quad 2t - t^*_{(R+1)\alpha}\right].$$

To ensure integer-valued subscripts, values for *R* such as 999, 9999 or 99999 are convenient.

Percentile:

An even simpler estimate:

$$\left[t^*_{(R+1)\alpha},\quad t^*_{(R+1)(1-\alpha)}\right].$$

BCa (adjusted percentile method):

This is a complex method, but somewhat more accurate than the simple and percentile bootstrap. Estimate a bias correction factor (called *z* in some texts):

$$w = \Phi^{-1}\left(\frac{\left|\left\{t_r^* < t\right\}\right|}{R+1}\right),$$

where Φ is the cumulative normal function and $|\cdot|$ is the number of elements in the set. Note we use strictly less than, unlike some sources. Then calculate a skewness correction factor:

$$a = \frac{\sum_{i=1}^{n}\left(t_{-i} - \overline{t_-}\right)^3}{6\left(\sum_{i=1}^{n}\left(t_{-i} - \overline{t_-}\right)^2\right)^{\frac{3}{2}}},$$

where t_{-i} is the statistic computed with value *i* removed (jackknifed), and $\overline{t_-}$ is the mean of the jackknifed values. With these values for w and z, compute adjusted CI endpoint values

$$a_1 = \Phi\left(w + \frac{w - 1.96}{1 - a(w - 1.96)}\right)$$

$$a_2 = \Phi\left(w + \frac{w + 1.96}{1 - a(w + 1.96)}\right),$$

where 1.96 is the approximate quantile for the normal distribution corresponding to a 95% CI (the actual value used is 1.959964). The bootstrapped confidence interval is

$$\left[t^*_{(R+1)a_1},\quad t^*_{(R+1)a_2}\right].$$

No interpolation is used if the index is not an integer.

Missing data: Supported by deletion.

One-sample tests

Tests for whether a single sample (single column of data) comes from a population with a given, often hypothetical, mean or median. For example, are a number of oxygen isotope values from sea shells (single sample) the same as average seawater composition (given mean)? The given test value must be typed in. In addition, single-case tests are used to test whether a single value comes from the same population as the given sample.

One-sample *t* test for given mean μ_0 (parametric)

Sample mean and standard deviation are estimated as described above under Univariate statistics. The 95% confidence interval for the difference in means is based on the standard error for the estimate of the mean, and the *t* distribution. Normal distribution is assumed. With *s* the estimate of the sample standard deviation, the confidence interval is

$$\left[|\bar{x}-\mu_0| - t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}}, \qquad |\bar{x}-\mu_0| + t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}}\right]$$

Here, *t* has *n*-1 degrees of freedom, and 1-α = 0.95 for a 95% confidence interval. The *t* test has null hypothesis

H_0: The sample is taken from a population with mean μ_0.

The test statistic is

$$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$$

One-sample Wilcoxon signed-rank test for given median *M* (nonparametric)

The one-sample Wilcoxon test has null hypothesis

H_0: The sample is taken from a population with median *M*.

All values equal to *M* are first removed by the program. Then the absolute values of the differences $|d_i|$ are ranked (R_i), with mean ranks assigned for ties. The sum of ranks for pairs where d_i is positive is W^+. The sum of ranks for pairs where d_i is negative is W^-. The reported test statistic is

$W = \max(W^+, W^-)$

(note that there are several other, equivalent versions of this test, reporting other statistics).

For large *n* (say *n*>10), the large-sample approximation to *p* can be used. This depends on the normal distribution of the test statistic *W*:

$$E(W) = \frac{n(n+1)}{4}$$

$$\mathrm{Var}(W) = \frac{n(n+1)(2n+1)}{24} - \frac{\sum_g f_g^3 - f_g}{48}$$

The last term is a correction for ties, where f_g is the number of elements in tie *g*. The resulting *z* is reported, together with the *p* value.

For *n*<13, an exact *p* value is computed, by complete enumeration of all possible reassignments (there are 2^n of them, e.g., 4096 for *n*=12). This is the preferred *p* value, if available.

Single-case tests

The single-case tests have null hypothesis

H_0: The given single value *y* is taken from the same population as the given sample.

Normal distribution is assumed. A simple *z* test is often used for this purpose, and is also provided by Past. However, the *z* test is inaccurate because it assumes that the mean and standard deviations are given exactly, whereas in reality, they are estimated from the sample. Therefore, Past also provides a modified *t* test (Sokal & Rohlf 1995; Crawford & Howell 1998):

$$t = \frac{y - \bar{x}}{s\sqrt{\frac{n+1}{n}}}$$

with s the sample standard deviation and *n*-1 degrees of freedom.

Binomial proportion

Expects binary data (0 or non-zero) in the given sample. The proportion of non-zeroes in the sample is compared with the given proportion, and confidence intervals are also reported.

The same test is provided in the Single proportion module, but there a binary data column is not required, only the observed proportion.

References

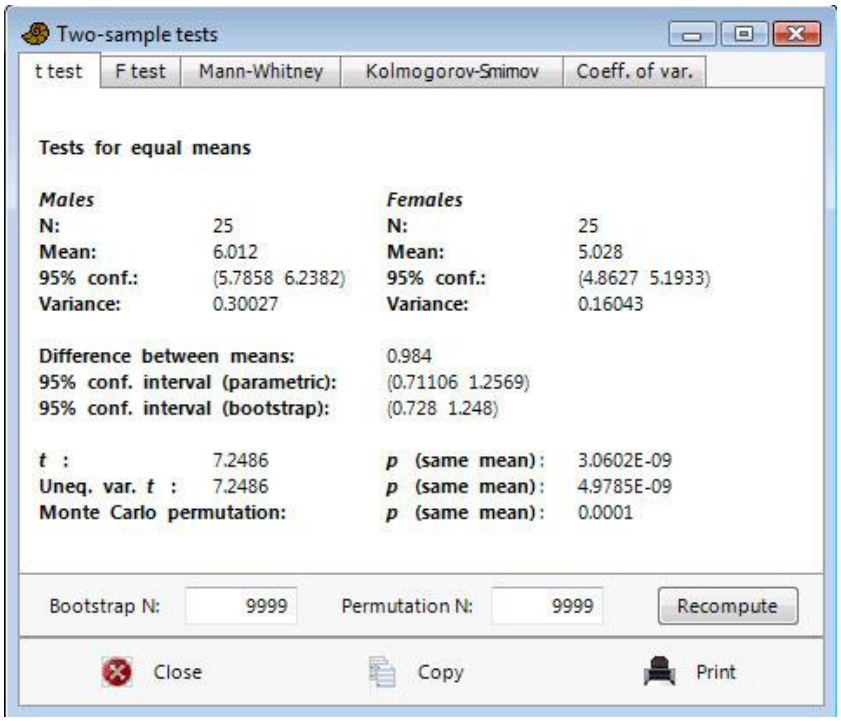
Crawford, J.R. & Howell, D.C. 1998. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist* 12:482-486.

Sokal, R.R. & Rohlf, J.F. 1995. Biometry. W.H. Freeman, San Francisco.

Two-sample tests

A number of classical statistics and tests for comparing two univariate samples, as given in two columns. It is also possible to specify the two groups using a single column of values and an additional Group column. Missing data are disregarded.

t test and related tests for equal means



Sample statistics

Means and variances are estimated as described above under Univariate statistics. The 95% confidence interval for the mean is based on the standard error for the estimate of the mean, and the *t* distribution. Normal distribution is assumed. With *s* the estimate of the standard deviation, the confidence interval is

$$\left[\bar{x} - t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}},\quad \bar{x} + t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}}\right].$$

Here, *t* has *n*-1 degrees of freedom, and 1-α = 0.95 for a 95% confidence interval.

The 95% confidence interval for the difference between the means accepts unequal sample sizes:

$$\left[\left|\overline{x-y}\right| - t_{(\alpha/2,df)}s_D,\quad \left|\overline{x-y}\right| + t_{(\alpha/2,df)}s_D\right],$$

where

$$SSE = \sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2$$

$$df = (n_1 - 1) + (n_2 - 1)$$

$$MSE = SSE / df$$

$$n_h = \frac{2}{1/n_1 + 1/n_2}$$

$$s_D = \sqrt{\frac{2MSE}{n_h}} \quad .$$

The confidence interval is computed for the larger mean minus the smaller, i.e. the center of the CI should always be positive. The confidence interval for the difference in means is also estimated by bootstrapping (simple bootstrap), with the given number of replicates (default 9999).

t test

The *t* test has null hypothesis

H_0: The two samples are taken from populations with equal means.

The *t* test assumes normal distributions and equal variances.

From the standard error s_D of the difference of the means given above, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{s_D} \quad .$$

Unequal variance t test

The unequal variance *t* test is also known as the Welch test. It can be used as an alternative to the basic *t* test when equal variances cannot be assumed. Note that this version of the *t* test is the default in R. The test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\mathrm{Var}(x)/n_1 + \mathrm{Var}(y)/n_2}} \quad .$$

The number of degrees of freedom is

$$df = \frac{\left[\frac{\mathrm{Var}(x)}{n_1} + \frac{\mathrm{Var}(y)}{n_2}\right]^2}{\frac{[\mathrm{Var}(x)/n_1]^2}{n_1 - 1} + \frac{[\mathrm{Var}(y)/n_2]^2}{n_2 - 1}} \quad .$$

Monte Carlo permutation test

The permutation test for equality of means uses the absolute difference in means as test statistic. This is equivalent to using the *t* statistic. The permutation test is non-parametric with few assumptions, but the two samples are assumed to be equal in distribution if the null hypothesis is

true. The number of permutations can be set by the user. The power of the test is limited by the sample size – significance at the $p<0.05$ level can only be achieved for $n>3$ in each sample.

Exact permutation test

As the Monte Carlo permutation test, but all possible permutations are computed. Only available if the sum of the two sample sizes is less than 27.

Bayes factor

The Jeffrey-Zellner-Siow (JZS) Bayes Factor is reported in favour of the alternative, i.e. it quantifies the evidence for the hypothesis of unequal means. It is calculated according to Rouder et al. (2009),

$$\mathrm{BF}_{10} = \frac{\int_0^\infty (1+Ngr^2)^{-1/2}\left(1+\frac{t^2}{\nu(1+Ngr^2)}\right)^{-(\nu+1)/2}(2\pi)^{-1/2}g^{-3/2}e^{-1/2g}dg}{\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}}$$

where $N = n_1 n_2 / (n_1+n_2)$ and $\nu = n_1 + n_2 - 2$ (degrees of freedom). The t value is the usual t statistic, and the r value is the Cauchy scale factor, fixed at $\sqrt{2}/2 = 0.707$. The integral is approximated numerically.

A BF value larger than 3 may be taken as "substantial" evidence for unequal means, but this is relative to the assumed "uninformative" prior (Rouder et al. 2009). Likewise, a BF value smaller than 1/3 may be taken as substantial evidence for equal means, but this rarely happens unless N is very large. BF>10 is reported as "strong evidence"; BF>30 as "very strong"; and BF>100 as "decisive evidence".

Cohen's d

Cohen's d (Cohen 1988) is a measure of effect size. Past calculates the classical version without bias correction (this usually does not matter):

$$d = \frac{\bar{x}-\bar{y}}{s}$$

where s is the pooled standard deviation

$$s = \sqrt{\frac{(n_1-1)\mathrm{Var}(x)+(n_2-1)\mathrm{Var}(y)}{n_1+n_2-2}}$$

The textual interpretation of the value for Cohen's d follows Sawilowsky (2009).

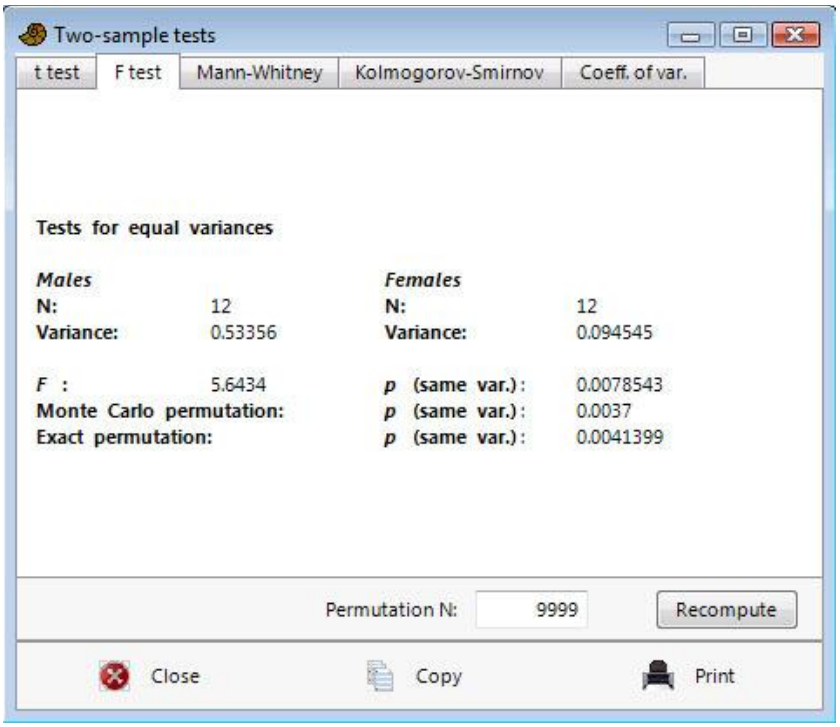
References

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences (2nd ed.)*. Academic Press, NY

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D. & Iverson, G. 2009. Bayesian ***t*** tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Revue* 16:225–237.

Sawilowsky, S. 2009. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods* 8:467–474.

F test for equal variances



The *F* test has null hypothesis

H_0: The two samples are taken from populations with equal variance.

Normal distribution is assumed. The *F* statistic is the ratio of the larger variance to the smaller. The significance is two-tailed, with n_1 and n_2 degrees of freedom.

Monte Carlo and exact permutation tests on the *F* statistic are computed as for the *t* test above.

Note: As usual, the permutation test assumes interchangeability. This assumption is broken if the population means are not equal. To partly alleviate this problem, Past subtracts the sample means before permutation. However, this hack is not ideal (Baker 1995).

Reference

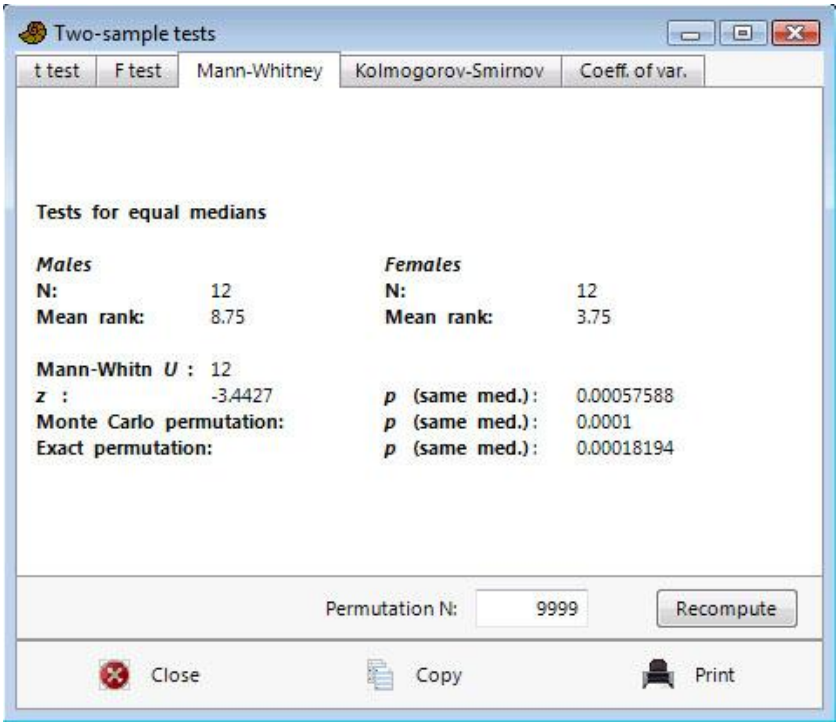
Baker, R. 1995. Two permutation tests of equality of variances. *Statistics and Computing* 5, 289-296.

Mann-Whitney test for statistic equality

The two-tailed (Wilcoxon) Mann-Whitney *U* test is a non-parametric test and does not assume normal distribution. The null hypothesis is

H_0: For randomly selected values *x* and *y* from the two populations, the probability of *x*>*y* is equal to the probability of *y*>*x*.

If the distributions are equal except for location, the null hypothesis can be interpreted as equal medians.



For each value in sample 1, count the number of values in sample 2 that are smaller than it (ties count 0.5). The total of these counts is the test statistic *U* (sometimes called *T*). If the value of *U* is smaller when reversing the order of samples, this value is chosen instead (it can be shown that $U_1+U_2=n_1n_2$).

The program computes an asymptotic approximation to *p* based on the normal distribution (two-tailed), which is only valid for large *n*. It includes a continuity correction and a correction for ties:

$$z = \frac{U - n_1 n_2/2 + 0.5}{\sqrt{\dfrac{n_1 n_2\left(n^3 - n - \sum_g f_g^3 - f_g\right)}{12n(n-1)}}}$$

where $n=n_1+n_2$ and f_g is the number of elements in tie *g*.

A Monte Carlo value based on the given number of random permutations (default 9999) is also given – the purpose of this is mainly as a control on the asymptotic value.

For $n_1+n_2<=26$ (e.g. 13 values in each group), an exact *p* value is given, based on all possible group assignments. If available, always use this exact value. For larger samples, the asymptotic approximation is quite accurate.

Vargha-Delaney A (effect size)

The Vargha-Delaney effect size (Vargha & Delaney 2000) is an estimate of the proportion P(X > Y) + 0.5*P(X = Y). Values close to 0.5 indicate small or no difference; close to 0 indicates X smaller than Y; close to 1 indicates X larger than Y.

Brunner-Munzel test for stochastic equality

The Brunner-Munzel test (Brunner & Munzel 2000; Karch 2021) has the same null hypothesis as Mann-Whitney:

H_0: For randomly selected values *x* and *y* from the two populations, the probability of *x*>*y* is equal to the probability of *y*>*x*.

However, the Brunner-Munzel test is less sensitive to difference in distribution and variance and is therefore possibly superior to Mann-Whitney. The Brunner-Munzel test usually gives slightly higher *p* values than Mann-Whitney. The equations used in Past were taken from Fagerland et al. (2011).

The "phat" value is an estimate of the proportion P(X < Y) + 0.5*P(X = Y). This also known as the Vargha-Delaney *A* effect size (see above).

If there is no overlap between the two samples, the Brunner-Munzel statistic is not defined. Past then sets the *p* value to 0.

A Monte Carlo value based on the given number of random permutations (default 9999) is also given (Neubert & Brunner 2007)

.

References

Brunner, E., Munzel, U. 2000. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation". *Biometrical Journal* 42:17–25.

Fagerland M.W., Sandvik L., Mowinckel P. 2011, Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables – Additional File 1: Test statistics. *BMC Medical Research Methodology* 11:44.

Karch J.D. 2021. Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Advances in Methods and Practices in Psychological Science* 2021;4(2).

Neubert, K., Brunner, E. 2007. A studentized permutation test for the non-parametric Behrens-Fisher problem". *Computational Statistics & Data Analysis* 51:5192–5204.

Vargha, A., Delaney. H.D. 2000. A critique and improvement of the *CL* common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* 25:101–252.

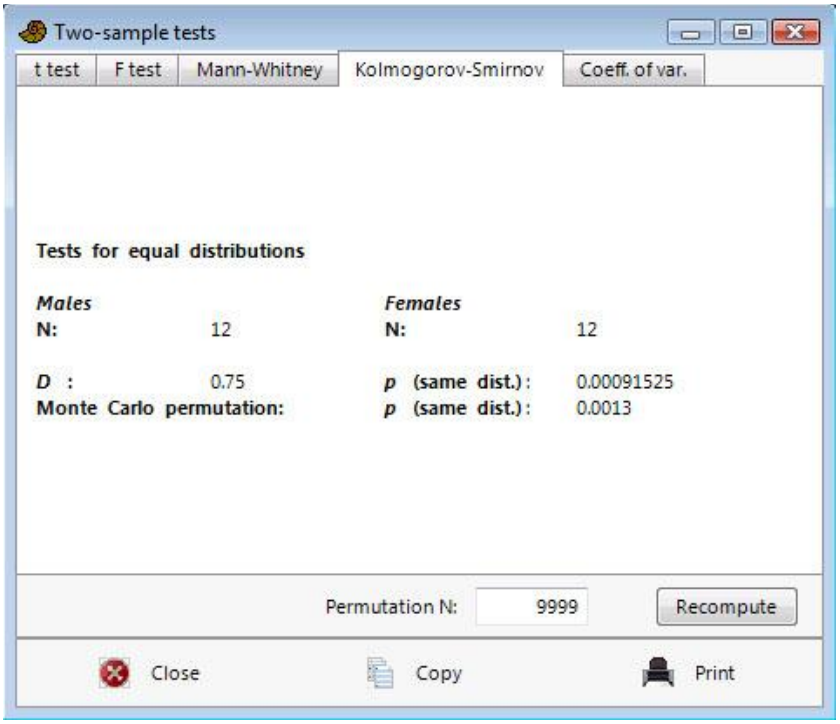
Mood's median test for equal medians

The median test is an alternative to the Mann-Whitney test. It truly tests for equal medians even for unequal distributions. The median test has low power, and the Mann-Whitney test is therefore usually preferable. However, there may be cases with strong outliers where the Mood's test may perform better.

The test simply counts the number of values in each sample that are above or below the pooled median, producing a 2x2 contingency table that is tested with a standard chi-squared test with two degrees of freedom, and with Yate's correction. For smaller data sets, a Fisher's exact test is also reported.

Kolmogorov-Smirnov test for equal distributions

The Kolmogorov-Smirnov test is a nonparametric test for overall equal distribution of two univariate samples. In other words, it does not test specifically for equality of mean, variance or any other parameter. The null hypothesis is H_0: The two samples are taken from populations with equal distribution.



In the version of the test provided by Past, both columns must represent samples. You cannot test a sample against a theoretical distribution (one-sample test).

The test statistic is the maximum absolute difference between the two empirical cumulative distribution functions:

$$D = \max_x \left| S_{N_1}(x) - S_{N_2}(x) \right|$$

The algorithm is based on Press et al. (1992), with significance estimated after Stephens (1970). Define the function

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2}.$$

With $N_e = N_1N_2/(N_1+N_2)$, the significance is computed as

$$p = Q_{KS}\left(\left[\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}\right]D\right).$$

The permutation test uses 10,000 permutations. Use the permutation *p* value for *N*<30 (or generally).

References

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 1992. Numerical Recipes in C. 2nd edition. Cambridge University Press.

Stephens, M.A. 1970. Use of the Kolmogorov-Smirnov, Cramer-von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society, Series B* 32:115-122.

Anderson-Darling test for equal distributions

The Anderson-Darling test is a nonparametric test for overall equal distribution of two univariate samples. It generally has higher power than the Kolmogorov-Smirnov test.

With two samples $x_1 .. x_n$ and $y_1 .. y_m$, the pooled sample size is $N = n+m$. The test statistic A^2_N can be computed according to Pettitt (1976):

$$A_N^2 = \frac{1}{mn}\sum_{i=1}^{N-1}\frac{(M_i N - ni)^2}{i(N-i)}$$

where M_i is the number of x's less than or equal to the *i*th smallest in the pooled sample. Past uses a slightly more complicated version of this equation, with better performance in the presence of ties (Scholz & Stephens 1987, eq. 6).

This statistic is transformed to a statistic called *Z* according to Scholz & Stephens (1987). For our case with *k*=2 samples, compute the variance of the statistic as follows:

$$H = \frac{1}{m} + \frac{1}{n}$$

$$h = \sum_{i=1}^{N-1}\frac{1}{i}$$

$$g = \sum_{i=1}^{N-2}\sum_{j=i+1}^{N-1}\frac{1}{(N-i)j}$$

$$a = 2(4g-6) + (10-6g)H - 4g + 6 = 4g - 6 + (10-6g)H$$

$$b = 4(2g-4) + 16h + (2g-14h-4)H - 8h + 4g - 6 = 12g + 8h - 22 + (2g-14h-4)H$$

$$c = 4(6h+2g-2) + 2(4h-4g+6) + (2h-6)H + 4h = 36h + 4 + (2h-6)H$$

$$d = 4(2h+6) - 8h = 6$$

$$\sigma_N^2 = \frac{aN^3 + bN^2 + cN + d}{(N-1)(N-2)(N-3)}$$

Then,

$$Z = \frac{A_N^2 - 1}{\rho_N}$$

The *p* value is computed by interpolation and extrapolation in Table 1 (*m*=1) of Scholz & Stephens (1987), using a curve fit to the von Bertalanffy model. The approximation is fairly accurate for

$p \leq 0.25$. For $p > 0.25$, the p values are estimated using a polynomial fit to values obtained by permutation. A p value based on Monte Carlo permutation with N=999 is also provided.

References

Pettitt, A.N. 1976. A two-sample Anderson-Darling rank statistic. *Biometrika* 63:161-168.

Scholz, F.W. & Stephens, M.A. 1987. K-sample Anderson–Darling tests. *Journal of the American Statistical Association* 82:918–924.

Epps-Singleton test for equal distributions

The Epps-Singleton test (Epps & Singleton 1986; Goerg & Kaiser 2009) is a nonparametric test for overall equal distribution of two univariate samples. It is typically more powerful than the Kolmogorov-Smirnov test, and unlike the Kolmogorov-Smirnov it can be used also for non-continuous (i.e. ordinal) data. The null hypothesis is H_0: The two samples are taken from populations with equal distribution.

The mathematics behind the Epps-Singleton test are complicated. The test is based on the Fourier transform of the empirical distribution function, called the empirical characteristic function (ECF). The ECF is generated for each sample and sampled at two points (t_1=0.4 and t_2=0.8, standardized for the pooled semi-interquartile range). The test statistic W_2 is based on the difference between the two sampled ECFs, standardized by their covariance matrices. A small-sample correction to W_2 is applied if both sample sizes are less than 25. The p value is based on the chi-squared distribution. For details, see Epps & Singleton (1986) and Goerg & Kaiser (2009).

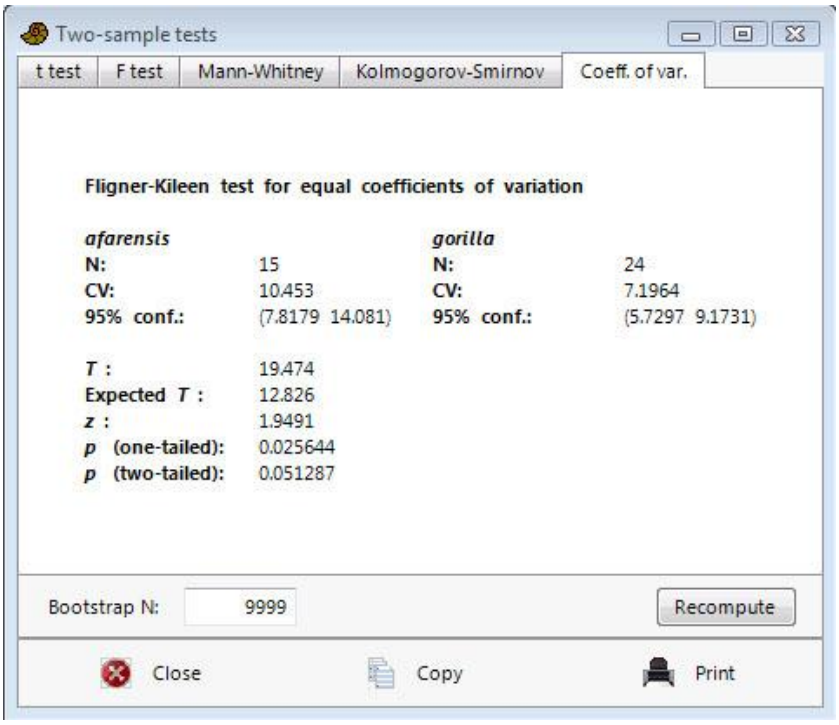
References

Epps, T.W. & Singleton, K.J. 1986. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation* 26:177–203.

Goerg, S.J. & Kaiser, J. 2009. Nonparametric testing of distributions – the Epps-Singleton two-sample test using the empirical characteristic function. *The Stata Journal* 9:454-465.

Coefficient of variation (Fligner-Kileen test)

This module tests for equal coefficient of variation in two samples.



The coefficient of variation (or relative variation) is defined as the ratio of standard deviation to the mean in percent, and is computed as:

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}}{\bar{x}} \cdot 100 .$$

The 95% confidence intervals are estimated by bootstrapping (simple bootstrap), with the given number of replicates (default 9999).

The null hypothesis if the statistical test is:

H_0: The samples were taken from populations with the same coefficient of variation.

If the given *p*(normal) is less than 0.05, equal coefficient of variation can be rejected. Donnelly & Kramer (1999) describe the coefficient of variation and review a number of statistical tests for the comparison of two samples. They recommend the Fligner-Killeen test (Fligner & Killeen 1976), as implemented in Past. This test is both powerful and is relatively insensitive to distribution. The following statistics are reported:

T: The Fligner-Killeen test statistic, which is a sum of transformed ranked positions of the smaller sample within the pooled sample (see Donnelly & Kramer 1999 for details).

E(*T*): The expected value for *T*.

z: The z statistic, based on *T*, Var(*T*) and E(*T*). Note this is a large-sample approximation.

p: The $p(H_0)$ value. Both the one-tailed and two-tailed values are given. For the alternative hypothesis of difference in either direction, the two-tailed value should be used. However, the Fligner-Killeen test has been used to compare variation within a sample of fossils with variation within a closely related modern species, to test for multiple fossil species (Donnelly & Kramer 1999). In this case the alternative hypothesis might be that CV is larger in the fossil population, if so then a one-tailed test can be used for increased power.

The screenshot above reproduces the example of Donnelly & Kramer (1999), showing that the relative variation within *Australopithecus afarensis* is significantly larger than in *Gorilla gorilla*. This could indicate that *A. afarensis* represents several species.

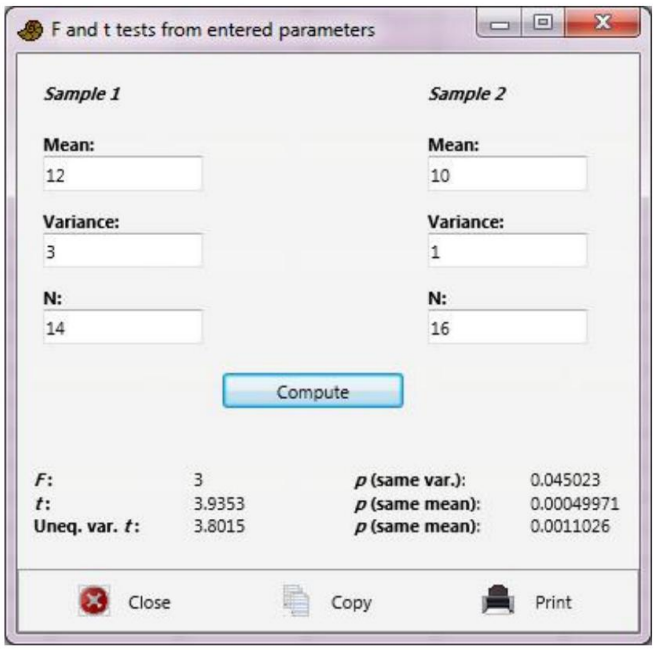
References

Donnelly, S.M. & Kramer, A. 1999. Testing for multiple species in fossil samples: An evaluation and comparison of tests for equal relative variation. *American Journal of Physical Anthropology* 108:507-529.

Fligner, M.A. & Killeen, T.J. 1976. Distribution-free two sample tests for scale. *Journal of the American Statistical Association* 71:210-213.

F and *t* tests from parameters

Sometimes publications give not the data, but values for sample size, mean and variance for two samples. These can be entered manually using the 'F and t from parameters' option in the menu. This module does not use any data from the spreadsheet.

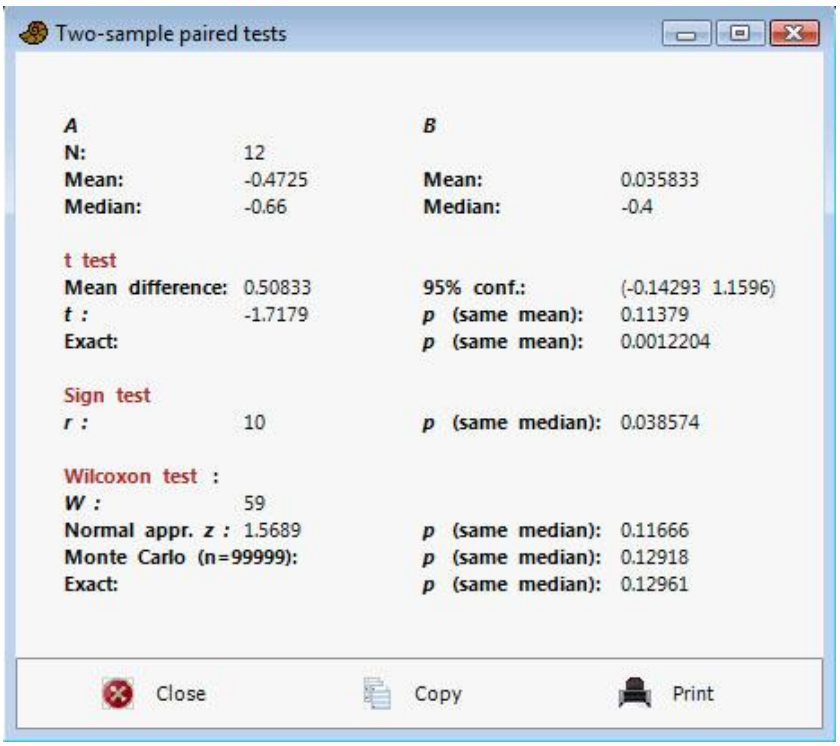


Two-sample paired tests (*t*, sign, Wilcoxon)

Three statistical tests (one parametric, two non-parametric) for two samples (columns) of univariate data. The data points are paired, meaning that the two values in each row are associated. For example, the test could be the for length of the left vs. the right arm in a number of people, or the diversity in summer vs. winter at a number of sites. Controlling for a "nuisance factor" (person, site) in this way increases the power of the test. The null hypothesis is:

H_0: The mean (*t* test) or median (sign test, Wilcoxon test) of the difference is zero.

All reported *p* values are two-tailed.



t test

Testing for mean difference equal to zero using the standard one-sample *t* test on the differences. With $d_i = x_i - y_i$, we have

$$s = \sqrt{\frac{1}{n-1}\sum\left(d_i - \bar{d}\right)^2},$$

$$t = \frac{\bar{d}}{s/\sqrt{n}}.$$

There are *n*-1 degrees of freedom. This test assumes normal distribution of the differences.

The exact version of the test calculates all possible group reassignments within pairs. It is only computed for less than 26 pairs.

Bayesian factor

The Jeffrey-Zellner-Siow (JZS) Bayes Factor is reported in favour of the alternative, i.e. it quantifies the evidence for the hypothesis of unequal means. For details, see the section on the (unpaired) two-sample *t* test above. The same equation is used, but with *N*-1 degrees of freedom.

Sign test

The sign (binomial) test simply counts the number of cases n_1 where $x_i>y_i$ and n_2 where $y_i>x_i$. The number max(n_1, n_2) is reported. The *p* value is exact, computed from the binomial distribution. The sign test may have lower power than the other paired tests, but makes few assumptions.

Wilcoxon signed rank test

A non-parametric rank test that does not assume normal distribution. The null hypothesis is no median shift (no difference).

All rows with zero difference are first removed by the program. Then the absolute values of the differences $|d_i|$ are ranked (R_i), with mean ranks assigned for ties. The sum of ranks for pairs where d_i is positive is W^+. The sum of ranks for pairs where d_i is negative is W^-. The reported test statistic is

$$W = \min(W^+, W^-)$$

(note that there are several other, equivalent versions of this test, reporting other statistics).

For large *n* (say *n*>10), the large-sample approximation to *p* can be used. This depends on the normal distribution of the test statistic *W*:

$$E(W) = \frac{n(n+1)}{4}$$

$$Var(W) = \frac{n(n+1)(2n+1)}{24} - \frac{\sum_g f_g^3 - f_g}{48}.$$

The last term is a correction for ties, where f_g is the number of elements in tie *g*. The resulting *z* is reported, together with the *p* value.

The Monte Carlo significance value is based on 99,999 random reassignments of values to columns, within each pair. This value will be practically identical to the exact *p* value.

For *n*<23, an exact *p* value is computed, by complete enumeration of all possible reassignments (there are 2^n of them, i.e. more than four million for *n*=22). This is the preferred *p* value, if available.

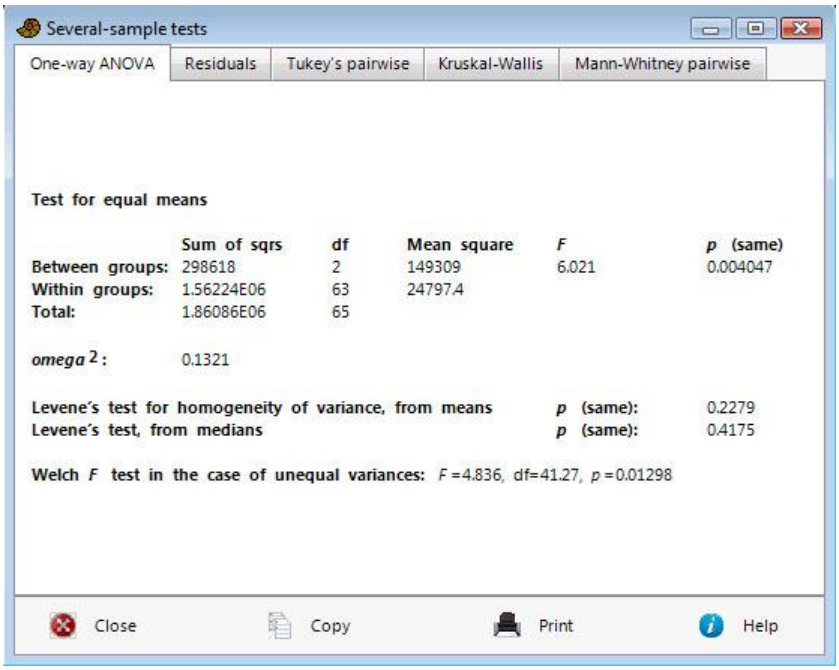
Missing data: Supported by deletion of the row.

Several-sample tests

One-way ANOVA and Kruskal-Wallis tests for equality of means or medians between several univariate samples, given in separate columns. It is also possible to specify the groups using a single column of values and an additional Group column. Missing data are supported by deletion.

One-way ANOVA

One-way ANOVA (analysis of variance) is a statistical procedure for testing the null hypothesis that several univariate samples are taken from populations with the same mean. The samples are assumed to be close to normally distributed and have similar variances. If the sample sizes are equal, these two assumptions are not critical. If the assumptions are strongly violated, the nonparametric Kruskal-Wallis test should be used instead.



ANOVA table

The between-groups sum of squares is given by:

$$\mathrm{SS}_{\mathrm{bg}} = \sum_g n_g \left(\bar{x}_g - \bar{x}_T\right)^2 ,$$

where n_g is the size of group g, and the means are group and total means. The between-groups sum of squares has an associated $\mathrm{df}_{\mathrm{bg}}$, the number of groups minus one.

The within-groups sum of squares is

$$\mathrm{SS}_{\mathrm{wg}} = \sum_g \sum_i \left(x_i - \bar{x}_g\right)^2$$

where the x_i are those in group g. The within-groups sum of square has an associated $\mathrm{df}_{\mathrm{wg}}$, the total number of values minus the number of groups.

The mean squares between and within groups are given by

$$\mathrm{MS}_{\mathrm{bg}} = \frac{\mathrm{SS}_{\mathrm{bg}}}{\mathrm{df}_{\mathrm{bg}}}$$

$$\mathrm{MS}_{\mathrm{wg}} = \frac{\mathrm{SS}_{\mathrm{wg}}}{\mathrm{df}_{\mathrm{wg}}}$$

Finally, the test statistic F is computed as

$$F = \frac{\mathrm{MS}_{\mathrm{bg}}}{\mathrm{MS}_{\mathrm{wg}}}$$

The p value is based on F with $\mathrm{df}_{\mathrm{bg}}$ and $\mathrm{df}_{\mathrm{wg}}$ degrees of freedom.

Random effects (Model II) ANOVA

For balanced, one-way ANOVA, the ANOVA table, F value and p value are the same for fixed-effects and random-effects ANOVA, so you can use the results for either type of ANOVA. If your design is of the random-effects type (i.e. the factor levels are taken at random from a larger population instead of being fixed by experiment), then you should also report the given variance due to random errors and the variance due to difference between groups (can become negative):

$$s^2 = \mathrm{MS}_{\mathrm{wg}}$$

$$s_g^2 = \frac{\mathrm{MS}_{\mathrm{bg}} - \mathrm{MS}_{\mathrm{wg}}}{n}$$

where n is the sample size for each group in the case of balanced design. For unbalanced design, set n to

$$n_0 = \frac{1}{G-1}\left(\sum_{i=1}^{G} n_i - \frac{\sum n_i^2}{\sum n_i}\right)$$

where G is the number of groups. The intraclass correlation coefficient ICC gives the proportion of variance due to group differences:

$$ICC = \frac{s_g^2}{s_g^2 + s^2}$$

Omega squared

The omega squared is a measure of effect size, varying from 0 to 1:

$$\omega^2 = \frac{\mathrm{SS_{bg}} - \mathrm{df_{bg}}\mathrm{MS_{wg}}}{\mathrm{SS_{total}} + \mathrm{MS_{wg}}}.$$

Levene's test

Levene's test for homogeneity of variance (homoskedasticity), that is, whether variances are equal as assumed by ANOVA, is also given. Two versions of the test are included. The original Levene's test is based on means. This version has more power if the distributions are normal or at least symmetric. The version based on medians has less power but is more robust to non-normal distributions. Note that this test can be used also for only two samples, giving an alternative to the *F* test for two samples described above.

Unequal-variance (Welch) ANOVA

If Levene's test is significant, meaning that you have unequal variances, you can use the unequal-variance (Welch) version of ANOVA, with the *F*, df and *p* values given.

Bayes factor

The Jeffrey-Zellner-Siow (JZS) Bayes Factor is reported in favour of the alternative, i.e. it quantifies the evidence for the hypothesis of unequal means. It is calculated according to Rouder et al. (2012).

Effect size (confidence intervals for the pairwise differences in means)

This table gives 95% confidence intervals for the differences in group means. If a confidence interval does not include zero, the difference may be considered significant, and it is marked in pink. These are “multiple-t” or “simultaneous” confidence intervals, which are wider than the confidence intervals for each pairwise difference in isolation. For the two groups i and *j*, the confidence interval is

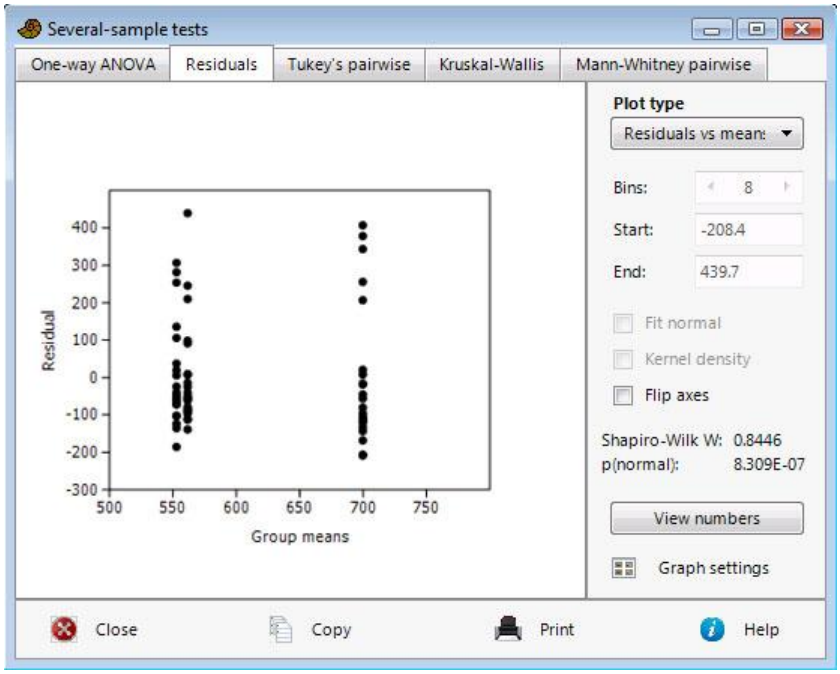
$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2m}\sqrt{MS_{wg}}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $t_{\alpha/2m}$ is the upper $\alpha/2m$ point of the *t* distribution with α=0.05, with *m* the number of confidence statements (total number of pairwise comparisons). The degrees of freedom is equal to the total number of values minus the number of groups.

Analysis of residuals

The "Residuals" tab shows properties of the residuals, in order to evaluate some assumptions of ANOVA such as normal and homoskedastic distribution of residuals.

The Shapiro-Wilk test for normal distribution is given, together with several common plots of residuals (normal probability, residuals vs. group means, and histogram).



Tukey's pairwise post-hoc tests

If the ANOVA shows significant inequality of the means (small *p*), you can go on to study the given table of "post-hoc" pairwise comparisons, based on the Tukey-Kramer test. The Studentized Range Statistic *Q* is given in the lower left triangle of the array, and the probabilities *p(equal)* in the upper right.

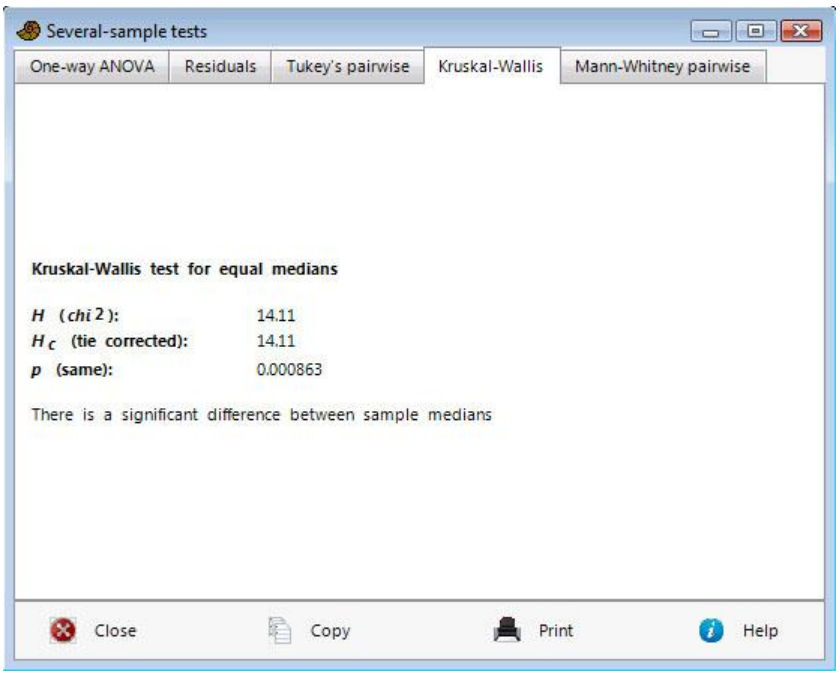
$$Q = \frac{\overline{X}_L - \overline{X}_S}{\sqrt{\frac{\mathrm{MS}_{\mathrm{wg}}}{n}}},$$

where X_L is the larger and X_S the smaller mean of the two samples being compared. If the sample sizes are not equal, their harmonic mean is used for *n*. Significance is estimated according to Copenhaver & Holland (1988), with $\mathrm{df}_{\mathrm{wg}}$ degrees of freedom. There is also an option to use the slightly less accurate method of Lund & Lund (1983) which was used in older versions of Past.

Kruskal-Wallis

The Kruskal-Wallis test is a non-parametric ANOVA, comparing the medians of several univariate groups (given in columns). It can also be regarded as a multiple-group extension of the Mann-Whitney test (Zar 1996). It does not assume normal distribution, but does assume equal-shaped distribution for all groups. The null hypothesis is

H_0: The samples are taken from populations with equal medians.



The test statistic H is computed as follows:

$$H = \frac{12}{n(n+1)}\left(\sum_g \frac{T_g^2}{n_g}\right) - 3(n+1),$$

where n_g is the number of elements in group g, n is the total number of elements, and T_g is the sum of ranks in group g.

The test statistic H_c is adjusted for ties:

$$H_c = \frac{H}{1 - \frac{\sum_i f_i^3 - f_i}{n^3 - n}},$$

where f_i is the number of elements in tie i.

With G the number of groups, the p value is approximated from H_c using the chi-square distribution with G-1 degrees of freedom. This is less accurate if any n_g<5.

Mann-Whitney pairwise post-hoc tests

Mann-Whitney pairwise test p values are given for all N_p=G(G-1)/2 pairs of groups. The asymptotic approximation described under the Mann-Whitney module is used. If samples are very small, it may be useful to run the exact test available in that module instead. Four different views are available for the symmetric table:

1. Raw p values, uncorrected significance: The p values from each individual pairwise test, marked in pink if p<0.05, not corrected for multiple testing.

2. Raw p values, sequential Bonferroni significance: The p values from each individual pairwise test are shown uncorrected for multiple testing. Significance (pink marking) is assessed by first evaluating the smallest p value, with Bonferroni correction for N_p pairs. If significant (pN_p<0.05) the next smallest p value is significant if p(N_p-1)<0.05, etc.

3. Bonferroni corrected p values: The values shown are $p' = pN_p$. Marked as significant if p'<0.05.

4. Mann-Whitney U: The test statistics.

Dunn's post-hoc

The Dunn's post hoc test (Dunn 1964) is a pairwise test often carried out after a significant Kruskal-Wallis test. It is an alternative to the pairwise Mann-Whitney. With T_g the sum of ranks within group g from the Kruskal-Wallis test, calculate for each group the average rank:

$$\bar{T}_g = \frac{T_g}{n_g}$$

To compare two groups A and B, calculate the z statistic

$$z_{\mathrm{AB}} = \frac{|\bar{T}_A - \bar{T}_B|}{\rho_{\mathrm{AB}}}$$

$$\rho_{\mathrm{AB}} = \sqrt{\left(\frac{n(n+1)}{12} - \frac{\sum_i f_i{}^3 - f_i}{12(n-1)}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

Here, n is the total sample size and f_i the number of elements in tie i, as in the Kruskal-Wallis test. See pairwise Mann-Whitney above for the various options in the table output (raw p values, raw p values with sequential Bonferroni, Bonferroni corrected p values, and the z statistics).

References

Copenhaver, M.D., Holland, B. 1988. Computation of the distribution of the maximum studentized range statistic with application to multiple significance testing of simple effects. *Journal of Statistical Computation and Simulation* 30:1-15.

Dunn, O.J. 1964. Multiple comparisons using rank sums. *Technometrics* 6:241-252.

Lund, R.E., Lund, J.R. 1983. Algorithm AS 190: Probabilities and upper quantiles for the studentized range. *Journal of the Royal Statistical Society C* 32:204-210.

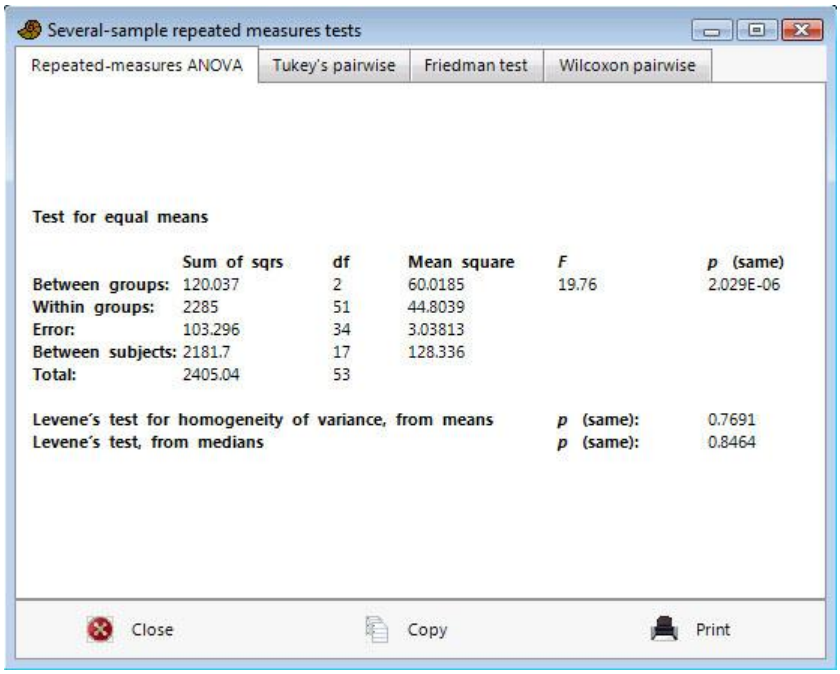
Rouder, J.N, Morey, R.D., Speckman, P.L., Province, J.M. 2012. Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology* 56:356-374.

Zar, J.H. 1996. Biostatistical analysis. 3rd ed. Prentice Hall.

Several-samples repeated measures tests

In repeated measures ANOVA, values in each row are observations on the same "subject". Repeated-measures ANOVA is the extension of the paired *t* test to several samples. Each column (sample) must contain the same number of values.

Missing values are not supported.



The procedure begins like the independent-samples one-way ANOVA above. In short,

$$\mathrm{SS}_{\mathrm{bg}} = \sum_g n\left(\bar{x}_g - \bar{x}_T\right)^2 ,$$

where *n* is the sample size. The associated $\mathrm{df}_{\mathrm{bg}}$ is the number of groups minus one.

$$\mathrm{SS}_{\mathrm{wg}} = \sum_g \sum_i \left(x_i - \bar{x}_g\right)^2$$

where the x_i are those in group *g*. The associated $\mathrm{df}_{\mathrm{wg}}$ is the total number of values minus the number of groups.

The between-subjects sum of squares is

$$\mathrm{SS}_{\mathrm{sub}} = \sum_i n\left(\bar{x}_i - \bar{x}_T\right)^2 ,$$

where the $\bar{x}_i$ are means of subject *i* across groups. The associated $\mathrm{df}_{\mathrm{sub}}$ is the number of subjects minus one.

The SS_{error} is simply $SS_{wg} - SS_{sub}$, with $df_{error} = df_{wg} - df_{sub}$.

The mean squares are then the sum squares divided by their respective degrees of freedom:

$$MS_{bg} = \frac{SS_{bg}}{df_{bg}}$$

$$MS_{wg} = \frac{SS_{wg}}{df_{wg}}$$

$$MS_{sub} = \frac{SS_{sub}}{df_{sub}}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}}.$$

Finally, the *F* ratio is MS_{bg} / MS_{error}, with df_{bg} and df_{error} degrees of freedom.

Sphericity estimates and corrections

An assumption of repeated measures ANOVA is *sphericity*, meaning equal variances of the differences between all combinations of groups. A statistic called *epsilon* approaches 1 for data meeting the sphericity assumption. For smaller values of epsilon, a correction can be applied to the degrees of freedom of the ***F*** test, providing a corrected *p* value for the ANOVA. PAST provides two versions of this procedure, Greenhouse-Geisser (Greenhouse & Geisser 1959) and Huynh-Feldt (Huynh & Feldt 1976).

Tukey's pairwise post-hoc tests

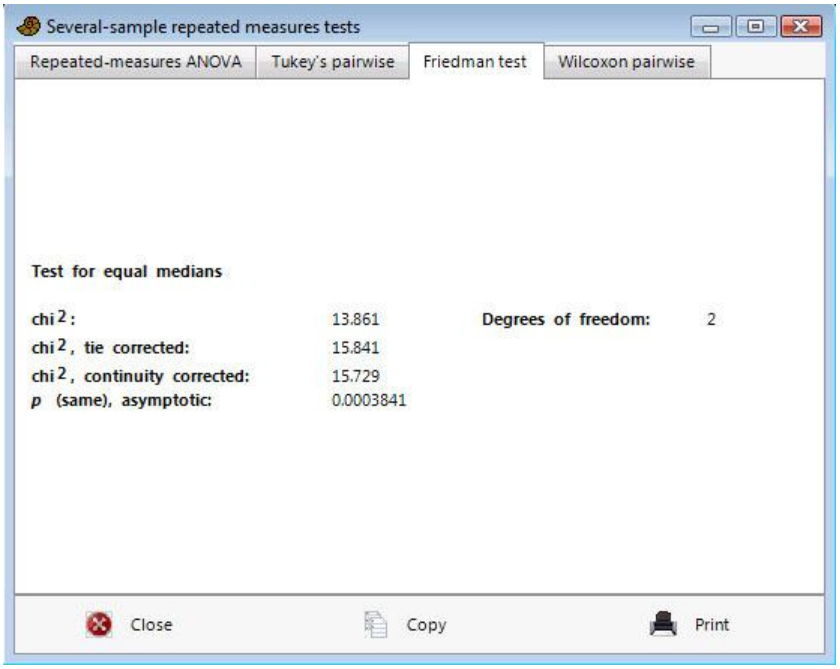
The "post-hoc" pairwise comparisons are based on the Tukey test. The Studentized Range Statistic *Q* is given in the lower left triangle of the array, and the probabilities *p(equal)* in the upper right.

$$Q = \frac{\overline{X}_L - \overline{X}_S}{\sqrt{\frac{MS_{error}}{n}}},$$

where X_L is the larger and X_S the smaller mean of the two samples being compared. There are df_{error} degrees of freedom.

Friedman test

The Friedman test is a non-parametric test for equality of medians in several repeated-measures univariate groups. It can be regarded as the non-parametric version of repeated-measures ANOVA, or the repeated-measures version of the Kruskal-Wallis test.



The Friedman test follows Bortz et al. (2000). The basic test statistic is

$$\chi^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} T_j^2 - 3n(k+1),$$

where n is the number of rows, k the number of columns and T_j the column sums of the data table.

The χ^2 value is then corrected for ties (if any):

$$\chi^2_{tie} = \frac{\chi^2}{1 - \frac{1}{nk(k^2-1)} \sum_{i=1}^{m} \left(t_i^3 - t_i\right)}$$

where m is the total number of tie groups and t_i are the numbers of values in each tie group.

For k=2, it is recommended to use one of the paired tests (e.g. sign or Wilcoxon test) instead. For small data sets where k=3 and n<10, or k=4 and n<8, the tie-corrected χ^2 value is looked up in a table of "exact" p values. When given, this is the preferred p value.

The asymptotic *p* value (using the χ^2 distribution with *k*-1 degrees of freedom) is fairly accurate for larger data sets. It is computed from a continuity corrected version of χ^2:

$$S = \sum_{j=1}^{k} \left(T_j - \frac{n(k+1)}{2} \right)^2$$

$$\chi^2 = \frac{12n(k-1)(S-1)}{n^2(k^3-k)+24}.$$

This χ^2 value is also corrected for ties using the equation above.

The post hoc tests are by simple pairwise Wilcoxon, exact for *n*<20, asymptotic for *n*>=20. These tests have higher power than the Friedman test.

References

Bortz, J., Lienert, G.A. & Boehnke, K. 2000. Verteilungsfreie Methoden in der Biostatistik. 2nd ed. Springer.

Greenhouse, S.W. & Geisser, S. 1959. On methods in the analysis of profile data. *Psychometrika* 24:95-112.

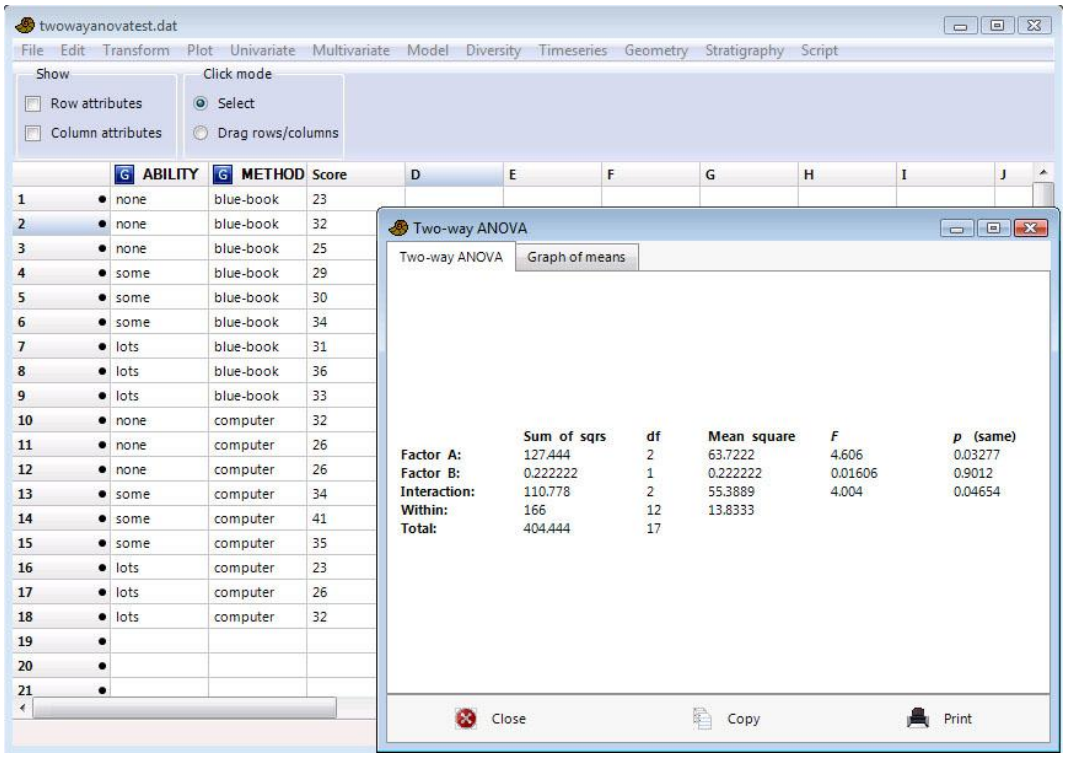
Huynh, H. & Feldt, L.S. 1976. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics* 1:69-82.

Two-way ANOVA

Two-way ANOVA (analysis of variance) tests the null hypotheses that several univariate samples have the same mean across each of two factors A and B, and that there are no dependencies (interactions) between factors. The samples are assumed to be close to normally distributed and have similar variances. If the sample sizes are equal, these two assumptions are not critical. The default analysis is a fixed-effect blocked design (the most usual case). There is also an option for random-effect blocked two-way ANOVA, and fixed-effect and random-effect, nested two-way ANOVA. There is no interaction term for the nested design.

Past supports unbalanced designs (unequal sample sizes for each factor combination). In this case, a "Model II" ANOVA is carried out (Fox 2016).

Three columns are needed: A group column (set data type to Group with 'Column attributes') with the levels for factor A, a group column with the levels for factor B, and a column of the corresponding measured values.



The algorithm uses weighted means.

Total sum of squares:

$$\mathrm{SS_T} = \sum_i (x_i - \bar{x})^2 ,$$

taken over all points. The associated degrees of freedom $\mathrm{df_T}$ is the total number of values minus one.

Within-group (error) sum of squares:

$$\mathrm{SS}_{\mathrm{wg}} = \sum_{g_1}\sum_{g_2}\sum_{i}\left(x_i - \bar{x}_{g_1 g_2}\right)^2 ,$$

where the x_i are those in group (level) g_1 for the first factor and g_2 for the second factor, and the mean is taken within the same group combination. The associated $\mathrm{df}_{\mathrm{wg}}$ is the total number of values minus the product of numbers of groups and columns.

The between-groups sum of squares $\mathrm{SS}_{\mathrm{bg}} = \mathrm{SS}_{\mathrm{T}} - \mathrm{SS}_{\mathrm{wg}}$ can be partitioned into three, namely the Factor A, Factor B and interaction terms.

$$\mathrm{SS}_{\mathrm{A}} = N_A \sum_{i}\left(\bar{x}_i - \bar{x}\right)^2 ,$$

where the sum is over levels of Factor A, and the two means are the level mean and the total mean, respectively. N_A is the number of levels of A. The degrees of freedom is $\mathrm{df}_{\mathrm{A}} = N_A - 1$. Similarly, for Factor B:

$$\mathrm{SS}_{\mathrm{B}} = N_B \sum_{j}\left(\bar{x}_j - \bar{x}\right)^2$$

where now the sum is over levels of Factor B. The degrees of freedom is $\mathrm{df}_{\mathrm{B}} = N_B - 1$.

The interaction sum of squares is $\mathrm{SS}_{\mathrm{AxB}} = \mathrm{SS}_{\mathrm{bg}} - \mathrm{SS}_{\mathrm{A}} - \mathrm{SS}_{\mathrm{B}}$, with $\mathrm{df}_{\mathrm{AxB}} = (N_A - 1)(N_B - 1)$ degrees of freedom.

For unbalanced designs (Model II), the computations are more complex, using dummy-variable multiple regressions (Fox 2016).

Mean squares MS are the sum of squares divided by their respective degrees of freedom.

Finally, the *F* ratios are

$F_A = \mathrm{MS}_{\mathrm{A}} / \mathrm{MS}_{\mathrm{wg}}$
$F_B = \mathrm{MS}_{\mathrm{B}} / \mathrm{MS}_{\mathrm{wg}}$
$F_{AxB} = \mathrm{MS}_{\mathrm{AxB}} / \mathrm{MS}_{\mathrm{wg}}$

Random effects

The random-effects ANOVA is computed as fixed effects ANOVA, except
$F_A = \mathrm{MS}_{\mathrm{A}} / \mathrm{MS}_{\mathrm{AxB}}$
$F_B = \mathrm{MS}_{\mathrm{B}} / \mathrm{MS}_{\mathrm{AxB}}$.

For random-effects ANOVA, the components of variance are computed as follows (only for balanced design, where *n* is the sample size for each level combination):
$\mathrm{var(A)} = (\mathrm{MS}_{\mathrm{A}} - \mathrm{var(err)} - n\ \mathrm{var(AxB)}) / (n\mathrm{N}_{\mathrm{B}})$
$\mathrm{var(B)} = (\mathrm{MS}_{\mathrm{B}} - \mathrm{var(err)} - n\ \mathrm{var(AxB)}) / (n\mathrm{N}_{\mathrm{A}})$
$\mathrm{var(AxB)} = (\mathrm{MS}_{\mathrm{AxB}} - \mathrm{var(err)}) / n$

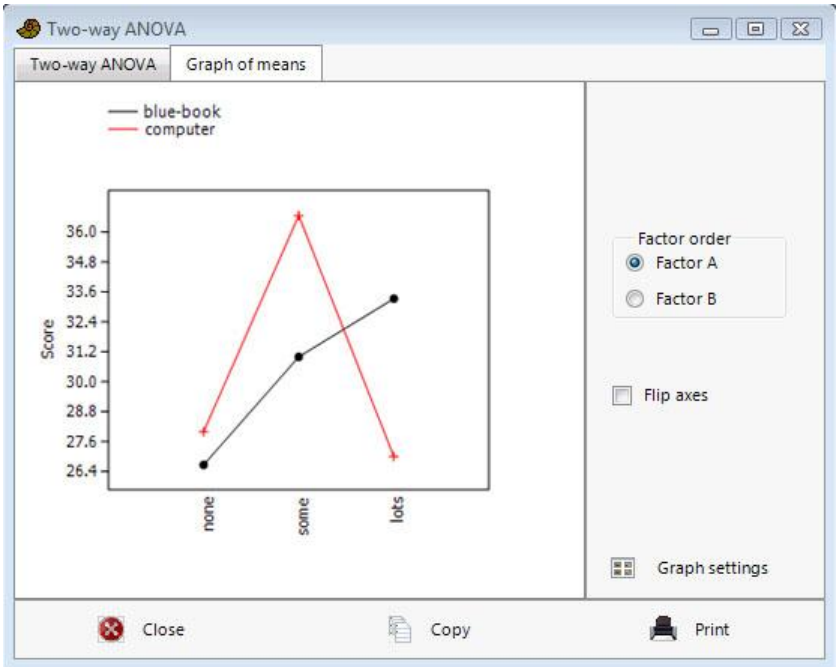
var(err) = MS_{wg}

Nested two-way ANOVA

The main factor (factor A) sum-of-squares SS_A, degrees of freedom df_A and mean square MS_A are calculated as above. For the fixed-effects model, F_A is calculated as above, while for random-effects it is $F_A=MS_A/MS_B$. The nested factor (factor B) is always random-effect.

Graph of means

The graph of means is a simple graphical device, traditionally used to see the effects of factors and their interaction for a two-way ANOVA. The means are shown with either the A levels or the B levels on the x axis, and the other levels as separate lines:



Tukey's post-hoc tests

Tukey's post-hoc tests are available for the two main factors, and for all unconfounded interactions. For the main factors, the Studentized Range Statistic *Q* is given in the lower left triangle of the array, and the probabilities *p*(equal) in the upper right. The interaction test uses the "adjusted k" value, accounting for the number of unconfounded comparisons, for up to 7 levels in each factor. For larger number of levels, the *k* value is not adjusted.

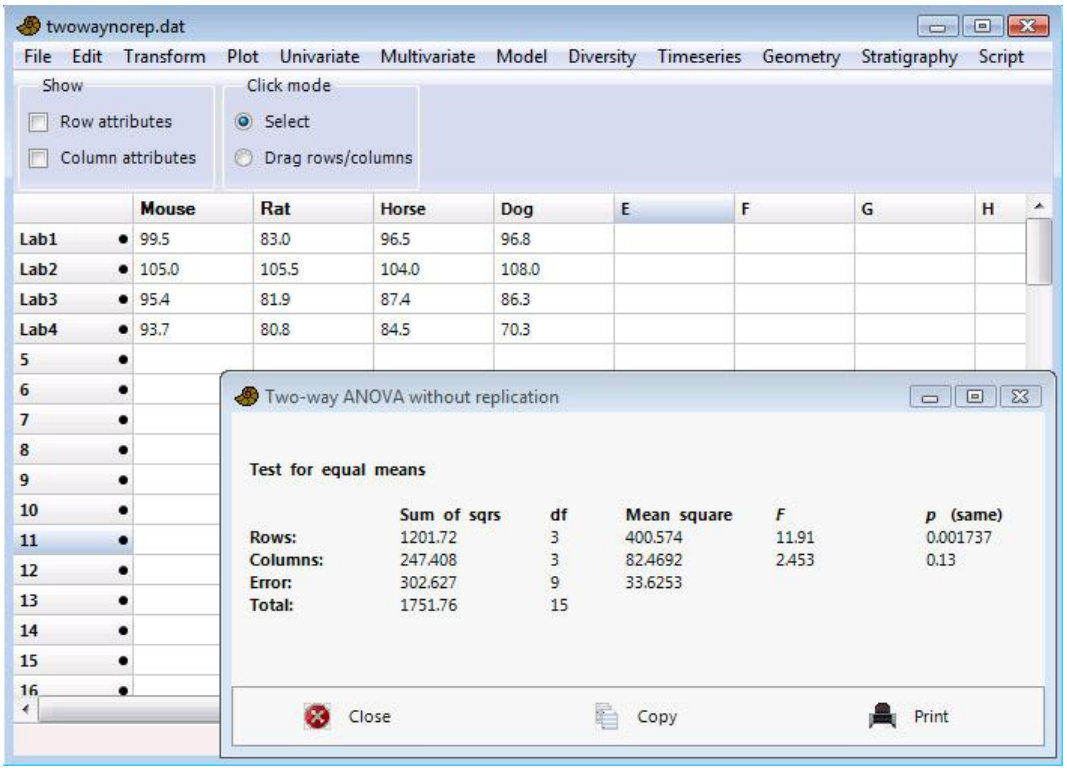
Missing values : Rows with missing values are deleted.

Reference

Fox, J. 2016. *Applied Regression Analysis and Generalized Linear Models, 3rd ed*. Sage Publications.

Two-way ANOVA without replication

Two-way ANOVA for testing the null hypotheses that several univariate samples have the same mean across each of two factors. This module expects only one observation for each combination of levels for the two factors. The input data format is a table where the first factor levels enter in rows, and the second factor level in columns, e.g. a table of veterinary lab results:



There is no interaction term.

The equations are given by Ireland (2010), pp. 130-131.

Reference

Ireland, C.R. 2010. Experimental Statistics for Agriculture and Horticulture. CABI, 352 pp.

Two-way repeated measures ANOVA

Three data columns are needed: A group column (set data type to Group with 'Column attributes') with the levels for factor A, a group column with the levels for factor B, a group column with identifiers for the cases (subjects) and a column of the corresponding measured values.

Each subject must have exactly one entry for each combination of levels. Therefore, if you have *M* levels for factor A, *N* levels for factor B and *S* subjects, you need exactly *M*x*N*x*S* rows in total.

Other functionality is similar to the two-way ANOVA module described above. Missing data are not supported!

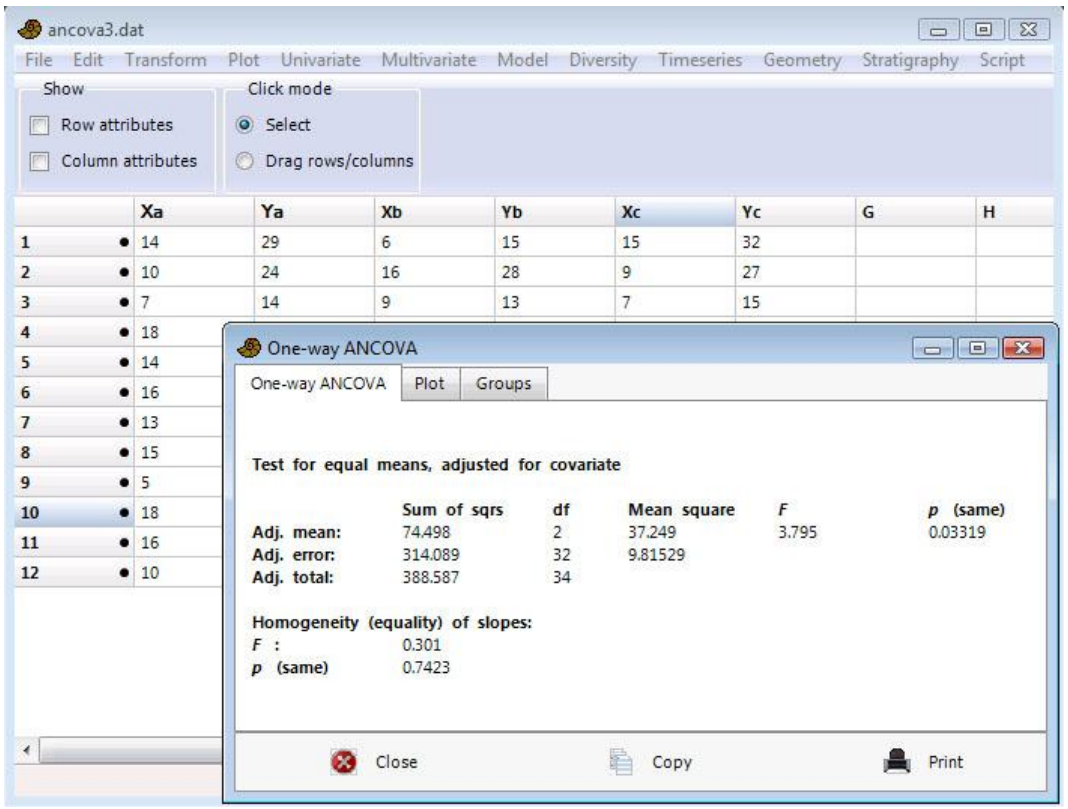
	Medicine	Year	Person	Effect
1	Aspirin	2015	Paul	40
2	Aspirin	2016	Paul	45
3	Aspirin	2015	Mary	30
4	Aspirin	2016	Mary	40
7	Paracet	2015	Paul	50
8	Paracet	2016	Paul	60
9	Paracet	2015	Mary	40
10	Paracet	2016	Mary	55
13	Placebo	2015	Paul	10
14	Placebo	2016	Paul	5
15	Placebo	2015	Mary	25
16	Placebo	2016	Mary	20

Example data formatted for two-way repeated measures ANOVA.

One-way ANCOVA

ANCOVA (Analysis of Covariance) tests for equality of means for several univariate groups, adjusted for covariance with another variate. ANCOVA can be compared with ANOVA, but has the added feature that for each group, variance that can be explained by a specified "nuisance" covariate (*x*) is removed. This adjustment can increase the power of the test substantially.

The program expects two or more pairs of columns, where each pair (group) is a set of correlated *x*-*y* data (means are compared for *y*, while *x* is the covariate). The example below uses three pairs (groups) a, b and c.



The Plot tab presents a scatter plot and linear regression lines for all the groups. The ANOVA-like summary table contains sum-of-squares etc. for the adjusted means (between-groups effect) and adjusted error (within-groups), together with an *F* test for the adjusted means. An *F* test for the equality of regression slopes (as assumed by the ANCOVA) is also given. In the example, equal adjusted means in the three groups can be rejected at $p<0.05$. Equality of slopes can not be rejected ($p=0.74$).

The Groups tab gives the summary statistics for each group (mean, adjusted mean and regression slope).

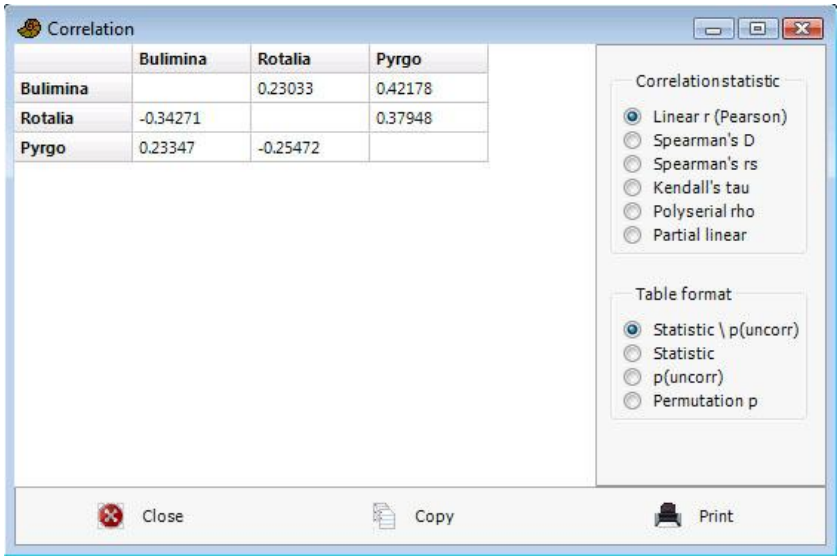
Assumptions include similar linear regression slopes for all groups, normal distributions, similar variance and sample sizes.

Missing data: *x*-*y* pairs with either *x* or *y* missing are disregarded.

Correlation table

Two or more columns are required. A matrix is presented with the correlations between all pairs of columns. In the 'Statistic \ p(uncorr)' table format, correlation values are given in the lower triangle of the matrix, and the two-tailed probabilities that the columns are uncorrelated are given in the upper. Both parametric and non-parametric coefficients and tests are available.

Missing data: Supported by pairwise deletion, except for partial correlation which uses mean value imputation.



Linear r (Pearson)

Pearson's *r* is the most commonly used parametric correlation coefficient. It is given by

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sqrt{(y_i - \bar{y})^2}}} .$$

The significance is computed using a two-tailed *t* test with *n*-2 degrees of freedom and

$$t = r\sqrt{\frac{n-2}{1-r^2}} .$$

Bayes factor

A Bayes factor for Pearson's *r* is computed according to "Jeffrey's integrated Bayes factor", eq. (30) in Ly et al. (2016):

$$BF_{10}^{j,i}(n,r) = \frac{\pi}{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}{}_2F_1\left(\frac{2n-3}{4};\ \frac{2n-1}{4};\ \frac{n+2}{2};\ r^2\right)$$

where ${}_2F_1$ is the Gaussian hypergeometric function (evaluated as a Taylor series; Section 4.2 in Pearson et al. 2016). Large values of the Bayes factor (>3) indicate evidence for correlation (positive or negative).

Spearman's D and rs

Spearman's (non-parametric) rank-order correlation coefficient is the linear correlation coefficient (Pearson's *r*) of the ranks. Following Press et al. (1992) it is computed as

$$r_s = \frac{1-\frac{6}{n^3-n}\left[D+\frac{1}{12}\sum_k\left(f_k^3-f_k\right)+\frac{1}{12}\sum_m\left(g_m^3-g_m\right)\right]}{\sqrt{\left(1-\frac{\sum_k\left(f_k^3-f_k\right)}{n^3-n}\right)\left(1-\frac{\sum_m\left(g_m^3-g_m\right)}{n^3-n}\right)}}.$$

Here, *D* is the sum squared difference of ranks (midranks for ties):

$$D = \sum_{i=1}^{n}\left(R_i - S_i\right)^2 .$$

The f_k are the numbers of ties in the *k*th group of ties among the R_i's, and the g_m are the numbers of ties in the *m*th group of ties among the S_i's.

For *n*>9, the probability of non-zero r_s (two-tailed) is computed using a *t* test with *n*-2 degrees of freedom:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} .$$

For small *n* this approximation is inaccurate, and for *n*<=9 the program therefore switches automatically to an exact test. This test compares the observed r_s to the values obtained from all possible permutations of the first column.

Kendall's tau

This non-parametric correlation coefficient is not in very common use. It is computed according to Press et al. (1992). All possible *N*(*N*-1)/2 pairs of bivariate data points are considered. If two pairs have the same direction in *x* as in *y* (*x* and *y* both decrease, or both increase), they are called *concordant*. If not, they are *discordant*. A tie in the *x*'s is called an *extra-x*, and a tie in the *y*'s is called an *extra-y*. Pairs with ties in both variables are discarded. The number of pairs in the four categories are counted. Then,

$$\tau = \frac{\text{concordant} - \text{discordant}}{\sqrt{\text{concordant} + \text{discordant} + \text{extray}}\sqrt{\text{concordant} + \text{discordant} + \text{extrax}}}.$$

The asymptotic test is based on Kendall's tau being approximately normal, with zero mean and

$$\text{var}(\tau) = \frac{4N + 10}{9N(N-1)}.$$

Polyserial correlation

This correlation is only carried out if the second column consists of integers with a range less than 100. It is designed for correlating a normally distributed continuous/interval variable (first column) with an ordinal variable (second column) that bins a normally distributed variable. For example, the second column could contain the numbers 1-3 coding for "small", "medium" and "large". There would typically be more "medium" than "small" or "large" values because of the underlying normal distribution of sizes.

Past uses the two-step algorithm of Olsson et al. (1982). This is more accurate than their "ad hoc" estimator, and nearly as accurate as the full multivariate ML algorithm. The two-step algorithm was chosen because of speed, allowing a permutation test (but only for *N*<100; not yet in Past 3). For larger *N* the given asymptotic test (log-ratio test) is accurate.

Partial linear correlation

Using this option, for each pair of columns, the linear correlation is computed while controlling for all the remaining columns. For example, with three columns A, B, C the correlation AB is controlled for C; AC is controlled for B; BC is controlled for A. The partial linear correlation can be defined as the correlation of the residuals after regression on the controlling variable(s). The significance is estimated with a *t* test with *n*-2-*k* degrees of freedom, where *k* is the number of controlling variables:

$$t = r\sqrt{\frac{n-2-k}{1-r^2}}$$

Phi coefficient

The phi coefficient (Lovell et al. 2015) was designed for compositional (relative) data such as percentages. The usual correlation coefficients can be misleading for such data. The coefficient measures the degree of proportionality; the smaller the value (close to zero), the more the variables exhibit a proportional relationship. Pairs of variables can show strong correlations but low proportionality when they are linearly related, but with a non-zero intercept term.

Each variable is first transformed using the centered logratio transformation

$$\text{clr}(\text{x}) = \left[\ln\frac{x_1}{g(\boldsymbol{x})}, \cdots, \ln\frac{x_N}{g(\boldsymbol{x})}\right]$$

where $g(\boldsymbol{x})$ is the geometric mean of the data vector. The proportionality between two transformed variables x and y is then computed by

$$\varphi = \frac{\mathrm{var}(\boldsymbol{x}/\boldsymbol{y})}{\mathrm{var}(\boldsymbol{x})}$$

where var($\boldsymbol{x}$) is the variance of the vector $\boldsymbol{x}$. Significance is computed by testing whether $\boldsymbol{x}+\boldsymbol{y}$ and $\boldsymbol{x}-\boldsymbol{y}$ are uncorrelated. Significance testing is however not recommended by Lovell et al. (2015).

Tetrachoric correlation

Tetrachoric correlation is appropriate when both variables are binary (0/1), but reflecting underlying quantities on a continuous scale. Past uses an accurate approximation due to Bonett & Price (2005).

Let f_{11} be the number of rows with the binary pair 1/1, f_{12} the number of 1/0, f_{21} the number of 0/1, and f_{22} is the number of rows with the binary pair 1/1. An estimator of the odds ratio of this contingency table is

$$\widehat{\omega} = \frac{(f_{11} + 0.5)(f_{22} + 0.5)}{(f_{12} + 0.5)(f_{21} + 0.5)}$$

We then calculate the sample proportions p from the frequencies f, after adding 0.5 to each f, i.e. $p_{11} = (f_{11}+0.5)/(N+2)$ etc. (we add 2 because the total number has been increased by adding 0.5 to each of the four f; this is not precisely specified by Bonett & Price). Then calculate $p_{1+} = p_{11}+p_{12}$ and $p_{+1} = p_{11}+p_{21}$. Also, $p_{\min}$ is the smallest marginal proportion (row or column sum in the p table). Then,

$$\hat{c} = \frac{1 - \frac{|p_{1+} - p_{+1}|}{5} - (0.5 - p_{\min})^2}{2}$$

and the estimator for the tetrachoric correlation coefficient is

$$\hat{\rho}^* = \cos\left(\frac{\pi}{1 + \widehat{\omega}^{\hat{c}}}\right)$$

A standard error of this estimate is calculated by eq. (9) in Bonett & Price (2005), and a p value is then estimated by a simple two-sided Z test. For small sample sizes, the permutation test calculated by Past is probably better.

Variance-covariance

The variance-covariance matrix (with the variances on the main diagonal) is computed in the usual way, for each column pair $\boldsymbol{x}$, $\boldsymbol{y}$:

$$\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1} \sum (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})$$

There is no statistical testing of variance-covariance.

Permutation tests

Monte Carlo permutation tests (*N*=9999) are available for all the correlation coefficients except polyserial and partial correlation, and the phi coefficient.

Correlation table plots

Plotting of the correlation table includes several options. The "Ellipses" function shows the correlation coefficients *r* as ellipses with major axis of unity, and minor axis *d* according to Schilling (1984):

$$r = \frac{1 - d^2}{1 + d^2}$$

The correlation table plot is not valid for the phi coefficient or the variance-covariance.

References

Bonett, D.G. & Price, R.M. 2005. Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics* 30:213-225.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S. & Bähler, J. 2015. Proportionality: A valid alternative to correlation for relative data. *PLoS Computational Biology* 11(3): e1004075

Ly, A., Verhagen, J., Wagenmakers, E.-J. 2016. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology* 72:19-32.

Olsson, U., F. Drasgow & N.J. Dorans. 1982. The polyserial correlation coefficient. *Psychometrika* 47:337-347.

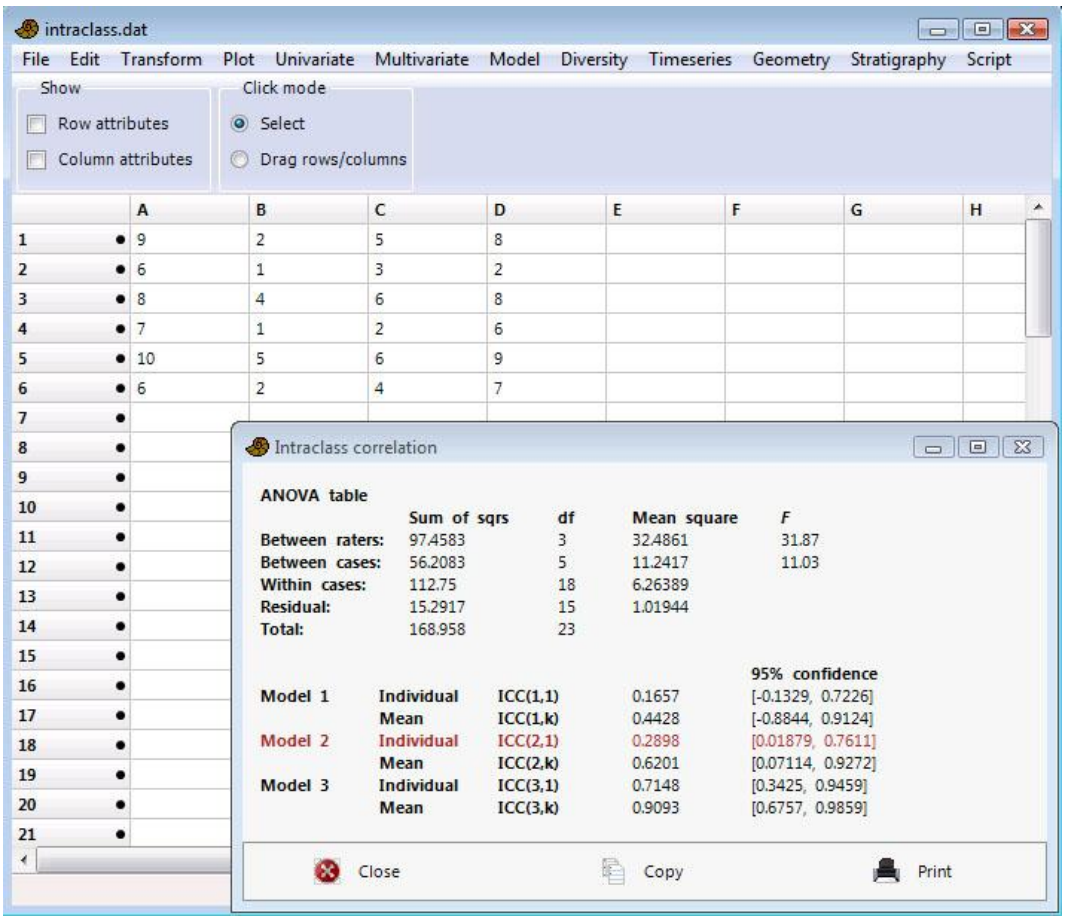
Pearson, J.W., Olver, S., Porter, M.A. 2016. Numerical methods for the computation of the confluent and Gauss hypergeometric functions. *Numerical Algorithms* DOI 10.1007/s11075-016-0173-0.

Press, W.H., S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. 1992. Numerical Recipes in C. Cambridge University Press.

Schilling, M.F. 1984. Some remarks on quick estimation of the correlation coefficient. *The American Statistician* 38:330.

Intraclass correlation

A typical use of the intraclass correlation coefficient (ICC) is to quantify *rater reliability*, i.e. level of agreement between several 'raters' measuring the same objects. It is a standard tool to assess measurement error. ICC=1 would indicate perfect reliability. Raters (or 'judges') go in columns, while the objects measured go in rows. In the example below there are four raters A-B, which have measured 6 objects.



Past follows the standard reference, Shrout and Fleiss (1979), which provides a number of different coefficients, referred to as ICC(m,k) where m is the model type. If k=1, the coefficient evaluates individual measurements (by a single rater); otherwise it evaluates the average measurement across raters. The models are

Model 1: the raters rating different objects are different, and randomly sampled from a larger set of raters

Model 2: the same raters rate all objects, and the raters are a subset of a larger set of raters.

Model 3: no assumptions about the raters.

The most commonly used ICC is ICC(2,1), which is therefore marked in red in Past.

The analysis is based on a two-way ANOVA without replication, as described elsewhere in this manual. Confidence intervals are parametric, following the equations of Shrout and Fleiss (1979). The

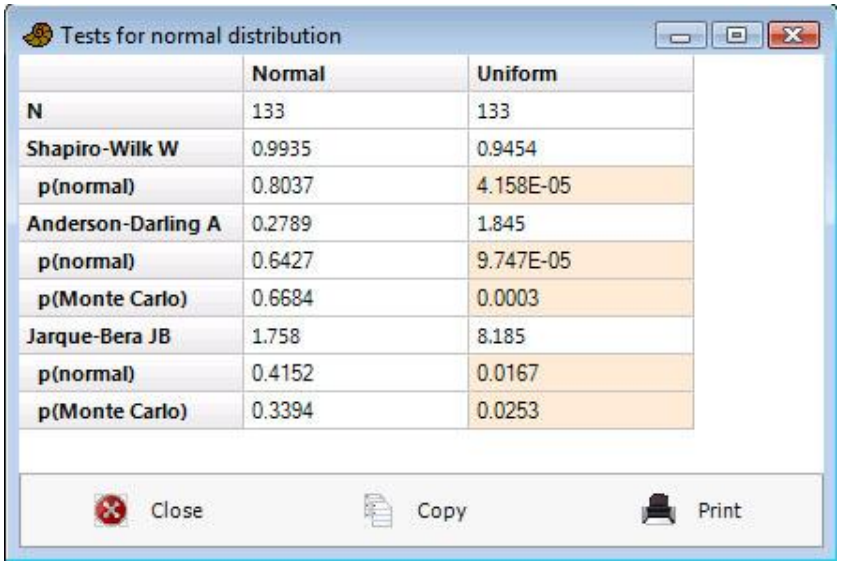
data in the example above are from the Shrout and Fleiss paper, the output from Past reproducing their results.

Reference

Shrout, P.E., Fleiss, J.L. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86:420-428.

Normality tests

Four statistical tests for normal distribution of one or several samples of univariate data, given in one or more separate columns or with a single data column and a group column. The data used below were generated by the normal and uniform random number generators in Past ('Evaluate expression' module).



Tests for normal distribution

	Normal	Uniform
N	133	133
Shapiro-Wilk W	0.9935	0.9454
p(normal)	0.8037	4.158E-05
Anderson-Darling A	0.2789	1.845
p(normal)	0.6427	9.747E-05
p(Monte Carlo)	0.6684	0.0003
Jarque-Bera JB	1.758	8.185
p(normal)	0.4152	0.0167
p(Monte Carlo)	0.3394	0.0253

Close Copy Print

For all four tests, the null hypothesis is

H_0: The sample was taken from a population with normal distribution.

If the given *p*(normal) is less than 0.05, normal distribution can be rejected (marked in pink). Of the given tests, the Shapiro-Wilk and Anderson-Darling are considered to be the more exact, and the Lilliefors and Jarque-Bera are given for reference. And even poorer test (four-bin chi-squared) was included in previous versions of Past. There is a maximum sample size of n=5000, while the minimum sample size is 3 (the tests will of course have extremely small power for such small n).

Remember the multiple testing issue if you run these tests on several samples – a Bonferroni or other correction may be appropriate.

Missing data: Supported by deletion.

Shapiro-Wilk test

The Shapiro-Wilk test (Shapiro & Wilk 1965) returns a test statistic W, which is small for non-normal samples, and a p value. The implementation is based on the standard code "AS R94" (Royston 1995), correcting an inaccuracy in the previous algorithm "AS 181" for large sample sizes.

Anderson-Darling test

The data X_i are sorted in ascending sequence, and normalized for mean and standard deviation:

$$Y_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}}.$$

With *F* the normal cumulative distribution function (CDF), the test statistic is

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F(Y_i) + \ln\left(1 - F(Y_{n+1-k})\right)\right].$$

Significance is estimated according to Stephens (1986). First, a correction for small sample size is applied:

$$A^{*2} = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right).$$

The *p* value is estimated as

$$p = \begin{cases} \exp\left(1.2937 - 5.709A^{*2} + 0.0186\left(A^{*2}\right)^2\right) & A^{*2} \geq 0.6 \\ \exp\left(0.9177 - 4.279A^{*2} - 1.38\left(A^{*2}\right)^2\right) & 0.34 < A^{*2} < 0.6 \\ 1 - \exp\left(-8.318 + 42.796a^{*2} - 59.938\left(A^{*2}\right)^2\right) & 0.2 < A^{*2} \leq 0.6 \\ 1 - \exp\left(-13.436 + 101.14a^{*2} - 223.73\left(a^{*2}\right)^2\right) & A^{*2} \leq 0.2 \end{cases}$$

Lilliefors test

The Lilliefors test is basically the same as the Kolmogorov-Smirnov test, comparing the sample distribution with a normal distribution with mean and variance estimated from the data. Because of this parameter estimation, significance must be computed in a different way than for Kolmogorov-Smirnov. In addition to the Monte Carlo procedure, Past reports a *p* value using an approximation to published tables given by Molin & Abdi (1998) and Abdi & Molin (2007):

b_2 = 0.08861783849346
b_1 = 1.30748185078790
b_0 = 0.37872256037043

$$A = \frac{-(b_1 + N) + \sqrt{(b_1 + N)^2 - 4b_2(b_0 - 1/L^2)}}{2b_2}$$

p = -0.37782822932809 + 1.67819837908004A - 3.02959249450445A^2
+ 2.80015798142101A^3 - 1.39874347510845A^4
+ 0.40466213484419A^5 - 0.06353440854207A^6
+ 0.00287462087623A^7 + 0.00069650013110A^8
+ 0.00011872227037A^9 + 0.00000575586834A^{10}

Jarque-Bera test

The Jarque-Bera test (Jarque & Bera 1987) is based on skewness *S* and kurtosis *K*. The test statistic is

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4} \right).$$

In this context, the skewness and kurtosis used are

$$S = \frac{1}{n} \frac{\sum (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \right)^3},$$

$$K = \frac{1}{n} \frac{\sum (x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \right)^4}.$$

Note that these equations contain simpler estimators than the G_1 and G_2 given in the Univariate summary statistics module, and that the kurtosis here will be 3, not zero, for a normal distribution.

Asymptotically (for large sample sizes), the test statistic has a chi-square distribution with two degrees of freedom, and this forms the basis for the *p* value given by Past. It is known that this approach works well only for large sample sizes, and Past therefore also includes a significance test based on Monte Carlo simulation, with 10,000 random values taken from a normal distribution.

References

Abdi, H. & Molin, P. 2007. Lilliefors/Van Soest's test of normality. In: Neil Salkind (Ed.) Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage.

Jarque, C. M. & Bera, A. K. 1987. A test for normality of observations and regression residuals. *International Statistical Review* 55:163–172.

Molin, P. & Abdi, H. 1998. New tables and numerical approximation for the KolmogorovSmirnov/Lilliefors/Van Soest test of normality. Technical report, University of Bourgogne.

Royston, P. 1995. A remark on AS 181: The *W*-test for normality. *Applied Statistics* 44:547-551.

Shapiro, S. S. & Wilk, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52:591–611.

Stephens, M.A. 1986. Tests based on edf statistics. Pp. 97-194 in D'Agostino, R.B. & Stephens, M.A. (eds.), Goodness-of-Fit Techniques. New York: Marcel Dekker.

Outlier tests

A single column of numbers is required. These tests provide objective procedures for detecting outliers in normally distributed data. The "single outlier" tests (Grubbs and Dixon) are designed to detect one outlier only, and should not be repeated for several outliers. The "multiple outlier" test (generalized ESD) attempts to detect multiple outliers, if present.

Grubbs test

The Grubbs test (also known as the Pearson-Hartley or the Extreme Studentized Deviate) tests for a single outlier. The basic reference is Grubbs (1950) but the actual reference for the *G* statistic as defined by e.g. Wikipedia and NIST is difficult to trace down. In any case, we define *G* as

$$G = \frac{\max|x_i - \bar{x}|}{s}$$

where $\bar{x}$ is the sample mean and *s* is the sample standard deviation. The critical value for *G* at a two-sided significance level of α is given by

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{\left(t_{\alpha/(2N),N-2}\right)^2}{N-2+\left(t_{\alpha/(2N),N-2}\right)^2}}$$

where $t_{\alpha/(2N),N-2}$ is the critical value of the *t* distribution with *N*-2 degrees of freedom and a significance level of $\alpha/(2N)$. Past solves this equation for α to give the *p* value for given *G* and *N*. Note that this is a two-sided test, testing for presence of an outlier at either end of the range (smallest or largest value). The Grubb's test is recommended for relatively large sample sizes (*N*>30). For smaller sample sizes, the Dixon test is preferable, although the two tests usually give similar results.

Dixon's test

The Dixon test (Dixon 1950) tests for a single outlier. It can only be used for small sample sizes (*N*<=30) but is then considered superior to the Grubbs test. The gap between the smallest (or largest) value and its adjacent value is compared to the total range, giving a test statistic *Q*. The calculation of *Q* depends on sample size as follows (the sample values *x* are sorted in ascending order):

Sample size *N*	Test statistic *Q*	At data minimum	At data maximum
$3 \le N \le 7$	r_{10}	$\frac{x_2 - x_1}{x_N - x_1}$	$\frac{x_N - x_{N-1}}{x_N - x_1}$
$8 \le N \le 10$	r_{11}	$\frac{x_2 - x_1}{x_{N-1} - x_1}$	$\frac{x_N - x_{N-1}}{x_N - x_2}$
$11 \le N \le 13$	r_{21}	$\frac{x_3 - x_1}{x_{N-1} - x_1}$	$\frac{x_N - x_{N-2}}{x_N - x_2}$
$14 \le N \le 30$	r_{22}	$\frac{x_3 - x_1}{x_{N-2} - x_1}$	$\frac{x_N - x_{N-2}}{x_N - x_3}$

The *p* value (two-tailed) is estimated by Monte Carlo simulation with 200,000 random, normally distributed samples of size *N* (it will vary slightly between runs).

Generalized ESD (Extreme Studentized Deviate) test

This procedure can detect more than one outlier. Moreover, it can detect outliers even when the one-outlier tests above do not report significance, because of so-called "masking".

The procedure starts by testing the most extreme value in the complete sample, giving a test statistic R_1 (= Grubbs *G*). This most extreme value is then removed from the sample and the procedure is repeated until 20% of the sample has been tested. The critical value R_{crit} (for significance at $p<0.05$) is adjusted for each iteration (Rosner 1983). Past marks the values for which $R>R_{crit}$ in pink. A *p* value is not calculated explicitly in Past.

Important: All of the most extreme values, up to the last value for which $R>R_{crit}$, are to be considered outliers, and are marked as such in Past. Quite often, the initial, most extreme values do not give $R>R_{crit}$ but they can still be outliers because of a significant value further down in the list. This is due to the masking effect. It looks odd, but is not a bug!

References

Dixon, W.J. 1950. Analysis of extreme values. *Annals of Mathematical Statistics* 21:488-506.

Grubbs, F. 1950. Sample criteria for testing outlying observations. *Annals of Mathematical Statistics* 21:27-58.

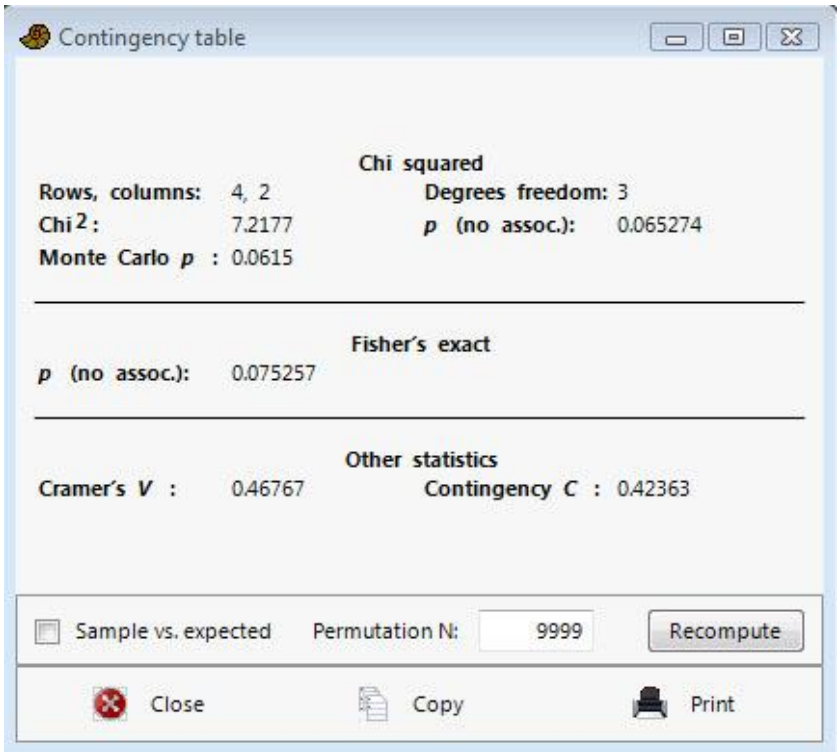
Rosner, B. 1983. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* 25:165-172.

Contingency table (chi^2 etc.)

This test expects a frequency table with numbers of elements in different categories (rows and columns). Rows represent the different states of one nominal variable, columns represent the states of another nominal variable, and cells contain the integer counts of occurrences of that specific state (row, column) of the two variables. The contingency table analysis then gives information on whether the two variables are associated. For example, this test can be used to compare two samples (columns) with the number of individuals in each taxon organized in the rows. You should be cautious about this test if any of the cells contain less than five individuals (see Fisher's exact test below).

The significance of association between the two variables is given, with *p* values from the chi-squared distribution and from a Monte Carlo (random) permutation test where all cells are random but with constant row and column sums.

For 2x2 tables, Yates's correction is applied.



The "Sample vs. expected" box should be ticked if you have two columns, and your second column consists of counts from a theoretical distribution (expected values) with zero sampling error, possibly non-integer. This is *not* a small-sample correction. In this case, only the chi-squared test is available.

See e.g. Brown & Rothery (1993) or Davis (1986) for details.

The *Fisher's exact test* is also given (two-tailed). When available, the Fisher's exact test will be superior to the chi-square. For large tables or large counts, the computation time can be prohibitive, and this test is not carried out. In such cases the parametric test is probably acceptable in any case. The procedure is complex and based on the network algorithm of Mehta & Patel (1986).

Two further measures of association are given. Both are transformations of chi-squared (Press et al. 1992). With *n* the total sum of counts, M the number of rows and N the number of columns:

Cramer's *V*: $$V = \sqrt{\frac{\chi^2}{n \min(M-1, N-1)}}$$

Contingency coefficient *C*: $$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Missing data not supported.

Residuals

If you get a significant association ($p<0.05$) in the chi-squared test, it may be of interest to see which of the cells contribute most strongly to the departure from the expected values under the null hypothesis of no association (post-hoc analysis). The table of residuals can show the following values for each cell:

Raw residuals: *O-E*, where *O* is the observed and *E* the expected value.

Standardized residuals: $(O-E)/\sqrt{E}$, standardizing for the magnitude of the expected value.

Adjusted residuals:

$$\text{adj_resid} = \frac{O - E}{\sqrt{E(1 - \text{RowMarginal}/n)}(1 - \text{ColumnMarginal}/n)}$$

where the RowMarginal is the row sum and ColumnMarginal is the column sum of the cell.

p values: The adjusted residuals are approximately normally distributed, meaning that values outside the two-sigma interval [-1.96, 1.96] can be considered significant at $p<0.05$, although the multiple testing problem applies. It is recommended to use the Bonferroni correction. Significant *p* values are marked in pink.

References

Brown, D. & P. Rothery. 1993. Models in biology: mathematics, statistics and computing. John Wiley & Sons.

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Mehta, C.R. & N.R. Patel. 1986. Algorithm 643: FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered r×c contingency tables. *ACM Transactions on Mathematical Software* 12:154-161.

Press, W.H., S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. 1992. Numerical Recipes in C. Cambridge University Press.

Cochran-Mantel-Haenszel test

Similar to the chi-squared test but testing several (repeated) 2x2 contingency tables simultaneously, in order to control for a nuisance factor (Mantel & Haenszel 1959). An example could be controlling for season when testing for the effect of a medicine that was used in repeated experiments through a year. The 2x2 tables must be given in consecutive rows in the spreadsheet (first, one 2x2 table, followed below by the next table, etc.).

Our algorithm follows McDonald (2014). Indexing the individual tables by k, we write one particular table as

$$\begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix}$$

With M the number of tables, and $n_k = a_k + b_k + c_k + d_k$, the chi-squared is calculated as

$$\chi^2_{MH} = \frac{\left[\left|\sum_{k=1}^{M}(a_k - (a_k + b_k)(a_k + c_k)/n_k)\right| - 0.5\right]^2}{\sum_{k=1}^{M}(a_k + b_k)\,(a_k + c_k)(b_k + d_k)(c_k + d_k)/(n_k^3 - n_k^2)}$$

Note that other, algebraically equivalent, forms are often given in the literature. The subtraction of 0.5 is a continuity correction, not always included in other software. This test statistic is distributed as $\chi 2$ with one degree of freedom.

In addition, the common odds ratio (equal to one for total independence) is calculated using the Mantel-Haenszel (1959) estimate:

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{M} a_k d_k / n_k}{\sum_{k=1}^{M} b_k c_k / n_k}$$

The confidence interval is calculated as [exp(ln θ_{MH} – 1.96 $\sqrt{V_{US}}$), exp(ln θ_{MH} + 1.96 $\sqrt{V_{US}}$)] where V_{US} is the variance of the logarithm of θ_{MH} (Robins et al. 1986).

Missing data not supported.

References

Mantel, N. & W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22:719-748.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

Robins, J., Breslow, N., Greenland, S. 1986. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 42:311-323.

Risk/odds

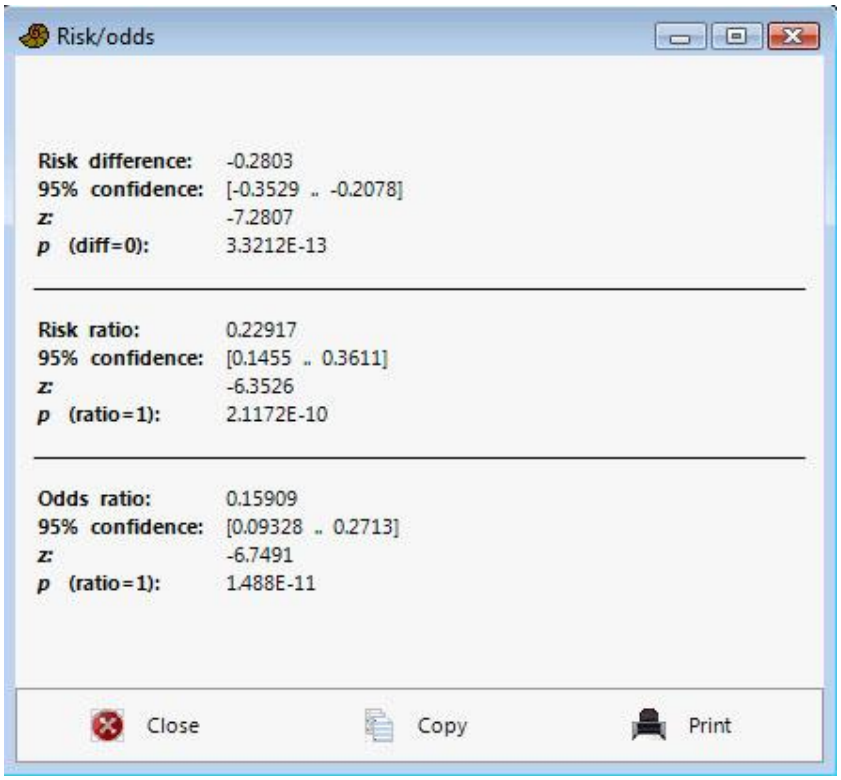
This module compares the counts of a binary outcome under two different treatments, with statistics that are in common use in medicine. The data are entered in a 2x2 table, with treatments in rows and counts of the two different outcomes in columns.

The following example shows the results of a vaccination trial on 460 patients:

	Got influenza	Did not get influenza
Vaccine	20	220
Placebo	80	140

In general, the data take the form

	Outcome 1	Outcome 2
Treatment 1	d_1	h_1
Treatment 2	d_0	h_0



Let $n_1=d_1+h_1$, $n_0=d_0+h_0$ and $p_1=d_1/n_1$, $p_0=d_0/n_0$. The statistics are then computed as follows:

Risk difference: $RD = p_1 - p_0$

95% confidence interval on risk difference (Pearson's chi-squared):

$$s_e = \sqrt{\frac{p_{|1}(1-p_1)}{n_1} + \frac{p_0(1-p_{0|})}{n_0}}$$

Interval: $RD - 1.96\, s_e$ to $RD + 1.96\, s_e$

Z test on risk difference (two-tailed):

$$z = \frac{RD}{s_e}$$

Risk ratio: $RR = p_1/p_0$

95% confidence interval on risk ratio ("delta method"):

$$s_e(\ln RR) = \sqrt{\frac{1}{d_1} - \frac{1}{n_1} + \frac{1}{d_0} - \frac{1}{n_0}}$$

$$EF = e^{1.96 s_e}$$

Interval: *RR* / *EF* to *RR* x *EF*

Z test on risk ratio (two-tailed):

$$z = \frac{\ln RR}{s_e}$$

Odds ratio: $OR = \frac{d_1/h_1}{d_0/h_0}$

95% confidence interval on odds ratio ("Woolfs's formula"):

$$s_e(\ln OR) = \sqrt{\frac{1}{d_1} + \frac{1}{h_1} + \frac{1}{d_0} + \frac{1}{h_0}}$$

$$EF = e^{1.96 s_e}$$

Interval: *OR* / *EF* to *OR* x *EF*

Note there is currently no continuity correction.

Missing data are not allowed and will give an error message.

Single proportion

A simple module for calculating the probability of an observed (sample) proportion (in the range 0-1) against a hypothetical proportion. No input data are required in the spreadsheet.

With *p* the sample proportion, *P* the hypothetical proportion, and *n* the sample size, we calculate the standard deviation

$$\sigma = \sqrt{\frac{P(1-P)}{n}}$$

Further, we calculate the *z* (normal distribution) statistic

$$z = \frac{p-P}{\sigma}$$

The (two-tailed) significance is calculated directly from *z* and the normal distribution.

The 95% confidence interval for the proportion is calculated using two different methods. The 'exact' interval is computed using the Clopper-Pearson method (Clopper and Pearson 1934) as

$$\left(1+\frac{n-x+1}{xF[1-\alpha/2\,;2x,2(n-x+1)]}\right)^{-1} < \theta < \left(1+\frac{n-x}{(x+1)F[\alpha/2\,;\;2(x+1),2(n-x)]}\right)^{-1}$$

where α = 0.05, *x* is the number of successes computed as round(*pn*), and $F(c;\, d_1, d_2)$ is the 1-*c* quantile for an *F* distribution with d_1 and d_2 degrees of freedom. The normal approximation is computed as

$$b = \sqrt{\frac{p(1-p)}{n}}$$

$$CI = (p-1.96b, p+1.96b)$$

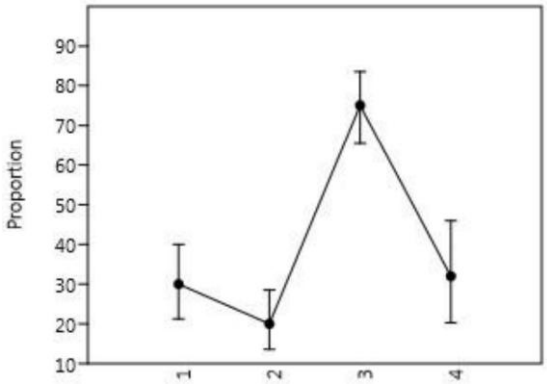
The normal approximation CI is more commonly used. The exact CI is more accurate for small *n*. For large *n* the two methods will give similar results.

Reference

Clopper, C. & Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26:404–413.

Multiple proportion confidence intervals

Accepts two columns of data, the first column with proportions given as percentages (0-100) and the second column containing sample sizes (*N*). The program will compute 95% confidence intervals for all the proportions using the Clopper-Pearson method (see above), and plot them.



Missing data are disregarded.

Ratio of counts confidence interval

This module calculates confidence intervals for ratios of counts. It is specially designed for environmental indices based on microfossil counts in paleontology, on the form $(a + c)/(b + c)$. This general formula covers the common cases a/b, where c=0, and $c/(b+c)$, where a=0. No data are required in the spreadsheet.

Two methods are provided, as described by Escarguel et al. (2019). The first is a Monte Carlo procedure based on a transformation of the normal distribution. The second is a straightforward bootstrap, with two versions (simple, which can produce negative values), and percentile.

For the Monte Carlo procedure, first note that $r = \frac{a+c}{b+c} = \frac{\frac{a+c}{T}}{\frac{b+c}{T}} = \frac{\frac{a}{T}+\frac{c}{T}}{\frac{b}{T}+\frac{c}{T}}$, where $T = a+b+c$.

Let e and g be the arcsine-transformed values $e = \sin^{-1}\sqrt{(a/T)}$ and $g = \sin^{-1}\sqrt{(c/T)}$. e and g should then be normally distributed variables with a sample standard deviation $s = \sqrt{(1/(4T))}$ (Sokal and Rohlf, 2011). Therefore, let e^* and g^* be normally distributed random variates with mean e and g, and standard deviation s. Then, e^* and g^* are back-transformed to proportion values $(a/T)^* = \sin^2 e^*$ and $(c/T)^* = \sin^2 g^*$. Finally, a Monte-Carlo estimate of r is calculated as $r^* = \frac{\left(\frac{a}{T}\right)^* + \left(\frac{c}{T}\right)^*}{1-\left(\frac{a}{T}\right)^*}$.

This procedure is reiterated a large number (say, 10,000) times, leading to a Monte-Carlo distribution of r, from which the 2.5 and 97.5 percentiles define the 95% confidence interval.

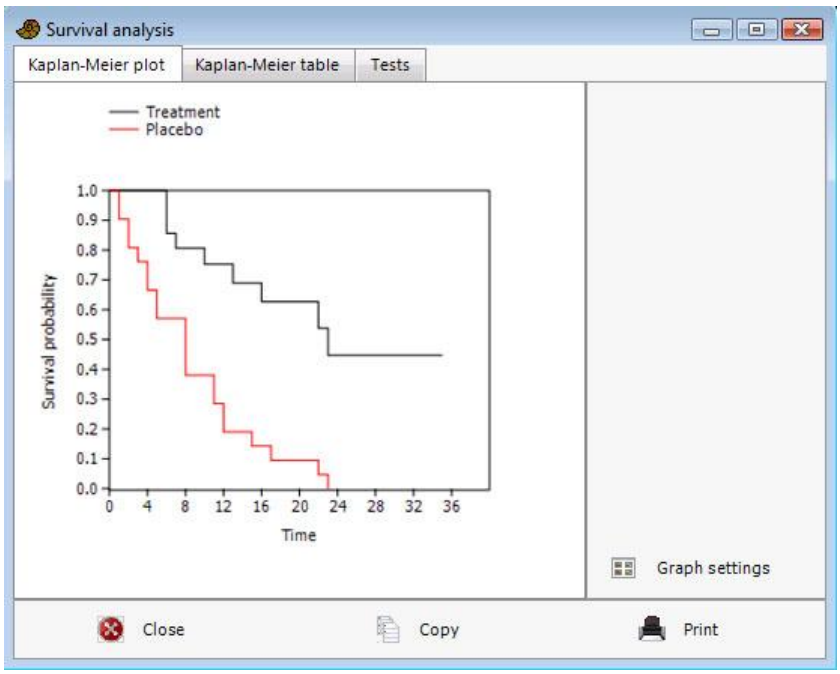
Reference

Suchéras-Marx, B., Escarguel, G., Ferreira, J. & Hammer, Ø. 2019. Statistical confidence intervals for relative abundances and abundance-based ratios: simple practical solutions for an old overlooked question. *Marine Micropaleontology* 151: 101751.

Survival analysis (Kaplan-Meier curves, log-rank test etc.)

Survival analysis for two groups (treatments) with provision for right censoring. The module draws Kaplan-Meier survival curves for the two groups and computes three different tests for equivalence. The program expects four columns. The first column contains times to failure (death) or censoring (failure not observed up to and including the given time) for the first group, the second column indicates failure (1) or censoring (0) for the corresponding individuals. The last two columns contain data for the second group. Failure times must be larger than zero.

The program also accepts only one treatment (given in two columns), or more than two treatments in consecutive pairs of columns, plotting one or multiple Kaplan-Meier curves. The statistical tests are only comparing the first two groups, however.



The Kaplan-Meier curves and the log-rank, Wilcoxon and Tarone-Ware tests are computed according to Kleinbaum & Klein (2005).

Average time to failure includes the censored data. Average hazard is number of failures divided by sum of times to failure or censorship.

The log-rank test is by chi-squared on the second group:

$$\chi^2 = \frac{(O_2 - E_2)^2}{\mathrm{var}(O_2 - E_2)} = \frac{\left(\sum_j (m_{2j} - e_{2j})\right)^2}{\sum_j \frac{n_{1j} n_{2j} (m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2 (n_{1j} + n_{2j} - 1)}}.$$

Here, n_{ij} is the number of individuals at risk, and m_{ij} the number of failures, in group i at distinct failure time j. The expected number of failures in group 2 at failure time j is

$$e_{2j} = \frac{n_{2j}\left(m_{1j} + m_{2j}\right)}{n_{1j} + n_{2j}}.$$

The chi-squared has one degree of freedom.

The Wilcoxon and Tarone-Ware tests are weighted versions of the log-rank test, where the terms in the summation formulas for O_2-E_2 and var(O_2-E_2) receive weights of n_j and $\sqrt{n_j}$, respectively. These tests therefore give more weight to early failure times. They are not in common use compared with the log-rank test.

This module is not strictly necessary for survival analysis without right censoring – the Mann-Whitney test may be sufficient for this simpler case.

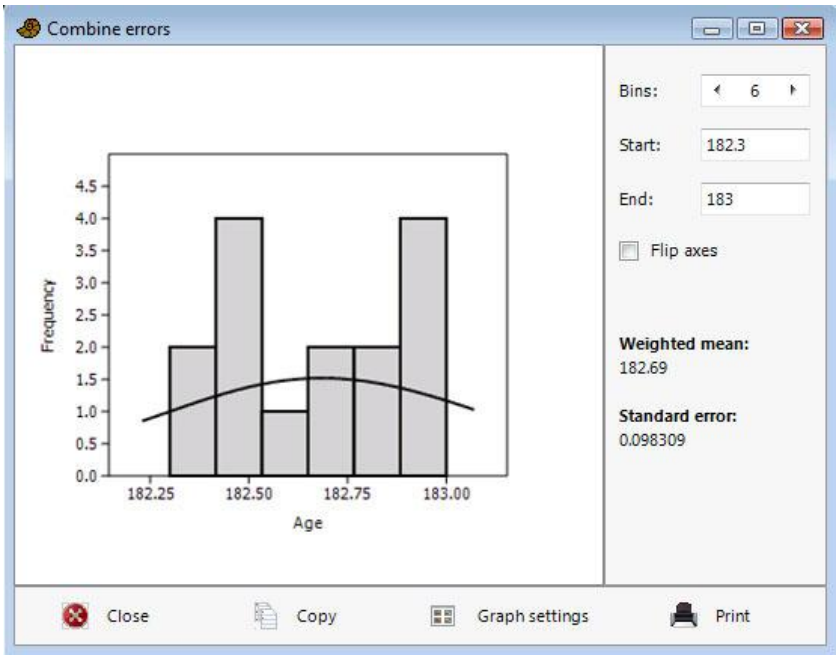
Missing data: Data points with missing value in one or both columns are disregarded.

Reference

Kleinbaum, D.G. & Klein, M. 2005. Survival analysis: a self-learning text. Springer.

Combine errors

A simple module for producing a weighted mean and its standard deviation from a collection of measurements with errors (one sigma). Expects two columns: the data *x* and their one-sigma errors *σ*. The sum of the individual gaussian distributions is also plotted.



The weighted mean and its standard deviation are computed as

$$\mu = \frac{\sum_i x_i / \sigma_i^2}{\sum_i 1/\sigma_i^2}, \qquad \sigma = \sqrt{\frac{1}{\sum_i 1/\sigma_i^2}}.$$

This is the maximum-likelihood estimator for the mean, assuming all the individual distributions are normal with the same mean.

Missing data: Rows with missing data in one or both columns are deleted.

Propagate errors

A simple calculator for propagation of errors. Two columns are required, with values and their standard deviations. The number of rows depends on the function *f* selected:

- ***f = A + B + …***: At least two rows (A and B) must be given. The variance is computed as

$$\sigma_f^2 = \sigma_A^2 + \sigma_A^2 + \cdots$$

- ***f* = A * B * …**: At least two rows (A and B) must be given:

$$\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 + \cdots \right]$$

- ***f* = A / B**: Exactly two rows (A and B) must be given:

$$\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 \right]$$

- ***f* = *A*^{a}**: A single row (A) must be given, in addition to the exponent a:

$$\sigma_f^2 \approx \left(\frac{f a \sigma_A}{A}\right)^2$$

- ***f* = *A*^{B}**: Exactly two rows (A and B) must be given:

$$\sigma_f^2 \approx f^2 \left[\left(\frac{B \sigma_A}{A}\right)^2 + (\ln(A)\sigma_B)^2 \right]$$

The standard deviation of *f* is computed as the square root of the variance. Note that any covariation between the given variables would produce additional terms in the equations above, but these are assumed to be zero.

Multivariate menu

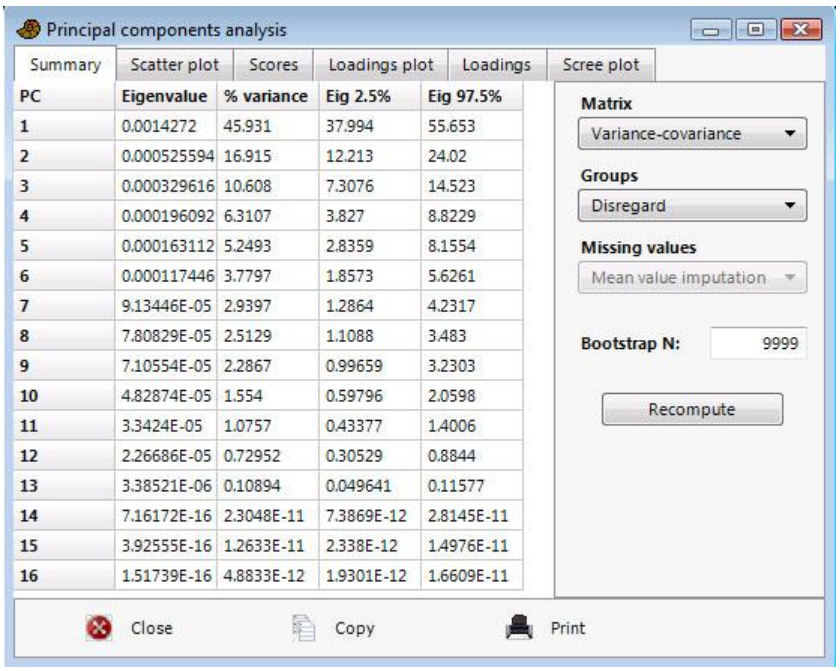
Principal components

Principal components analysis (PCA) finds hypothetical variables (components) accounting for as much as possible of the variance in your multivariate data (Davis 1986, Harper 1999, Legendre & Legendre 1998). These new variables are linear combinations of the original variables. PCA may be used for reduction of the data set to only two variables (the two first components), for plotting purposes. One might also hypothesize that the most important components are correlated with other underlying variables. For morphometric data, this might be size, while for ecological data it might be a physical gradient (e.g. temperature or depth).

The input data is a matrix of multivariate data, with items in rows and variates in columns.

The PCA routine finds the eigenvalues and eigenvectors of the variance-covariance matrix or the correlation matrix, with the SVD algorithm. Use variance-covariance if all variables are measured in the same units (e.g. centimetres). Use correlation (normalized var-covar) if the variables are measured in different units; this implies normalizing all variables using division by their standard deviations. The eigenvalues give a measure of the variance accounted for by the corresponding eigenvectors (components). The percentages of variance accounted for by these components are also given. If most of the variance is accounted for by the first one or two components, you have scored a success, but if the variance is spread more or less evenly among the components, the PCA has in a sense not been very successful.

In the example below (landmarks from gorilla skulls), component 1 is strong, explaining 45.9% of variance. The bootstrapped confidence intervals are not shown unless the 'Bootstrap N' value is non-zero.



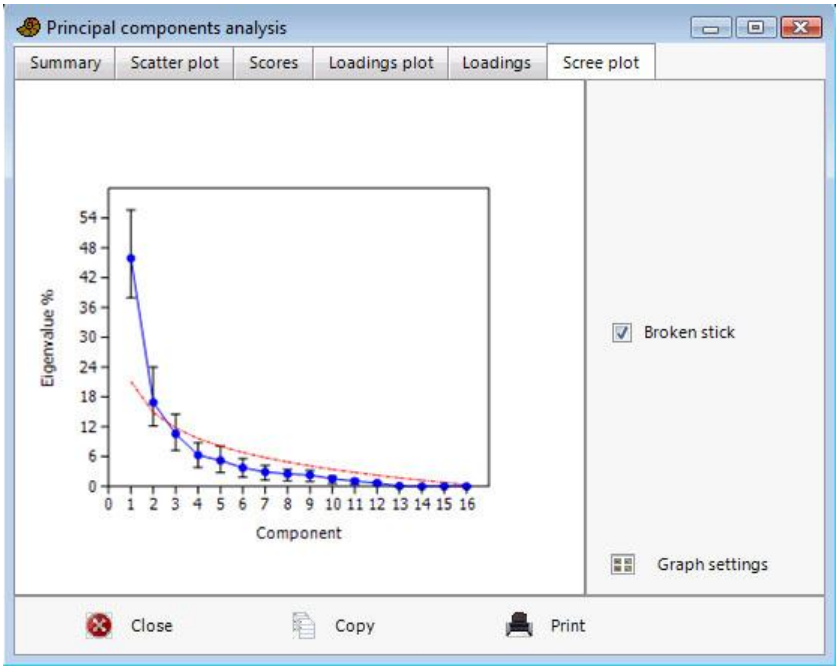
PC	Eigenvalue	% variance	Eig 2.5%	Eig 97.5%
1	0.0014272	45.931	37.994	55.653
2	0.000525594	16.915	12.213	24.02
3	0.000329616	10.608	7.3076	14.523
4	0.000196092	6.3107	3.827	8.8229
5	0.000163112	5.2493	2.8359	8.1554
6	0.000117446	3.7797	1.8573	5.6261
7	9.13446E-05	2.9397	1.2864	4.2317
8	7.80829E-05	2.5129	1.1088	3.483
9	7.10554E-05	2.2867	0.99659	3.2303
10	4.82874E-05	1.554	0.59796	2.0598
11	3.3424E-05	1.0757	0.43377	1.4006
12	2.26686E-05	0.72952	0.30529	0.8844
13	3.38521E-06	0.10894	0.049641	0.11577
14	7.16172E-16	2.3048E-11	7.3869E-12	2.8145E-11
15	3.92555E-16	1.2633E-11	2.338E-12	1.4976E-11
16	1.51739E-16	4.8833E-12	1.9301E-12	1.6609E-11

Groups: If groups are specified with a group column, the PCA can optionally be carried out *within-group* or *between-group*. In within-group PCA, the average within each group is subtracted prior to eigenanalysis, essentially removing the differences between groups. In between-group PCA, the eigenanalysis is carried out on the group means (i.e. the items analysed are the groups, not the rows). For both within-group and between-group PCA, the PCA scores are computed using vector products with the original data.

Supplementary variables: It is possible to include one or more initial columns containing additional supplementary variables for the analysis. These variables are not included in the ordination. The correlation coefficients between each supplementary variable and the PCA scores are presented as vectors from the origin (triplot). The lengths of the vectors are arbitrarily scaled to make a readable plot, so only their directions and relative lengths should be considered.

Row-wise **bootstrapping** is carried out if a positive number of bootstrap replicates (e.g. 1000) is given in the 'Bootstrap N' box. The bootstrapped components are re-ordered and reversed according to Peres-Neto et al. (2003) to increase correspondence with the original axes. 95% bootstrapped confidence intervals are given for the eigenvalues.

The **'Scree plot'** (simple plot of eigenvalues) may also indicate the number of significant components. After this curve starts to flatten out, the components may be regarded as insignificant. 95% confidence intervals are shown if bootstrapping has been carried out. The eigenvalues expected under a random model (Broken Stick) are optionally plotted - eigenvalues under this curve may represent non-significant components (Jackson 1993).

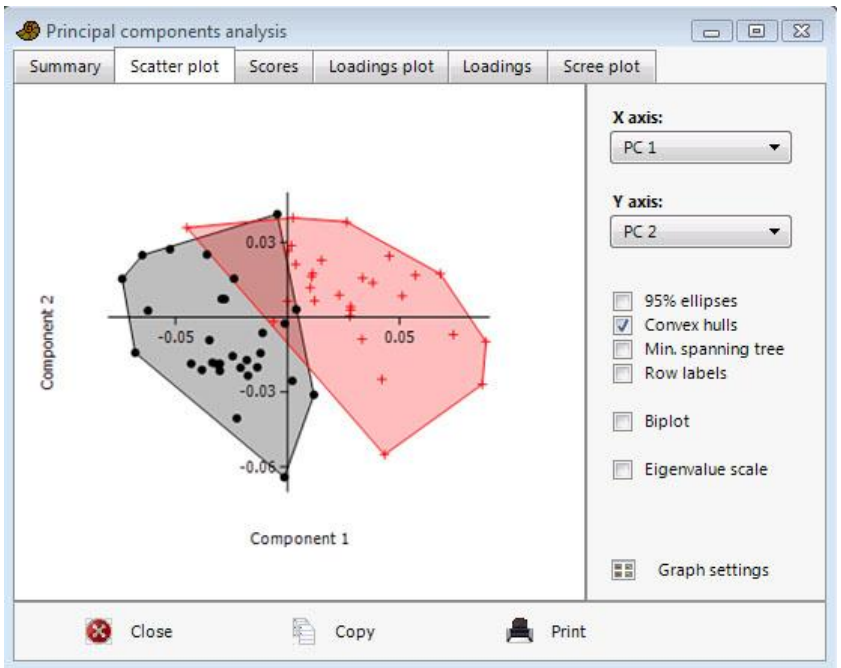


In the gorilla example above, the eigenvalues for the 16 components (blue line) lie above the broken stick values (red dashed line) for the first two components, although the broken stick is inside the 95% confidence interval for the second component.

The **scatter plot** shows all data points (rows) plotted in the coordinate system given by two of the components. If you have groups, they will be shown with different symbols and colours. The Minimal Spanning Tree is the shortest possible set of lines connecting all points. This may be used as a visual

aid in grouping close points. The MST is based on an Euclidean distance measure of the original data points, and is most meaningful when all variables use the same unit. The 'Biplot' option shows a projection of the original axes (variables) onto the scattergram. This is another visualisation of the PCA loadings (coefficients) - see below.

If the "Eigenval scale" is ticked, the data points will be scaled by $1/\sqrt{d_k}$, and the biplot eigenvectors by $\sqrt{d_k}$ - this is the correlation biplot of Legendre & Legendre (1998). If not ticked, the data points are not scaled, while the biplot eigenvectors are normalized to equal length (but not to unity, for graphical reasons) - this is the distance biplot.



The **loadings plot** shows to what degree your different original variables (given in the original order along the x axis) contribute to the different components (as chosen in the radio button panel). These component loadings are important when you try to interpret the 'meaning' of the components. The 'Coefficients' option gives the PC coefficients, while 'Correlation' gives the correlation between a variable and the PC scores. If bootstrapping has been carried out, 95% confidence intervals are shown (only for the Coefficients option).

Sphericity tests

Bartlett's sphericity test (Bartlett 1951) tests the null hypothesis that the points are sampled from a spherical distribution. If so, PCA will not be able to provide a useful reduction of dimensionality. The *p* value from this test should ideally be <0.05 (significant departure from sphericity). For PCA on the correlation matrix, Past uses the chi-squared approximation

$$\chi^2 = -\left(M - 1 - \frac{1}{6}(2N + 5)\right) \ln \det \boldsymbol{R}$$

where M is the number of points, N the number of variables, and **R** the correlation matrix. The determinant is calculated as the product of eigenvalues. Bartlett (1951) used M instead of M-1, but later authors have tended to the latter as a bias correction. The χ^2 has $N(N-1)/2$ degrees of freedom.

Past also provides a Bartlett's sphericity test for PCA on the var-covar matrix **V**, following Bartlett (1951). However, this test is not in common use, and its properties are not quite clear. It assumes equal variances in the variables. Use with caution:

$$\chi^2 = -\left(M - \frac{1}{6}(2N+1) + 2/N\right) \ln \det \boldsymbol{V}$$

with $(N+2)(N-1)/2$ degrees of freedom. Before calculating the determinant, all eigenvalues are divided by their mean.

The *Kaiser-Meyer-Olkin (KMO)* measure (Kaiser 1970), also known as Measure of Sampling Adequacy (MSA), is only supported for PCA on the correlation matrix.

The inverse of the correlation matrix $\mathbf{P} = \mathbf{R}^{-1}$ is used to calculate the partial correlation matrix **A**, with elements

$$A_{ij} = -\frac{P_{ij}}{\sqrt{P_{ii}P_{jj}}}$$

The KMO is then calculated as

$$KMO = \frac{\sum_i \sum_{j \neq i} R_{ij}^2}{\sum_i \sum_{j \neq i} R_{ij}^2 + \sum_i \sum_{j \neq i} A_{ij}^2}$$

KMO < 0.5 is reported as "unacceptable"; 0.5 ≤ KMO < 0.7 is reported as "mediocre"; 0.7 ≤ KMO <0.8 is "good"; 0.8 ≤ KMO < 1 is "excellent".

Missing data can be handled by one of two methods:

1. *Mean value imputation*: Missing values are replaced by their column average. Not recommended.
2. *Iterative imputation*: Missing values are inititally replaced by their column average. An initial PCA run is then used to compute regression values for the missing data. The procedure is iterated until convergence. This is usually the preferred method, but can cause some overestimation of the strength of the components (see Ilin & Raiko 2010).

References

Bartlett, M.S. 1951. The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika* 38:337-344.

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Harper, D.A.T. (ed.). 1999. Numerical Palaeobiology. John Wiley & Sons.

Ilin, A. & T. Raiko. 2010. Practical approaches to Principal Component Analysis in the presence of missing values. *Journal of Machine Learning Research* 11:1957-2000.

Jackson, D.A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74:2204-2214.

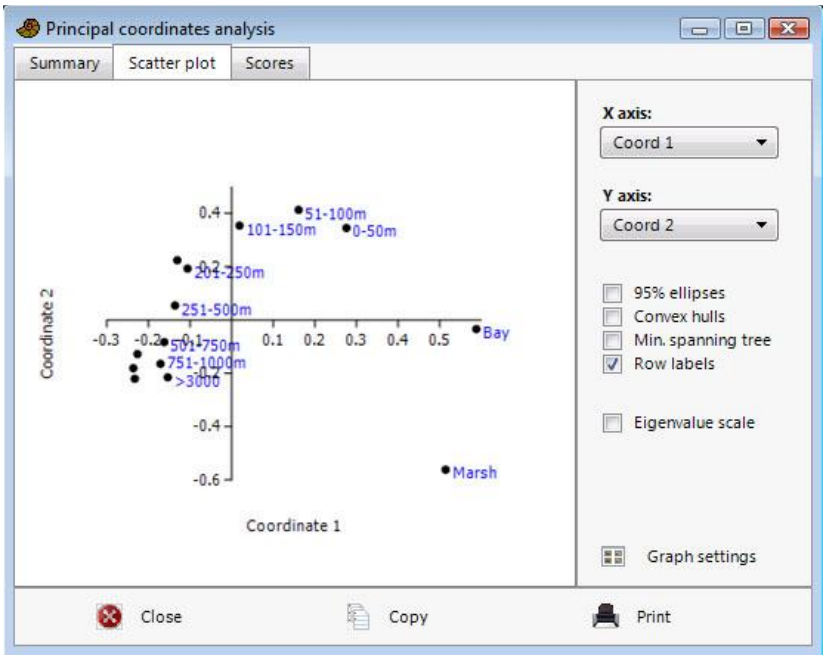
Kaiser, H.F. 1970. A second generation little jiffy. *Psychometrika* 35:401-415.

Legendre, P. & L. Legendre. 1998. Numerical Ecology, 2nd English ed. Elsevier, 853 pp.

Peres-Neto, P.R., D.A. Jackson & K.M. Somers. 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology* 84:2347-2363.

Principal coordinates

Principal coordinates analysis (PCoA) is another ordination method, also known as Metric Multidimensional Scaling. The algorithm is from Davis (1986).



The PCoA routine finds the eigenvalues and eigenvectors of a matrix containing the distances or similarities between all data points. The Gower measure will normally be used instead of Euclidean distance, which gives results similar to PCA. An additional eleven distance measures are available - these are explained under Cluster Analysis. The eigenvalues, giving a measure of the variance accounted for by the corresponding eigenvectors (coordinates) are given for the first four most important coordinates (or fewer if there are fewer than four data points). The percentages of variance accounted for by these components are also given.

The similarity/distance values are raised to the power of *c* (the "Transformation exponent") before eigenanalysis. The standard value is *c*=2. Higher values (4 or 6) may decrease the "horseshoe" effect (Podani & Miklos 2002).

The **scatter plot** allows you to see all your data points (rows) plotted in the coordinate system given by the PCoA. If you have colored (grouped) rows, the different groups will be shown using different symbols and colours. The "Eigenvalue scaling" option scales each axis using the square root of the eigenvalue (recommended). The minimal spanning tree option is based on the selected similarity or distance index in the original space.

Missing data is supported by pairwise deletion (not for the Raup-Crick, Rho or user-defined indices).

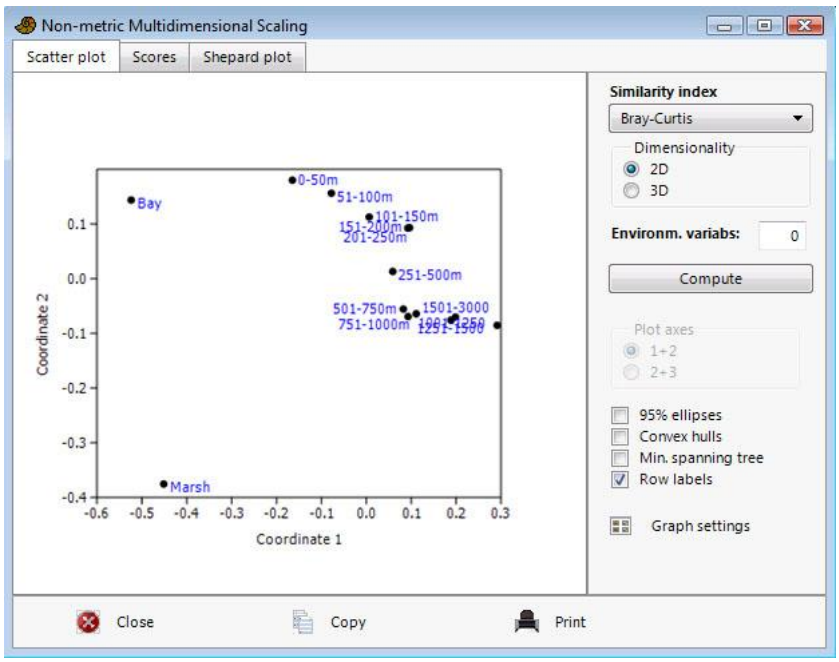
References

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Podani, J. & I. Miklos. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology* 83:3331-3343.

Non-metric MDS

Non-metric multidimensional scaling is based on a distance matrix computed with any of 21 supported distance measures, as explained under Similarity and Distance Indices below. The algorithm then attempts to place the data points in a two- or three-dimensional coordinate system such that the ranked differences are preserved. For example, if the original distance between points 4 and 7 is the ninth largest of all distances between any two points, points 4 and 7 will ideally be placed such that their euclidean distance in the 2D plane or 3D space is still the ninth largest. Non-metric multidimensional scaling intentionally does not take absolute distances into account.



The program may converge on a different solution in each run, depending upon the initial conditions. Each run is actually a sequence of 11 trials, from which the one with smallest stress is chosen. One of these trials uses PCoA as the initial condition, the others are random. The solution is automatically rotated to the major axes (2D and 3D).

The algorithm implemented in PAST, which seems to work very well, is based on a new approach developed by Taguchi and Oono (2005).

The minimal spanning tree option is based on the selected similarity or distance index in the original space.

Environmental variables: It is possible to include one or more initial columns containing additional "environmental" variables for the analysis. These variables are not included in the ordination. The correlation coefficients between each environmental variable and the NMDS scores are presented as vectors from the origin. The lengths of the vectors are arbitrarily scaled to make a readable biplot, so only their directions and relative lengths should be considered.

Column scores: The columns can be included in the NMDS plot as weighted averages of the row scores, as in Correspondence Analysis. The weighting uses the raw data values, and therefore does

not honour the choice of similarity index. However, it seems to work well for e.g. ecological data, allowing the plotting of species together with samples (sites).

Shepard plot: This plot of obtained versus observed (target) ranks indicates the quality of the result. Ideally, all points should be placed on a straight ascending line (x=y). The R^2 values are the coefficients of determination between distances along each ordination axis and the original distances (perhaps not a very meaningful value but is reported by other NMDS programs so is included for completeness).

Variance: Lists the Euclidean distances to the group centroids in the NMDS scatter plot. The data points are arranged in columns according to group. These values can be used e.g. in the study of disparity.

Missing data is supported by pairwise deletion (not for the Raup-Crick, Rho and user-defined indices). For environmental variables, missing values are not included in the computation of correlations.

Reference

Taguchi, Y.-H., Oono, Y. 2005. Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics* 21:730-40.

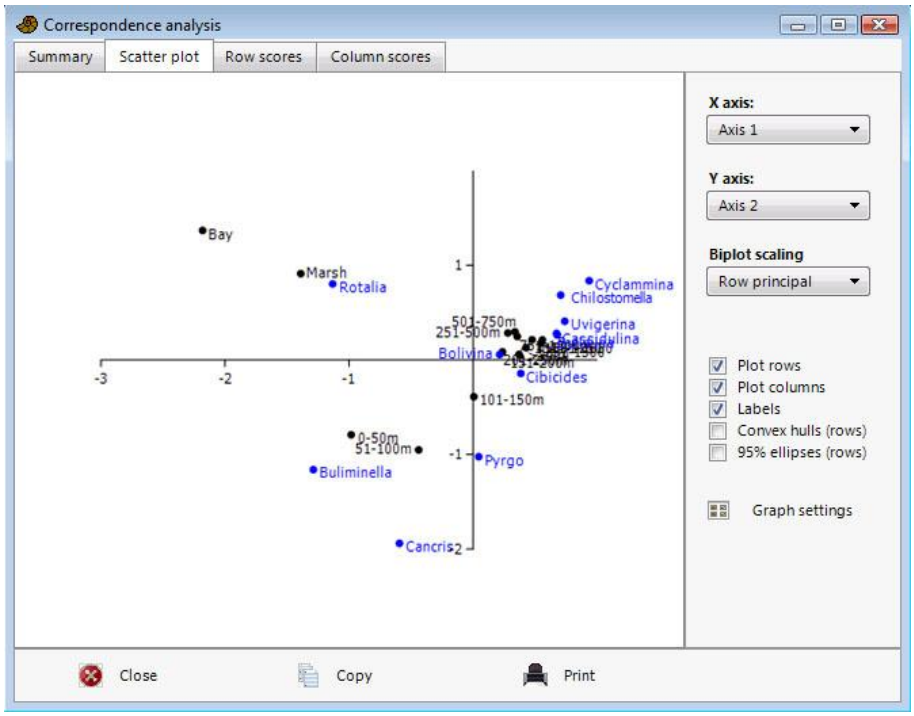
Correspondence analysis

Correspondence analysis (CA) is yet another ordination method, somewhat similar to PCA but for *counted data* (Legendre & Legendre 1998). For comparing associations (columns) containing counts of taxa, or counted taxa (rows) across associations, CA is the more appropriate algorithm. Also, CA is more suitable if you expect that species have unimodal responses to the underlying parameters, that is they favor a certain range of the parameter, becoming rare for lower and higher values (this is in contrast to PCA, which assumes a linear response).

The CA routine finds the eigenvalues and eigenvectors of a matrix containing the Chi-squared distances between all rows (or columns, if that is more efficient – the result is the same). The algorithm follows Greenacre (2010), with SVD. The eigenvalue, giving a measure of the similarity accounted for by the corresponding eigenvector, is given for each eigenvector. The percentages of similarity accounted for by these components are also given.

The *scatter plot* allows you to see all your data points (rows) plotted in the coordinate system given by the CA. If you have grouped rows, the different groups can be shown using separate convex hulls and concentration ellipses.

In addition, the variables (columns, associations) can be plotted in the same coordinate system (Q mode), optionally including the column labels. If your data are 'well behaved', taxa typical for an association should plot in the vicinity of that association.



Relay plot: This is a composite diagram with one plot per column. The plots are ordered according to CA column scores. Each data point is plotted with CA first-axis row scores on the vertical axis, and the original data point value (abundance) in the given column on the horizontal axis. This may be most

useful when samples are in rows and taxa in columns. The relay plot will then show the taxa ordered according to their positions along the gradients, and for each taxon the corresponding plot should ideally show a unimodal peak, partly overlapping with the peak of the next taxon along the gradient (see Hennebert & Lees 1991 for an example from sedimentology).

Scalings: Three different scalings of the row and column scores are available. In the row principal and column principal scalings, the distances between rows, between columns, and between rows and columns are all proportional to chi-squared distance. However, this will show rows and columns in two different scales, which can be difficult to read. The symmetric scaling uses the same scale for rows and columns, but chi-squared distance between rows and columns is not maintained.

Row and column contributions: These tables report the proportion (0-1) of contribution of each row or column point to the inertia of each axis. In the literature, these values are referred to as absolute contributions, or CTR. For row *i* and axis *j*, they are calculated as

$$\omega_{ij} = \frac{r_i}{\lambda_j^2 F_{ij}^2}$$

where r_i is the row sum divided by total count, λ_j is the eigenvalue for the axis, and F are the row scores. Absolute column contributions are calculated similarly, using column sums and scores.

Also, the contributions of each axis to the inertia of each point (relative contribution, COR) are given:

$$\sigma_{ij} = \frac{F_{ij}^2}{\sum_l F_{il}^2}$$

Supplementary variables: It is possible to include one or more initial columns containing additional supplementary variables for the analysis. These variables are not included in the ordination. The correlation coefficients between each supplementary variable and the CA scores are presented as vectors from the origin (triplot). The lengths of the vectors are arbitrarily scaled to make a readable plot, so only their directions and relative lengths should be considered.

Missing data is supported by column average substitution.

References

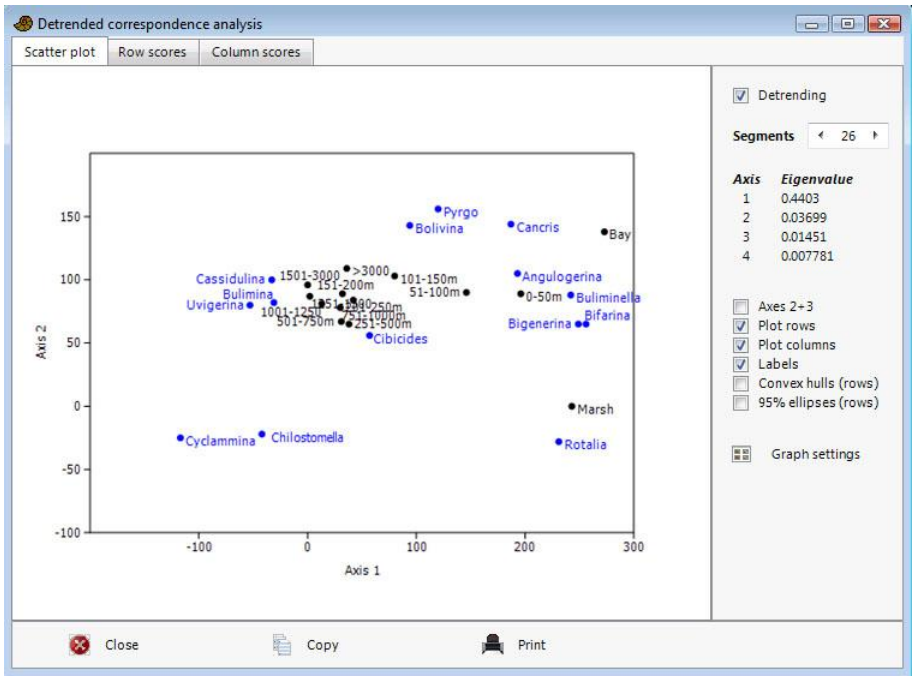
Greenacre, M. 2010. Biplots in practice. Fundación BBVA, 237 pp.

Hennebert, M. & A. Lees. 1991. Environmental gradients in carbonate sediments and rocks detected by correspondence analysis: examples from the Recent of Norway and the Dinantian of southwest England. *Sedimentology* 38:623-642.

Legendre, P. & L. Legendre. 1998. Numerical Ecology, 2nd English ed. Elsevier, 853 pp.

Detrended correspondence analysis

The Detrended Correspondence (DCA) module uses the same algorithm as Decorana (Hill & Gauch 1980), with modifications according to Oxanen & Minchin (1997). It is specialized for use on 'ecological' data sets with abundance data; samples in rows, taxa in columns.



Eigenvalues for the four ordination axes are given as in CA, indicating their relative importance in explaining the spread in the data.

Detrending is a sort of normalization procedure in two steps. The first step involves an attempt to 'straighten out' points lying in an arch, which is a common occurrence. The second step involves 'spreading out' the points to avoid clustering of the points at the edges of the plot. Detrending may seem an arbitrary procedure, but can be a useful aid in interpretation.

The units on the axes are *standard deviations x 100*.

Missing data is supported by column average substitution.

References

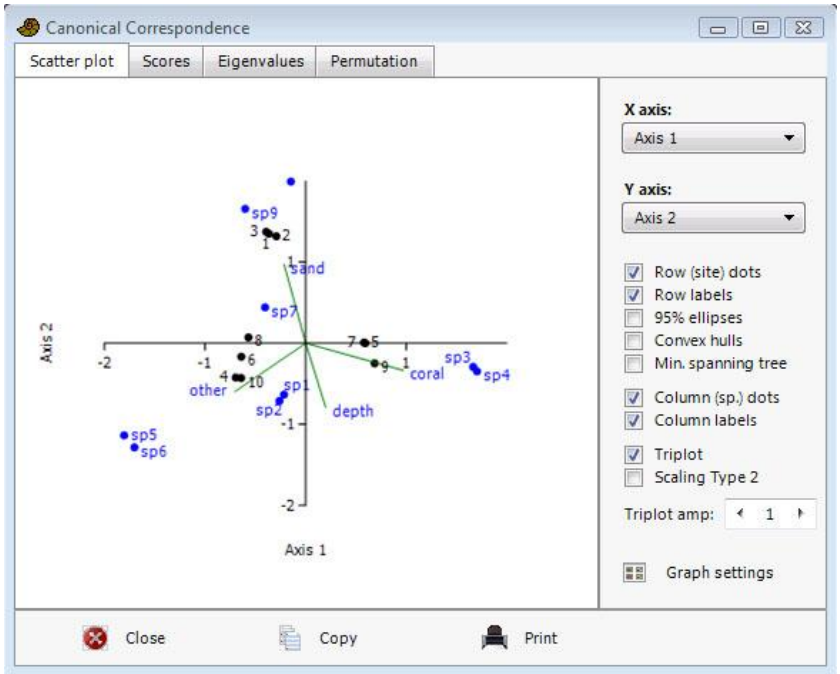
Hill, M.O. & H.G. Gauch Jr. 1980. Detrended Correspondence analysis: an improved ordination technique. *Vegetatio* 42:47-58.

Oxanen, J. & P.R. Minchin. 1997. Instability of ordination results under changes in input data order: explanations and remedies. *Journal of Vegetation Science* 8:447-454.

Canonical correspondence

Canonical Correspondence Analysis (Legendre & Legendre 1998) is correspondence analysis of a site/species matrix where each site has given values for one or more environmental variables (temperature, depth, grain size etc.). The ordination axes are linear combinations of the environmental variables. CCA is thus an example of direct gradient analysis, where the gradient in environmental variables is known *a priori* and the species abundances (or presence/absences) are considered to be a response to this gradient.

Each site should occupy one row in the spreadsheet. The environmental variables should enter in the first columns, followed by the abundance data (the program will ask for the number of environmental variables).



The implementation in PAST follows the eigenanalysis algorithm given in Legendre & Legendre (1998). The ordinations are given as site scores - fitted site scores are presently not available. Environmental variables are plotted as correlations with site scores. Both scalings (type 1 and 2) of Legendre & Legendre (1998) are available. Scaling 2 emphasizes relationships between species.

Missing values are supported by column average substitution.

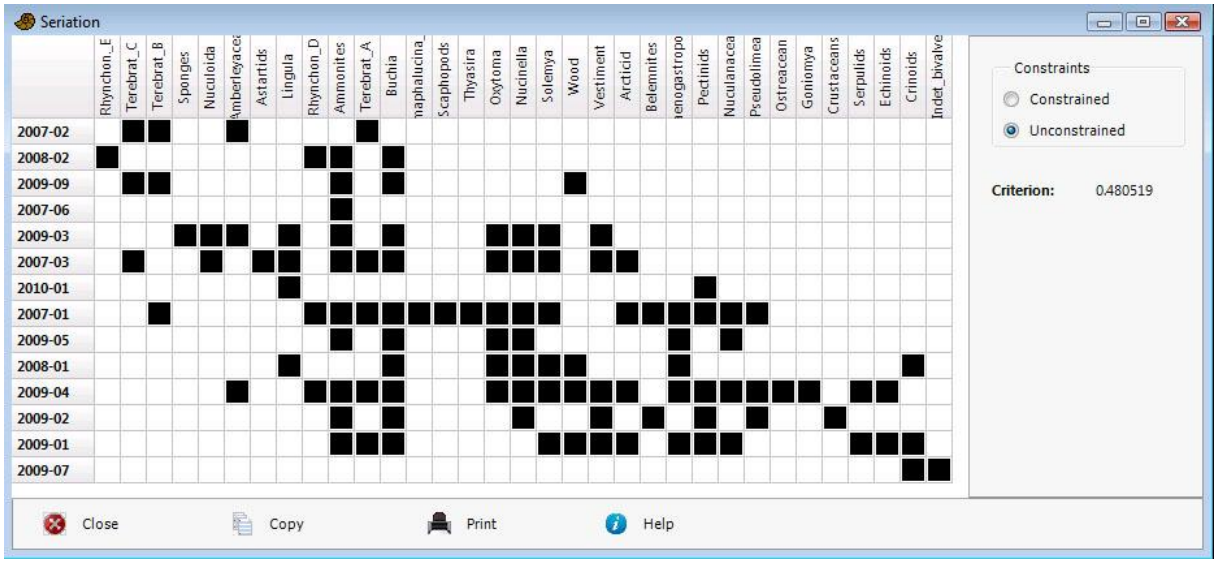
Mystery rows: Rows can contain missing values ('?') for all environmental variables. These rows, which must be placed at the bottom of the data matrix, are not included in the CCA analysis itself, but their site scores are estimated using the CCA vectors and included in the biplot. This could be used, for example, when environmental variables are only known for a modern data set but not for "fossil" (downcore) samples.

Reference

Legendre, P. & L. Legendre. 1998. Numerical Ecology, 2nd English ed. Elsevier, 853 pp.

Seriation

Seriation of an absence-presence (0/1) matrix using the algorithm described by Brower & Kile (1988). This method is typically applied to an association matrix with taxa (species) in the rows and samples in the columns. For constrained seriation (see below), columns should be ordered according to some criterion, normally stratigraphic level or position along a presumed faunal gradient.



The seriation routines attempt to reorganize the data matrix such that the presences are concentrated along the diagonal. There are two algorithms: Constrained and unconstrained optimization. In constrained optimization, only the rows (taxa) are free to move. Given an ordering of the columns, this procedure finds the 'optimal' ordering of rows, that is, the ordering of taxa which gives the prettiest range plot. Also, in the constrained mode, the program runs a 'Monte Carlo' simulation, generating and seriating 30 random matrices with the same number of occurences within each taxon, and compares these to the original matrix to see if it is more informative than a random one (this procedure is time-consuming for large data sets).

In the unconstrained mode, both rows and columns are free to move.

Missing data are treated as absences.

Reference

Brower, J.C. & K.M. Kile. 1988. Seriation of an original data matrix as applied to palaeoecology. *Lethaia* 21:79-93.

CABFAC factor analysis

This module implements the classical Imbrie & Kipp (1971) method of factor analysis and environmental regression (CABFAC and REGRESS, see also Klovan & Imbrie 1971).

The program asks whether the first column contains environmental data. If not, a simple factor analysis with Varimax rotation will be computed on row-normalized data.

If environmental data are included, the factors will be regressed onto the environmental variable using the second-order (parabolic) method of Imbrie & Kipp, with cross terms. PAST then reports the RMA regression of original environmental values against values reconstructed from the transfer function. Different methods for cross-validation (leave-one-out and *k*-fold) are available. You can also save the transfer function as a text file that can later be used for reconstruction of palaeoenvironment (see below). This file contains:

- Number of taxa
- Number of factors
- Factor scores for each taxon
- Number of regression coefficients
- Regression coefficients (second- and first-order terms, and intercept)

Missing values are supported by column average substitution.

References

Imbrie, J. & N.G. Kipp. 1971. A new micropaleontological method for quantitative paleoclimatology: Application to a late Pleistocene Caribbean core. In: The Late Cenozoic Glacial Ages, edited by K.K. Turekian, pp. 71-181, Yale Univ. Press, New Haven, CT.

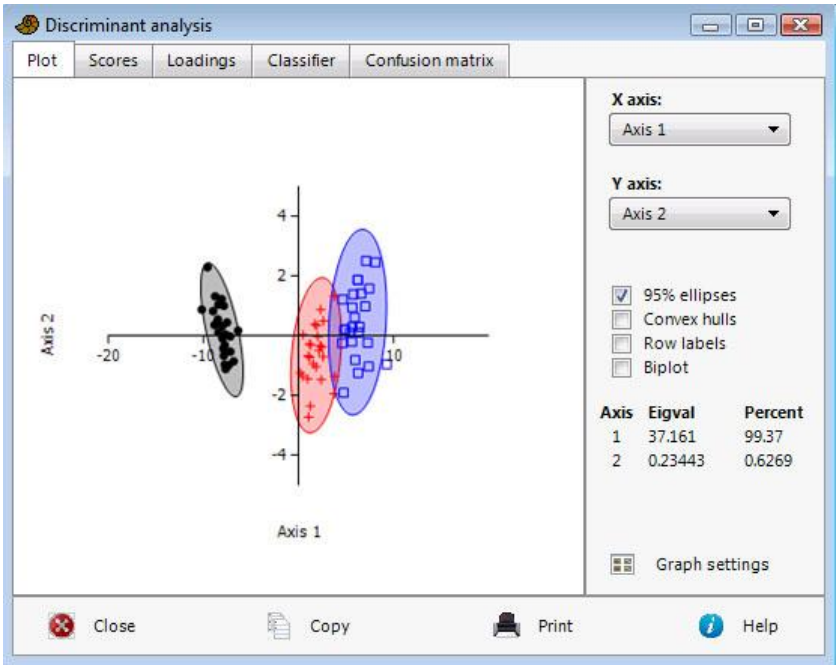
Klovan, J.E. & J. Imbrie. 1971. An algorithm and FORTRAN-IV program for large scale Q-mode factor analysis and calculation of factor scores. *Mathematical Geology* 3:61-77.

Discriminant analysis

This module provides discriminant analysis for two or more groups (the latter is sometimes called Canonical Variates Analysis). The groups must be specified with a group column.

A scatter plot of specimens along the first two canonical axes produces maximal and second to maximal separation between all groups. The axes are linear combinations of the original variables as in PCA, and eigenvalues indicate amount of variation explained by these axes. If only two groups are given, a histogram is plotted instead.

Missing data supported by column average substitution.



Classifier

Classifies the data, assigning each point to the group that gives minimal Mahalanobis distance to the group mean. The Mahalanobis distance is calculated from the pooled within-group covariance matrix, giving a linear discriminant classifier. The given and estimated group assignments are listed for each point. In addition, group assignment is cross-validated by a leave-one-out cross-validation (jackknifing) procedure.

Mystery specimens: Rows with unknown group, i.e. '?' in the group column, are not included in the discriminant analysis itself, but will be classified. In this way, it is possible to classify new specimens that are not part of the training set.

Confusion matrix

A table with the numbers of points in each given group (rows) that are assigned to the different groups (columns) by the classifier. Ideally each point should be assigned to its respective given group, giving a diagonal confusion matrix. Off-diagonal counts indicate the degree of failure of classification.

Computational details

Different softwares use different versions of CVA. The computations used by Past are given below.

Let **B** be the given data, with *n* items in rows and *k* variates in columns, centered on the grand means of columns (column averages subtracted). Let *g* be the number of groups, n_i the number of items in group *i*. Compute the *g*x*k* matrix **X** of weighted means of within group residuals, for group *i* and variate *j*

$$\mathbf{X}_{ij} = \sqrt{n_i}\,\overline{\mathbf{B}}_{ij},$$

where $\overline{\mathbf{B}}_{ij}$ is a column average within group *i*. Compute $\mathbf{B}_2$ from **B** by centering within groups. Now compute **W** and the normalized, pooled, within-group covariance matrix $\mathbf{W}_{\mathbf{cov}}$:

$$\mathbf{W} = \mathbf{B}_2'\mathbf{B}_2$$

$$\mathbf{W}_{\mathrm{cov}} = \frac{1}{n-g}\mathbf{W}.$$

e and **U** are the eigenvalues and eigenvectors of **W**; $\mathbf{e_c}$ and $\mathbf{U_c}$ are the eigenvalues and eigenvectors of $\mathbf{W}_{\mathbf{cov}}$. Then,

$$\mathbf{Z}'\mathbf{Z} = \mathrm{diag}(1/\mathbf{e})\mathbf{U}'\mathbf{X}'\mathbf{X}\mathbf{U}\,\mathrm{diag}(1/\mathbf{e}).$$

a and **A** are the eigenvalues and eigenvectors of **Z'Z**. We take only the first *g*-1 eigenvectors (columns of **A**), as the rest will be zero. The canonical variates are now

$$\mathbf{C} = \mathbf{U}\,\mathrm{diag}(1/\mathbf{e_c})\mathbf{A}.$$

The CVA scores are then **BC**. Reification of variables can be done along vectors $\mathbf{W}_{\mathbf{cov}}\mathbf{C}$.

Two-block PLS

Two-block Partial Least squares can be seen as an ordination method comparable with PCA, but with the objective of maximizing covariance between two sets of variates on the same rows (specimens, sites). For example, morphometric and environmental data collected on the same specimens can be ordinated in order to study covariation between the two.

The program will ask for the number of columns belonging to the first block. The remaining columns will be assigned to the second block. There are options for plotting PLS scores both within and across blocks, and PLS loadings.

The algorithm follows Rohlf & Corti (2000). Permutation tests and biplots are not yet implemented.

Partition the *nxp* data matrix **Y** into $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (the two blocks), with p_1 and p_2 columns. The correlation or covariance matrix **R** of **Y** can then be partitioned as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}.$$

The algorithm proceeds by singular value decomposition of the matrix $\mathbf{R}_{12}$ of correlations across blocks:

$$\mathbf{R}_{12} = \mathbf{F}_1 \mathbf{D} \mathbf{F}_2^{\mathbf{t}}.$$

The matrix **D** contains the singular values λ_i along the diagonal. $\mathbf{F}_1$ contains the loadings for block 1, and $\mathbf{F}_2$ the loadings for block 2 (cf. PCA).

The "Squared covar %" is a measure of the overall squared covariance between the two sets of variables, in percent relative to the maximum possible (all correlations equal to 1) (Rohlf & Corti p. 741). The "% covar" of axes are the amounts of covariance explained by each PLS axis, in percents of the total covariance. They are calculated as $100\frac{\lambda_i^2}{\sum \lambda_i^2}$.

Missing data supported by column average substitution.

Reference

Rohlf, F.J. & M. Corti. 2000. Use of two-block partial least squares to study covariation in shape. *Systematic Biology* 49:740-753.

Redundancy analysis (RDA)

Redundancy Analysis (Legendre & Legendre 1998) takes as input a site/data matrix where each site has given values for one or more environmental/explanatory variables as well as a number of response (dependent) variables. The ordination axes are linear combinations of the explanatory (independent) variables. RDA can be thought of as a canonical version of PCA, i.e. with axes constrained by explanatory variables.

Each site should occupy one row in the spreadsheet. The explanatory variables should enter in the first columns, followed by the response data (the program will ask for the number of explanatory variables).

The implementation in PAST follows Legendre & Legendre (1998). The ordinations can be shown as site scores or fitted site scores. Explanatory variables are plotted as correlations with site scores. Both scalings (type 1 and 2) of Legendre & Legendre (1998) are available. The scores can be manually scaled with the "Amplitude" controls for a clearer plot (these factors should be reported together with the plot).

Missing values are supported by column average substitution.

Mystery rows: Rows can contain missing values ('?') for all explanatory variables. These rows, which must be placed at the bottom of the data matrix, are not included in the RDA analysis itself, but their site scores are estimated using the RDA vectors and included in the biplot. This could be used, for example, when explanatory variables are only known for a modern data set but not for "fossil" (downcore) samples. Mystery rows are only reported for unfitted site scores.

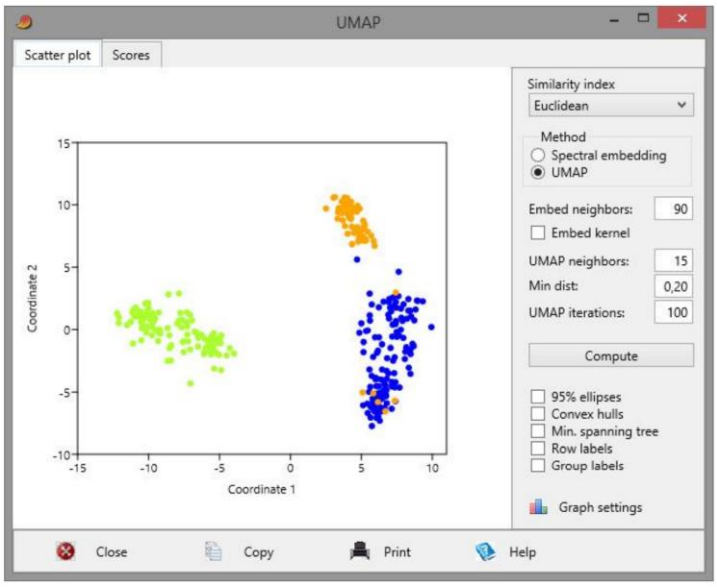
Reference

Legendre, P. & L. Legendre. 1998. Numerical Ecology, 2nd English ed. Elsevier, 853 pp

Nonlinear ordination (UMAP)

PAST includes three methods for nonlinear ordination (embedding): spectral embedding, UMAP, and ISOMAP. The first two are found in the UMAP module, the last in the ISOMAP module. These methods can be compared with e.g. Principal Coordinates Analysis, and are based on a distance measure. However, they can pick up groups having complicated shapes in high-dimensional space, possibly interfingering with other groups. These methods can be very good at identifying groups and gradients (often "too good", causing overfitting), but can also behave quite erratically, depending on analysis parameters.

The input is a multivariate data set with variables in columns, optionally with a group column for showing given *groups* (this does not influence the ordination). A distance measure must be chosen (default Euclidean).



Spectral embedding

The spectral embedding algorithm basically follows Belkin & Niyogi (2003). First, a graph is constructed where two points *i* and *j* are connected if and only if *i* is among the *n* nearest neighbours of *j*, or *j* is among the *n* nearest neighbours of *i*. The number *n* can be chosen by the user ("Embed neighbors"), and it can substantially influence the ordination. Larger *n* will tend to produce larger, less concentrated groups.

A matrix **W** is constructed, with W_{ij}=1 if *i* and *j* are connected (adjacency matrix). If the "Embed kernel" option is selected, the values in this matrix are additionally scaled by the kernel function

$$W_{ij} = e^{\frac{\|x_i - x_j\|^2}{t}}$$

where the norm is according to the selected distance measure. The parameter *t* is fixed at 1/50 of the largest distance between any pairs of points. Then compute the "graph Laplacian" **L** = **D** – **W**, where **D** is a diagonal matrix containing the row (or column, as **W** is symmetric) sums of **W**. Past uses the normalized graph Laplacian,

$$\mathbf{L}' = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$$

The scores on the two ordination axes are given by the two eigenvectors corresponding to the second smallest and third smallest eigenvalue.

UMAP

UMAP (McInnes et al. 2018) is a modern method for nonlinear ordination. The implementation in Past also builds on a description by Oskolkov (2019). The procedure is complex, for details see the references.

The implementation in Past uses a spectral embedding as starting condition, with parameters as given above. The "UMAP neighbors" (*k*) and the "Minimum distance" (min_dist) parameters can change the ordination quite a lot.

References

Belkin, M., Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15:1373-1396.

McInnes, L., Healy, J., Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. https://arxiv.org/abs/1802.03426.

Oskolkov, N. 2019. How to program UMAP from scratch. https://towardsdatascience.com/how-to-program-umap-from-scratch-e6eff67f55fe

Nonlinear ordination (ISOMAP)

ISOMAP (Tenenbaum et al. 2000) is a relatively simple but often effective method for nonlinear dimensionality reduction. The implementation in Past follows these steps:

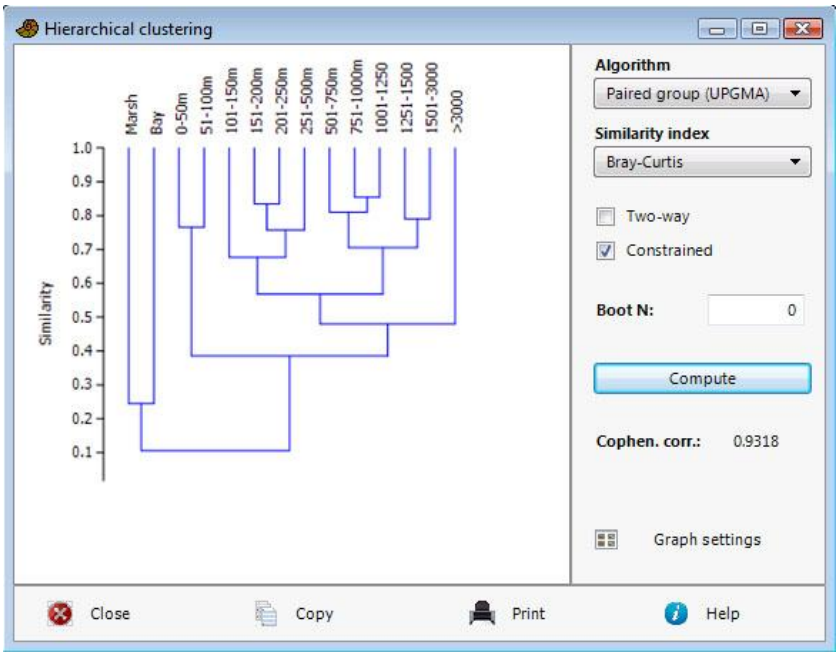
1. Find the *k* closest neighbors of each point, using any distance measure. The result will strongly depend on the value of *k*, which must be specified by the user.
2. Produce a graph where two vertices are connected if they are among their closest neighbors.
3. Find the shortest paths between all vertices in the graph using the Floyd-Warshall algorithm. These path lengths represent "geodesic" distances along the manifold.
4. Compute a standard PCoA (metric multidimensional scaling) on the geodesic distances.

Reference

Tenenbaum, J.B., de Silva, V., Langford, J.C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319-2323.

Classical cluster analysis

The hierarchical clustering routine produces a 'dendrogram' showing how data points (rows) can be clustered. For 'R' mode clustering, putting weight on groupings of taxa, taxa should go in rows. It is also possible to find groupings of variables or associations (Q mode), by entering taxa in columns. Switching between the two is done by transposing the matrix (in the Edit menu).



Three different algorithms are available:

- Unweighted pair-group average (UPGMA). Clusters are joined based on the average distance between all members in the two groups.
- Single linkage (nearest neighbor). Clusters are joined based on the smallest distance between the two groups.
- Ward's method. Clusters are joined such that increase in within-group variance is minimized.
- Complete linkage. Clusters are joined based on the largest distance within groups.

One method is not necessarily better than the other, though single linkage is not recommended by some. It can be useful to compare the dendrograms given by the different algorithms, to informally assess the robustness of the clusters.

For Ward's method, a Euclidean distance measure is inherent to the algorithm. For the other methods, the distance matrix can be computed using 24 different indices, as described under the 'Similarity and distance indices' section.

Missing data: The cluster analysis algorithm can handle missing data, coded with question marks (?). This is done using pairwise deletion, meaning that when distance is calculated between two points, any variables that are missing are ignored in the calculation. For Raup-Crick, missing values are treated as absence. Missing data are not supported for Ward's method, nor for the Rho similarity measure.

Two-way clustering: The two-way option allows simultaneous clustering in R-mode and Q-mode.

Stratigraphically constrained clustering: This option will allow only adjacent rows or groups of rows to be joined during the agglomerative clustering procedure. May produce strange-looking (but correct) dendrograms.

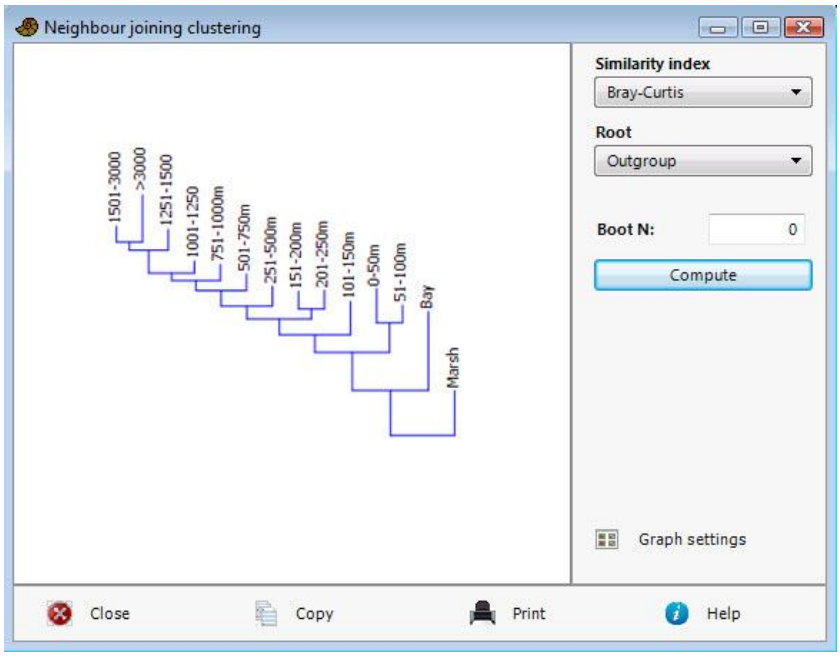
Group-constrained clustering: This option will only allow joining of clusters within the given groups. May produce strange-looking (but correct) dendrograms.

Bootstrapping: If a number of bootstrap replicates is given (e.g. 100), the columns are subjected to resampling. Press Enter after typing to update the value in the "Boot N" number box. The percentage of replicates where each node is still supported is given on the dendrogram.

Note on Ward's method: PAST produces Ward's dendrograms identical to those made by Stata, but somewhat different from those produced by Statistica. The reason for the discrepancy is unknown. Constrained clustering does not work.

Neighbour joining

Neigbour joining clustering (Saitou & Nei 1987) is an alternative method for hierarchical cluster analysis. The method was originally developed for phylogenetic analysis, but may be superior to UPGMA also for ecological data. In contrast with UPGMA, two branches from the same internal node do not need to have equal branch lengths. A phylogram (unrooted dendrogram with proportional branch lengths) is given.



Distance indices and bootstrapping are as for other cluster analysis (above). To run the bootstrap analysis, type in the number of required bootstratp replicates (e.g. 1000, 10000) in the "Boot N" box and press Enter to update the value.

Negative branch lengths are forced to zero, and transferred to the adjacent branch according to Kuhner & Felsenstein (1994).

The tree is by default rooted on the last branch added during tree construction (this is not midpoint rooting). Optionally, the tree can be rooted on any row in the data matrix, as selected in the Root menu.

Missing data supported by pairwise deletion.

References

Kuhner, M.K. & J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11:459-468

Saitou, N. & M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

K-means clustering

K-means clustering (e.g. Bow 1984) is a non-hierarchical clustering method. The number of clusters to use is specified by the user, usually according to some hypothesis such as there being two sexes, four geographical regions or three species in the data set.

The cluster assignments are initially random, using the Forgy method. In an iterative procedure, items are then moved to the cluster which has the closest cluster mean, and the cluster means are updated accordingly. This continues until items are no longer "jumping" to other clusters. The result of the clustering is to some extent dependent upon the initial, random ordering, and cluster assignments may therefore differ from run to run. This is not a bug, but normal behaviour in k-means clustering.

The cluster assignments may be copied and pasted back into the main spreadsheet, and corresponding colors (symbols) assigned to the items using the 'Numbers to colors' option in the Edit menu.

Clustering statistics

WGSS: The Within-Group (or Within-Cluster) sum of squares; will decrease with cluster separationand with the number of clusters.

F: Ratio of Between-Group to Within-Group sum of squares; will increase with cluster separation and number of clusters.

Variance explained (percent): Ratio of Between-Group to total sum of squares, multiplied with 100.

Average silhouette: The average silhouette value over all objects (see below).

Silhouette plot and table

The silhouette plot (Rousseeuw 1987) gives an indication of how well each object has been classified, on a scale from -1 to 1, where 1 means a perfectly appropriate assignment to a cluster; -1 means the object would have been better placed in another cluster; 0 means the object is on the boundary between two clusters.

Missing data supported by column average substitution.

References

Bow, S.-T. 1984. Pattern recognition. Marcel Dekker, New York.

Rousseeuw, P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Computational and Applied Mathematics* 20:53–65.

K-medoids clustering

K-medoids clustering (Kaufman & Rousseeuw 1990) can be compared to k-means clustering, and requires the user to select the number of clusters. Unlike k-means, the clusters are centered on a point in the data set, rather than a cluster mean. Also, importantly, k-medoids allows any distance measure to be used, making it useful for e.g. ecological and genetic data.

The algorithm in Past follows the original PAM method described by Kaufman & Rousseeuw (1990).

Reference

Kaufman, L. & Rousseeuw, P.J. 1990. Partitioning around medoids (program PAM). Ch. 2 in *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

DBSCAN clustering

DBSCAN clustering (Ester et al. 1996) is a non-hierarchical clustering method. Unlike K-means or K-medoids, the number of clusters does not need to be pre-defined. Any distance measure can be used. DBSCAN is a popular clustering method in the machine learning community. It does not always perform better than K-medoids but can trace out non-linear cluster boundaries.

The "MinPts" parameter is the minimum number of points required to form a dense region. By default it is set to the number of data dimensions dim*2, or N/4, whichever is smaller. It may be increased for large or noisy data sets.

The "Epsilon" parameter is the neighborhood radius. By default, it is set to 0.1 times the maximal pairwise distance in the data set. The value can be optimized using the K-distance graph.

In the output table, cluster assignment is indicated by a positive integer. The value -1 indicates a "noise point" that is not be assigned to any group.

K-distance graph: This graph shows the distance from each point to its MinPts-1 nearest neighbor, sorted from smallest to largest. You should look for a break in the slope of this curve. The corresponding "K-distance" may be used as a value for the parameter Epsilon.

Reference

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.

K-nearest neighbors classifier

K-nearest neighbours (Fix & Hodges 1951) is a simple but effective supervised machine learning algorithm. A multivariate data set with a group column is supplied, where one part (the training set) is assigned to groups, while the other part (the test set) has question marks in the group column. The rows in the test set are then assigned to groups (classification) according to their similarity with rows in the training set. Any distance measure can be used.

For each test row, the group is selected by a majority vote among the nearest *k* neighbors. The value of *k* is a parameter selected by the user. Small *k* may lead to overfitting, while large *k* may lead to oversmoothing. In addition, each vote among the nearest neighbors may be weighted according to distance *d* from the test point. Four weighting schemes are available in Past:

- No weighting. This is usually a poor choice (Gou et al. 2012) but is included for comparison.
- Inverse weighting, $w = 1/d$. This is the default weighting in some programs, but it is usually not optimal.
- Linear weighting (Dudani 1976) is a standard weighting that usually performs well:

$$w = \frac{d_{max} - d}{d_{max} - d_{min}}$$

- Weighting due to Gou et al. (2012), a modification of linear weighting that can perform slightly better:

$$w = \frac{d_{max} - d}{d_{max} - d_{min}} \times \frac{d_{max} + d_{min}}{d_{max} + d}$$

Jackknifing

Past automatically performs a jackknifing (cross-validation) procedure on the training set. One row is removed from the training set at a time and classified using the remaining training data. The percentage of correctly classified rows is reported. This may be used to indicate good choices for *k* and weighting.

References

Dudani, S.A. 1976. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics* 6:325-327.

Fix, E. Hodges, J.L. 1951. Discriminatory analysis, nonparametric discrimination: consistency properties. *Technical Report No. 4, USAF School of Aviation Medicine, Randolf Field Texas*, 238–247.

Gou, J., Du, L., Zhang, Y., Xiong, T. 2012. A new distance-weighted k-nearest neighbor classifier. *Journal of Information & Computational Science* 9:1429–1436

Naïve Bayes classifier

Naïve Bayes (Hand & Yu 2001) is a simple supervised machine learning algorithm, like k-nearest neighbors. A multivariate data set with a group column should be supplied, where one part (the training set) is assigned to groups, while the other part (the test set) has question marks in the group column. The rows in the test set are then assigned to groups (classification) according to their similarity with rows in the training set.

The *Gaussian model* is used for continuous data. For this model, all within-group variances for all variables must be non-zero. Therefore, a value is added to all zero-variances. This value is set to 0.1 times the variance of the given variable across all rows in the test and training sets. Classification is usually not very good when this rule is applied.

The *multinomial model* is used for count data. Laplacian smoothing is used.

The *Bernoulli* model is used for binary data, with smoothing.

The multinomial and Bernoulli models are implemented according to Manning et al. (2008).

The "flat prior" option assigns the same value to the prior for all group assignments, instead of being proportional to sample size.

Jackknifing cross-validation is implemented as for K-nearest neighbors.

Missing values supported by column average substitution.

References

Hand, D.J., Yu, K. 2001. Idiot's Bayes – not so stupid after all? *International Statistical Review* 69:385–398.

Manning, C.D., Raghavan, P., Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Decision tree classification

A multivariate data set with a group column should be supplied, where one part (the training set) is assigned to groups, while the other part (the test set) has question marks in the group column. The rows in the test set are then assigned to groups (i.e., classified) according to their similarity with rows in the training set.

The algorithm is similar to CART algorithm (Breiman et al. 1984) but with some modifications.

Initial tree and pre-pruning

Tree construction starts with a root node Q_m containing all samples (rows) in the training set. All possible splitting variables and splitting values are traversed, searching for minimal impurity, defined as

$$G(Q_m) = \frac{n_m^{left}}{n_m} H\left(Q_m^{left}\right) + \frac{n_m^{right}}{n_m} H\left(Q_m^{right}\right)$$

where n_m is the number of samples in a node, while *left* and *right* refer to the two subnodes in the tree that would result from the given splitting variable and value. We use *Gini impurity* for each subnode:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

where p_{mk} is the proportion of samples of class k in the node. After the optimal splitting rule has been found in this way, the sample set is assigned to the new subnodes accordingly, and the procedure is repeated recursively for each subnode.

The recursion stops when there is a single sample in the node, or all samples in the node are assigned to the same class. These termination criteria lead to a fully resolved tree, where each sample is assigned to a class unambiguously. This can lead to overfitting, giving poorer quality classification of new samples. The user can therefore set a stricter stopping rule ("pre-pruning") by choosing a value N>=2 for the minimum number of rows In a node to allow it to split, and/or the maximal depth of the tree (the large default value of 10 will usually give a fully resolved tree).

Post-pruning

The initial tree computed above can be too detailed, leading to overfitting in the classification. The tree can therefore be pruned further, using the "minimal cost-complexity" procedure of Breiman et al. (1984), also formulated by Mingers (1989). For an internal node t, considered for pruning, we calculate the error cost

$$R(t) = r(t)p(t)$$

where r is the proportion of samples in the node that will be incorrectly classified if the node is pruned (the error rate), and p is the total proportion of samples in the node relative to the complete training set. Also compute the error cost for the subtree if the node is not pruned, as

$$R(T_t) = \sum_i R(i)$$

where i are the leaves of the subtree T_t rooted on t. Finally, compute the complexity parameter

$$\alpha = \frac{R(t) - R(T_t)}{N_T - 1}$$

where N_T is the number of leaves in the subtree. The node in the tree with the smallest α is selected for pruning, and the procedure continues until α is greater than the "Pruning alpha" value. Setting this value to zero will disable the post-pruning entirely.

Classification

Classification of a sample follows the decision rules of the tree down to a leaf node. The class is then decided by a majority vote among the class memberships of the training samples in the node. The "probability" value (a slight misnomer) is the proportion of samples of the majority class in the node.

The complete decision tree is shown in the "Rules" tab. Each node in the tree has two subnodes; the first subnode is selected if the splitting rule is true, otherwise the second subnode is selected. You can click on a node to hide its subtree.

Cross validation

The decision tree process is automatically cross-validated by the leave-one-out (jackknifing) method.

Missing values are treated as follows. During *training*, for each internal tree node, the optimal splitting variable and its splitting point are chosen using only non-missing values. Then, the test set samples in the node are assigned to the left or right subnode by this splitting rule only if they have the appropriate non-missing values. The remaining samples in the node (those with missing values for the splitting variable) are assigned to the subnode that acquired the most samples in the first step. During *classification*, the decision tree is traversed as usual, except that when a decision cannot be made because of a missing value, the subnode with the larger number of samples is chosen.

References

Breiman, L., Friedman, J., Olshen, R., Stone, C. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Mingers, J. 1989. An empirical comparison of pruning methods for decision tree induction. *Machine Learning* 4:227–243.

Random forest classifier

Random forest classification (Breiman 2001) in Past uses the same algorithm as decision trees except there is no pre- or post-pruning – each tree is grown to its tips. To decrease sensitivity to noise in the training set, a large number of randomized trees are generated, and classification is based on a majority vote among these trees. There are two levels of randomization:

1. The rows in the training set are sampled randomly with replacement (bootstrapped). This is also known as *bagging*.
2. During training, for every node, not all of the N variables are considered when searching for the optimal splitting rule. Instead, a number $M < N$ of variables is selected at random. Setting $M = N$ will disable this step. Typically, M is set to $N/2$ or smaller - $\sqrt{N}$ is sometimes suggested.

Reference

Breiman, L. 2001. Random forests. *Machine Learning* 45:5–32.

Parsimony (cladistic)

Warning: the cladistics package in PAST is fully operational, but lacking in comprehensive functionality. The PAST cladistics package is adequate for education and initial data exploration, but for more 'serious' work we recommend a specialized program for systematics (Paup, TNT etc.).

The module requires a character matrix with taxa in rows, outgroup in first row.

Algorithms are from Kitching et al. (1998) and Felsenstein (2004).

Character states should be coded using integers in the range 0 to 255, or with the letters c, a, g, t, u (upper or lower case). The first taxon is treated as the outgroup, and will be placed at the root of the tree.

Missing values are coded with a question mark (?). Please note that PAST does not collapse zero-length branches. Because of this, missing values can lead to a proliferation of equally shortest trees *ad nauseam*, many of which are in fact equivalent.

Algorithms

There are four algorithms available for finding short trees:

Exhaustive

The exhaustive algorithm evaluates all possible trees. Like the branch-and-bound algorithm it will necessarily find all shortest trees, but it is very slow. The maximum number of taxa allowed for exhaustive search is 12, for which more than 600 million trees are evaluated. The only advantage over branch-and-bound is the plotting of tree length distribution. This histogram may indicate the 'quality' of your matrix, in the sense that there should be a tail to the left such that few short trees are 'isolated' from the greater mass of longer trees (but see Kitching et al. 1998 for critical comments on this).

Branch-and-bound

The branch-and-bound algorithm is guaranteed to find all shortest trees. The total number of shortest trees is reported, but a maximum of 10000 trees are saved. The branch-and-bound algorithm can be very time consuming for data sets with more than 16 taxa.

Heuristic, nearest neighbour interchange

This heuristic algorithm adds taxa sequentially in the order they are given in the matrix, to the branch where they will give least increase in tree length. After each taxon is added, all nearest neighbour trees are swapped to try to find an even shorter tree.

Like all heuristic searches, this one is much faster than the algorithms above and can be used for large numbers of taxa, but is not guaranteed to find all or any of the most parsimonious trees. To decrease the likelihood of ending up on a suboptimal local minimum, a number of reorderings can be specified. For each reordering, the order of input taxa will be randomly permuted and another heuristic search attempted.

Note: Because of the random reordering, the trees found by the heuristic searches will normally be different each time. To reproduce a search exactly, you will have to start the parsimony module again from the menu, using the same value for "Random seed". This will reset the random number generator to the seed value.

Heuristic, subtree pruning and regrafting

This algorithm (SPR) is similar to the one above (NNI), but with a more elaborate branch swapping scheme: A subtree is cut off the tree, and regrafting onto all other branches in the tree is attempted in order to find a shorter tree. This is done after each taxon has been added, and for all possible subtrees. While slower than NNI, SPR will often find shorter trees.

Heuristic, tree bisection and reconnection

This algorithm (TBR) is similar to the one above (SPR), but with an even more complete branch swapping scheme. The tree is divided into two parts, and these are reconnected through every possible pair of branches in order to find a shorter tree. This is done after each taxon is added, and for all possible divisions of the tree. TBR will often find shorter trees than SPR and NNI, at the cost of longer computation time.

Character types

Unordered

Characters are reversible and unordered, meaning that all changes have equal cost. This is the criterion with fewest assumptions and is therefore generally preferable.

Ordered

Characters are reversible and ordered, meaning that 0->2 costs more than 0->1, but has the same cost as 2->0.

Bootstrap

Bootstrapping is performed when the 'Bootstrap replicates' value is set to non-zero. The specified number of replicates (typically 100 or even 1000) of your character matrix are made, each with randomly weighted characters. The bootstrap value for a group is the percentage of replicates supporting that group. A replicate supports the group if the group exists in the majority rule consensus tree of the shortest trees made from the replicate.

Warning: Specifying 1000 bootstrap replicates will clearly give a thousand times longer computation time than no bootstrap! Exhaustive search with bootstrapping is unrealistic and is not allowed.

Cladogram plotting

All shortest (most parsimonious) trees can be viewed, up to a maximum of 10000 trees. If bootstrapping has been performed, a bootstrap value is given at the root of the subtree specifying each group.

Character states can also be plotted onto the tree, as selected by the 'Character' menu. This character reconstruction is unique only in the absence of homoplasy. In case of homoplasy, character changes are placed as close to the root as possible, favouring one-time acquisition and later reversal of a character state over several independent gains (so-called *accelerated transformation*).

The 'Phylogram' option allows plotting of trees where the length of vertical lines (joining clades) is proportional to branch length.

Consistency index

The per-character consistency index (ci) is defined as m/s, where m is the minimum possible number of character changes (steps) on any tree, and s is the actual number of steps on the current tree. This index hence varies from one (no homoplasy) and down towards zero (a lot of homoplasy). The ensemble consistency index CI is a similar index summed over all characters.

Retention index

The per-character retention index (ri) is defined as $(g-s)/(g-m)$, where m and s are as for the consistency index, while g is the maximal number of steps for the character on any cladogram (Farris 1989). The retention index measures the synapomorphy on the tree and varies from 0 to 1.

Note that in the present version, the retention index is only correctly calculated when using unordered characters.

Consensus tree

The consensus tree of all shortest (most parsimonious) trees can also be viewed. Two consensus rules are implemented: Strict (groups must be supported by all trees) and majority (groups must be supported by more than 50% of the trees).

Bremer support (decay index)

The Bremer support for a clade is the number of extra steps you need to construct a tree (consistent with the characters) where that clade is no longer present. There are reasons to prefer this index rather than the bootstrap value. PAST does not compute Bremer supports directly, but for smaller data sets it can be done 'manually' as follows:

- Perform parsimony analysis with exhaustive search or branch-and-bound. Take note of the clades and the length N of the shortest tree(s) (for example 42). If there are more than one shortest tree, look at the strict consensus tree. Clades which are no longer found in the consensus tree have a Bremer support value of 0.
- In the box for 'Longest tree kept', enter the number N+1 (43 in our example) and perform a new search.
- Additional clades which are no longer found in the strict consensus tree have a Bremer support value of 1.
- For 'Longest tree kept', enter the number N+2 (44) and perform a new search. Clades which now disappear in the consensus tree have a Bremer support value of 2.
- Continue until all clades have disappeared.

References

Farris, J.S. 1989. The retention index and the rescaled consistency index. *Cladistics* 5:417-419.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates.

Kitching, I.J., P.L. Forey, C.J. Humphries & D.M. Williams. 1998. Cladistics. Oxford University Press.

Phylogenetically independent contrasts

Phylogenetically independent contrasts (Felsenstein 1985; Garland et al. 1992) is a method for removing phylogenetic information from a character dataset, e.g. morphometric data. The purpose of this is to reduce the phylogenetic interdependence of data points, which can violate the assumptions of many statistical tests.

The module requires one or more columns of character data. In addition, the "tree collection" in Past must contain one or more trees. The taxon names in the trees must match the row names in the spreadsheet.

References

Felsenstein, J. 1985. Phylogenies and the comparative method. *American Naturalist* 125:1-15.

Garland Jr., T., Harvey, P.H., Ives, A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology* 41:18-32.

Phylogenetic generalized least squares (PGLS)

Bivariate linear regression (for two data columns), taking into account the expected phylogenetic covariation from a Brownian model for evolution (Grafen 1989; Pagel 1999; Symonds & Blomberg 2014).

The "tree collection" in Past must contain one or more trees with branch lengths (otherwise, branch lengths of 1 are assumed). The taxon names in the trees must match the row names in the spreadsheet. IMPORTANT: The first row in the data is not used, as it is assumed to refer to an outgroup. This (dummy) outgroup must be included in the tree – the branch length below the outgroup is not used. This is to ensure compatibility with the parsimony trees generated by Past, which are always rooted on the first row.

Standard errors on regression parameters are calculated according to Smaers & Rohlf (2016).

The PGLS var-covar matrix is transformed according to the parameter lambda (Symonds & Blomberg 2014), which can be selected as 0 (equivalent to OLS regression if the tree is ultrametric; a weighted OLS otherwise), or 1 (full PGLS) or the maximum-likelihood estimate based on regression residuals (Revell 2010).

The maximum-likelihood estimates for lambda (and their optimization profiles) are calculated for the first data column only (x), the second data column only (y), a combination of x and y (Freckleton et al. 2002) or the regression residuals (Revell 2010).

References

Freckleton, R.P., Harvey, P.H., Pagel, M. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. *American Naturalist* 160:712-726.

Grafen, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society B* 326:119–157.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

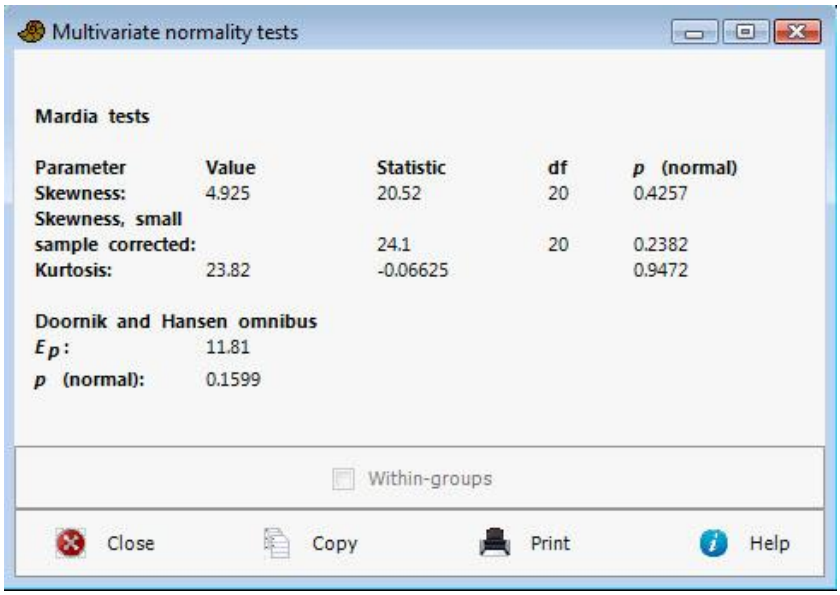
Revell, L.J. 2010. Phylogenetic signal and linear regression on species data. *Methods in Ecology & Evolution* 1:319–329.

Smaers, J.B., Rohlf, F.J. 2016. Testing species' deviation from allometric predictions using the phylogenetic regression. *Evolution* 70:1145–1149.

Symonds, M.R.E., Blomberg, S.P. 2014. A primer on phylogenetic least squares. Pp. 105-130 in: L. Z. Garamszegi (ed.), Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology. Springer-Verlag.

Multivariate normality

Multivariate normality is assumed by a number of multivariate tests. PAST computes Mardia's multivariate skewness and kurtosis, with tests based on chi-squared (skewness) and normal (kurtosis) distributions. A powerful omnibus (overall) test due to Doornik & Hansen (1994) is also given. If at least one of these tests show departure from normality (small *p* value), the distribution is significantly non-normal. Sample size should be reasonably large (>50), although a small-sample correction is also attempted for the skewness test.



Missing data supported by column average substitution.

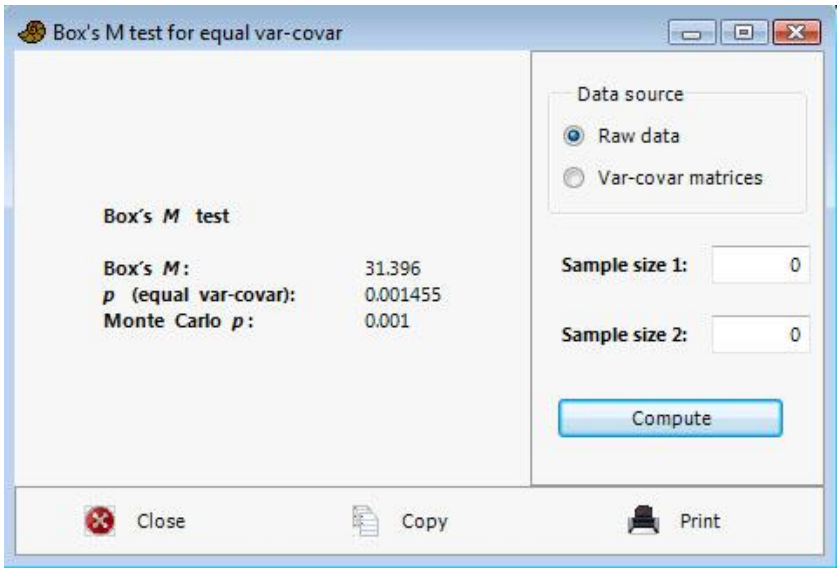
References

Doornik, J.A. & H. Hansen. 1994. An omnibus test for univariate and multivariate normality. W4&91 in Nuffield Economics Working Papers.

Mardia, K.V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 36:519-530.

Box's *M*

Test for the equivalence of the covariance matrices for two multivariate samples marked with a group column. This is a test for homoscedasticity, as assumed by MANOVA. You can use either two original multivariate samples from which the covariance matrices are automatically computed, or two specified variance-covariance matrices. In the latter case, you must also specify the sizes (number of individuals) of the two samples.



The Box's *M* statistic is given together with a significance value based on a chi-square approximation. Note that this test is supposedly very sensitive. This means that a high *p* value will be a good, although informal, indicator of equality, while a highly significant result (low *p* value) may in practical terms be a somewhat too sensitive indicator of inequality.

The statistic is computed as follows – note this equals the "-2 ln *M*" of some texts (Rencher 2002).

$$M = (n-2)\ln|\mathbf{S}| - (n_1 - 1)\ln|\mathbf{S_1}| - (n_2 - 1)\ln|\mathbf{S_2}|,$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are the covariance matrices, **S** is the pooled covariance matrix, $n=n_1+n_2$ and $|\bullet|$ denotes the determinant.

The Monte Carlo test is based on 999 random permutations.

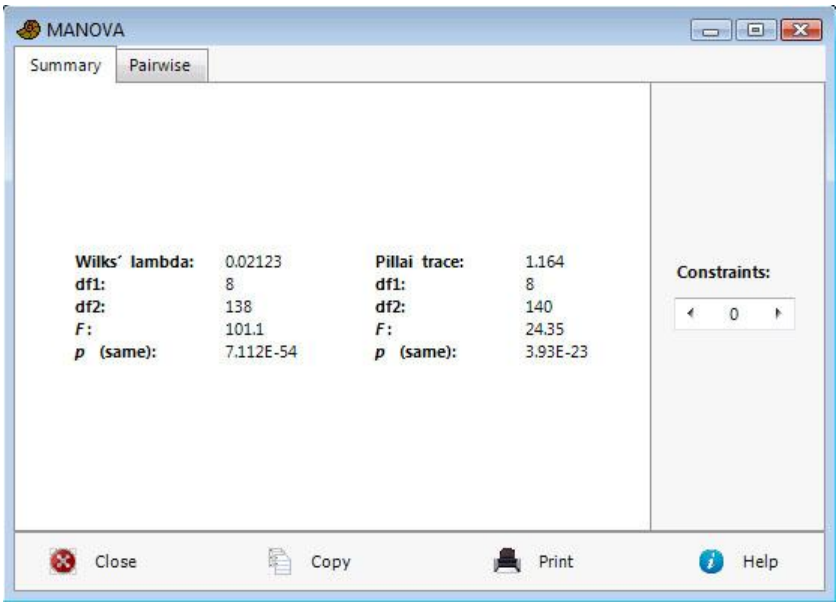
Missing data supported by column average substitution.

Reference

Rencher, A.C. 2002. Methods of multivariate analysis, 2nd ed. Wiley.

MANOVA

One-way MANOVA (Multivariate ANalysis Of VAriance) is the multivariate version of the univariate ANOVA, testing whether two or more groups (specified with a group column) have the same multivariate mean.



Two statistics are provided: Wilk's lambda with its associated Rao's *F* and the Pillai trace with it's approximated *F*. Wilk's lambda is probably more commonly used, but the Pillai trace may be more robust.

Number of constraints: For correct calculation of the *p* values, the number of dependent variables (constraints) must be specified. It should normally be left at 0, but for Procrustes fitted landmark data use 4 (for 2D) or 6 (for 3D).

Pairwise comparisons (post-hoc): If the MANOVA shows significant overall difference between groups, the analysis can proceed by pairwise comparisons. In PAST, the post-hoc analysis is simple, by pairwise Hotelling's tests. The following values can be displayed in the table:

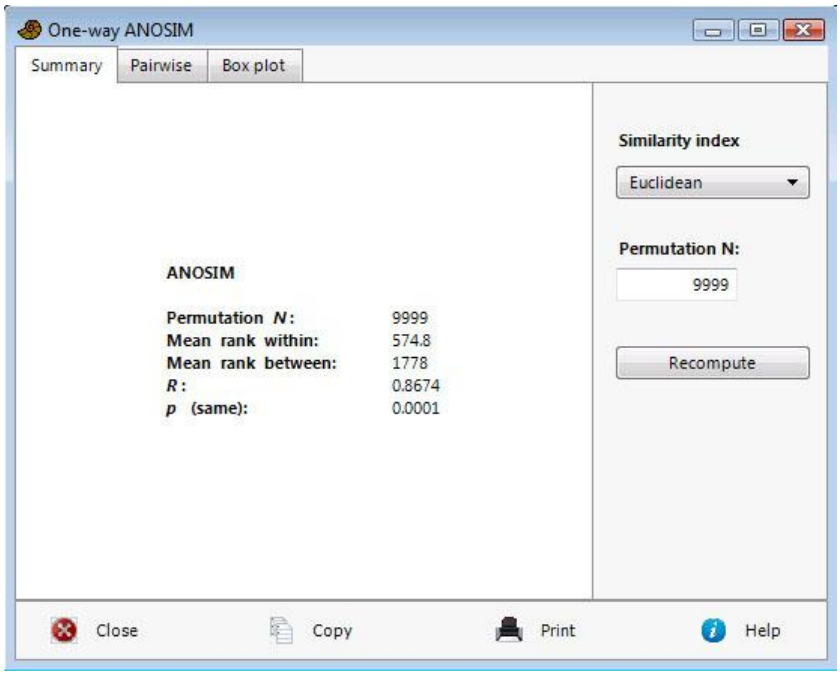
- Hotelling's *p* values, not corrected for multiple testing. Marked in pink if significant ($p<0.05$).
- The same *p* values, but significance (pink) assessed using the sequential Bonferroni scheme.
- Bonferroni corrected *p* values (multiplied by the number of pairwise comparisons). The Bonferroni correction gives very little power.
- Squared Mahalanobis distances.

Note: These pairwise comparisons use the within-group covariance matrix pooled over all groups participating in the MANOVA. They may therefore give slightly other results than if only two of the groups are selected for analysis.

Missing data supported by column average substitution.

One-way ANOSIM

ANOSIM (ANalysis Of Similarities) is a non-parametric test of significant difference between two or more groups, based on any distance measure (Clarke 1993). The distances are converted to ranks. ANOSIM is normally used for taxa-in-samples data, where groups of samples are to be compared. Items go in rows, variates in columns, and groups should be specified with a group column as usual.



In a rough analogy with ANOVA, the test is based on comparing distances between groups with distances within groups. Let r_b be the mean rank of all distances between groups, and r_w the mean rank of all distances within groups. The test statistic R is then defined as

$$R = \frac{r_b - r_w}{N(N-1)/4}.$$

Large positive R (up to 1) signifies dissimilarity between groups. The one-tailed significance is computed by permutation of group membership, with 9,999 replicates (can be changed).

Pairwise ANOSIMs between all pairs of groups are provided as a post-hoc test. Significant comparisons (at $p<0.05$) are shown in pink. The optional Bonferroni correction multiplies the p values with the number of comparisons. This correction is very conservative (produces large p values). The sequential Bonferroni option does not output corrected p values, but significance is decided based on step-down sequential Bonferroni, which is slightly more powerful than simple Bonferroni.

Missing data supported by pairwise deletion (not for the Raup-Crick, Rho and user-defined indices).

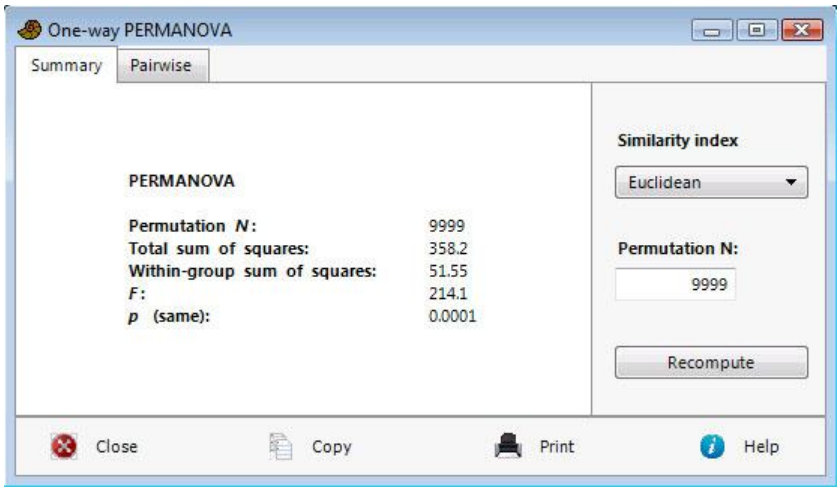
Reference

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18:117-143.

One-way PERMANOVA

PerMANOVA (Permutational MANOVA, also known as NPMANOVA) is a non-parametric test of significant difference between two or more groups, based on any distance measure (Anderson 2001). PerMANOVA is normally used for ecological taxa-in-samples data, where groups of samples are to be compared, but may also be used as a general non-parametric MANOVA.

Items go in rows, variates in columns, and groups should be specified with a group column.



PerMANOVA calculates an *F* value in analogy with ANOVA. In fact, for univariate data sets and the Euclidean distance measure, PerMANOVA is equivalent to ANOVA and gives the same *F* value.

The significance is computed by permutation of group membership, with 9,999 replicates (can be changed by the user).

Repeated measures

A repeated measures (blocked) ANOVA is carried out with the "Repeated measures" box is ticked. In this case, each group must have the same number of rows, and all the rows in each group must be consecutive. The first row in each group is then assumed to belong to the first subject (block), the second row in each group belongs to the second subject, etc. The calculations are analogous to the univariate repeated measures ANOVA. First, the within-subjects sum of squares SS_{ws} is calculated in the same way as the within-group sum of squares above, but with terms taken only when I and j are within the same subject. The between-subjects sum of squares is then $SS_{bs} = SS_T - SS_{ws}$. Finally, the error sum of squares $SS_{err} = SS_{wg} - SS_{bs}$. The *F* value is calculated as described for the several- sample repeated-measures tests in the Univariate menu.

The permutations for the *p* value are carried out only within subjects.

Pairwise tests

Pairwise PerMANOVAs between all pairs of groups are provided as a post-hoc test. Significant comparisons (at $p<0.05$) are shown in pink. The Bonferroni correction shown in the upper triangle of the matrix multiplies the *p* values with the number of comparisons. This correction is very conservative (produces large *p* values).

Missing data supported by pairwise deletion.

Reference

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32-46.

Two-way ANOSIM

The two-way ANOSIM in PAST uses the crossed design (Clarke 1993). For more information see one-way ANOSIM, but two group columns are required. There must be several rows (replication) for each combination of group levels.

Reference

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18:117-143.

Two-way ANOSIM without replication

Input data as for two-way ANOSIM above, i.e. two group columns are required. There must be exactly one row (no replication) for each combination of group levels.

Reference

Clarke, K.R. & Warwick, R.M. 1994. Similarity-based testing for community pattern: the two-way layout with no replication. *Marine Biology* 118:167-176.

Two-way PERMANOVA

The two-way NPMANOVA (Anderson, 2001) in PAST uses the crossed design. The algorithm follows Anderson (2001) with a modification to allow unbalanced designs (each combination of factors can contain different numbers of values). For more information see one-way NPMANOVA, but two group columns are required (as for two-way ANOSIM).

Reference

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26:32-46.

Test for multivariate dispersion (PERMDISP)

Tests for equal dispersion ("spread") in two or more groups of multivariate data (PERMDISP; Anderson 2005).

The procedure starts from a standard Principal Coordinates Analysis of the complete data set (all groups), using the selected distance measure. In the PCoA space, calculate the Euclidean distance z_i from each point i to its group centroid. These distances are subjected to a standard one-way, univariate ANOVA. The significance (*p* value) is estimated by a permutation test.

PAST also includes pairwise tests across groups. These tests are based on the original PCoA of the complete data set (not PCoA on the reduced data set with only two groups).

A box plot of dispersions is provided, showing the z values for each group.

Reference

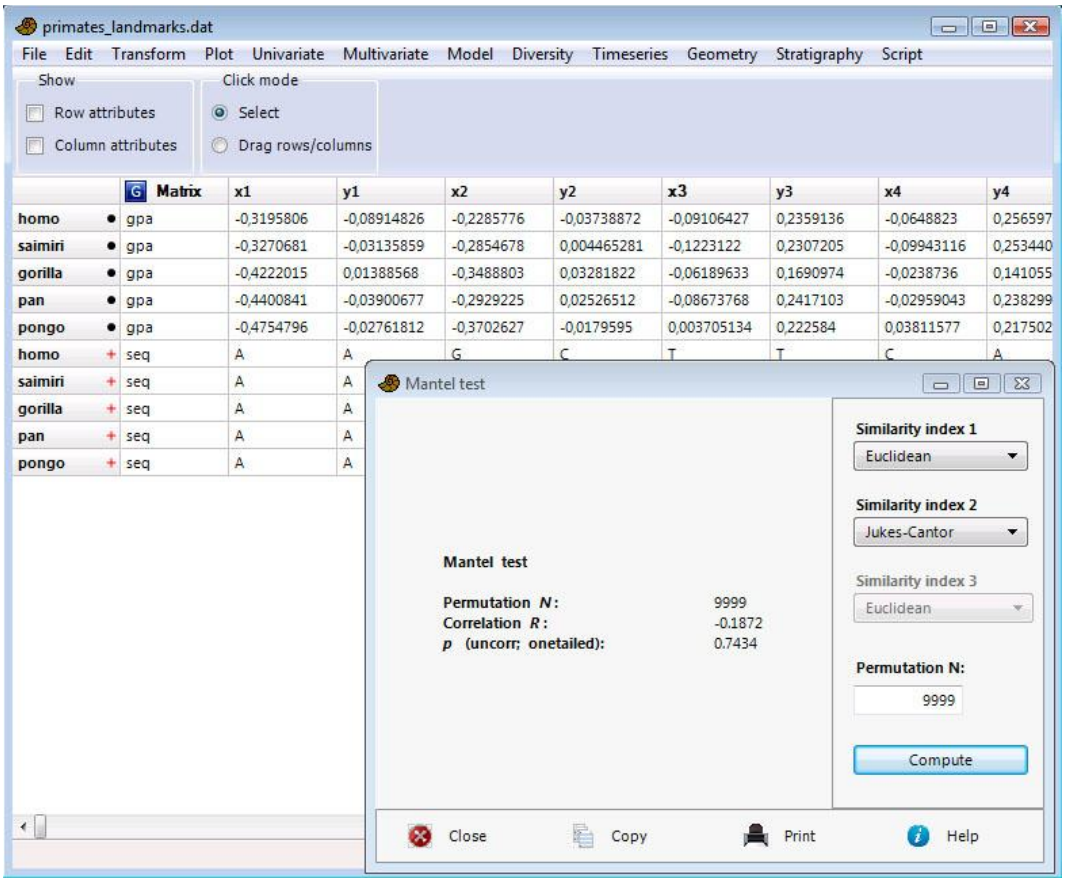
Anderson, M.J. 2005. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 62:245-253.

Mantel test and partial Mantel test

The Mantel test (Mantel 1967, Mantel & Valand 1970) is a permutation test for correlation between two distance or similarity matrices. In PAST, these matrices can also be computed automatically from two sets of original data. The first matrix must be above the second matrix in the spreadsheet, and the rows be specified as two groups (with a group column). The two matrices must have the same number of rows. If they are distance or similarity matrices, they must also have the same number of columns.

The *R* value is simply the Pearson's correlation coefficient between all the entries in the two matrices (because the matrices are symmetric it is only necessary to correlate the lower triangles). It ranges from -1 to +1. The permutation test compares the original *R* to *R* computed in e.g. 9999 random permutations. The reported *p* value is one-tailed.

In the example below, the first matrix (gpa) consists of Procrustes-fitted landmark coordinates from primate skulls, while the second matrix (seq) contains sequence data from the same primates. The user has selected the Euclidean measure for the first matrix, and Jukes-Cantor for the second. The two data sets seem to be negatively correlated (*R*=-0.19), and there is no significant positive correlation (the test is one-tailed). In other words, there is no correlation between morphology and genetics.



Partial Mantel test

It is possible to add a third matrix **C** below the two matrices **A** and **B** as described above. This matrix must be marked as above, and contain the same number of rows as **A** and **B**. A separate similarity measure can then be selected for this matrix. If such a third matrix is included, the program will carry out a partial Mantel test for the correlation of **A** and **B**, controlling for similarities given in **C** (Legendre & Legendre 1998). Only matrix **A** is permutated, and the *R* value is computed as

$$R(\mathbf{AB} \bullet \mathbf{C}) = \frac{R(\mathbf{AB}) - R(\mathbf{AC})R(\mathbf{BC})}{\sqrt{1 - R(\mathbf{AC})^2}\sqrt{1 - R(\mathbf{BC})^2}}$$

where *R*(**AB**) is the correlation coefficient between **A** and **B**.

References

Legendre, P. & L. Legendre. 1998. Numerical Ecology, 2nd English ed. Elsevier, 853 pp.

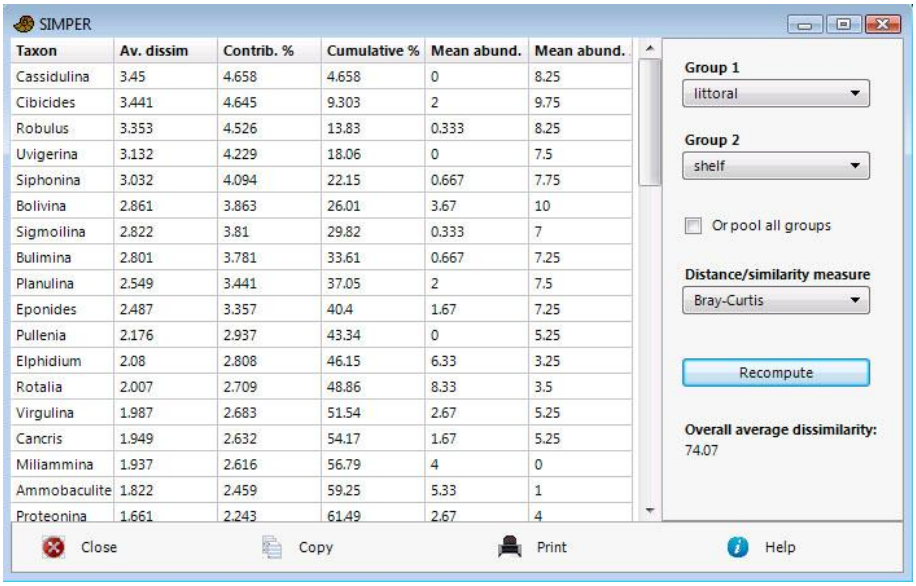
Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209-220.

Mantel, N. & R.S. Valand 1970. A technique of nonparametric multivariate analysis. *Biometrics* 26:547-558.

SIMPER

SIMPER (Similarity Percentage) is a simple method for assessing which taxa are primarily responsible for an observed difference between groups of samples (Clarke 1993). The overall significance of the difference is often assessed by ANOSIM. The Bray-Curtis similarity measure (multiplied with 100) is most commonly used with SIMPER, but the Euclidean, cosine and chord measures can also be used.

If more than two groups are selected, you can either compare two groups (pairwise) by choosing from the lists of groups, or you can pool all samples to perform one overall multi-group SIMPER. In the latter case, all possible pairs of samples are compared using the Bray-Curtis measure. The overall average dissimilarity is computed using all the taxa, while the taxon-specific dissimilarities are computed for each taxon individually.



Taxon	Av. dissim	Contrib. %	Cumulative %	Mean abund.	Mean abund.
Cassidulina	3.45	4.658	4.658	0	8.25
Cibicides	3.441	4.645	9.303	2	9.75
Robulus	3.353	4.526	13.83	0.333	8.25
Uvigerina	3.132	4.229	18.06	0	7.5
Siphonina	3.032	4.094	22.15	0.667	7.75
Bolivina	2.861	3.863	26.01	3.67	10
Sigmoilina	2.822	3.81	29.82	0.333	7
Bulimina	2.801	3.781	33.61	0.667	7.25
Planulina	2.549	3.441	37.05	2	7.5
Eponides	2.487	3.357	40.4	1.67	7.25
Pullenia	2.176	2.937	43.34	0	5.25
Elphidium	2.08	2.808	46.15	6.33	3.25
Rotalia	2.007	2.709	48.86	8.33	3.5
Virgulina	1.987	2.683	51.54	2.67	5.25
Cancris	1.949	2.632	54.17	1.67	5.25
Miliammina	1.937	2.616	56.79	4	0
Ammobaculite	1.822	2.459	59.25	5.33	1
Proteonina	1.661	2.243	61.49	2.67	4

Samples go in rows, grouped with a group column, and taxa in columns. In the output table, taxa are sorted in descending order of contribution to group difference. The last three columns show the mean abundance in each of the groups.

Missing data supported by column average substitution.

Reference

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18:117-143.

Indicator species analysis (IndVal)

An alternative to SIMPER for identifying species indicative of given groups of sites (Dufrene & Legendre 1997). Requires abundances (counts) with samples (sites) in rows, taxa in columns. Also a group column with at least two groups specified.

For each species *i* in group *j*, define the specificity as

$$A_{ij} = N_{ij}/N_i$$

where N_{ij} is the mean number of individuals of species *i* across sites in group *j*, and N_i is the sum of the mean numbers of individuals of species *i* over all groups. If $N_i = 0$, we set A_{ij} to 0.

Similarly, define the fidelity as

$$B_{ij} = Nsites_{ij}/Nsites_j$$

where $Nsites_{ij}$ is the number of sites in group *j* where species *i* is present, and $Nsites_j$ is the total number of sites in group *j*.

The indicator value of species *i* in group *j* is then a value from 0 to 100 (percentage):

$$INDVAL_{ij} = 100A_{ij}B_{ij}$$

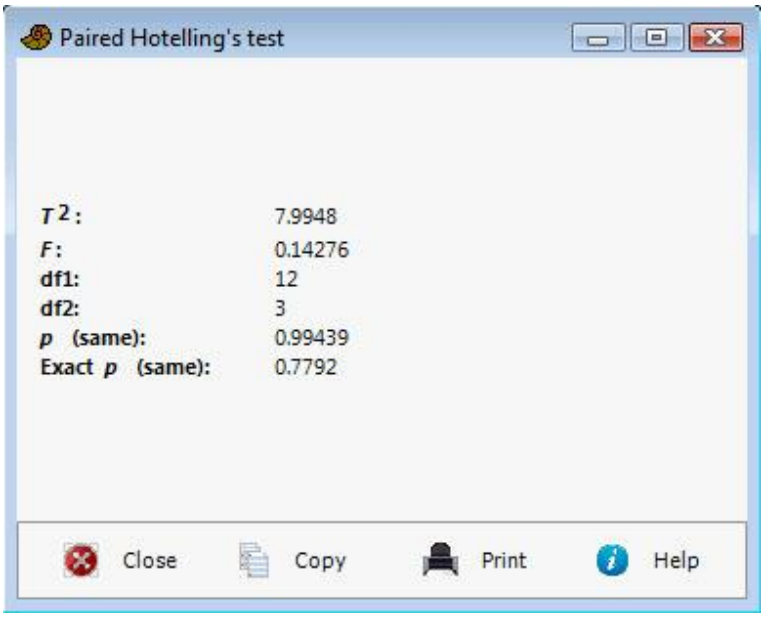
The statistical significances (*p* values) of the indicator values are estimated by 9999 random reassignments (permutations) of sites across groups. The *p* values can optionally be Bonferroni corrected (multiplied with the total number of indicator values, this is highly conservative).

Reference

Dufrene, M. & P. Legendre. 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs* 67:345-366.

Paired Hotelling

The paired Hotelling's test expects two groups of multivariate data, marked with a group column. Rows within each group must be consecutive. The first row of the first group is paired with the first row of the second group, the second row is paired with the second, etc.



With *n* the number of pairs and *p* the number of variables:

$$\mathbf{Y}_i = \mathbf{X}_{1i} - \mathbf{X}_{2i}$$

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_i \mathbf{Y}_i$$

$$\mathbf{S}_y = \frac{1}{n-1}\sum_i (\mathbf{Y}_i - \bar{\mathbf{y}})(\mathbf{Y}_i - \bar{\mathbf{y}})^{\mathrm{T}}$$

$$T^2 = n\bar{\mathbf{y}}^{\mathrm{T}}\mathbf{S}_{\mathrm{y}}^{-1}\bar{\mathbf{y}}$$

$$F = \frac{n-p}{p(n-1)}T^2$$

The *F* has *p* and *n-p* degrees of freedom.

For *n*<=16, the program also calculates an exact *p* value based on the T^2 statistic evaluated for all possible permutations.

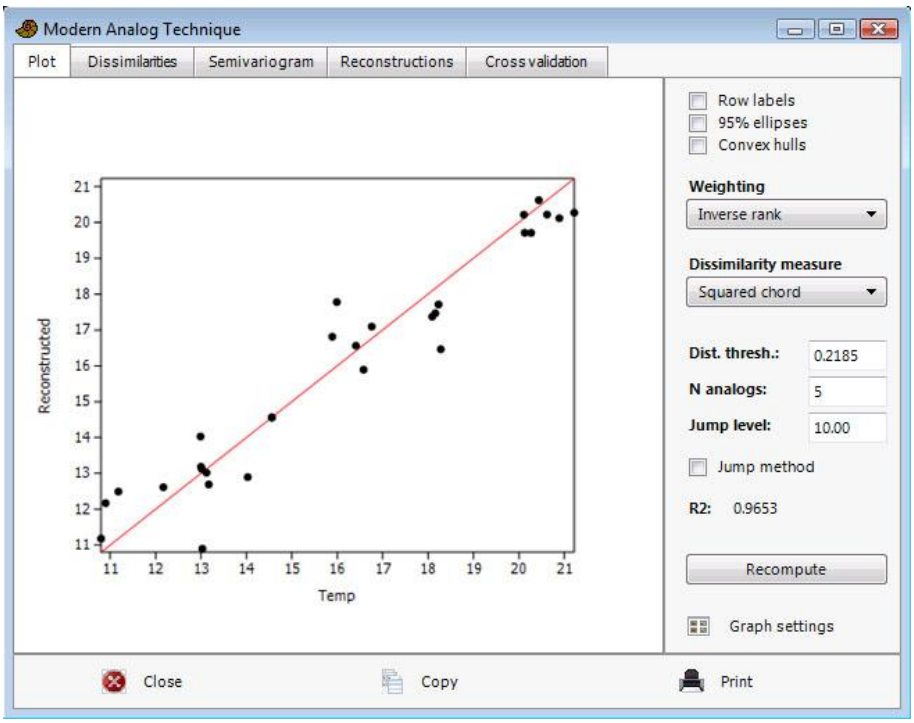
Missing data supported by column average substitution.

Modern Analog Technique

The Modern Analog Technique is a calibration method for reconstructing a past environmental parameter (e.g. temperature) from faunal associations. It works by finding modern sites with faunal associations close to those in downcore samples. Environmental data from the modern sites are then used to estimate the environment downcore.

The (single) environmental variable, usually temperature, enters in the first column, and taxa in consecutive columns. All the modern sites, with known values for the environmental variable, go in the first rows, followed by all the downcore samples (these should have question marks in the environmental column).

The plot on the first tab shows all the modern samples, with the observed temperature (for example) versus the MAT reconstructed temperature using leave-one-out cross-validation (jackknifing).



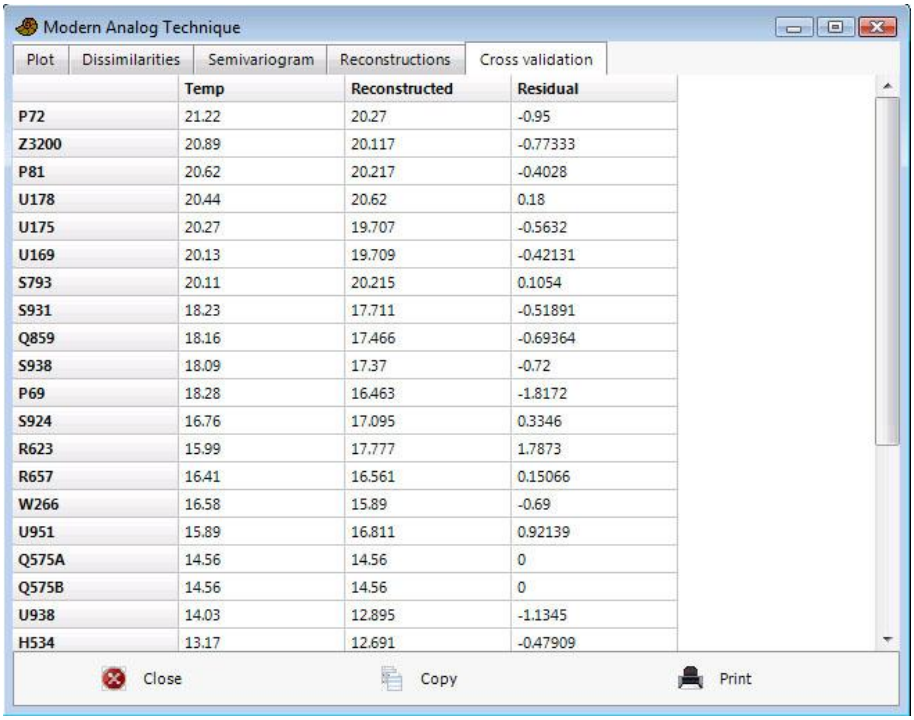
Parameters to set:

- Weighting: When several modern analogs are linked to one downcore sample, their environmental values can be weighted equally, inversely proportional to faunal distance, or inversely proportional to ranked faunal distance.
- Distance measure: Several distance measures commonly used in MAT are available. "Squared chord" has become the standard choice in the literature.
- Distance threshold: Only modern analogs closer than this threshold are used. A default value is given, which is the tenth percentile of distances between all sample pairs in the modern data. The "Dissimilarity distribution" histogram may be useful when selecting this threshold.
- N analogs: This is the maximum number of modern analogs used for each downcore sample.
- Jump method (on/off): For each downcore sample, modern samples are sorted by ascending distance. When the distance increases by more than the selected percentage, the subsequent modern analogs are discarded.

Note that one or more of these options can be disabled by entering a large value. For example, a very large distance threshold will never apply, so the number of analogs is decided only by the "N analogs" value and optionally the jump method.

Cross validation

The scatter plot and R^2 value show the results of a leave-one-out (jackknifing) cross-validation within the modern data. The *y=x* line is shown in red. This only partly reflects the "quality" of the method, as it gives little information about the accuracy of downcore estimation.



	Temp	Reconstructed	Residual
P72	21.22	20.27	-0.95
Z3200	20.89	20.117	-0.77333
P81	20.62	20.217	-0.4028
U178	20.44	20.62	0.18
U175	20.27	19.707	-0.5632
U169	20.13	19.709	-0.42131
S793	20.11	20.215	0.1054
S931	18.23	17.711	-0.51891
Q859	18.16	17.466	-0.69364
S938	18.09	17.37	-0.72
P69	18.28	16.463	-1.8172
S924	16.76	17.095	0.3346
R623	15.99	17.777	1.7873
R657	16.41	16.561	0.15066
W266	16.58	15.89	-0.69
U951	15.89	16.811	0.92139
Q575A	14.56	14.56	0
Q575B	14.56	14.56	0
U938	14.03	12.895	-1.1345
H534	13.17	12.691	-0.47909

Dissimilarity distribution

A histogram of all distances in the core-top (modern) data.

Semivariogram

Shows a semivariogram of variance in the environmental variable as a function of faunal difference. Several semivariogram models can be fitted. This type of plot is familiar from spatial geostatistics, but is also useful for MAT because it gives a good impression of the degree of “noise” in the faunal data with respect to environmental prediction.

Reconstructions

Reconstruction of the paleoenvironmental values using MAT.

Weighted averaging calibration (WA)

Like the Modern Analogue Technique, WA is a method for reconstructing past environmental parameters (temperature, pH) from a fossil assemblage, based on a training set of modern samples.

The (single) environmental variable, usually temperature, enters in the first column, and taxa in consecutive columns. All the modern sites, with known values for the environmental variable, go in the first rows, followed by all the downcore samples (these should have question marks in the environmental column).

The plot on the first tab shows all the modern samples, with the observed temperature (for example) versus the reconstructed temperature. This does not use cross-validation. The RMSE (Root Mean Square Error) is based on these values.

WA in Past follows Birks et al. (1990), with RMA deshrinking.

Reference

Birks, H.J.B., Line, J.M., Juggins, S., Stevenson, A.C., ter Braak, C.J.F. 1990. Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society of London B* 327:263-278.

Weighted averaging partial least squares (WA-PLS)

Like the Modern Analogue Technique and weighted averaging, WA-PLS is a method for reconstructing past environmental parameters (temperature, pH) from a fossil assemblage, based on a training set of modern samples. First described by ter Braak & Juggins (1993) and ter Braak et al. (1993), WA-PLS is considered by some as the overall most accurate calibration method.

The (single) environmental variable, usually temperature, enters in the first column, and taxa in consecutive columns. All the modern sites, with known values for the environmental variable, go in the first rows, followed by all the downcore samples (these should have question marks in the environmental column).

The plot on the first tab shows all the modern samples, with the observed temperature (for example) versus the reconstructed temperature. This does not use cross-validation. The RMSE (Root Mean Square Error) is based on these values.

The method is also cross-validated with the leave-one-out procedure (jackknifing), which is basis for the RMSEP value (Root Mean Square Error of Prediction). The number of PLS components should be set in order to minimize the RMSEP. Sometimes the minimal value is obtained already with only one component, when the method is equivalent to two-way weighted averaging (WA).

Individual RMSEP values for each downcore sample are estimated with bootstrapping (2000 replicates). The bootstrapping procedure is somewhat complex and follows Birks et al. (1990). Approximate 95% confidence intervals can be obtained as the predicted value plus/minus 1.96*RMSEP.

The implementation of WA-PLS in Past is based on the algorithm of ter Braak & Juggins (1993), described below with some additional details and comments.

For the modern (training) set, we have x_i the value of the measured environmental parameter in site *i*, and the *n* x *m* matrix **Y** with y_{ik} the abundance of taxon *k* in site *i*. There are *n* sites and *m* taxa. Moreover, a '+' replacing a subscript means summation over that subscript.

Step 0

Subtract the weighted mean from the environmental variable:

$$x_i = x_i - \sum_i y_{i+} x_i / y_{++}$$

Step 1

Take the centered environmental variable x_i as initial site scores r_i.

Do steps 2 to 7 for each PLS component *p*:

Step 2

Calculate new species scores u_k* by weighted averaging of the site scores:

$$u_k^* = \sum_i y_{ik} r_i / y_{+k}$$

Step 3

Calculate new site scores r_i by weighted averaging of the species scores:

$$r_i = \sum_k y_{ik} u_k^* / y_{i+}$$

Step 4

For the first PLS component, go to step 5. For second and higher components, make the new site scores r_i uncorrelated with previous components by orthogonalization, according to ter Braak (1987), Table 5.2b.

Step 5

Take the site scores r_i and the species scores u_k* as the new PLS component consisting of two vectors $\mathbf{r}^p$ and $\mathbf{u}^p$. *Note 1*: In the original algorithm, site scores are standardized in Step 5. In the Past implementation, this standardization is not carried out, in order to facilitate reconstruction for new samples (ter Braak, pers. comm. 2019). *Note 2*: The species scores $\mathbf{u}^p$ are saved as well as the site scores, as part of the PLS component.

Step 6

Do a weighted multiple regression of x_i on the components **r** obtained so far using weights y_{i+} / y_{++}. The regression coefficients are a_0 … a_p. Take the fitted values as current estimates $\hat{x}_i$ (as shown in the plot and used for calculating RMSE). Go to Step 2 with the residuals of the regression as the new site scores r_i .

Reconstruction

After Steps 2-6 have been iterated the specified number of times, a full PLS model has been constructed. Reconstruction of the environmental variable x_0 from a new sample y_{0k} is then computed as follows (in addition it must be remembered to add back the mean value subtracted in Step 0).

First calculate updated species optima:

$$\hat{u}_{k=} a_0 + \sum_p a_p u_k^p$$

Then the reconstructed x_0 is calculated as the weighted sum

$$x_0 = \sum_k y_{0k} \hat{u}_k / y_{0+}$$

References

Birks, H.J.B., Line, J.M., Juggins, S., Stevenson, A.C., ter Braak, C.J.F. 1990. Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society B* 327: 263-278.

ter Braak, C.J.F. 1987. Ordination. Pp. 91-173 in: Jongman, R.H.G., ter Braak, C.J.F., van Tongeren, O.F.R. (eds), Data Analysis in Community and Landscape Ecology, Pudoc.

ter Braak, C.J.F., Juggins, S., Birks, H.J.B., van der Voet, H. 1993. Weighted Averaging Partial Least Squares regression (WA-PLS): definition and comparison with other methods for species-

environment calibration. Pp. 525-560 in: Patil, G.P. & Rao, C.R. (eds), Multivariate Environmental Statistics, Elsevier.

ter Braak, C.J.F., Juggins, S. 1993. Weighted averaging partial least squares (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiology* 269/270: 485-502.

Similarity and distance indices

Computes a number of similarity or distance measures between all pairs of rows. The data can be univariate or (more commonly) multivariate, with variables in columns. The results are given as a symmetric similarity/distance matrix. This module is rarely used, because similarity/distance matrices are usually computed automatically from primary data in modules such as PCO, NMDS, cluster analysis and ANOSIM in Past.

Euclidean

Basic Euclidean distance (the value is adjusted for missing data).

$$d_{jk} = \sqrt{\sum_i \left(x_{ji} - x_{ki}\right)^2}.$$

Gower

A distance measure that averages the difference over all variables, each term normalized for the range of that variable:

$$d_{jk} = \frac{1}{n}\sum_i \frac{\left|x_{ji} - x_{ki}\right|}{\max_s x_{si} - \min_s x_{si}}.$$

The Gower measure is similar to Manhattan distance (see below) but with range normalization. When using mixed data types (see below), this is the default measure for continuous and ordinal data.

Chord

Euclidean distance between normalized vectors. Commonly used for abundance data. Can be written as

$$d_{jk} = \sqrt{2 - 2\frac{\sum_i x_{ji} x_{ki}}{\sqrt{\sum_i x_{ji}^2 \sum_i x_{ki}^2}}}.$$

Manhattan

The sum of differences in each variable:

$$d_{jk} = \sum_i \left|x_{ji} - x_{ki}\right|.$$

Bray-Curtis

Bray-Curtis is a popular similarity index for abundance data. Past calculates Bray-Curtis similarity as follows:

$$d_{jk} = 1 - \frac{\sum_i \left| x_{ji} - x_{ki} \right|}{\sum_i \left(x_{ji} + x_{ki} \right)}.$$

This is algebraically equivalent to the form given originally by Bray and Curtis (1957):

$$d_{jk} = 2 \frac{\sum_i \min\left(x_{ji}, x_{ki} \right)}{\sum_i \left(x_{ji} + x_{ki} \right)}.$$

Many authors operate with a Bray-Curtis distance, which is simply 1-*d.*

Cosine

The inner product of abundances each normalised to unit norm, i.e. the cosine of the angle between the vectors.

$$d_{jk} = \frac{\sum_i x_{ji} x_{ki}}{\sqrt{\sum_i x_{ji}^2} \sqrt{\sum_i x_{ki}^2}}.$$

Morisita

For abundance data.

$$\lambda_1 = \frac{\sum_i x_{ji} \left(x_{ji} - 1 \right)}{\sum_i x_{ji} \left(\sum_i x_{ji} - 1 \right)}$$

$$\lambda_2 = \frac{\sum_i x_{ki} \left(x_{ki} - 1 \right)}{\sum_i x_{ki} \left(\sum_i x_{ki} - 1 \right)}$$

$$d_{jk} = \frac{2 \sum_i x_{ji} x_{ki}}{\left(\lambda_1 + \lambda_2 \right) \sum_i x_{ji} \sum_i x_{ki}}.$$

Horn

Horn's overlap index for abundance data (Horn 1966).

$$N_j = \sum_i x_{ji}$$

$$N_k = \sum_i x_{ki}$$

$$d_{jk} = \frac{\sum_i \left[\left(x_{ji} + x_{ki} \right) \ln \left(x_{ji} + x_{ki} \right) \right] - \sum_i x_{ji} \ln x_{ji} - \sum_i x_{ki} \ln x_{ki}}{\left(N_j + N_k \right) \ln \left(N_j + N_k \right) - N_j \ln N_j - N_k \ln N_k}.$$

Mahalanobis

A distance measure taking into account the covariance structure of the data. With **S** the variance-covariance matrix:

$$d_{jk} = \sqrt{\left(\mathbf{x_j} - \mathbf{x_k} \right)^T \mathbf{S}^{-1} \left(\mathbf{x_j} - \mathbf{x_k} \right)}.$$

Correlation

The complement 1-*r* of Pearson's *r* correlation across the variables:

$$d_{jk} = 1 - \frac{\sum_i \left(x_{ji} - \bar{x}_j \right) \left(x_{ki} - \bar{x}_k \right)}{\sqrt{\sum_i \left(x_{ji} - \bar{x}_j \right)^2 \sqrt{\left(x_{ki} - \bar{x}_k \right)^2}}}.$$

Taking the complement makes this a distance measure. See also the Correlation module, where Pearson's *r* is given directly and with significance tests.

Rho

The complement 1-r_s of Spearman's rho, which is the correlation coefficient of ranks. See also the Correlation module, where rho is given directly and with significance tests.

Dice

Also known as the Sorensen coefficient. For binary (absence-presence) data, coded as 0 or 1 (any positive number is treated as 1). The Dice similarity puts more weight on joint occurences than on mismatches.

When comparing two rows, a match is counted for all columns with presences in both rows. Using *M* for the number of matches and *N* for the total number of columns with presence in just one row, we have

$$d_{jk} = 2M / (2M+N).$$

Jaccard

A similarity index for binary data. With the same notation as given for Dice similarity above, we have

$$d_{jk} = M / (M+N).$$

Kulczynski

A similarity index for binary data. With the same notation as given for Dice similarity above (with N_1 and N_2 referring to the two rows), we have

$$d_{jk} = \frac{\frac{M}{M+N_1} + \frac{M}{M+N_2}}{2}.$$

Ochiai

A similarity index for binary data, comparable to the cosine similarity for other data types:

$$d_{jk} = \sqrt{\frac{M}{M+N_1}\frac{M}{M+N_2}}.$$

Simpson

The Simpson index is defined simply as M / N_{min}, where N_{min} is the smaller of the numbers of presences in the two rows. This index treats two rows as identical if one is a subset of the other, making it useful for fragmentary data.

Raup-Crick

Raup-Crick index for absence-presence data. This index (Raup & Crick 1979) uses a randomization (Monte Carlo) procedure, comparing the observed number of species ocurring in both associations with the distribution of co-occurrences from 1000 random replicates from the pool of samples.

Hamming

Hamming distance for categorical data as coded with integers (or sequence data coded as CAGT). The Hamming distance is the number of differences (mismatches), so that the distance between (3,5,1,2) and (3,7,0,2) equals 2. In PAST, this is normalised to the range [0,1], which is known to geneticists as "p-distance".

Jukes-Cantor

Distance measure for genetic sequence data (CAGT). Similar to *p* (or Hamming) distance, but takes into account probability of reversals:

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

Kimura

The Kimura 2-parameter distance measure for genetic sequence data (CAGT). Similar to Jukes-Cantor distance but takes into account different probabilities of nucleotide transitions vs. transversions (Kimura 1980). With *P* the observed proportion of transitions and *Q* the observed number of transversions, we have

$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q).$$

Tajima-Nei

Distance measure for genetic sequence data (CAGT). Similar to Jukes-Cantor distance, but does not assume equal nucleotide frequencies.

Tamura

Distance measure for genetic sequence data (CAGT). An extension of the Kimura 2-parameter distance, handling unequal transition/transversion probability, but also taking into account a possible bias in the G+C frequency. With *P* and *Q* as for Kimura distance, and *h* = 2θ(1- θ) where θ is the G+C frequency (0-1):

$$d = -h\ln\left(1 - \frac{P}{h} - Q\right) - \frac{1}{2}(1-h)\ln(1-2Q)$$

Geographical

Distance in meters along a great circle between two points on the Earth's surface. Exactly two variables (columns) are required, with latitudes and longitudes in decimal degrees (e.g. 58 degrees 30 minutes North is 58.5). Coordinates are expected in the WGS84 datum, and distance is calculated with respect to the WGS84 ellipsoid. Use of other datums will result in very slight errors.

The accuracy of the algorithm used (Vincenty 1975) is on the order of 1 mm with respect to WGS84.

User-defined similarity

Expects a symmetric similarity matrix rather than original data. No error checking!

User-defined distance

Expects a symmetric distance matrix rather than original data. No error checking!

Mixed

This option requires that data types have been assigned to columns (see *Entering and manipulating data*). A pop-up window will ask for the similarity/distance measure to use for each datatype. These will be combined using an average weighted by the number of variates of each type. The default choices correspond to those suggested by Gower, but other combinations may well work better. The "Gower" option is a range-normalised Manhattan distance.

All-zeros rows: Some similarity measures (Dice, Jaccard, Simpson etc.) are undefined when comparing two all-zero rows. To avoid errors, especially when bootstrapping sparse data sets, the similarity is set to zero in such cases.

Missing data: Most of these measures treat missing data (coded as '?') by pairwise deletion, meaning that if a value is missing in one of the variables in a pair of rows, that variable is omitted from the computation of the distance between those two rows. The exceptions are rho distance, using column average substitution, and Raup-Crick, which does not accept missing data.

References

Bray, J.R. & J.T. Curtis. 1957. An ordination of the upland forest communities of Southern Wisconsin. Ecological Monographs 27:325-349.

Horn, H.S. 1966. Measurement of overlap in comparative ecological studies. *American Naturalist* 100:419-424.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120.

Raup, D. & R.E. Crick. 1979. Measurement of faunal similarity in paleontology. *Journal of Paleontology* 53:1213-1227.

Simpson, G.G. 1943. Mammals and the nature of continents. *American Journal of Science* 241:1-31.

Vincenty, T. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review* 176:88-93.

Genetic sequence stats

A number of simple statistics on genetic sequence (DNA or RNA) data. The module expects a number of rows, each with a sequence. The sequences are expected to be aligned and of equal length including gaps (coded as '?'). Some of these statistics are useful for selecting appropriate distance measures elsewhere in Past.

Total length:	The total sequence length, including gaps, of one sequence
Polymorphic sites:	The number of positions with variable states
Average gap:	The number of gap positions, averaged over all sequences
Average A, T/U, C, G:	The average number of positions containing each nucleotide
Average *p* distance:	The *p* distance between two sequences, averaged over all pairs of sequences. The *p* (or Hamming) distance is defined as the proportion of unequal positions
Average Jukes-Cantor *d*:	The Jukes-Cantor *d* distance between two sequences, averaged over all pairs of sequences. $d = -3\ln(1 - 4p/3)/4$, where *p* is the *p* distance
Maximal Jukes-Cantor *d*:	Maximal Jukes-Cantor distance between any two sequences
Average transitions (*P*):	Average number of transitions (a↔g, c↔t, i.e. within purines, pyrimidines)
Average transversions (*Q*):	Average number of transversions (a↔t, a↔c, c↔g, t↔g, i.e. across purines, pyrimidines)
***R*=*P*/*Q*:**	The transition/transversion ratio

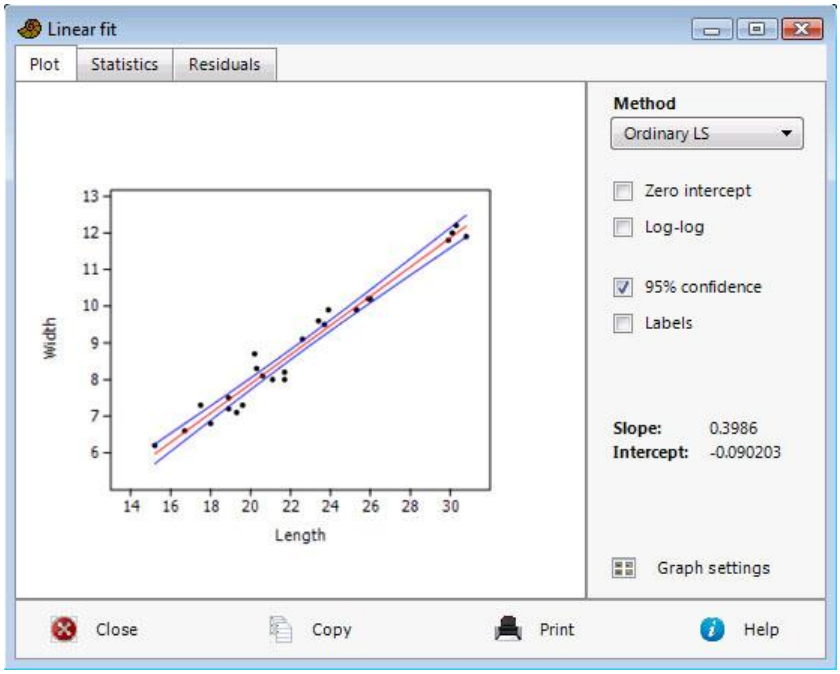
Missing data: Treated as gaps.

Model menu

Linear, bivariate

If two columns are selected, they represent *x* and *y* values, respectively. If one column is selected, it represents *y* values, and *x* values are taken to be the sequence of positive integers (1,2,...). A straight line *y=ax+b* is fitted to the data. Several bivariate data sets can be regressed in the same plot, and their slopes compared, by giving an even number of columns, each pair of columns being one x-y set. Finally, new values can be predicted by entering the *x* value but giving a '?' for the *y* value.

There are five different algorithms available: Ordinary Least Squares (OLS), Reduced Major Axis (RMA), Major Axis (MA), Robust, and Prais-Winsten. OLS regression assumes the *x* values are fixed, and finds the line which minimizes the squared errors in the *y* values. Use this if your *x* values have very little error associated with them. RMA and MA try to minimize both the *x* and the *y* errors. RMA/MA fitting, standard error estimation and slope comparison are according to Warton *et al*. (2006).

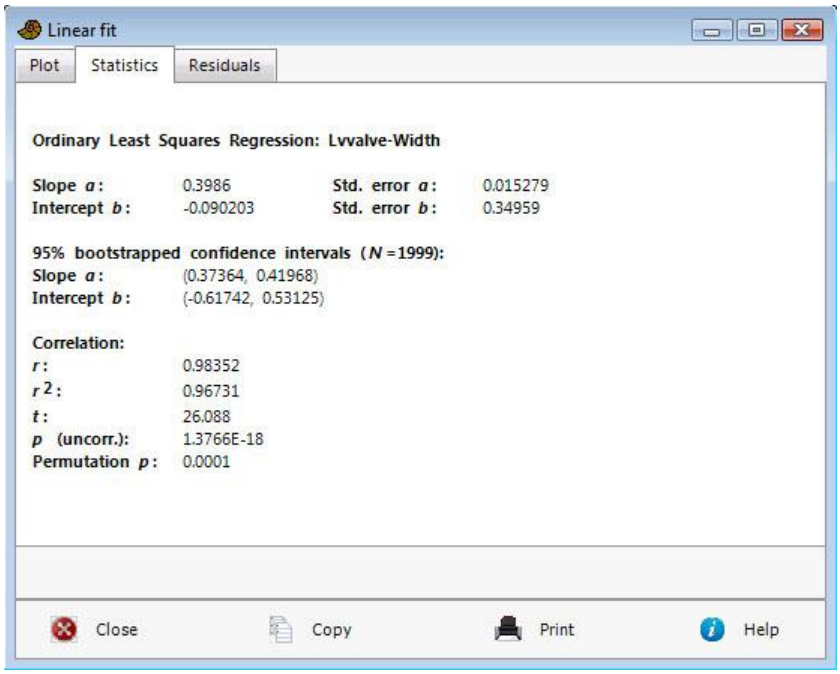


The "Robust" method is an advanced Model I (fixed *x* values) regression which is robust to outliers. It sometimes gives strange results, but can be very successful in the case of "almost" normally distributed errors but with some far-off values. The algorithm is "Least Trimmed Squares" based on the "FastLTS" code of Rousseeuw & Driessen (1999). Parametric error estimates are not available, but Past gives bootstrapped confidence intervals on slope and intercept (beware – this is extremely slow for large data sets).

Prais-Winsten regression (e.g. Wooldridge 2012, ch. 12) is appropriate for data with serially correlated residuals, typically time series. The fitted model is a sum of a linear function and an AR(1) autoregressive process with autocorrelation rho. An iterative procedure is used, with a tolerance on rho of 0.001 and a maximum of 10 iterations. Bootstrapping is not carried out as it would violate the serial correlation.

Both *x* and *y* values can be log-transformed (base 10), in effect fitting your data to the 'allometric' function $y=10^b x^a$. An *a* value around 1 indicates that a straight-line ('isometric') fit may be more applicable.

The values for *a* and *b*, their errors, Pearson's *r* correlation, and the probability that the columns are *not* correlated are given. Note the r^2 is simply the Pearson's coefficient squared – it does not adjust for regression method.



The calculation of standard errors for slope and intercept assumes normal distribution of residuals and independence between the variables and the variance of residuals. If these assumptions are strongly violated, it is preferable to use the bootstrapped 95 percent confidence intervals (1999 replicates).

The permutation test on correlation (r^2) uses 9,999 replicates.

Confidence band for the regression

In OLS regression (not RMA/MA/Robust/Prais-Winsten), a 95 percent "Working-Hotelling" confidence band for the fitted line is available. The confidence band is calculated as

$$\mathrm{CI} = b + ax \pm t_{0.05/2,n-2}\sqrt{SE_{reg}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

where the squared sum of residuals $SE_{reg}^2 = \sum(y_i - b - ax_i)^2$.

When the intercept is forced to zero, the confidence band is calculated as

$$\mathrm{CI} = ax \pm t_{0.05/2,n-1}\sqrt{SE_{reg}^2 \frac{x^2}{\sum x_i^2}}\ .$$

Confidence band for the forecast (prediction)

In OLS regression, a 95 percent confidence band for forecasting is also given. The confidence band is calculated as follows. First calculate the standard error of the estimate and the standard error of the estimate of the mean:

$$SEE = \sqrt{\frac{\sum(y_i - b - ax_i)^2}{n-2}}$$

$$SE_{mean}(x) = \frac{SEE}{\sqrt{n}}\sqrt{1 + \frac{(x-\bar{x})^2}{\frac{1}{n}\sum(x_i - x)^2}}$$

Then,

$$CI = b + ax \pm t_{\frac{0.05}{2},n-2}\sqrt{SEE^2 + \left(SE_{mean}(x)\right)^2}$$

Zero intercept

Forces the regression line through zero. This has implications also for the calculation of slope and the standard error of the slope. All five methods handle this option.

Residuals

The Residuals window reports the distances from each data point to the regression line, in the *x* and *y* directions. Only the latter is of interest when using ordinary linear regression rather than RMA or MA. The residuals can be copied back to the spreadsheet and inspected for normal distribution and independence between independent variable and residual variance (homoskedasticity).

Durbin-Watson test

The Durbin-Watson test for positive autocorrelation of residuals in *y* (violating an assumption of OLS regression) is given in the Residuals window. The test statistic varies from zero (total positive autocorrelation) through 2 (zero autocorrelation) to 4 (negative autocorrelation). For *n*<=400, an exact *p* value for no positive autocorrelation is calculated using the PAN algorithm (Farebrother 1980, with later corrections). The test is not accurate when using the Zero intercept option.

Breusch-Pagan test

The Breusch-Pagan test for heteroskedasticity, i.e. nonstationary variance of residuals (violating an assumption of OLS regression) is given in the Residuals window. The test statistic is $LM = nr^2$ where r is the correlation coefficient between the x values and the squared residuals. It is asymptotically distributed as χ^2 with one degree of freedom. The null hypothesis of the test is homoskedasticity.

Exponential functions

Your data can be fitted to an exponential function $y=e^b e^{ax}$ by first log-transforming just your y column (in the Transform menu) and then performing a straight-line fit.

Prediction (forecasting)

Rows with a '?' for the y value will be included in the table under the 'Prediction' tab. The predicted y value is calculated for the given x, together with a 95% prediction interval calculated as above (confidence band for the forecast). If the 'log-log' option was selected, the back-transformed prediction and interval will also be given for convenience. Note that this prediction is only strictly valid for the OLS model, but will be approximately correct also for the RMA and MA models.

RMA equations

Slope

$$a = \mathrm{sign}(r)\sqrt{\frac{\sum(y-\bar{y})^2}{\sum(x-\bar{x})^2}}\ .$$

Standard error on $a = \mathrm{abs}(a)\sqrt{\frac{1-r^2}{n-2}}$.

Intercept $b = \bar{y} - a\bar{x}$.

Standard error on $b = \frac{s_r^2}{n} + \bar{x}^2 s_a^2$, where s_r is the estimate of standard deviation of residuals and s_a is the standard error on slope.

For zero intercept (b=0), set $\bar{x} = 0$ and $\bar{y} = 0$ for the calculation of slope and its standard error (including the calculation of r therein), and use n-1 instead of n-2 for the calculation of standard error.

Missing data: Supported by row deletion.

References

Farebrother, R.W. 1980. Pan's procedure for the tail probabilities of the Durbin-Watson statistic. *Applied Statistics* 29:224–227.

Rousseeuw, P.J. & van Driessen, K. 1999. Computing LTS regression for large data sets. Institute of Mathematical Statistics Bulletin.

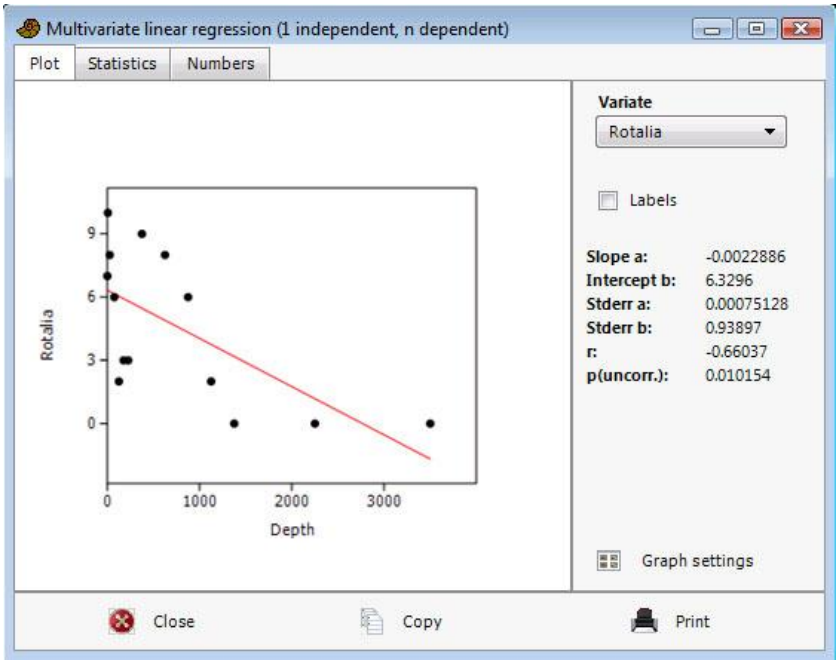
Warton, D.I., Wright, I.J., Falster, D.S. & Westoby, M. 2006. Bivariate line-fitting methods for allometry. *Biological Review* 81:259-291.

Wooldridge, J.M. 2012. Introductory Econometrics – a Modern Approach (5th ed.). South-Western Cengage Learning.

Linear, multivariate (one independent, n dependent)

When you have one independent variate and several dependent variates, you can fit each dependent variate separately to the independent variate using simple linear regression. This module makes the process more convenient by having a scroll button going through each dependent variate.

The module expects two or more columns of measured data, with the independent in the first column and the dependents in consecutive columns.



In addition, an overall MANOVA test of multivariate regression significance is provided. The Wilks' lambda test statistic is computed as the ratio of determinants

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|},$$

where **E** is the error (residuals) sum of squares and crossproducts, and **H** is the hypothesis (predictions) sum of squares and crossproducts. The Rao's *F* statistic is computed from the Wilks' lambda and subjected to a one-tailed *F* test (see 'Linear, n independent, n dependent' below).

Missing data supported by column average substitution.

Regression for geometric morphometrics

For Procrustes-fitted landmarks or Elliptic Fourier coefficients as the dependent variables, see the Geometry menu for regression with visualization of shape change.

Linear, multiple (one dependent, n independent)

Requires two or more columns of measured data, with the dependent in the first column and the independents in consecutive columns.

The program will present the multiple correlation coefficient *R* and R^2, together with the "adjusted" R^2 and an overall ANOVA-type significance test.

With SSR the regression sum of squares, SSE the error (residuals) sum of squares, *n* the number of points and *k* the number of independent variates, we have R^2=SSR/SST,

$$R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1},$$

$$F = \frac{SSR/k}{SSE/(n-k-1)}.$$

The coefficients (intercept, and slope for each independent variate) are presented with their estimated standard errors and *t* tests.

Missing data supported by column average substitution.

Linear, multivariate multiple (m independent, n dependent)

Requires two or more columns of measured data, with the dependent variables in the first column(s) and the independents in consecutive columns. The program will ask for the number of dependent variables. The output consists of four main parts.

Overall MANOVA

An overall test of multivariate regression significance. The Wilks' lambda test statistic is computed as the ratio of determinants

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|},$$

where **E** is the error (residuals) sum of squares and crossproducts, and **H** is the hypothesis (predictions) sum of squares and crossproducts.

The Rao's *F* statistic is computed from the Wilks' lambda. With *n* the number of rows, *p* the number of dependent variables and *q* the number of independent variables, we have:

$$m = n - q - 1 - \frac{1}{2}(p - q + 1)$$

$$\tau = \begin{cases} \sqrt{\dfrac{p^2q^2 - 4}{p^2 + q^2 - 5}} & \text{if } p^2 + q^2 - 5 > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$F = \frac{1 - \Lambda^{1/\tau}}{\Lambda^{1/\tau}} \cdot \frac{m\tau + 1 - pq/2}{pq}$$

The *F* test has pq and $m\tau + 1 - pq/2$ degrees of freedom.

Tests on independent variables

The test for the overall effect of each independent variable (on all dependent variables) is based on a similar design as the overall MANOVA above, but comparing the residuals of regression with and without the independent variable in question.

Tests on dependent variables

See 'Linear, n independent, one dependent' above for details of the ANOVA tests for the overall effect of all independent variables on each dependent.

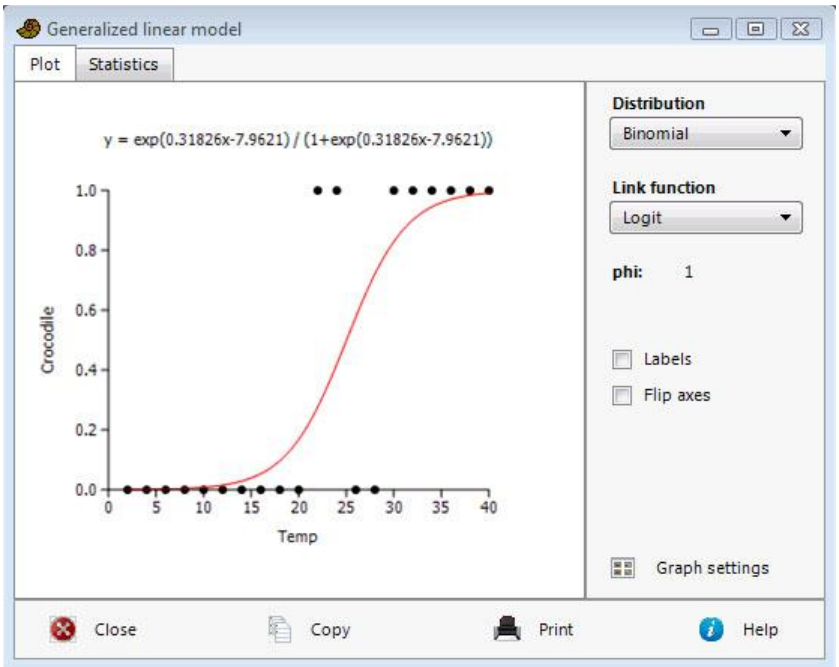
Regression coefficients and statistics

The complete set of coefficients and their significances for all combinations of independent and dependent variables.

Missing data supported by column average substitution.

Generalized Linear Model

This module computes a basic version of the Generalized Linear Model, for a single explanatory variable. It requires two columns of data (independent and dependent variables). For proportion data, a third column with total counts can be included.



GLM allows non-normal distributions, and also "transformation" of the model through a link function. Some particularly useful combinations of distribution and link function are:

Normal distribution and the identity link: This is equivalent to ordinary least squares linear regression.

Normal distribution and the reciprocal link: Fit to the function $y=1/(ax+b)$.

Normal or gamma distribution and the log link: Fit to the function $y=\exp(ax+b)$.

Poisson distribution and the log link: Fit to the function $y=\exp(ax+b)$, for counts

Binomial distribution and the logit link: Logistic regression for a binary response variable (see figure above). A third data column with sample sizes (total count) can be included, if so then the binomial distribution can be used for proportion data (0-1).

Technical details

The program uses the Iteratively Reweighted Least Squares (IRLS) algorithm for maximum likelihood estimation.

The dispersion parameter φ, which is used only for the inference, not the parameter estimation, is fixed at φ =1 for the Poisson and binomial distributions, and estimated using Pearson's chi-square for the normal and gamma distributions.

The log-likelihood *LL* is computed from the deviance *D* by $LL = -\frac{D}{2\varphi}$.

The deviance is computed as follows:

Normal: $$D = \sum_i (y_i - \mu_i)^2$$

Gamma: $$D = 2\sum_i \left[-\ln \frac{y_i}{\mu_i} + \frac{y_i - \mu_i}{\mu_i} \right]$$

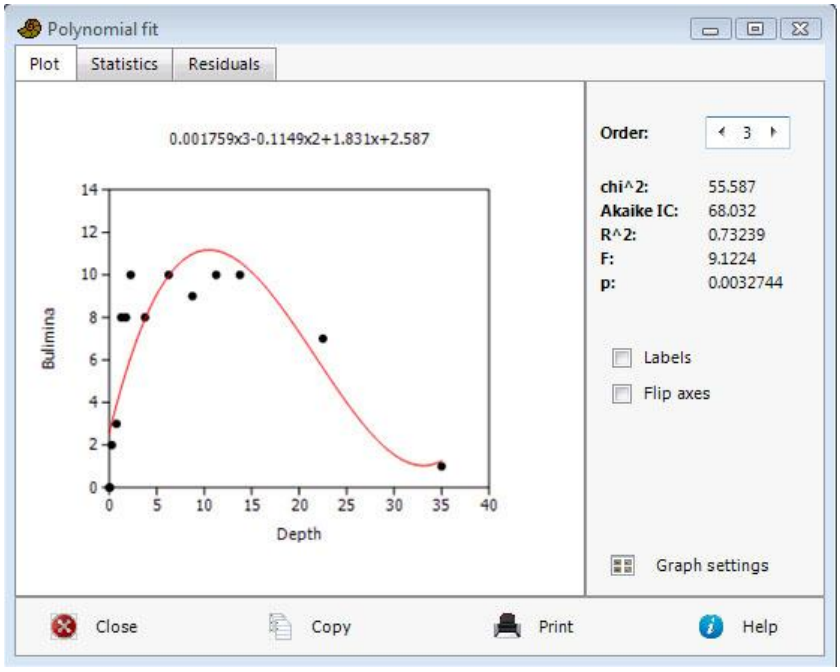
Binomial: $$D = 2\sum_i \left[y_i \ln \frac{y_i}{\mu_i} + (1 - y_i) \ln \frac{1 - y_i}{1 - \mu_i} \right]$$ (the first term defined as zero if y_i=0)

Poisson: $$D = 2\sum_i \left[y_i \ln \frac{y_i}{\mu_i} - (y_i - \mu_i) \right]$$

The *G* statistic is the difference in *D* between the full model and an additional GLM run where only the intercept is fitted. *G* is approximately chi-squared with one degree of freedom, giving a significance for the slope.

Polynomial regression

Two columns must be selected (*x* and *y* values). A polynomial of up to the fifth order is fitted to the data. The algorithm is based on a least-squares criterion and singular value decomposition (Press et al. 1992), with mean and variance standardization for improved numerical stability.



The polynomial is given by

$$y = a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 .$$

The chi-squared value is a measure of fitting error - larger values mean poorer fit. The Akaike Information Criterion has a penalty for the number of terms. The AIC should be as low as possible to maximize fit but avoid overfitting.

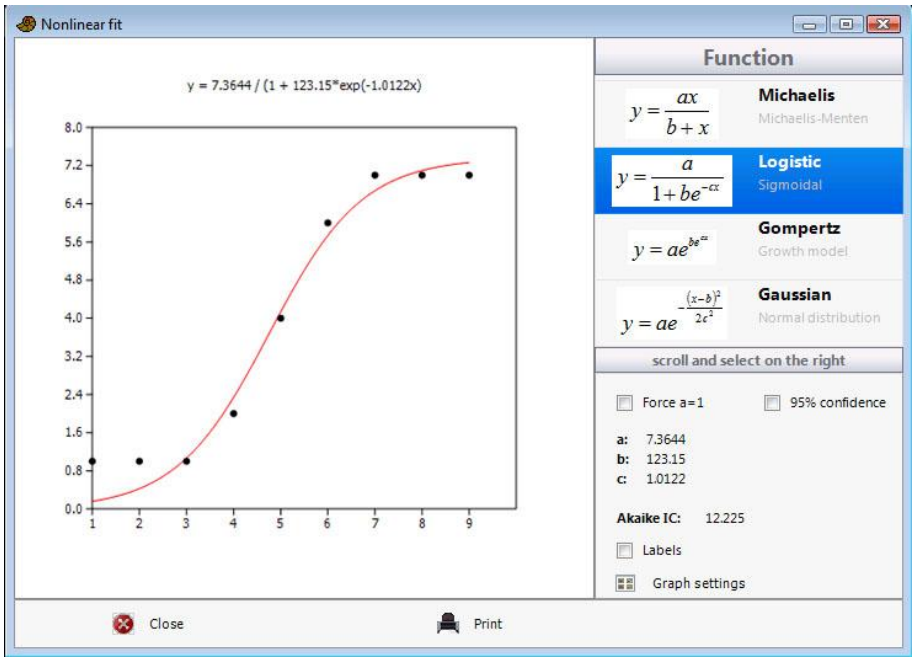
R^2 is the coefficient of determination, or proportion of variance explained by the model. Finally, a *p* value, based on an *F* test, gives the significance of the fit.

Reference

Press, W.H., S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. 1992. Numerical Recipes in C. Cambridge University Press.

Nonlinear

Attempts to fit two columns of x-y data to a number of nonlinear equations, using least squares. Select a function name from the list. To see more functions, grab a function name and drag up and down to scroll.



The 95% confidence intervals are based on 1999 bootstrap replicates.

Fitting to a nonlinear function can be a bit tricky. For most of the functions, Past uses an educated guess for the parameters, followed by Levenberg-Marquardt optimization.

The Akaike Information Criterion (AIC) may aid in the selection of model. Lower values for the AIC imply a better fit, adjusted for the number of parameters.

Linear

$$y = ax + b$$

Included for comparison with the nonlinear functions. Fitting by ordinary least squares regression. The "Zero constant" option will set *b*=0.

Quadratic

$$y = ax^2 + bx + c$$

Included for reference. Fitting by least-squares and SVD (the equation is linear in its coefficients). The "Zero constant" option will set *c*=0. See also the Polynomial Model module.

Power

$$y = ax^b + c$$

The usual power law equation. Initial guess by log-log transformation and linear regression (i.e. *c* = 0), followed by nonlinear optimization. The "Zero constant" option will set *c*=0.

Exponential

$$y = ae^{bx} + c$$

Initial guess by linearization (log-transforming *y*), followed by nonlinear optimization. The "Zero constant" option will set *c*=0. See also the Generalized Linear Model module.

Von Bertalanffy

$$y = a\left(1 - be^{-cx}\right)$$

This equation is used for modelling growth of multi-celled animals (Brown & Rothery 1993). It is sometimes given in a slightly different form:

$$y = L_\infty\left(1 - e^{-K(x-t_0)}\right)$$

It is easy to see that $L_\infty = a$, *K* = *c* and $t_0 = (\ln b)/c$.

The value of *a* is first estimated by the maximal value of *y*, and *b* and *c* using a straight-line fit to a linearized model. Finally nonlinear optimization.

Michaelis-Menten

$$y = \frac{ax}{b + x}$$

The Michaelis-Menten curve can make accurate fits to rarefaction curves, and may therefore (somewhat controversially) be used for extrapolating these curves to estimate biodiversity (Colwell & Coddington 1994). It is also an important model equation for chemical kinetics.

The algorithm uses maximum-likelihood estimators for the so-called Eadie-Hofstee transformation (Raaijmakers 1987; Colwell & Coddington 1994), followed by nonlinear optimization.

Logistic

$$y = \frac{a}{1 + be^{-cx}}$$

A sigmoidal (S-shaped) curve. The logistic equation can model growth with saturation (Brown & Rothery 1993), and was used by Sepkoski (1984) to describe the proposed stabilization of marine diversity in the late Palaeozoic.

The value of *a* is first estimated by the maximal value of *y*, and *b* and *c* using a straight-line fit to a linearized model. Finally nonlinear optimization. See also the Generalized Linear Model module.

Gompertz

$$y = ae^{be^{cx}}$$

Initial estimate is computed using regression on a linearized model , followed by nonlinear optimization.

Gaussian

$$y = ae^{-\frac{(x-b)^2}{2c^2}}$$

The 'bell curve' with mean *b* and standard deviation *c*. Initial guess of *a* by maximal value of *y*, *b* by weighted mean, and *c*=1, followed by nonlinear optimization.

Hill's equation (4-parameter logistic)

$$y = d + \frac{a-d}{1+\left(\frac{x}{b}\right)^c}$$

This sigmoidal function is often used to model dosage-response data. *d* is the minimum and *a* the maximum asymptote. *b* is the dosage at which 50% of subjects show the response (the IC_{50} value), while *c* is the "Hill slope". The "Zero constant" option will set *d*=0.

(In previous versions of Past, a slightly different form was used, with *b*/*x* instead of *x*/*b* and consequently *c* had opposite sign).

References

Brown, D. & P. Rothery. 1993. Models in biology: mathematics, statistics and computing. John Wiley & Sons.

Colwell, R.K. & J.A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B* 345:101-118.

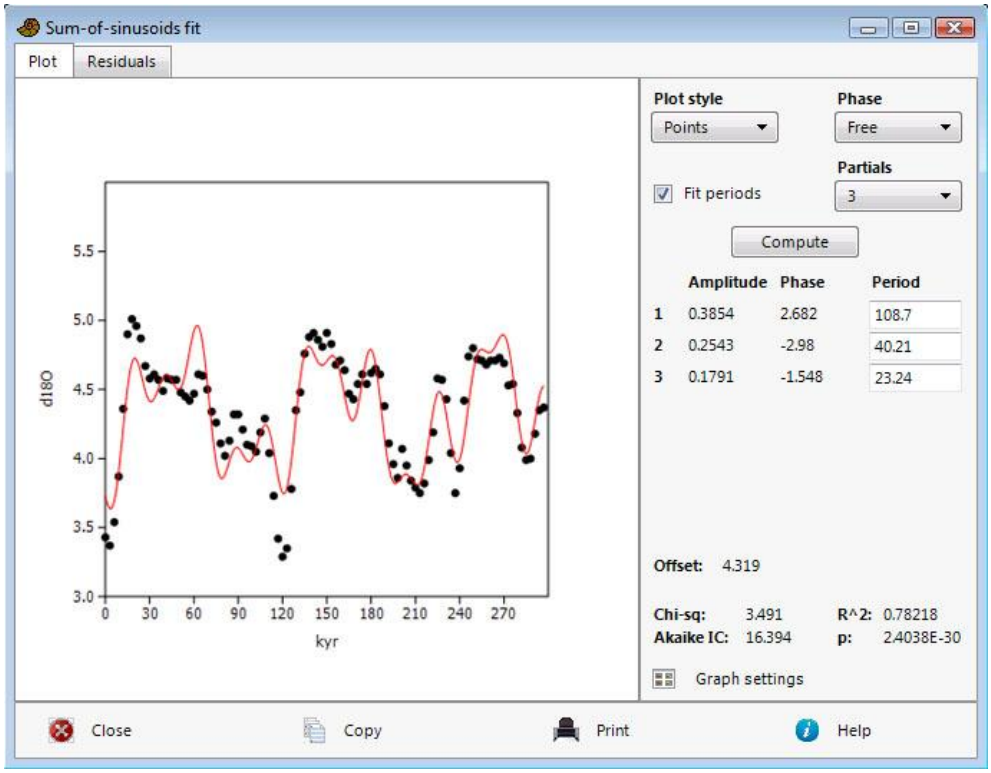
Raaijmakers, J.G.W. 1987. Statistical analysis of the Michaelis-Menten equation. *Biometrics* 43:793-803.

Sepkoski, J.J. 1984. A kinetic model of Phanerozoic taxonomic diversity. *Paleobiology* 10:246-267.

Sinusoidal regression

Two columns must be selected (*x* and *y* values). A sum of up to eight sinusoids with periods specified by the user, but with unknown amplitudes and phases, is fitted to the data. This can be useful for modeling periodicities in time series, such as annual growth cycles or climatic cycles, usually in combination with spectral analysis. The algorithm is based on a least-squares criterion and singular value decomposition. By default, the periods are set to the range of the *x* values, and harmonics (1/2, 1/3, 1/4, 1/5, 1/6, 1/7 and 1/8 of the fundamental period). These values can be changed, and need not be in harmonic proportion.

The "Fit periods" option will sequentially optimize the period of each sinusoid (over the full meaningful range from one period to the Nyquist frequency), after subtracting all previously fitted sinusoids. This is a simple example of the "Matching pursuit" algorithm. The algorithm is slow but robust and will fairly reliably find the global optimum.



The chi-squared value is a measure of fitting error - larger values mean poorer fit. The Akaike Information Criterion has a penalty for the number of sinusoids (the equation used assumes that the periods are estimated from the data). The AIC should be as low as possible to maximize fit but avoid overfitting.

R^2 is the coefficient of determination, or proportion of variance explained by the model. Finally, a *p* value, based on an *F* test, gives the significance of the fit.

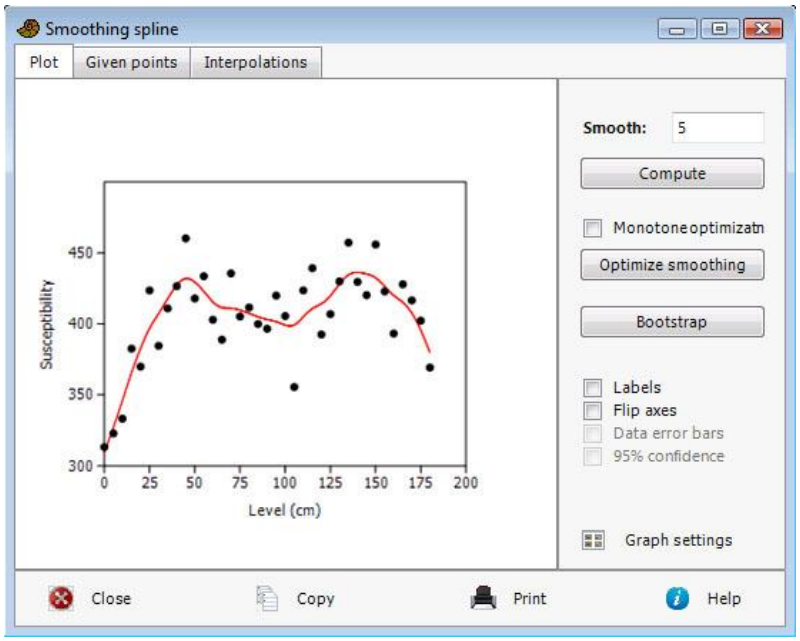
It is not meaningful to specify periodicities that are smaller than two times the typical spacing of data points.

Each sinusoid is given by $y=a*cos(2*pi*(x-x_0) / T - p)$, where a is the amplitude, T is the period and p is the phase. x_0 is the first (smallest) x value. An overall constant offset (mean) is also given.

There are also options to enforce a pure sine or cosine series, i.e. with fixed phases.

Smoothing spline

Two columns must be selected (*x* and *y* values). The data are fitted to a smoothing spline, which is a sequence of third-order polynomials continuous up to the second derivative. A typical application is the construction of a smooth curve going through a noisy data set. The algorithm follows de Boor (2001). Sharp jumps in your data can give rise to oscillations in the curve, and you can also get large excursions in regions with few data points. Multiple data points at the same *X* value are collapsed to a single point by weighted averaging and calculation of a combined standard deviation.



An optional third column specifies standard deviations on the data points. These are used for weighting the data. If unspecified, they are all set to 10% of the standard deviation of the *Y* values.

The smoothing value set by the user is a normalized version of the smoothing factor of de Boor (default 1). Larger values give smoother curves. A value of 0 will start a spline segment at every point. Clicking "Optimize smoothing" will calculate an "optimal" smoothing by a cross validation procedure.

"View given points" gives a table of the given data points *X*, *Y* and stdev(*Y*), the corresponding *Y* values on the spline curve (ys) and the residuals. The chi-squared test for each point may be used to identify outliers. The final column suggests an stdev(*Y*) value to use if forcing the p value to 0.5.

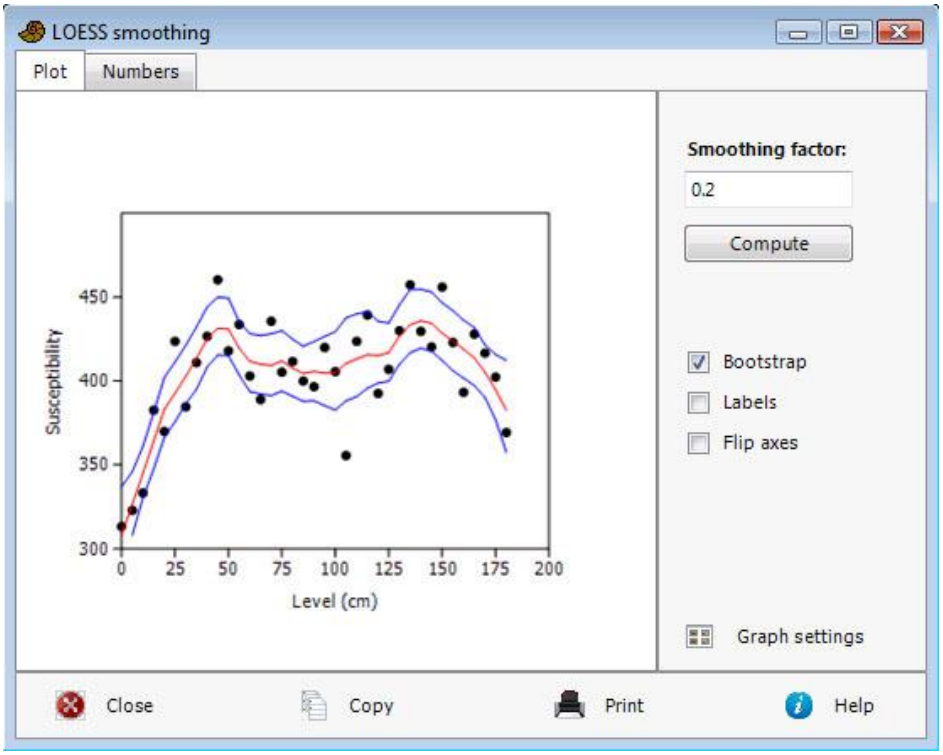
An optional fourth input column (if used then the third column must also be filled with stdev values) may contain a different number of values from the previous columns. It contains *X* values to be used for interpolation between the data points. Optional columns 5-7 contain lower and upper limits for *X* values (rectangular distribution) and standard deviation for *Y* values (normal distribution), to be used by bootstrapping (Monte Carlo) simulation providing error bars for the interpolated values. These functions are included mainly for computing boundary ages for the geological time scale.

Reference

de Boor, Carl. 2001. A practical guide to splines. Springer.

LOESS smoothing

Two columns *must* be selected (*x* and *y* values). The algorithm used is "LOWESS" (LOcally WEighted Scatterplot Smoothing; Cleveland 1979, 1981), with its recommended default parameters (including two robustness iterations). Given a number of points *n* and a smoothing parameter *q* specified by the user, the program fits the *nq* points around each given point to a straight line, with a weighting function decreasing with distance. The new smoothed point is the value of the fitted linear function at the original *x* position.



The *Bootstrap* option will estimate a 95% confidence band for the curve based on 999 random replicates. In order to retain the structure of the interpolation, the procedure uses resampling of residuals rather than resampling of original data points.

The *Optimize smoothing* option will suggest an "optimal" smoothing factor by minimizing the sum of squared prediction errors through a 5-fold cross validation. The smoothing factor is constrained to the range 0.15 to 0.95. The resulting value will tend towards the lower limit (0.15) for smooth (serially autocorrelated) data. As LOESS is primarily a technique for improving visual interpretation, the purpose of such automatic selection of the smoothing parameter is debatable.

LOESS or smoothing spline?

This is almost a matter of taste. Compare the curves above, for the same dataset. The spline often gives a more aesthetically pleasing curve because of its continuous derivatives but can suffer from overshooting near sharp bends in the data.

References

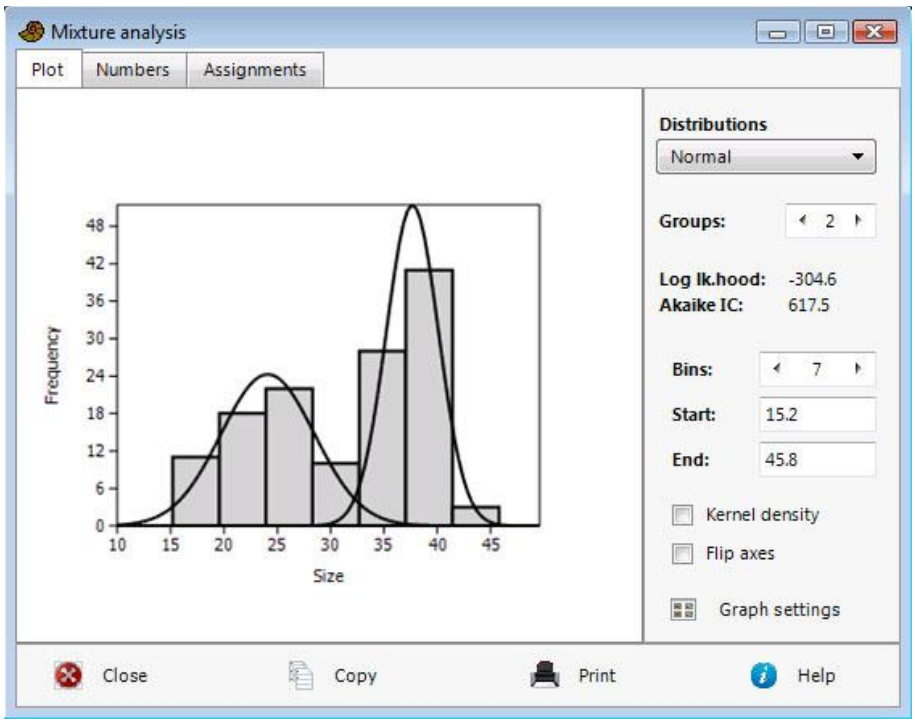
Cleveland, W.S. 1979. Robust locally weighted fitting and smoothing scatterplots. *Journal of the American Statistical Association* 74:829-836.

Cleveland, W.S. 1981. A program for smoothing scatterplots by robust locally weighted fitting. *The American Statistician* 35:54.

Mixture analysis

Mixture analysis is a maximum-likelihood method for estimating the parameters (mean, standard deviation and proportion) of two or more univariate normal distributions, based on a pooled sample. The program can also estimate mean and proportion of exponential, Poisson, and lognormal distributions. For example, the method can be used to study differences between sexes (two groups), or several species, or size classes, when no independent information about group membership is available.

The program expects one column of univariate data, assumed to be taken from a mixture of normally distributed populations (or exponential or Poisson). In the example below, sizes of two brachiopod samples have been pooled into one sample. The means, standard deviations and proportions of the two original samples have been almost perfectly recovered.



PAST uses the EM algorithm (Dempster et al. 1977), which can get stuck on a local optimum. The procedure is therefore automatically run 20 times, each time with new, random starting positions for the means. The starting values for standard deviation are set to *s/G*, where *s* is the pooled standard deviation and *G* is the number of groups. The starting values for proportions are set to 1/*G*. The user is still recommended to run the program a few times to check for stability of the solution ("better" solutions have less negative log likelihood values).

The Akaike Information Criterion (AIC; Akaike 1974) is calculated with a small-sample correction:

$$\mathrm{AICc} = 2k - 2\ln L + \frac{2k(k+1)}{n-k-1}$$

where *k* is the number of parameters, *n* the number of data points and *L* the likelihood of the model given the data. A minimal value for AIC indicates that you have chosen the number of groups that produces the best fit without overfitting.

It is possible to assign each of the data points to one of the groups with a maximum likelihood approach. This can be used as a non-hierarchical clustering method for univariate data. The "Assignments" button will open a window where the value of each probability density function is given for each data point. The data point can be assigned to the group that shows the largest value.

Missing data: Supported by deletion.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716-723.

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society, Series B* 39:1-38.

Abundance models

This module can be used for plotting taxon abundances in descending rank order on a linear or logarithmic (Whittaker plot) scale, or number of species in abundance octave classes (as shown when fitting to log-normal distribution). Taxa go in rows, samples (often only one) in columns. It can also fit the data to one of four different standard abundance models:

- Geometric, where the 2nd most abundant species should have a taxon count of *k*<1 times the most abundant, the 3rd most abundant a taxon count of *k* times the 2nd most abundant etc. for a constant *k*. With n_i the count of the *i*th most abundant taxon, we have $n_i = n_1 k^{i-1}$. This will give a straight descending line in the Whittaker plot. Fitting is by simple linear regression of the log abundances.
- Log-series, with two parameters *alpha* and *x*. The fitting algorithm is from Krebs (1989). The number of species with *n* individuals (this equation does not translate directly to the Whittaker plot representation):

$$S_n = \frac{\alpha x^n}{n}$$

- Broken stick (MacArthur 1957). There are no free parameters to be fitted in this model. With S_{tot} the total number of species and n_{tot} the total number of individuals:

$$n_i = \frac{n_{tot}}{S_{tot}} \sum_{j=0}^{S_{tot}-i} \frac{1}{S_{tot} - j}.$$

- Log-normal. The fitting algorithm is from Krebs (1989). The logarithm (base 10) of the fitted mean and variance are given. The *octaves* refer to power-of-2 abundance classes:

Octave	Abundance
1	1
2	2-3
3	4-7
4	8-15
5	16-31
6	32-63
7	64-127
...	...

A significance value based on chi-squared is given for each of these models, but the power of the test is not the same for the four models and the significance values should therefore not be compared. It is important, as always, to remember that a high *p* value can not be taken to imply a good fit. A low value does however imply a bad fit. Also note that the chi-squared tests in Past do not seem to correspond with some other software, possibly because Past use counts rather than the log-transformed values in the Whittaker plots.

References

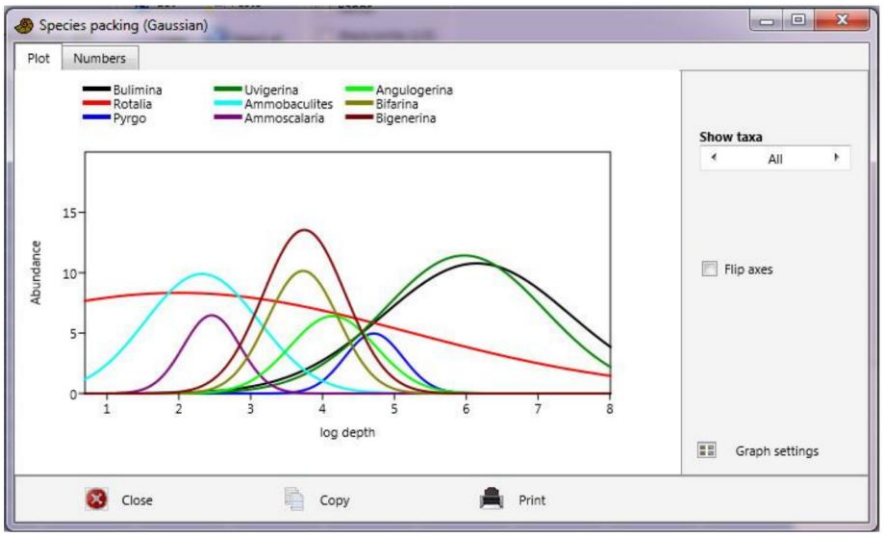
Krebs, C.J. 1989. Ecological Methodology. Harper & Row, New York.

MacArthur, R.H. 1957. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences, USA* 43:293-295.

Species packing (Gaussian)

This module fits Gaussian response models to species abundances along a gradient, for one or more species. The fitted parameters are optimum (average), tolerance (standard deviation) and maximum.

The module requires one first column of environmental measurements in samples (e.g. temperature), and one or more additional columns of abundance data (taxa in columns).

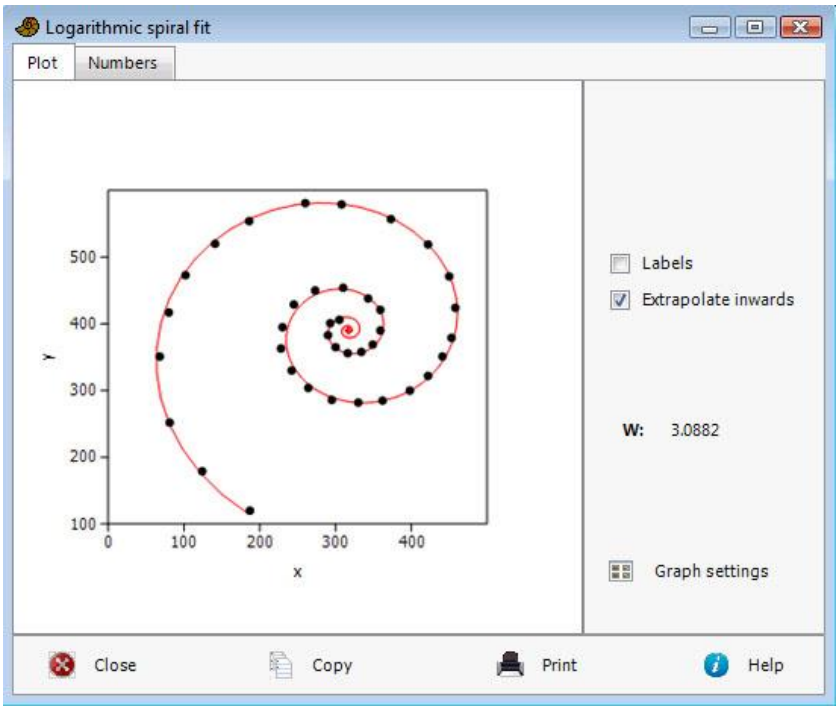


The algorithm is the same as for the Gaussian function in the nonlinear regression module: Initial estimation of optimum and tolerance based on the weighted average, followed by a nonlinear optimization by the Levenberg-Marquardt method.

Note that the R^2 value (indicating goodness of fit) can become negative – this is not a bug but means that the Gaussian fits worse than the sample mean.

Fit spiral

Fits a set of points in the plane to a logarithmic or Archimedes spiral. Useful for characterizing e.g. mollusc shells, teeth, claws and horns. Requires two columns of coordinates (x and y). The points must be given in sequence, either inwards or outwards. Left-handed and right-handed spirals are both acceptable.



The fitted logarithmic spiral in polar coordinates: $r = ae^{b\theta}$. The scale *a* and the exponent *b* are given, together with the estimated center point, marked with a red cross. The whorl expansion rate *W* (factor increase in radius per whorl) is calculated from *b* as $W = e^{2\pi b}$.

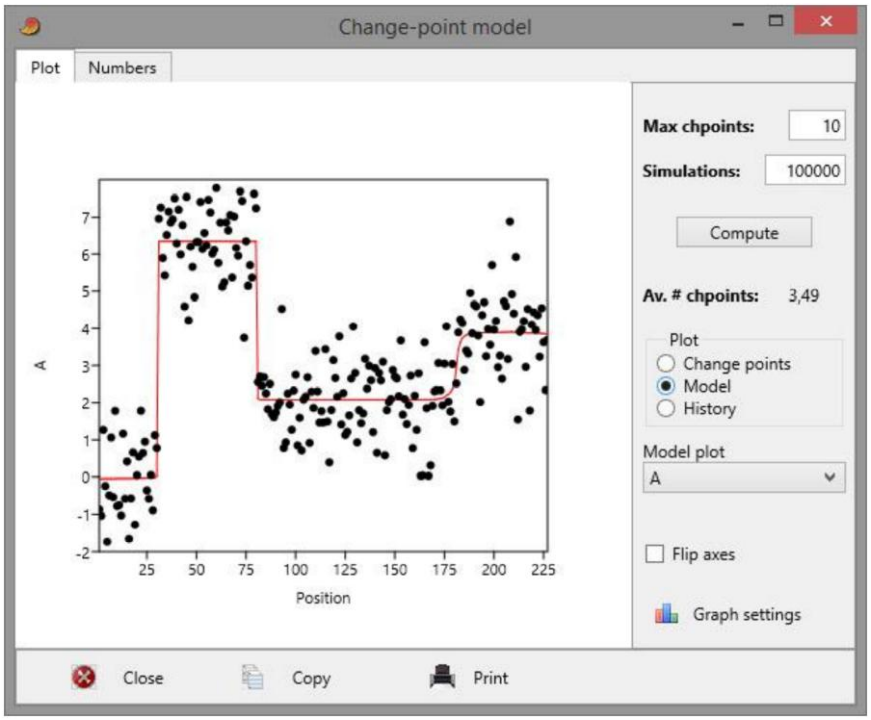
The center position is estimated by nonlinear optimization and the spiral itself by linearization and regression.

Changepoint modeling

This module suggests positions of abrupt change (changepoints) in a time series, with constant values between the changepoints. The input data should be a single column with a series of numbers, or multiple columns with multivariate data collected at the same points in time or stratigraphy. An example application is the detection of breaks in multivariate geochemical data through a sediment core. The module implements the method described by Gallagher et al. (2011).

The algorithm is Bayesian, "transdimensional" Markov chain Monte Carlo (MCMC). It produces not a single set of model parameters, but a large number as samples ("simulations") from the probability distribution.

For multiple-column data sets, note that each column is weighted equally, as mean and standard deviation are automatically normalized away prior to modeling.



Max chpoints: The maximal number of changepoints. This can often be left at the default, 10, unless you want to allow a larger or enforce a smaller number of changepoints. After analysis, the actual average number of changepoints (across simulations) is reported.

Simulations: The number of MCMC iterations, default 100,000. This includes the so-called burn-in, which is the initial number of simulations before the algorithm hopefully converges and data start to be collected. The number of burn-in iterations is fixed at 20,000. The "History" curve (see below) should be inspected to see if the number of simulations should be increased. For noisy data, it may be necessary to increase the number of simulations to a million or more, giving long computation times.

Changepoints plot: Shows a histogram of the positions of changepoints across all simulations.

Model plot: Shows the average changepoint model as a red curve superimposed on the given data points. If most simulations agree on the changepoint positions, this will be a step-like curve (i.e. constant between changepoints). Variance (i.e. uncertainty) in changepoint positions will give a more rounded appearance. For multivariate data, you can select the plotted variable in the drop-down menu.

History: Shows the model log likelihood as a function of iteration number. Ideally, this curve should start at some large negative value, and quickly increase to a relatively stable value, varying as unstructured noise around a mean. The end of the burn-in is shown as a vertical line. If the log likelihood does not seem to stabilize, the number of simulations may have to be increased.

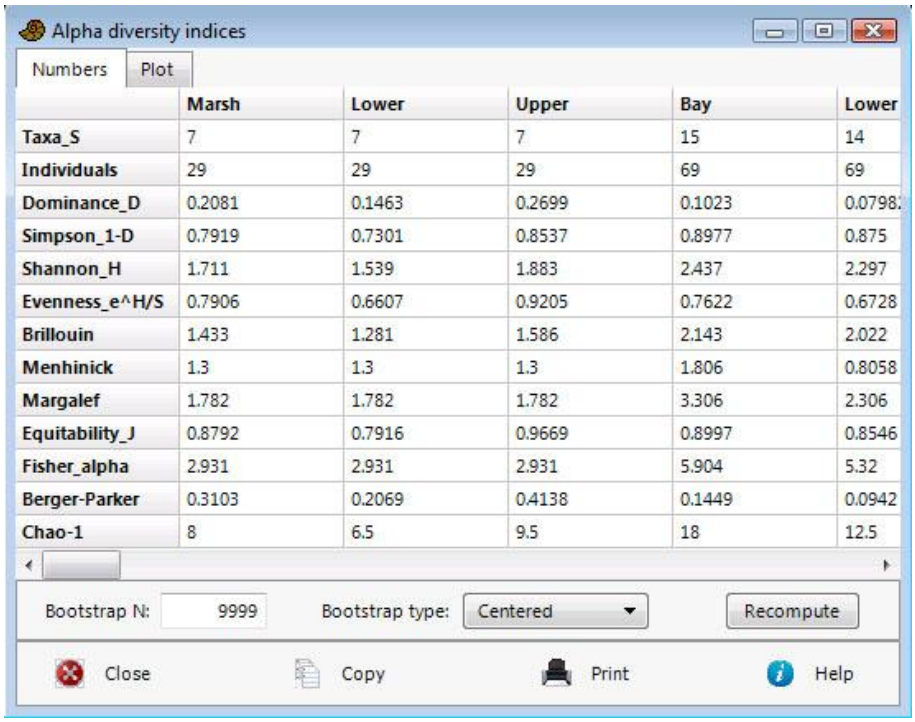
Missing values are treated by linear interpolation before the analysis.

Reference

Gallagher, K., Bodin, T., Sambridge, M., Weiss, D., Kylander, M., Large, D. 2011. Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models. *Earth and Planetary Science Letters* 311:182-194.

Diversity menu

Alpha diversity indices



	Marsh	Lower	Upper	Bay	Lower
Taxa_S	7	7	7	15	14
Individuals	29	29	29	69	69
Dominance_D	0.2081	0.1463	0.2699	0.1023	0.0798
Simpson_1-D	0.7919	0.7301	0.8537	0.8977	0.875
Shannon_H	1.711	1.539	1.883	2.437	2.297
Evenness_e^H/S	0.7906	0.6607	0.9205	0.7622	0.6728
Brillouin	1.433	1.281	1.586	2.143	2.022
Menhinick	1.3	1.3	1.3	1.806	0.8058
Margalef	1.782	1.782	1.782	3.306	2.306
Equitability_J	0.8792	0.7916	0.9669	0.8997	0.8546
Fisher_alpha	2.931	2.931	2.931	5.904	5.32
Berger-Parker	0.3103	0.2069	0.4138	0.1449	0.0942
Chao-1	8	6.5	9.5	18	12.5

These statistics apply to association data, where number of individuals are tabulated in rows (taxa) and possibly several columns (samples). The available statistics are as follows, for each sample:

- Number of taxa (*S*)

- Total number of individuals (*n*)

- Dominance = 1-Simpson index. Ranges from 0 (all taxa are equally present) to 1 (one taxon dominates the community completely).

  $D = \sum_i \left(\frac{n_i}{n}\right)^2$ where n_i is number of individuals of taxon *i*.

  If the "Unbiased" option is selected, an alternative form of *D* is computed:

  $$D = \sum_i \frac{n_i(n_i - 1)}{n(n-1)}$$

- Simpson index 1-*D*. Measures 'evenness' of the community from 0 to 1. Note the confusion in the literature: Dominance and Simpson indices are often interchanged!

- Shannon index (entropy). A diversity index taking into account the number of individuals as well as number of taxa. Varies from 0 for communities with only a single taxon to high values for communities with many taxa, each with few individuals.

$$H = -\sum_i \frac{n_i}{n} \ln \frac{n_i}{n}$$

If the "Unbiased" option is selected, *H* is computed with a bias correction

$H_u = H + (S\text{-}1)/(2n)$.

If the "Unbiased" and "Use ACE for S" options are selected, the ACE species richness estimator is used instead of *S* in the bias correction. This corresponds to the Bias-corrected MLE (MLE_bc) estimator for Shannon's index given by Chao & Shen (2003).

If the "log2" option is selected, the Shannon index is reported with the logarithm to base 2.

- Buzas and Gibson's evenness: e^H/S

- Brillouin's index:

$$HB = \frac{\ln(n!) - \sum_i \ln(n_i!)}{n}$$

- Menhinick's richness index: $\frac{S}{\sqrt{n}}$

- Margalef's richness index: $(S\text{-}1) / \ln(n)$

- Equitability. Shannon diversity *H* divided by the logarithm of number of taxa. This measures the evenness with which individuals are divided among the taxa present. This index is computed without bias correction for *H* (because *S* is not corrected correspondingly, this could produce values larger than 1 which is outside the standard range for equitability).

- Fisher's alpha - a diversity index, defined implicitly by the formula $S=a*\ln(1+n/a)$ where *S* is number of taxa, *n* is number of individuals and *a* is the Fisher's alpha.

- Berger-Parker dominance: simply the number of individuals in the dominant taxon relative to *n*.

- Chao1: An estimate of total species richness (Chao 1984). Version without bias correction:

$$S_{Chao1} = S + \left(\frac{n-1}{n}\right)\frac{F_1^2}{2F_2}$$

where F_1 is the number of singleton species and F_2 the number of doubleton species. If F_2=0, we use

$$S_{Chao1} = S + \left(\frac{n-1}{n}\right)\frac{F_1(F_1-1)}{2}$$

When the "Unbiased" option is selected, the equation is

$$S_{Chao1} = S + \left(\frac{n-1}{n}\right)\frac{F_1(F_1-1)}{2(F_2+1)}$$

unless $F_1 = 1$, in which case we revert to the version without bias correction.

- iChao1: "Improved" Chao1 estimator (Chiu et al. 2014), taking into account also the numbers F_3 and F_4 of species observed 3 and 4 times:

$$S_{iChao1} = S_{Chao1} + \frac{n-3}{4n}\frac{F_3}{F_4} \times \max\left[F_1 - \frac{n-3}{2(n-1)}\frac{F_2F_3}{F_4}, 0\right]$$

If $F_4 = 0$, we use F_4=1 to avoid division by zero (Chiu et al. 2014).

- ACE: Abundance-based Coverage Estimator (Chao & Lee 1992):
S_{rare} is the number of species with 10 or less individuals. S_{abund} is the number of species with more than 10 individuals ($S = S_{rare} + S_{abund}$). n_{rare} is the number of individuals among the rare species.

$$C_{ACE} = 1 - \frac{F_1}{n_{rare}} \qquad \text{(sample cover estimate)}$$

$$\gamma^2_{ACE} = \max\left[\frac{S_{rare}}{C_{ACE}}\frac{\sum_{k=1}^{10} k(k-1)F_k}{n_{rare}(n_{rare}-1)}, 0\right] \qquad \text{(coefficient of variation)}$$

$$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{F_1}{C_{ACE}}\gamma^2_{ACE}$$

- Squares: Richness estimator (Alroy, 2018), designed to be more accurate than Chao-1 when abundance distributions are even:

$$S_{sq} = S + \frac{F_1^2}{n^2 - F_1 S}\sum_{i=1}^{S} n_i^2$$

Some of these indices are explained in Harper (1999).

Confidence intervals

Approximate confidence intervals for all these indices can be computed with a bootstrap procedure. The given number of random samples (default 9999) is produced, each with the same total number of individuals as in the original sample. For each individual in the random sample, the taxon is chosen with probabilities proportional to the original abundances. A 95 percent confidence interval is then calculated. Note that the diversity in the replicates will often be less than, and never larger than, the pooled diversity in the total data set – this bias can optionally be "fixed" by centering the confidence interval on the original value.

Alternatively, analytical estimates of the 95% confidence interval are available for some of the indices:

- Dominance and Simpson index: Confidence interval estimated as [*D* - 1.96√(var *D*), *D* + 1.96√(var *D*)], with the variance var *D* computed as in the "Diversity t test" module (see below).

- Shannon index: Confidence interval estimated as [*H* - 1.96√(var *H*), *H* + 1.96√(var *H*)], with the variance var *H* computed as in the "Diversity t test" module (see below).

These analytical confidence intervals for the Dominance/Simpson and the Shannon indices correspond well to the centred bootstrap intervals, at least for large *n*.

- Fisher alpha: Following Fisher et al. (1943), we use

$$\operatorname{var}\alpha = \frac{\alpha^3(N+\alpha)^2 \ln\frac{2n+\alpha}{n+\alpha} - \alpha n}{(Sn + S\alpha - N\alpha)^2}$$

  The 95% confidence interval is then [α - 1.96√(var α), α + 1.96√(var α)]. This analytical confidence interval tends to be considerably wider than the bootstrapped CI. The reason for the discrepancy is unknown.

- Chao1: The computations are a bit complex. Past uses equations given in Colwell (2013). For F_1>0 and F_2>0, and for the version without bias correction, we use

$$\operatorname{var}(S_{Chao1}) = F_2\left[\frac{1}{2}\frac{n-1}{n}\left(\frac{F_1}{F_2}\right)^2 + \left(\frac{n-1}{n}\right)^2\left(\frac{F_1}{F_2}\right)^3 + \frac{1}{4}\left(\frac{n-1}{n}\right)^2\left(\frac{F_1}{F_2}\right)^4\right]$$

  For the bias-corrected version:

$$\operatorname{var}(S_{Chao1}) = \left(\frac{n-1}{n}\right)\frac{F_1(F_1-1)}{2(F_2+1)} + \left(\frac{n-1}{n}\right)^2\frac{F_1(2F_1-1)^2}{4(F_2+1)^2} + \left(\frac{n-1}{n}\right)^2\frac{F_1^2F_2(F_1-1)^2}{4(F_2+1)^4}$$

  If F_2 =0 but F_1 > 1, we use for both versions:

$$\operatorname{var}(S_{Chao1}) = \left(\frac{n-1}{n}\right)\frac{F_1(F_1-1)}{2} + \left(\frac{n-1}{n}\right)^2\frac{F_1(2F_1-1)^2}{4} - \left(\frac{n-1}{n}\right)^2\frac{F_1^4}{4S_{Chao1}}$$

  We then calculate $T = S_{Chao1} - S$ and

$$K = \exp\left(1.96\sqrt{\ln\left(1 + \frac{\operatorname{var}(S_{Chao1})}{T^2}\right)}\right)$$

  The 95% confidence interval is [$S + T/K$, $S + TK$].

  Otherwise (F_1=0, or F_2=0 and F_1=1), we use

$$\mathrm{var}(S_{Chao1}) \approx \sum_{i\geq 1} F_i\left(e^{-i} - e^{-2i}\right) - \frac{1}{n}\left(\sum_{i\geq 1} ie^{-i}F_i\right)^2$$

$$P = \frac{\sum_{i\geq 1} F_i e^{-i}}{S}$$

95% confidence interval: [S/(1-P) – 1.96√(var(S_{Chao1}))/(1-P), S/(1-P) – 1.96√(var(S_{Chao1}))/(1-P)]. If the lower bound is smaller than S, it is set to S.

Bootstrapped comparison of diversity indices in two samples is provided in the Compare diversities module.

References

Alroy, J. 2018. Limits to species richness in terrestrial communities. *Ecology Letters* 21:1781–1789.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11:265-270.

Chao, A., Lee, S.-M. 1992. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* 87:210–217.

Chao, A., Shen, T.-J. 2003. Nonparametric estimation of Shannon's diversity index when there are unseen species in sample. *Environmental and Ecological Statistics* 10:429-443.

Chiu, C.-H., Wang, Y.-T., Walther, B.A., Chao, A. 2014. An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula. *Biometrics* 70:671-682.

Colwell, R. K. 2013. EstimateS: Statistical estimation of species richness and shared species from samples. Version 9. User's Guide and application published at: http://purl.oclc.org/estimates.

Fisher, R.A., Corbet, A.S., Williams, C.B. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* 12:42-58.

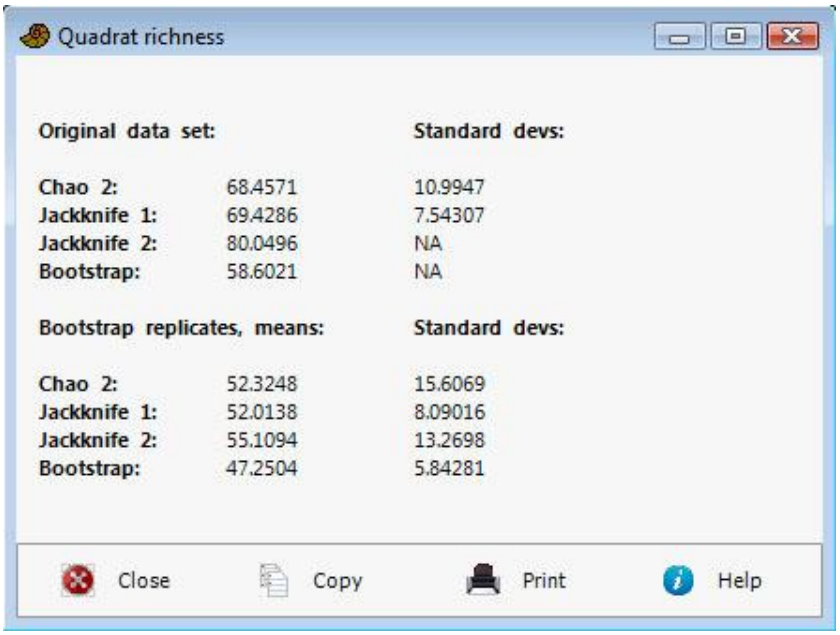
Harper, D.A.T. (ed.). 1999. Numerical Palaeobiology. John Wiley & Sons.

Quadrat richness

Requires two or more columns, each containing presence/absence (1/0) of different taxa down the rows (positive abundance is treated as presence).

Four non-parametric species richness estimators are included in PAST: Chao 2, first- and second-order jackknife, and bootstrap. All of these require presence-absence data in two or more sampled quadrats of equal size. Colwell & Coddington (1994) reviewed these estimators and found that the Chao2 and the second-order jackknife performed best.

The output from Past is divided into two panels. First, the richness estimators and their analytical standard deviations (only for Chao2 and Jackknife1) are computed from the given set of samples. Then the estimators are computed from 1000 random resamplings of the samples with replacement (bootstrapping), and their means and standard deviations are reported. In other words, the standard deviations reported here are bootstrap estimates, and not based on the analytical equations.



Chao2

The Chao2 estimator (Chao 1987) is calculated as in EstimateS version 8.2.0 (Colwell 2009), with bias correction:

$$\hat{S}_{Chao2} = S_{obs} + \left(\frac{m-1}{m}\right)\frac{Q_1(Q_1-1)}{2(Q_2+1)}$$

where S_{obs} is the total observed number of species, m the number of samples, Q_1 the number of uniques (species that occur in precisely one sample) and Q_2 the number of duplicates (species that occur in precisely two samples).

If Q_1>0 and Q_2>0, variance is estimated as

$$\hat{\text{var}}\left(\hat{S}_{Chao2}\right)=\left(\frac{m-1}{m}\right)\frac{Q_1\left(Q_1-1\right)}{2\left(Q_2+1\right)}+\left(\frac{m-1}{m}\right)^2\frac{Q_1\left(2Q_1-1\right)^2}{4\left(Q_2+1\right)^2}+\left(\frac{m-1}{m}\right)^2\frac{Q_1{}^2Q_2\left(Q_1-1\right)^2}{4\left(Q_2+1\right)^4}.$$

If Q_1>0 but Q_2=0:

$$\hat{\text{var}}\left(\hat{S}_{Chao2}\right)=\left(\frac{m-1}{m}\right)\frac{Q_1\left(Q_1-1\right)}{2}+\left(\frac{m-1}{m}\right)^2\frac{Q_1\left(2Q_1-1\right)^2}{4}-\left(\frac{m-1}{m}\right)^2\frac{Q_1{}^4}{4\hat{S}_{Chao2}}.$$

If Q_1=0:

$$\hat{\text{var}}\left(\hat{S}_{Chao2}\right)=S_{obs}e^{-M/S_{obs}}\left(1-e^{-M/S_{obs}}\right),$$

where M is the total number of occurrences of all species in all samples.

Jackknife 1

First-order jackknife (Burnham & Overton 1978, 1979; Heltshe & Forrester 1983):

$$\hat{S}_{jack1}=S_{obs}+\left(\frac{m-1}{m}\right)Q_1.$$

$$\hat{\text{var}}\left(\hat{S}_{jack1}\right)=\left(\frac{m-1}{m}\right)\left(\sum_{j=0}^{S}j^2f_j-\frac{Q_1{}^2}{m}\right),$$

where f_j is the number of samples containing j unique species.

Jackknife 2

Second-order jackknife (Smith & van Belle 1984):

$$\hat{S}_{jack2}=S_{obs}+\frac{Q_1\left(2m-3\right)}{m}-\frac{Q_2\left(m-2\right)^2}{m\left(m-1\right)}.$$

No analytical estimate of variance is available.

Bootstrap

Bootstrap estimator (Smith & van Belle 1984):

$$\hat{S}_{boot}=S_{obs}+\sum_{k=1}^{S_{obs}}\left(1-p_k\right)^m,$$

where p_k is the proportion of samples containing species k. No analytical estimate of variance is available.

References

Burnham, K.P. & W.S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65:623-633.

Burnham, K.P. & W.S. Overton. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology* 60:927-936.

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43, 783-791.

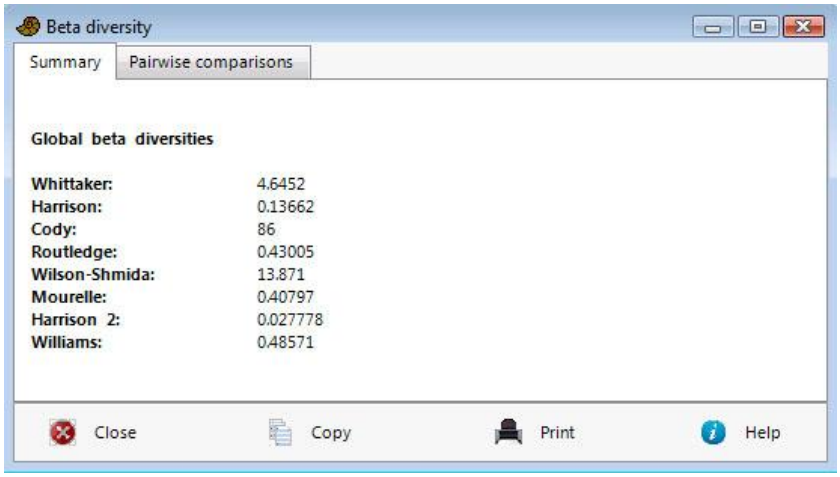
Colwell, R.K. & J.A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society (Series B) 345:101-118.

Heltshe, J. & N.E. Forrester. 1983. Estimating species richness using the jackknife procedure. *Biometrics* 39:1-11.

Smith, E.P. & G. van Belle. 1984. Nonparametric estimation of species richness. *Biometrics* 40:119-129.

Beta diversity

Two or more rows (samples) of presence-absence (0/1) data, with taxa in columns.



The beta diversity module in Past can be used for any number of samples (not limited to only two samples). The eight measures available are described in Koleff et al. (2003):

Past	Koleff et al.	Equation	Ref.
Whittaker	b_w	$\frac{S}{\overline{\alpha}}-1$	Whittaker (1960)
Harrison	b_{-1}	$\frac{\frac{S}{\overline{\alpha}}-1}{N-1}$	Harrison et al. (1992)
Cody	b_c	$\frac{g(H)+l(H)}{2}$	Cody (1975)
Routledge	b_l	$\log_{10}(T)-\left[\frac{1}{T}\sum_i e_i \log_{10}(e_i)\right]-\left[\frac{1}{T}\sum_i \alpha_i \log_{10}(\alpha_i)\right]$	Routledge (1977)
Wilson-Shmida	b_t	$\frac{g(H)+l(H)}{2\overline{\alpha}}$	Wilson & Shmida (1984)
Mourelle	b_{me}	$\frac{g(H)+l(H)}{2\overline{\alpha}(N-1)}$	Mourelle & Ezcurra (1997)
Harrison 2	b_{-2}	$\frac{\frac{S}{\alpha_{\max}}-1}{N-1}$	Harrison et al. (1992)
Williams	b_{-3}	$1-\frac{\alpha_{\max}}{S}$	Williams (1996)

S: total number of species; $\overline{\alpha}$: average number of species; *N*: number of samples; *g*(*H*): total gain of species along gradient (samples ordered along columns); *l*(*H*): total loss of species; e_i: number of samples containing species *i*; *T*: total number of occurrences.

References

Harrison, S., S.J. Ross & J.H. Lawton. 1992. Beta diversity on geographic gradients in Britain. Journal of Animal Ecology 61:151-158.

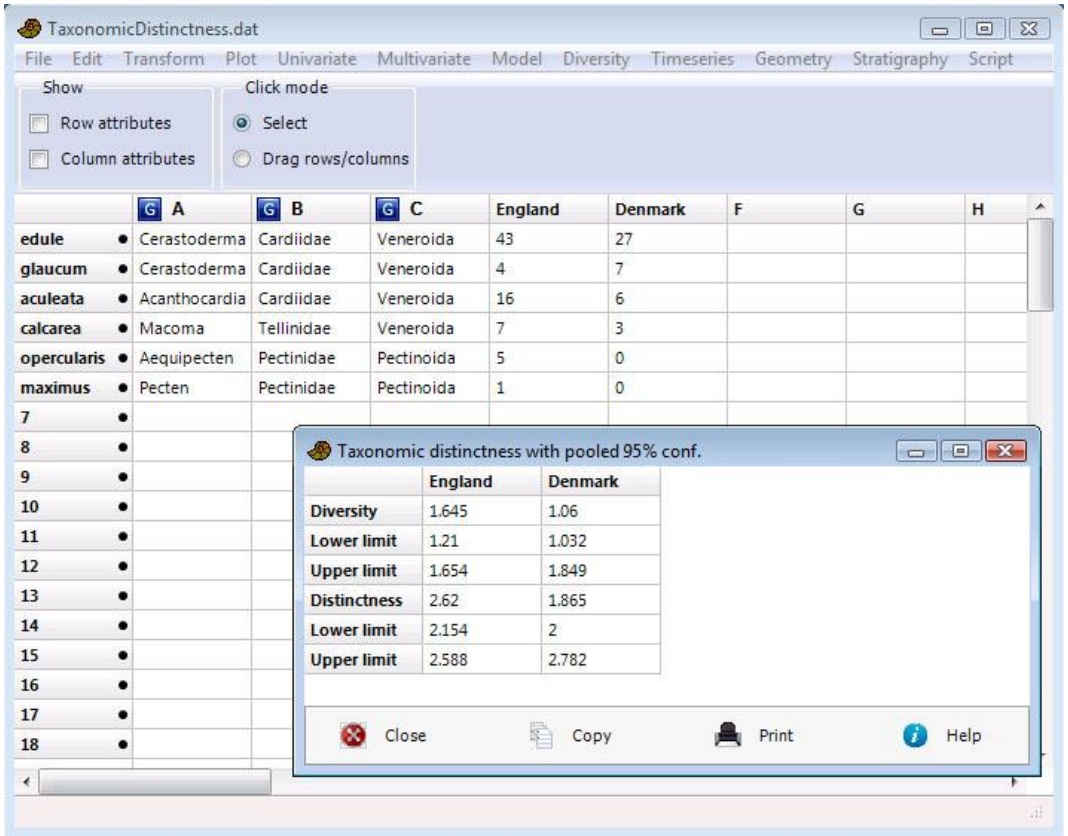
Koleff, P., K.J. Gaston & J.J. Lennon. 2003. Measuring beta diversity for presence-absence data. *Journal of Animal Ecology* 72:367-382.

Routledge, R.D. 1977. On Whittaker's components of diversity. *Ecology* 58:1120-1127.

Whittaker, R.H. 1960. Vegetation of the Siskiyou mountains, Oregon and California. *Ecological Monographs* 30:279-338.

Taxonomic

Requires one or more columns (samples), each containing counts of individuals of different taxa down the rows. In addition, one or more group columns with names of genera/families etc. (see below).



	A	B	C	England	Denmark	F	G	H
edule	Cerastoderma	Cardiidae	Veneroida	43	27			
glaucum	Cerastoderma	Cardiidae	Veneroida	4	7			
aculeata	Acanthocardia	Cardiidae	Veneroida	16	6			
calcarea	Macoma	Tellinidae	Veneroida	7	3			
opercularis	Aequipecten	Pectinidae	Pectinoida	5	0			
maximus	Pecten	Pectinidae	Pectinoida	1	0			

	England	Denmark
Diversity	1.645	1.06
Lower limit	1.21	1.032
Upper limit	1.654	1.849
Distinctness	2.62	1.865
Lower limit	2.154	2
Upper limit	2.588	2.782

Taxonomic diversity and taxonomic distinctness as defined by Clarke & Warwick (1998), including confidence intervals computed from 10,000 random replicates taken from the pooled data set (all samples). Note that the "global list" of Clarke & Warwick is not entered directly, but is calculated internally by pooling (summing) the given samples.

These indices depend on taxonomic information also above the species level, which has to be entered for each species as follows. Species names go in the name column (leftmost; in the Row attributes), genus names in the first group column, family in second group column etc., up to four group columns. Of course you can substitute for other taxonomic levels as long as they are in ascending order. Species counts for the samples follow in the columns thereafter.

Taxonomic diversity in one sample is given by (note other, equivalent forms exist):

$$\Delta = \frac{\sum\sum_{i<j} w_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j + \sum_i x_i (x_i - 1)/2},$$

where the w_{ij} are weights such that w_{ij} = 0 if *i* and *j* are the same species, w_{ij} = 1 if they are the same genus, etc. The *x* are the abundances.

Taxonomic distinctness:

$$\Delta^* = \frac{\sum\sum_{i<j} w_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j}.$$

For presence-absence data, taxonomic diversity and distinctness will be valid but equal to each other.

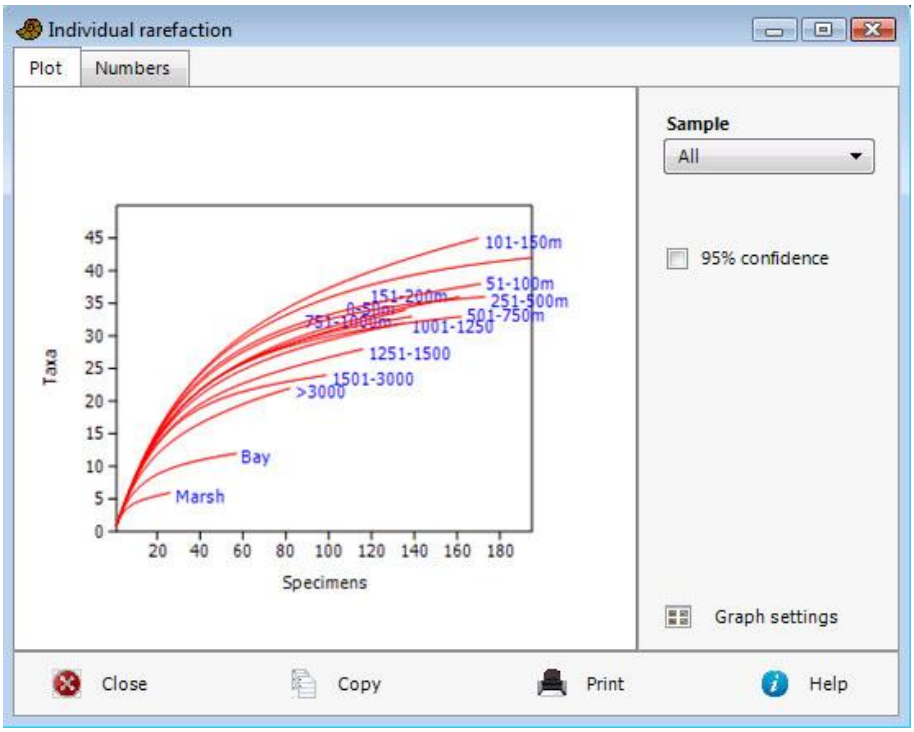
Reference

Clarke, K.R. & Warwick, R.M. 1998. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology* 35:523-531.

Individual rarefaction

For comparing taxonomical diversity in samples of different sizes. Requires one or more columns of counts of individuals of different taxa (each column must have the same number of values). When comparing samples: Samples should be taxonomically similar, obtained using standardised sampling and taken from similar 'habitat'.

Given one or more columns of abundance data for a number of taxa, this module estimates how many taxa you would expect to find in a sample with a smaller total number of individuals. With this method, you can compare the number of taxa in samples of different size. Using rarefaction analysis on a large sample, you can read out the number of expected taxa for any smaller sample size (including that of the *smallest* sample). The algorithm is from Krebs (1989), using a log Gamma function for computing combinatorial terms. An example application in paleontology can be found in Adrain et al. (2000).



Let N be the total number of individuals in the sample, s the total number of species, and N_i the number of individuals of species number i. The expected number of species $E(S_n)$ in a sample of size n and the variance $V(S_n)$ are then given by

$$E(S_n) = \sum_{i=1}^{s} \left[1 - \frac{\binom{N - N_i}{n}}{\binom{N}{n}} \right]$$

$$V(S_n) = \sum_{i=1}^{s} \left[\frac{\binom{N-N_i}{n}}{\binom{N}{n}} \left(1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right) \right]$$
$$+ 2\sum_{j=2}^{s}\sum_{i=1}^{j-1} \left[\frac{\binom{N-N_i-N_j}{n}}{\binom{N}{n}} - \frac{\binom{N-N_i}{n}\binom{N-N_j}{n}}{\binom{N}{n}\binom{N}{n}} \right]$$

Standard errors (square roots of resampling variances) are given by the program. In the graphical plot, these standard errors are converted to 95 percent confidence intervals.

Unconditional variance

The classical rarefaction variance given above is called conditional variance. It is conditional on the reference sample, and will reduce to zero for $S_n = s$. In contrast, Colwell et al. (2012) described an unconditional rarefaction variance estimate that will not reduce to zero at the end of the rarefaction curve. This method is also available in Past.

There are two models for individual rarefaction described by Colwell et al. (2012), the multinomial model (classical rarefaction) and the Poisson model (Coleman rarefaction). The two methods give quite similar results. The "industry standard" rarefaction software, EstimateS, somewhat incongruously computes $E(S_n)$ according to the multinomial equation (eq. (4) in Colwell et al., equivalent to the equation given above), while $V(S_n)$ uses the Poisson formulation (eq. 7 in Colwell et al.), according to the EstimateS manual. This approach is followed in Past for compatibility with EstimateS. The computation also requires an estimate for total (sampled and unsampled) species richness. The Chao1 estimator is used for this (cf. Colwell et al. 2012).

Rarefaction of Simpson and Shannon indices

In addition to rarefaction of species richness, Past also includes rarefaction of the Simpson *D* and Shannon *H* diversity indices, following Chao et al. (2014). In the forms $1/D$ and e^H, these are special cases of so-called Hill numbers, which is a family of diversity indices. Past reports the rarefaction curves in these forms, for consistency with Chao et al. (2014), but for convenience the Shannon index can also be reported as the conventional *H*. Past does not yet compute confidence intervals for these rarefaction curves (which would require bootstrapping).

Let *Xi* be the number of individuals of the *i*th species that are observed in the sample, $i = 1, 2, \ldots, S$. Let f_k be the number of species represented by exactly *k* individuals in the sample, $k = 0, 1, \ldots, n$. For smaller sample sizes ($m<n$), and $k>0$, the estimator for f_k is (eq. 7 in Chao et al. 2014)

$$\hat{f}_k = \sum_{X_i \geq k} \frac{\binom{X_i}{k}\binom{n-X_i}{m-k}}{\binom{n}{m}}$$

The estimator for the Hill number with q=1, which is equivalent to the exponential e^H of the Shannon index, is (eq. 10a in Chao et al. 2014)

$$^{1}\widehat{D}(m) = \exp\left[\sum_{k=1}^{m}\left(-\frac{k}{m}\ln\frac{k}{m}\right)\hat{f}_k(m)\right]$$

The estimator for the Hill number with q=2, which is equivalent to the inverse $1/D$ of the Simpson index, is (eq. 11b in Chao et al. 2014)

$$^{2}\widehat{D}(m) = \frac{1}{\sum_{k=1}^{m}\left(\frac{k}{m}\right)^{2}\hat{f}_k(m)}$$

Missing values: Treated as zero.

References

Adrain, J.M., S.R. Westrop & D.E. Chatterton. 2000. Silurian trilobite alpha diversity and the end-Ordovician mass extinction. *Paleobiology* 26:625-646.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. & Ellison, A.M. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84:45-67.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-L., Mao, C.X., Chazdon, R.L. & Longino, J.T. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* 5:3-21.

Krebs, C.J. 1989. Ecological Methodology. Harper & Row, New York.

Shareholder Quorum Subsampling (SQS)

Shareholder Quorum Subsampling was introduced by John Alroy (2010). Chao and Jost (2012) provided analytical solutions for an almost equivalent method, which they called "Coverage-based rarefaction". Past follows the algorithm of Alroy (2010).

Like rarefaction, SQS can be used to standardize species counts across samples of different size, but standardizing on a fixed "coverage" rather than a fixed sample size. Coverage is defined as the proportion of individuals in the population that are represented by the species recovered in the sample. If all species are recovered in the sample, coverage is 1. If only one species is recovered, but 50% of the individuals in the population belong to this species, coverage is 0.5. SQS gives a more fair sampling of communities with different evenness than classical rarefaction, which suffers from a "compression effect" where differences in richness are artificially dampened. SQS is therefore gradually replacing classical rarefaction in the literature.

The module expects one or more columns of count data. The following parameters can be selected:

Quorum: The desired subsampled coverage level (0-1). This value should be larger than 0.4. A value of e.g. 0.9 might be better, but for small samples this coverage may not be achieved, giving an error message. The quorum value is not critical: Larger values will give higher SQS species richness, but this is of little consequence for the comparison of several samples.

Trials: The number of subsampling runs. Higher values give more exact results, but takes longer.

Ignore singletons and *Ignore dominant*: Disregard species with abundance 1; disregard the single most common species. The possible advantage of this is slightly unclear.

The module outputs the original sample size (N), the number of observed species (*S_obs*), and Good's u which is a simple estimator of sample coverage: $u = 1 - f_1/N$, where f_1 is the number of singletons (Chao & Jost 2012 give an improved estimator), in addition to the SQS richness.

References

Alroy, J. 2010. Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. *Palaeontology* 53:1211–1235.

Chao, A., Jost, L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 93:2533-2547.

Sample rarefaction (Mao's tau)

Sample rarefaction requires a matrix of presence-absence data (abundances treated as presences), with taxa in rows and samples in columns. Sample-based rarefaction (also known as the species accumulation curve) is applicable when a number of samples are available, from which species richness is to be estimated as a function of number of samples. PAST implements the analytical solution known as "Mao's tau", with standard deviation. In the graphical plot, the standard errors are converted to 95 percent confidence intervals. See Colwell et al. (2004) for details.

With H samples and S_{obs} the total number of observed species, let s_j be the number of species found in j samples, such that s_1 is the number of species found in exactly one sample, etc. The total number of species expected in $h \leq H$ samples is then

$$\tilde{\tau}(h) = S_{obs} - \sum_{j=1}^{H} \alpha_{jh} s_j \ .$$

The combinatorial coefficients α are

$$\alpha_{jh} = \begin{cases} \dfrac{(H-h)!(H-j)!}{(H-h-j)!H!} & \text{for } j+h \leq H \\ 0 & \text{for } j+h > H \end{cases} .$$

These coefficients are computed via a log Gamma function. The variance estimator is

$$\tilde{\sigma}^2 = \sum_{j=1}^{H} \left(1-\alpha_{jh}\right)^2 s_j - \frac{\tilde{\tau}^2(h)}{\tilde{S}},$$

where $\tilde{S}$ is an estimator for the unknown total species richness. Following Colwell et al. (2004), a Chao2-type estimator is used. For s_2>0,

$$\tilde{S} = S_{obs} + \frac{(H-1)s_1^2}{2Hs_2} .$$

For s_2=0,

$$\tilde{S} = S_{obs} + \frac{(H-1)s_1(s_1-1)}{2H(s_2+1)} .$$

For modeling and extrapolating the curve using the Michaelis-Menten equation, use the Copy Data button, paste to a new Past spreadsheet, and use the nonlinear fitting module in the Model menu.

Past also includes extrapolation of the sample rarefaction curve, up to 4H samples, using equation 18 in Colwell et al. (2012). Confidence intervals for the extrapolation is not yet implemented.

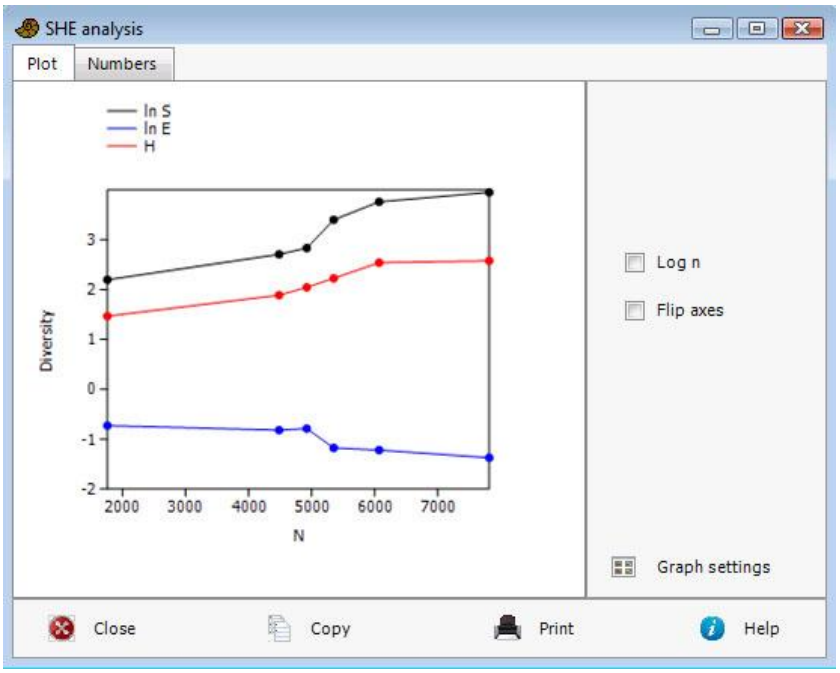
References

Colwell, R.K., C.X. Mao & J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* 85:2717-2727.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L., Longino, J.T. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* 5:3–21

SHE analysis

SHE analysis (Hayek & Buzas 1997, Buzas & Hayek 1998) requires a matrix of integer abundance data (counts), with taxa in rows and samples in columns. The program calculates log species abundance (ln *S*), Shannon index (*H*) and log evenness (ln $E = H - \ln S$) for the first sample. Then the second sample is added to the first, and the process continues. The resulting cumulative SHE profiles can be interpreted ecologically. If the samples are taken not from one homogenous population but across a gradient or up a stratigraphic section, breaks in the curve may be used to infer discontinuities (e.g. biozone boundaries).



References

Buzas, M.A. & L.-A. C. Hayek. 1998. SHE analysis for biofacies identification. *The Journal of Foraminiferal Research* 28:233-239.

Hayek, L.-A. C. & M.A. Buzas. 1997. Surveying natural populations. Columbia University Press.

Diversity permutation test

Expects two columns of abundance data with taxa down the rows.This module computes a number of diversity indices for two samples, and then compares the diversities using random permutations. 999 random matrices with two columns (samples) are generated, each with the same row and column totals as in the original data matrix.

Diversity permutation test

	0-50m	51-100m	Perm p(eq)
Taxa S	34	38	0.331
Individuals	136	171	0
Dominance	0.04336	0.0369	0.104
Shannon H	3.274	3.422	0.104
Evenness e^H/S	0.7772	0.8058	0.525
Simpson indx	0.9566	0.9631	0.104
Menhinick	2.915	2.906	0.961
Margalef	6.717	7.196	0.492
Equitability J	0.9285	0.9406	0.411
Fisher alpha	14.55	15.15	0.862
Berger-Parker	0.06618	0.05848	0.552

Close Copy Print Help

Diversity *t* test

Comparison of the Shannon and Simpson diversities in two samples. Then Shannon *t* test is described by e.g. Hutcheson (1970), Poole (1974), Magurran (1988). This is an alternative to the randomization test available in the *Diversity permutation test* module. Requires two columns of abundance data with taxa down the rows.

The **Shannon index** here include a bias correction and may diverge slightly from the uncorrected estimates calculated elsewhere in PAST, at least for small samples. With p_i the proportion (0-1) of taxon *i*, *S* the number of taxa and *N* the number of individuals, the estimator of the index is

$$H' = -\sum_{i=1}^{S} p_i \ln p_i - \frac{S-1}{2N}$$ (note that the second term is incorrect in Magurran 1988).

The variance of the estimator is

$$\operatorname{Var} H' = \frac{\sum p_i (\ln p_i)^2 - \left[\sum (p_i \ln p_i)\right]^2}{N} + \frac{S-1}{2N^2}.$$

The *t* test statistic is given by

$$t = \frac{H_1' - H_2'}{\sqrt{\operatorname{Var} H_1' + \operatorname{Var} H_2'}}.$$

The degrees of freedom for the *t* test is

$$df = \frac{\left(\operatorname{Var} H_1' + \operatorname{Var} H_2'\right)^2}{\frac{\left(\operatorname{Var} H_1'\right)^2}{N_1} + \frac{\left(\operatorname{Var} H_2'\right)^2}{N_2}}.$$

The **Simpson index** (dominance) has estimated variance (Brower et al. 1998):

$$\operatorname{Var} D = \frac{4N(N-1)(N-2)\sum p_i^3 + 2N(N-1)\sum p_i^2 - 2N(N-1)(2N-3)\left(\sum p_i^2\right)^2}{N^2(N-1)^2}.$$

References

Brower, J.E., Zar, J.H., von Ende, C.N. 1998. *Field and Laboratory Methods for General Ecology.* McGraw-Hill, Boston.

Hutcheson, K. 1970. A test for comparing diversities based on the Shannon formula. *Journal of Theoretical Biology* 29:151-154.

Magurran, A. 1988. Ecological Diversity and Its Measurement. Princeton University Press.

Poole, R.W. 1974. An introduction to quantitative ecology. McGraw-Hill, New York.

Diversity profiles

This module requires one or more columns of abundance data with taxa down the rows. The main purpose is to compare diversities in several samples.

The validity of comparing diversities across samples can be criticized because of arbitrary choice of diversity index. One sample may for example contain a larger number of taxa, while the other has a larger Shannon index. A number of diversity indices may be compared to make sure that the diversity ordering is robust. A formal way of doing this is to define a family of diversity indices, dependent upon a single continuous parameter (Tothmeresz 1995).

PAST uses the exponential of the so-called Renyi index, which depends upon a parameter α. For α=0, this function gives the total species number. α=1 (in the limit) gives an index proportional to the Shannon index, while α=2 gives an index which behaves like the Simpson index.

$$\exp(H_\alpha) = \exp\left(\frac{1}{1-\alpha}\ln\sum_{i=1}^{S} p_i^\alpha\right)$$

The program can plot several such diversity profiles together. If the profiles cross, the diversities are non-comparable. The bootstrapping option (giving a 95% confidence interval) is based on 2000 replicates.

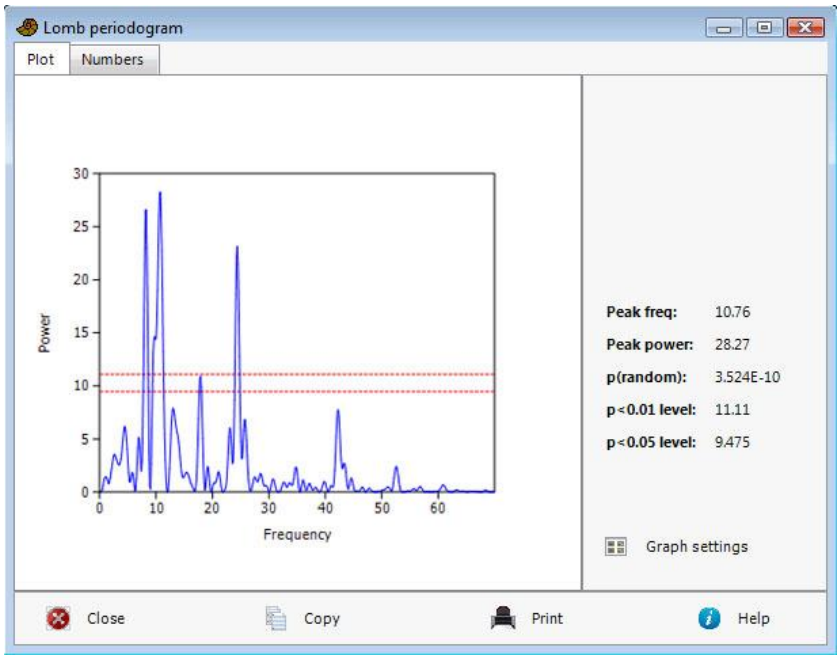
Reference

Tothmeresz, B. 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science* 6:283-290.

Time series menu

Simple periodogram

Since palaeontological data are often unevenly sampled, Fourier-based methods can be difficult to use. PAST therefore includes the Lomb periodogram for unevenly sampled data (Press et al. 1992), with time values given in the first column and dependent values in the second column. If only one column is selected, even spacing of one unit between data points is assumed. The Lomb periodogram should then give similar results as the FFT. The data are automatically detrended prior to analysis.



The frequency axis is in units of *1/(x unit)*. If for example, your *x* values are given in millions of years, a frequency of 0.1 corresponds to a period of 10 million years. The power axis is in units proportional to the square of the amplitudes of the sinusoids present in the data. Also note that the frequency axis extends to very high values. If your data are evenly sampled, the upper half of the spectrum is a mirror image of the lower half, and is of little use. If some of your regions are closely sampled, the algorithm may be able to find useful information even above the half-point (Nyquist frequency).

The highest peak in the spectrum is presented with its frequency and power value, together with a probability that the peak could occur from random data. The 0.01 and 0.05 significance levels ('white noise lines') are shown as red dashed lines.

The example above shows a spectral analysis of a foram oxygen isotope record from 1 Ma to Recent, with an even spacing of 0.003 Ma (3 ka). There are periodicities at frequencies of about 9 (split peak), 25 and 43 Ma^{-1}, corresponding to periods of 111 ka, 40 ka and 23 ka – clearly orbital forcing.

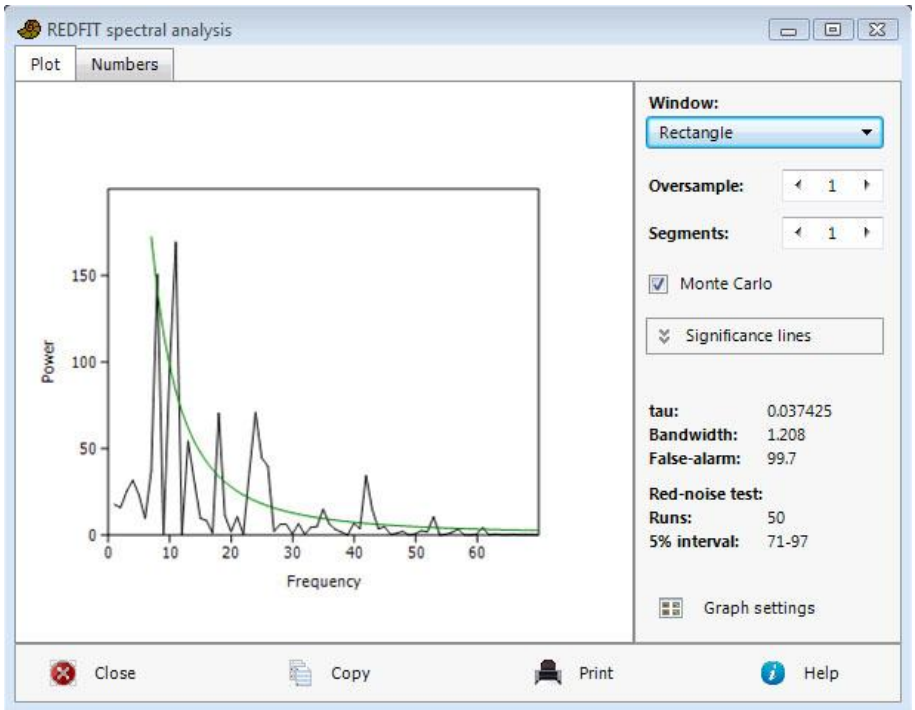
Reference

Press, W.H., S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. 1992. Numerical Recipes in C. Cambridge University Press.

REDFIT spectral analysis

This module is an implementation of the REDFIT procedure of Schulz and Mudelsee (2002). It is a more advanced version of the simple Lomb periodogram described above. REDFIT includes an option for "Welch overlapped segment averaging", which implies splitting the time series into a number of segments, overlapping by 50%, and averaging their spectra. This reduces noise but also reduces spectral resolution. In addition, the time series is fitted to an AR(1) red noise model which is usually a more appropriate null hypothesis than the white noise model described above. The given "false-alarm lines" are based on both parametric approximations (chi2) and Monte Carlo (using 1000 random realizations of an AR(1) process).

The input must be in the form of two columns with time and data values, or one column of equally-spaced data values. The data are automatically detrended. The fitting to AR(1) implies that the data must have the correct time direction (in contrast with the simple spectrogram above where the time direction is arbitrary). The time values are expected to be ages before present. If not, it will be necessary to give them negative signs.



The frequency oversampling value controls the number of points along the frequency axis (but having many points does not increase frequency resolution!). Increasing the number of segments will reduce noise, but also decrease the resolution. The window function influences the trade-off between spectral resolution and attenuation of side lobes.

The (average) tau value is the characteristic time scale (the parameter of the AR model). The bandwidth is the spectral resolution given as the width between the -6dB points.

The fit to an AR(1) model can be assessed using the runs value and its 5% acceptance interval. This test is only available with Monte Carlo on, oversampling=1, segments=1, window=rectangular.

In addition to a fixed set of false-alarm levels (80%, 90%, 95% and 99%), the program also reports a "critical" false-alarm level (False-al) that depends on the segment length (Thomson 1990).

Important: Because of long computation time, the Monte Carlo simulation is not run by default, and the Monte Carlo false-alarm levels are therefore not available. When the Monte Carlo option is enabled, the given spectrum may change slightly because the Monte Carlo results are then used to compute a "bias-corrected" version (see Schulz and Mudelsee 2002).

Missing values supported.

References

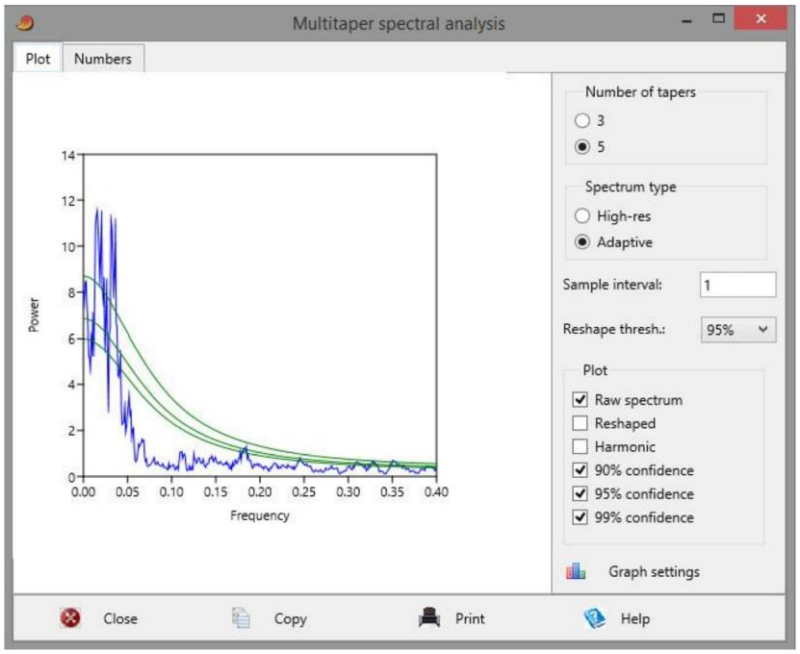
Schulz, M. & M. Mudelsee. 2002. REDFIT: estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. *Computers & Geosciences* 28:421-426.

Thomson, D.J. 1990. Time series analysis of Holocene climate data. *Philosophical Transactions of the Royal Society of London, Series A* 330:601-616.

Multitaper spectral analysis

In traditional spectral estimation, the data are often "windowed" (multiplied with a bell-shaped function) to reduce spectral leakage. In the multitaper method, several orthogonal window functions are applied, and the results combined. The resulting spectrum has low leakage, low variance, and retains information from the beginning and end of the time series. Also, statistical testing can take advantage of the multiple spectral estimates. A possible disadvantage is reduced spectral resolution.

Requires evenly spaced data, given in one column. The data are not detrended.



The implementation is based on Mann & Lees (1996) and Fortran code by Michael Mann. It includes a red-noise model based on a "reshaped" spectrum, i.e. after removing and interpolating peaks.

The number of tapers can be set to 3 or 5 for different trade-offs between variance and resolution.

Spectrum type: The two algorithms give very similar results, but the "adaptive" option is recommended by Mann & Lees (1996).

Sample interval: This value only affects the scaling of the frequency axis.

Reshape threshold: This value affects how strong a spectral peak must be to count as a harmonic component, to be removed by the reshaping procedure.

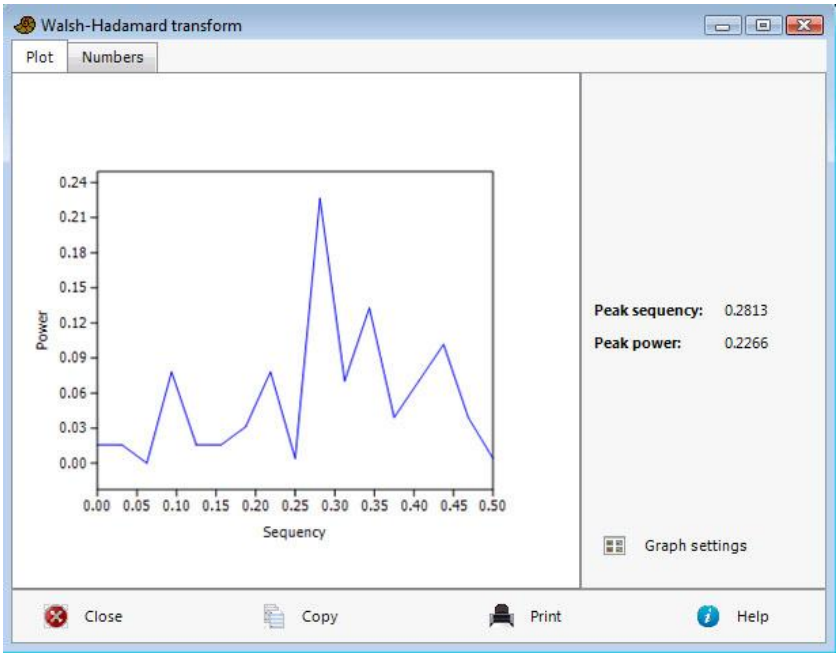
Note: Past reproduces examples in Mann & Lees (1996), but some other implementations seem to give higher levels for the confidence (significance) lines. The reason for this is unknown.

Reference

Mann, M.E. & Lees. J. 1996. Robust estimation of background noise and signal detection in climatic time series. *Climatic Change* 33:409-445.

Walsh transform

The Walsh transform is a type of spectral analysis (for finding periodicities) of binary or ordinal data. It assumes even spacing of data points, and expects one column of binary (0/1) or ordinal (integer) data.



The normal methods for spectral analysis are perhaps not optimal for binary data, because they decompose the time series into sinusoids rather than "square waves". The Walsh transform may then be a better choice, using basis functions that flip between -1 and +1. These basis functions have varying "frequencies" (number of transitions divided by two), known as *sequencies*. In PAST, each pair of even ("cal") and odd ("sal") basis functions is combined into a power value using cal^2+sal^2, producing a "power spectrum" that is comparable to the Lomb periodogram.

In the example above, compare the Walsh periodogram (top) to the Lomb periodogram (bottom). The data set has 0.125 periods per sample. Both analyses show harmonics.

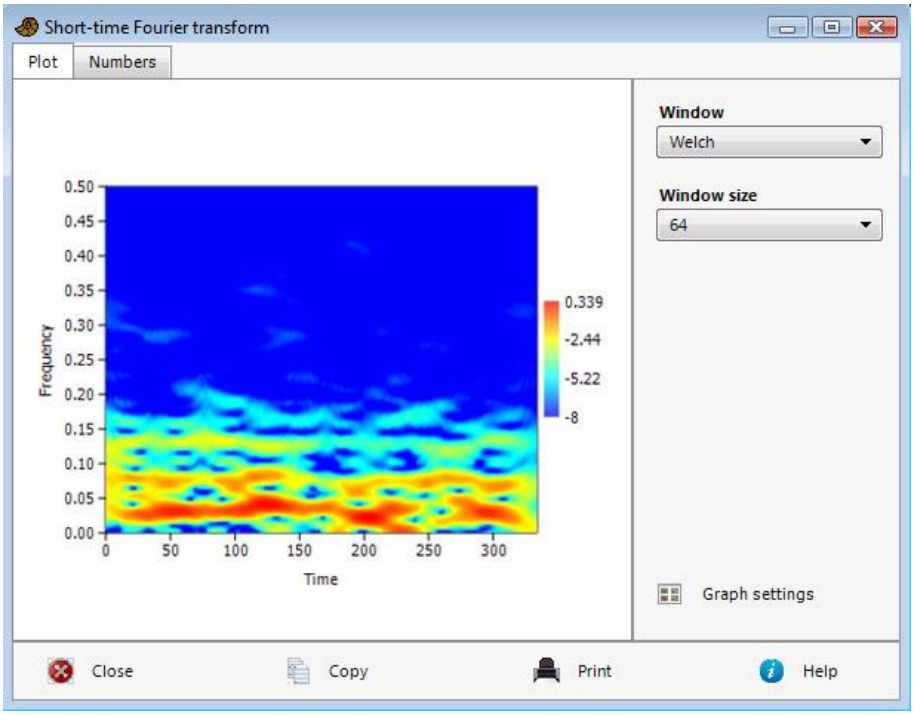
The Walsh transform is slightly exotic compared with the Fourier transform, and the results must be interpretated cautiously. For example, the effects of the duty cycle (percentage of ones versus zeros) are somewhat difficult to understand.

In PAST, the data values are pre-processed by multiplying with two and subtracting one, bringing 0/1 binary values into the -1/+1 range optimal for the Walsh transform. The data are zero-padded to the next power of 2 if necessary, as required by the method.

Evolutive Fourier transform

Spectral analysis using the Fourier transform (FFT), but dividing the signal into a sequence of overlapping windows, which are analysed individually. This allows development of the spectrum in time, in contrast with the global analysis provided by the other spectral analysis modules. Sample position is shown on the *x* axis, frequency (in periods per sample) on the *y* axis, and power on a logarithmic scale as colour or grey scale.

The Short-time Fourier Transform (STFT) can be compared with wavelet analysis, but with a linear frequency scale and with constant time resolution independent of frequency.



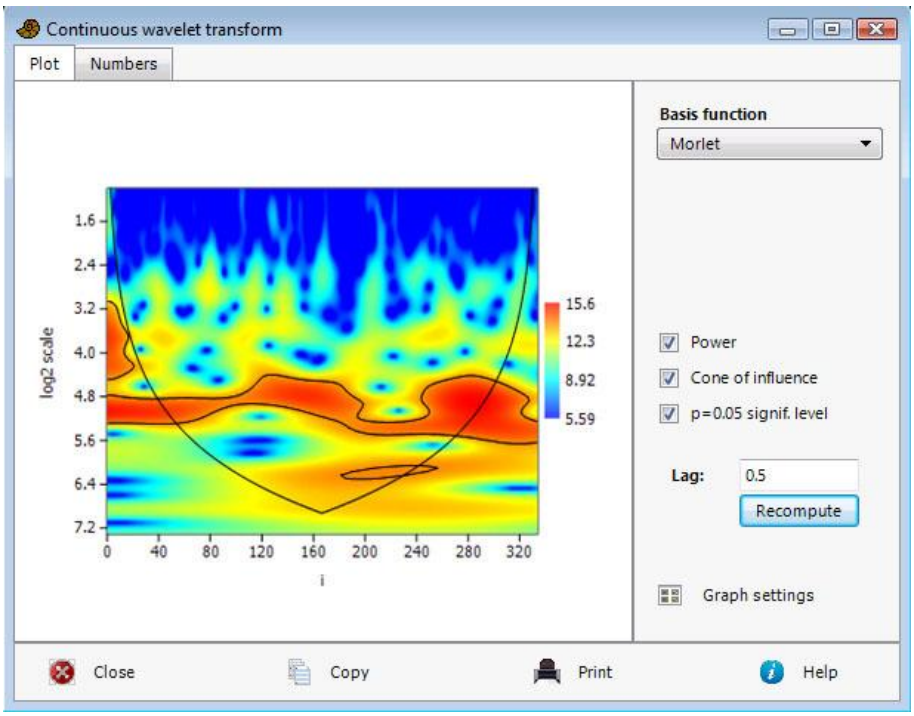
The window size controls the trade-off between resolution in time and frequency; small windows give good time resolution but poor frequency resolution. Windows are zero-padded by a factor eight to give a smoother appearance of the diagram along the frequency axis. The window functions (Rectangle, Welch, Hanning, Hamming, Blackman-Harris, multitaper with 3, 4 or 5 tapers) give different trade-offs between frequency resolution and sideband rejection.

Missing values are treated using linear interpolation before analysis.

Wavelet transform

Inspection of time series at different scales. Requires one column of ordinal or continuous data with even spacing of points.

The continuous wavelet transform (CWT) is an analysis method where a data set can be inspected at small, intermediate and large scales simultaneously. It can be useful for detecting periodicities at different wavelengths, self-similarity and other features. The vertical axis in the plot is a logarithmic size scale (base 2), with the signal observed at a scale of only two consecutive data points at the top, and at a scale of one fourth of the whole sequence at the bottom. One unit on this axis corresponds to a doubling of the size scale. The top of the figure thus represents a detailed, fine-grained view, while the bottom represents a smoothed overview of longer trends. Signal power (or more correctly squared correlation strength with the scaled mother wavelet) is shown with a grayscale or in colour.



The shape of the mother wavelet can be set to Morlet (wavenumber 6), Paul (4th order) or DOG (Derivative Of Gaussian, 2nd or 6th derivative). The Morlet wavelet usually performs best.

The example above is based on a foram oxygen isotope record from 1 Ma to Recent, with an even spacing of 0.003 Ma (3 ka). A band can be seen at a scale of about 2^5=32 samples, or about 100 ka. A weaker band around $2^{3.7}$=13 samples corresponds to a scale of about 40 ka. These are orbital periodicities. In contrast with the "bulk" spectral analysis, the scalogram makes visible changes in strength and frequency over time.

The so-called "cone of influence" can be plotted to show the region where boundary effects are present.

The 'Sample interval' value can be set to a value other than 1. This will only influence the scaling of the labels on the *x* and *y* axes.

The algorithm is based on fast convolution of the signal with the wavelet at different scales, using the FFT.

Significance test: The significance level corresponding to p=0.05 can be plotted as a contour (chi-squared test according to Torrence & Compo 1998). The "Lag" value, as given by the user, specifies the null hypothesis. Lag=0 specifies a white-noise model. Values 0<Lag<1 specifies a red-noise model with the given MA(1) autocorrelation coefficient. It can be estimated using the ARMA module in the Time menu (specify zero AR terms and one MA term and note the MA value in the Coefficients tab).

If the "Power" option is deselected, the program will show only the real part of the scalogram (not squared). This shows the signal in the time domain, filtered at different scales.

In the 'View numbers' window, each row shows one scale, with sample number (position) along the columns.

The 'Filter' tab shows the time series at one scale value, as power values if the 'Power' option is selected in the main tab, or real parts if not. This, in effect, works as a bandpass filter.

The wavelet transform was used by Prokoph et al. (2000) for illustrating cycles in diversity curves for planktic foraminifera. The code in Past is based on Torrence & Compo (1998).

Missing values are treated using linear interpolation before analysis.

References

Prokoph, A., A.D. Fowler & R.T. Patterson. 2000. Evidence for periodicity and nonlinearity in a high-resolution fossil record of long-term evolution. *Geology* 28:867-870.

Torrence, C. & G.P. Compo. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* 79:61-78.

Wavelets for unequal spacing

Inspection of time series at different scales. Requires two input columns, containing time and data values.

This module is similar to the Wavelet Transform module, but accepts unequally spaced data. It does not (yet) provide a significance test, and because it does not use the FFT it is considerably slower. Also, the frequency axis is linear, not logarithmic. The algorithm is based on the Weighted Wavelet z-transform (WWZ) of Foster (1996) and the Fortran implementation of Templeton (2004). Please note that this module does not do *magic* – in intervals with little data the wavelet analysis will not be informative, especially at high frequencies.

The *c* parameter of Foster (1996) is fixed at $c = 1/72$, slightly higher than the recommended $c = 1/8\pi$.

This value is chosen because it corresponds to a wavenumber of 6 as used by the Wavelet Transform module.

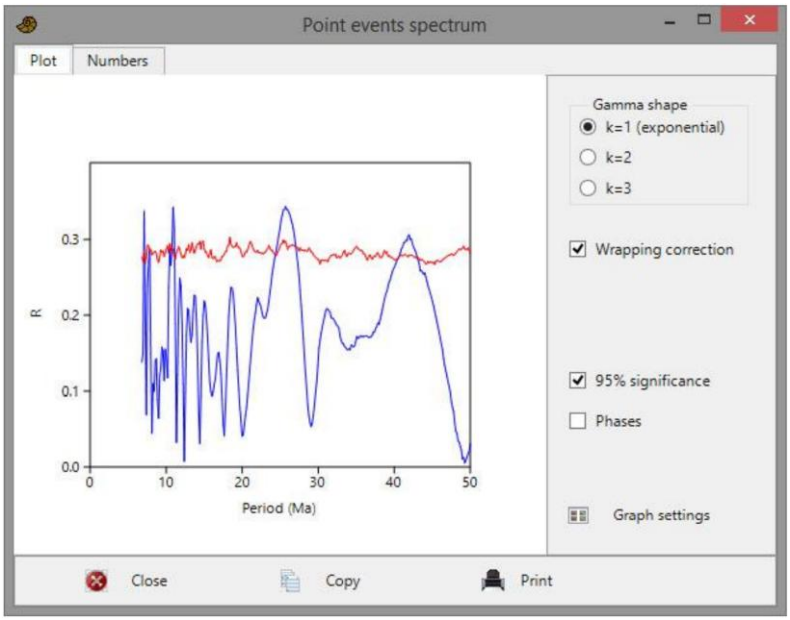
References

Foster, G. 1996. Wavelets for period analysis of unevenly sampled time series. *The Astronomical Journal* 112:1709-1729.

Templeton, M. 2004. Time-series analysis of variable star data. *The Journal of the American Association of Variable Star Observers* 32:41-54.

Point events spectrum

This module, using the "circular spectral analysis" method (e.g. Lutz 1985) is used to search for periodicity in point event series such as earthquakes, volcanic eruptions, and mass extinctions (e.g. Rampino & Caldeira 2015). A single column of event times (e.g. dates of eruptions in millions of years) is required. The event times do not need to be in sequential order.



The method works by wrapping the time line around a circle with a circumference corresponding to a trial period *P*. If points are *P*-periodic, they will cluster at a certain angle corresponding to the phase.

The ages of events t_i are converted to angles a_i:

$$a_i = \frac{2\pi t_i}{P} \bmod 2\pi$$

As in directional statistics, the mean sines and cosines are computed and converted to a mean vector magnitude (Rayleigh statistic) *R* and a phase t_0:

$$S = \frac{1}{N}\sum \sin a_i$$

$$C = \frac{1}{N}\sum \cos a_i$$

$$R = \sqrt{S^2 + C^2}$$

$$t_0 = \frac{P}{2\pi}\tan^{-1}\frac{S}{C}$$

(taken to the correct quadrant)

R and t_0 are computed for *P* ranging from the average waiting time up to 1/3 of the total duration of the series, giving a full spectrum.

A 95% significance line for *R* (in red) is computed by a Monte Carlo procedure with 1,000 replicates. Random event times are computed by a gamma distribution for waiting times. The shape parameter should be set to *k*=1 (i.e. exponential distribution) for a null model with no interactions between events (Poisson process). If closely spaced points are expected to be rare, you can set *k*=2 or *k*=3.

Wrapping correction: Lutz (1985) described a correction for non-integer number of wraps causing some points to be over-represented. This correction, optional in Past, gives a jagged appearance of the spectral curve and seems to work best for relatively large numbers of points (*N*>20).

Harmonics: This method is as plagued by harmonics as traditional Fourier analysis. A spectral peak for a period *P* will be accompanied by strong peaks also on harmonics, i.e. at *P*/2, *P*/3 etc. It is important to take this into account when interpreting the spectrum.

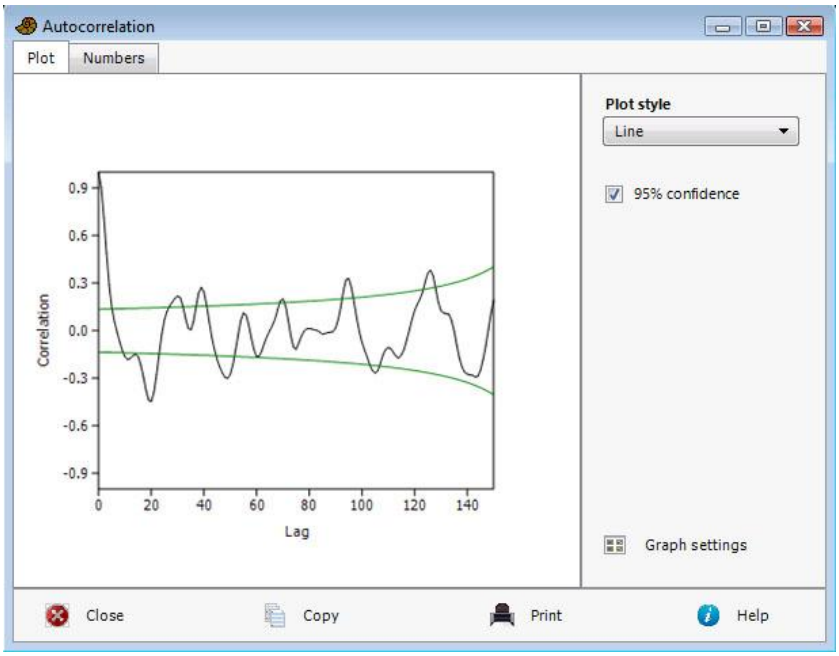
References

Lutz, T.M. 1985. The magnetic reversal record is not periodic. *Nature* 317:404-407.

Rampino, M.R. & K. Caldeira. 2015. Periodic impact cratering and extinction levels over the last 260 million years. *Monthly Notices of the Royal Astronomical Society* 454:3480-3484.

Autocorrelation

Autocorrelation (Davis 1986) is carried out on separate column(s) of *evenly sampled* temporal/stratigraphic data. Lag times τ up to $n/2$, where n is the number of values in the vector, are shown along the x axis (positive lag times only - the autocorrelation function is symmetrical around zero). A predominantly zero autocorrelation signifies random data - periodicities turn up as peaks.



The "95 percent confidence interval" option will draw lines at

$$\pm 1.76\sqrt{\frac{1}{n-\tau+3}}$$

after Davis (1986). This is the confidence interval for random, independent points (white noise). There are two issues: White noise is an unrealistic null model, and the confidence interval is only strictly valid at each *individual* lag (multiple testing problem).

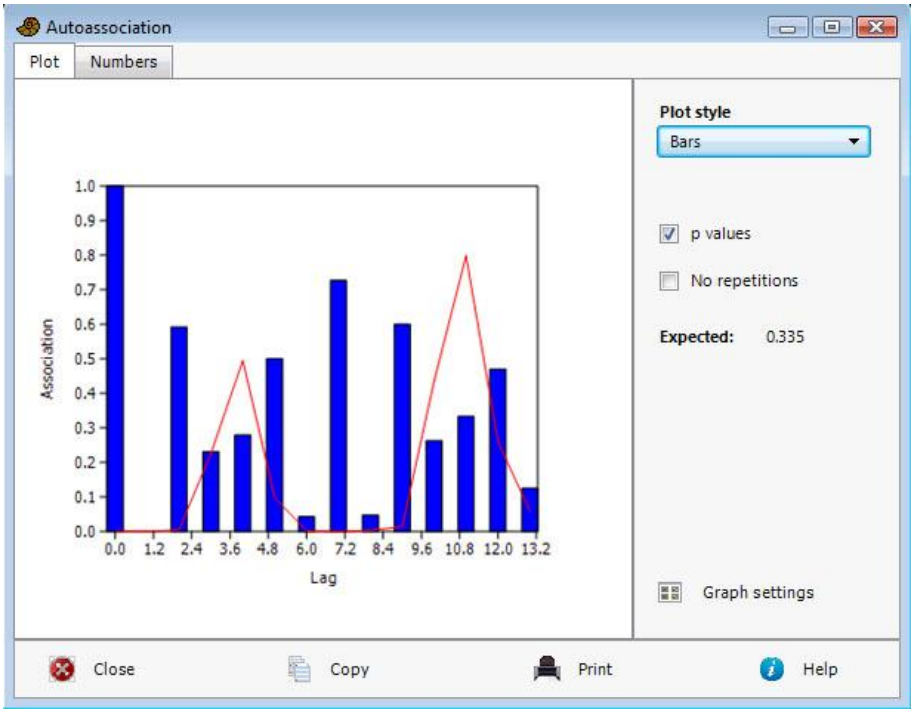
Missing data supported.

Reference

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Autoassociation

Autoassociation is analogous to autocorrelation, but for a sequence of binary or nominal data coded as integer numbers.



For each lag, the autoassociation value is simply the ratio of matching positions to total number of positions compared. The expected autoassociation value (0.335 in the example above) for a random sequence is (Davis 1986)

$$P = \frac{\sum_{k=1}^{m} X_k^2 - n}{n^2 - n}$$

where n is the total number of positions, m is the number of distinct states (3 in the example above) and X_k is the number of observations in state k.

For non-zero lags, a P value is computed from the overlapping positions only, and the expected number of matches is then given by $E=nP$. This is compared with the observed number of matches O to produce a χ^2 with 1 degree of freedom:

$$\chi^2 = \frac{(O - E - 1/2)^2}{E} + \frac{(O' - E' - 1/2)^2}{E'}$$

with $O'=n-O$ and $E'=n(1-P)$ the observed and expected number of mismatches. Note the Yates' correction. The resulting p values (two-tailed) can be shown as a function of lag.

The multiple testing issue arises for the set of p values.

The test above is not strictly valid for “transition” sequences where repetitions are not allowed (the sequence in the example above is of this type). In this case, select the “No repetitions” option. The *p* values will then be computed by an exact test, where all possible permutations without repeats are computed and the autoassociation compared with the original values (on-tailed). This test will take a long time to run for $n>30$, and the option is not available for $n>40$.

Missing data supported.

Reference

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Cross-correlation

Cross-correlation (Davis 1986)is carried out on two column(s) of *evenly sampled* temporal/stratigraphic data. The x axis shows the displacement of the second column with respect to the first, the y axis the correlation between the two time series for a given displacement. The "p values" option will draw the significance of the correlation, after Davis (1986).

For two time series **x** and **y**, the cross-correlation value at lag time *m* is

$$r_m = \frac{\sum (x_i - \bar{x})(y_{i-m} - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_{i-m} - \bar{y})^2}} .$$

The summations and the mean values are only taken over the parts where the sequences overlap for a given lag time.

The equation shows that for positive lags, **x** is compared with a **y** that has been delayed by *m* samples. A high correlation value at positive lags thus means that features in **y** are leading, while **x** lags behind. For negative lags, features in **x** are leading. A reminder of this is given by the program.

The *p* value for a given *m* is given by a *t* test with *n*-2 degrees of freedom, with *n* the number of samples that overlap:

$$t = r_m \sqrt{\frac{n-2}{1-r_m^2}} .$$

It is important to note that this test concerns *one particular m*. Plotting *p* as a function of all *m* raises the issue of multiple testing – *p* values smaller than 0.05 are expected for 5% of lag times even for completely random (uncorrelated) data sets.

In the example above, the “earthquakes” data seem to lag behind the “injection” data with a delay of 0-2 samples (months in this case), where the correlation values are highest. The *p* values (red curve) indicates significance at these lags. Curiously, there also seems to be significance for negative correlation at large positive and negative lags.

Missing data supported.

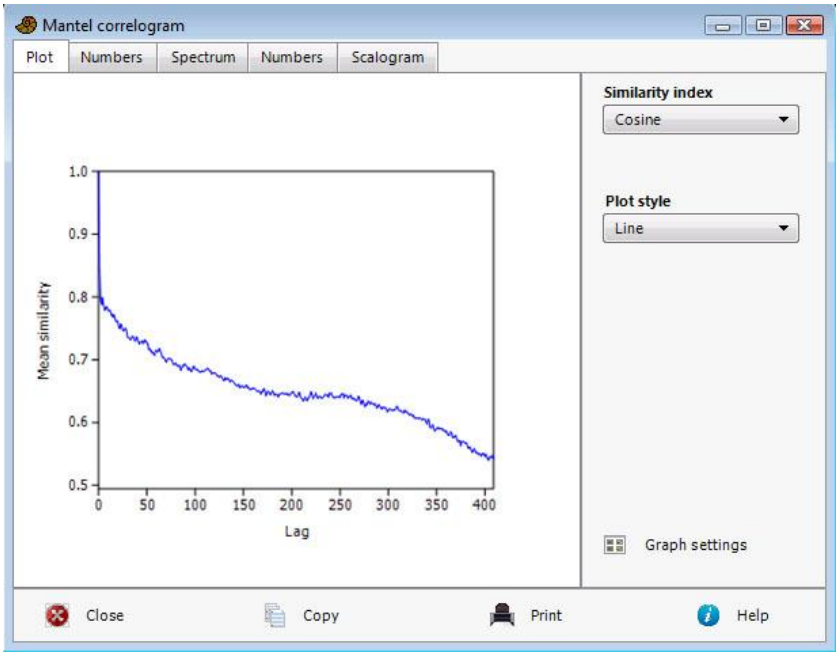
Reference

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Mantel correlogram (and periodogram)

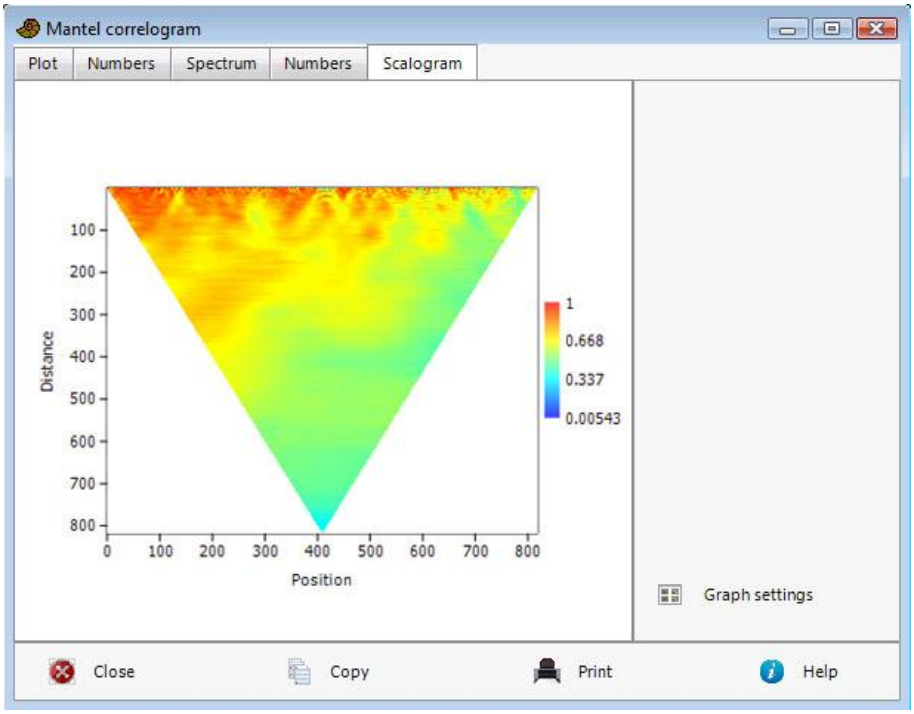
This module expects several rows of multivariate data, one row for each sample. Samples are assumed to be evenly spaced in time.

The Mantel correlogram (e.g. Legendre & Legendre 1998) is a multivariate extension to autocorrelation, based on any similarity or distance measure. The Mantel correlogram in PAST shows the average similarity between the time series and a time lagged copy, for different lags.



The Mantel periodogram is a power spectrum of the multivariate time series, computed from the Mantel correlogram (Hammer 2007).

The Mantel scalogram is an experimental plotting of similarities between all pairs of points along the time series. The apex of the triangle is the similarity between the first and last point. The base of the triangle shows similarities between pairs of consecutive points.



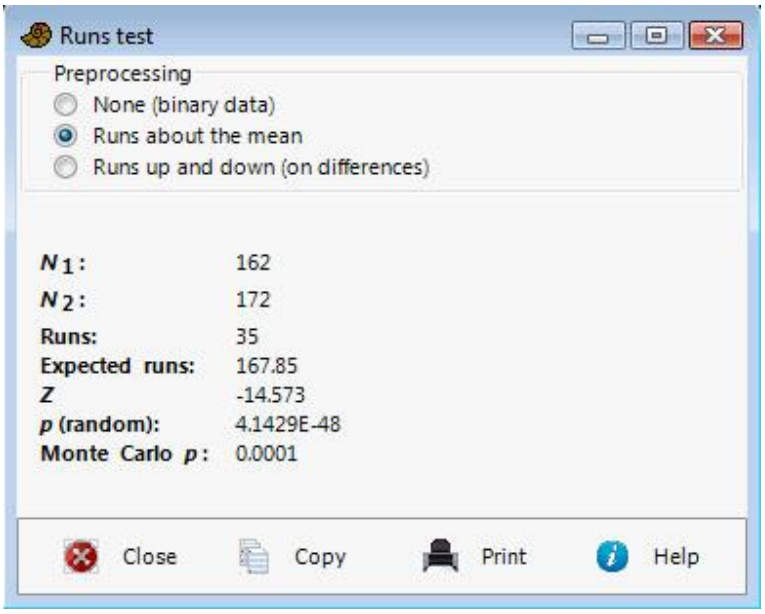
References

Hammer, Ø. 2007. Spectral analysis of a Plio-Pleistocene multispecies time series using the Mantel periodogram. *Palaeogeography, Palaeoclimatology, Palaeoecology* 243:373-377.

Legendre, P. & L. Legendre. 1998. Numerical Ecology, 2nd English ed. Elsevier, 853 pp.

Runs test

The runs test is a non-parametric test for randomness in a sequence of values such as a time series. Non-randomness may include such effects as autocorrelation, trend and periodicity. The module requires one column of data, which are internally converted to 0 ($x \leq 0$) or 1 ($x > 0$).



The test is based on a dichotomy between two values ($x \leq 0$ or $x > 0$). It counts the number of runs (groups of consecutive equal values) and compares this to a theoretical value. The runs test can therefore be used directly for sequences of binary data. There are also options for "runs about the mean" (the mean value subtracted from the data prior to testing), or "runs up and down" (the differences from one value to the next taken before testing).

With n the total number of data points, n_1 the number of points ≤0 and n_2 the number of points >0, the expected number of runs in a random sequence, and the variance, are

$$E(R) = \frac{n + 2n_1 n_2}{n}.$$

$$Var(R) = \frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}.$$

With the observed number of runs R, a z statistic can be written as

$$z = \frac{R - E(R)}{\sqrt{Var(R)}}.$$

The resulting two-tailed p value is not accurate for $n<20$. A Monte Carlo procedure is therefore also included, based on 10,000 random replicates using the observed n, n_1 and n_2.

Mann-Kendall trend test

A non-parametric test for trend. Requires a single column of data. Missing values are deleted, and n adjusted accordingly. The procedure follows Gilbert (1987).

Data x_1, ... x_n are assumed to be ordered in sequence of collection time, or in spatial sequence. Define the indicator function

$$\operatorname{sgn} x = \begin{cases} 1, \text{if } x > 0 \\ 0, \text{if } x = 0 \\ 1, \text{if } x > 0 \end{cases}$$

The *S* statistic is calculated by summing over all pairs of values:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_j - x_i)$$

S will be negative for a negative trend, zero for no trend, and positive for an increasing trend.

For $n \leq 10$, the *p* value is taken from a table of exact values (Gilbert 1987). For $n > 10$, a normal approximation is used, as follows.

Determine the total number of groups of ties *g* and the number of tied values t_j within each group, in the sorted sequence. Then estimate the standard deviation of *S* by

$$SD = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} t_j(t_j - 1)(2t_j + 5)\right]}$$

The *Z* statistic is then

$$Z = \frac{|S| - 1}{SD} \operatorname{sgn} S$$

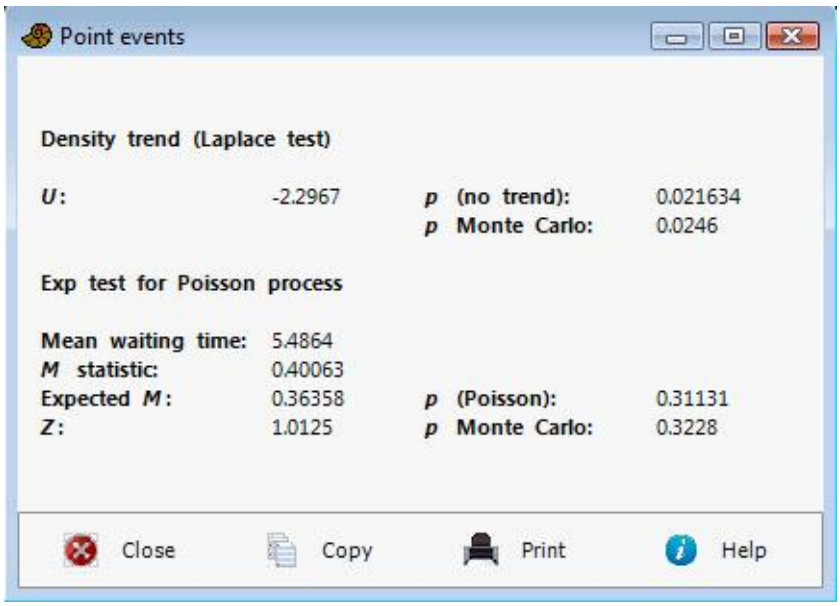
which is used to calculate *p* from the cumulative normal distribution as usual. The subtraction of 1 is a continuity correction.

Reference

Gilbert, R.O. 1987. Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold, New York.

Point events

Expects one column containing times of events (e.g. earthquakes or clade divergences) or positions along a line (e.g. a transect). The times do not have to be in increasing order.



Density trend (Laplace test)

The "Laplace" test for a trend in density (intensity) is described by Cox & Lewis (1978). It is based on the test statistic

$$U = \frac{\bar{t} - \frac{L}{2}}{L\sqrt{\frac{1}{12n}}}$$

where $\bar{t}$ is the mean event time, n the number of events and L the length of the interval. L is estimated as the time from the first to the last event, plus the mean waiting time. U is approximately normally distributed with zero mean and unit variance under the null hypothesis of constant intensity. This is the basis for the given p value.

If $p<0.05$, a positive U indicates an increasing trend in intensity (decreasing waiting times), while a negative U indicates a decreasing trend. Note that if a trend is detected by this test, the sequence is not stationary and the assumptions of the exp test below are violated.

Exp test for Poisson process

The exp test (Prahl 1999) for a stationary Poisson process (random, independent events) is based on the set of n waiting times Δt_i between successive events in the sorted sequence . The test statistic is:

$$M = \frac{1}{n}\sum_{\Delta t_i < T}\left(1 - \frac{\Delta t_i}{T}\right)$$

where T is the mean waiting time. M will tend to zero for a regularly spaced (overdispersed) sequence, and to 1 for a highly clustered sequence. For the null hypothesis of a Poisson process, M is asymptotically normally distributed with mean $1/e - \alpha/n$ and standard deviation $\beta/\sqrt{n}$, where α=0.189 and β=0.2427. This is the basis for the given z test.

In summary, if p<0.05 the sequence is not Poisson. You can then inspect the M statistic; if smaller than the expected value this indicates regularity, if higher it indicates clustering.

For both tests, p values are also estimated by Monte Carlo simulation with 9999 random data sets.

References

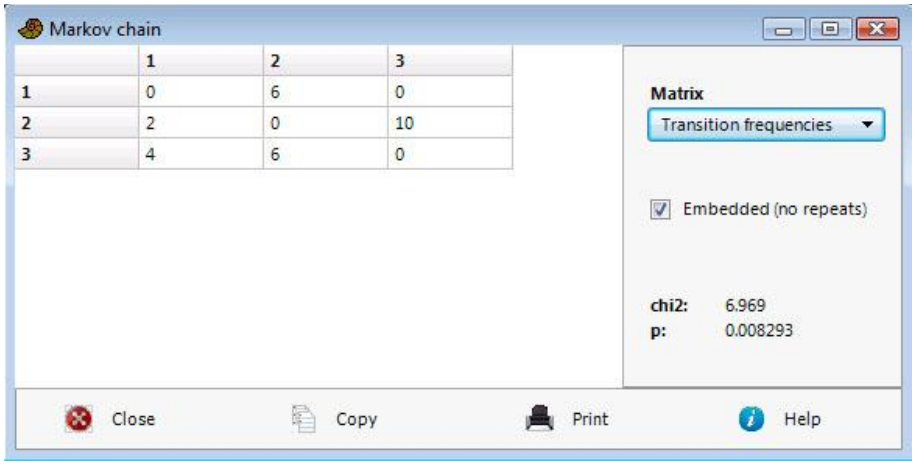
Cox, D. R. & P. A. W. Lewis. 1978. *The Statistical Analysis of Series of Events*. Chapman and Hall, London.

Prahl, J. 1999. A fast unbinned test on event clustering in Poisson processes. *Arxiv, Astronomy and Astrophysics* September 1999.

Markov chain

This module requires a single column containing a sequence of nominal data coded as integer numbers. For example, a stratigraphic sequence where 1 means limestone, 2 means shale and 3 means sand. A transition matrix containing counts or proportions (probabilities) of state transitions is displayed. The "from"-states are in rows, the "to"-states in columns.

It is also possible to specify several columns, each containing one or more state transitions (two numbers for one transition, *n* numbers for a sequence giving *n*-1 transitions).



The chi-squared test reports the probability that the data were taken from a system with random proportions of transitions (i.e. no preferred transitions). The transitions with anomalous frequencies can be identified by comparing the observed and expected transition matrices.

The "Embedded (no repeats)" option should be selected if the data have been collected in such a way that no transitions to the same state are possible (data points are only collected when there is a change). The transition matrix will then have zeroes on the diagonal.

The algorithms, including an iterative algorithm for embedded Markov chains, are according to Davis (1986).

Reference

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

ARMA (and intervention analysis)

Analysis and removal of serial correlations in time series, and analysis of the impact of an external disturbance ("intervention") at a particular point in time. Assumes stationary time series, except for a single intervention. Requires one column of equally spaced data.

This powerful but somewhat complicated module implements maximum-likelihood ARMA analysis, and a minimal version of Box-Jenkins intervention analysis (e.g. for investigating how a climate change might impact biodiversity).

By default, a simple ARMA analysis without interventions is computed. The user selects the number of AR (autoregressive) and MA (moving-average) terms to include in the ARMA difference equation. The log-likelihood and Akaike information criterion are given. Select the numbers of terms that minimize the Akaike criterion, but be aware that AR terms are more "powerful" than MA terms. Two AR terms can model a periodicity, for example.

The main aim of ARMA analysis is to remove serial correlations, which otherwise cause problems for model fitting and statistics. The residual should be inspected for signs of autocorrelation, e.g. by copying the residual from the numerical output window back to the spreadsheet and using the autocorrelation module. Note that for many paleontological data sets with sparse data and confounding effects, proper ARMA analysis (and therefore intervention analysis) will be impossible.

The program is based on the likelihood algorithm of Melard (1984), combined with nonlinear multivariate optimization using simplex search.

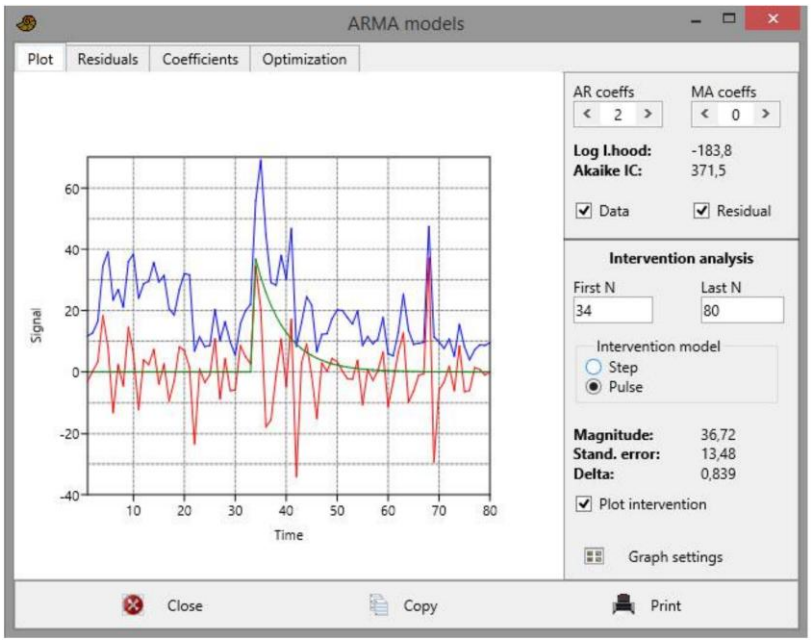
Intervention analysis proceeds as follows. First, carry out ARMA analysis on only the samples preceding the intervention, by typing the last pre-intervention sample number in the "last samp" box. It is also possible to run the ARMA analysis only on the samples following the intervention, by typing the first post-intervention sample in the "first samp" box, but this is not recommended because of the post-intervention disturbance. Also tick the "Intervention" box to see the optimized intervention model.

The analysis follows Box and Tiao (1975) in assuming an "indicator function" u(i) that is either a unit step or a unit pulse, as selected by the user. The indicator function is transformed by an AR(1) process with a parameter delta, and then scaled by a magnitude (note that the magnitude given by PAST is the coefficient on the transformed indicator function: first do $y(i) = delta*y(i-1)+u(i)$, then scale y by the magnitude). The algorithm is based on ARMA transformation of the complete sequence, then a corresponding ARMA transformation of y, and finally linear regression to find the magnitude. The parameter *delta* is optimized by exhaustive search over [0,1].

For small impacts in noisy data, *delta* may end up on a sub-optimum. Try both the step and pulse options, and see what gives smallest standard error on the magnitude. Also, inspect the "delta optimization" data, where standard error of the estimate is plotted as a function of delta, to see if the optimized value may be unstable.

The Box-Jenkins model can model changes that are abrupt and permanent (step function with *delta*=0, or pulse with *delta*=1), abrupt and non-permanent (pulse with *delta*<1), or gradual and permanent (step with *delta*<0).

Be careful with the standard error on the magnitude - it will often be underestimated, especially if the ARMA model does not fit well. For this reason, a *p* value is deliberately not computed (Murtaugh 2002).



The example data set (blue curve) is Sepkoski's curve for percent extinction rate on genus level, interpolated to even spacing at ca. 5 million years. The largest peak is the Permian-Triassic boundary extinction. The user has specified an ARMA(2,0) model. The residual is plotted in red. The user has specified that the ARMA parameters should be computed for the points before the P-T extinction at time slot 37, and a pulse-type intervention. The analysis seems to indicate a large time constant (delta) for the intervention, with an effect lasting into the Jurassic.

References

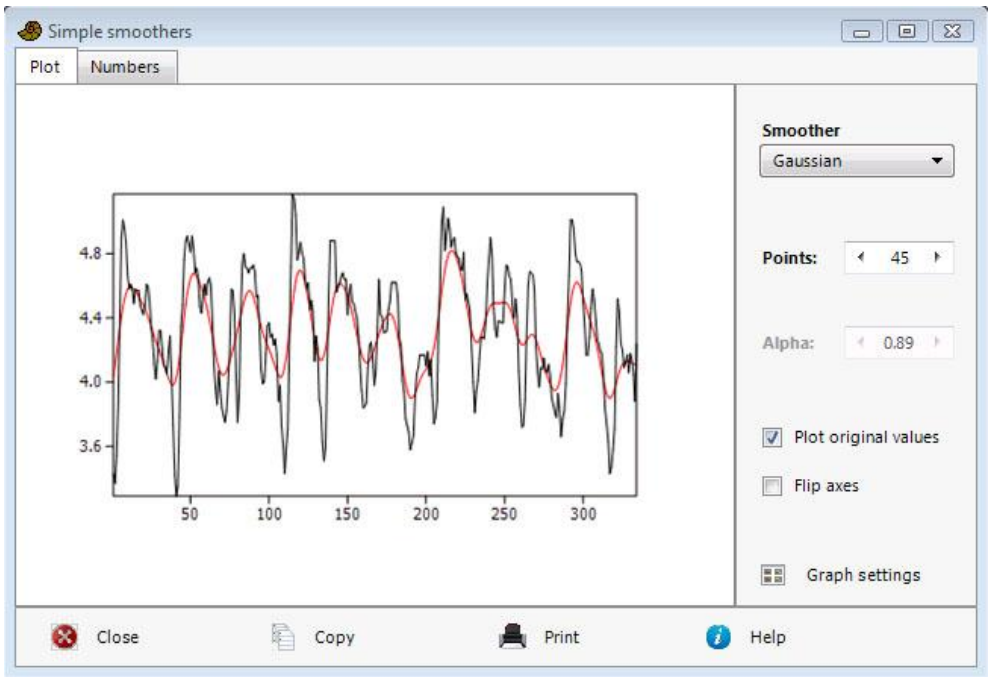
Box, G.E.P. & G.C. Tiao. 1975. Intervention analysis with applications to economic and environental problems. *Journal of the American Statistical Association* 70:70-79.

Melard, G. 1984. A fast algorithm for the exact likelihood of autoregressive-moving average models. *Applied Statistics* 33:104-114.

Murtaugh, P.A. 2002. On rejection rates of paired intervention analysis. *Ecology* 83:1752-1761.

Simple smoothers

A set of simple smoothers for a single column of evenly spaced data. See also the spline and LOESS smoothers in the Model menu. Missing data are supported.



Moving average

Simple *N*-point, centered moving average (*N* must be odd). Commonly used, but has unfortunate properties such as a non-monotonic frequency response.

Gaussian

Weighted moving average using a Gaussian kernel with standard deviation set to 1/5 of the window size (of *N* points). This is an overall good method.

Moving median

Similar to moving average but takes the median instead of the mean. This method is more robust to outliers but can produce a "blocky" appearance.

AR 1 (exponential)

Recursive (autoregressive) filter, $y_i = \alpha y_{i-1} + (1-\alpha)x_i$ with α a smoothing coefficient from 0 to 1. This corresponds to weighted averaging with exponentially decaying weights. Gives a phase delay and also a transient in the beginning of the series. Included for completeness.

Savitzky-Golay

The Savitzky-Golay method (Savitzky & Golay 1964) implements least-squares fit to a polynomial inside a moving window of size *n* points. Second-order (m = 2) and fourth-order (m = 4) polynomials

are included. These are "optimal" smoothers in the sense that they preserve all moments up to *m*. This tends to give better preservation of peak values and peak widths than other smoothers. The implementation in Past is based on Press et al. (1992).

Non-local means

Non-local means is a relatively new, powerful smoothing method, mostly used for image denoising but also effective for time series (Tracey & Miller 2012). It is an averaging method like the moving average and Gaussian methods, but the average is taken not over neighboring points but over points in similar regions, which can be far away. This tends to preserve peaks and transitions better than local averaging. In Past, the size of the local regions (patch size) can be selected; it could be set to e.g. *N* = 7 or *N* = 13. The search radius is fixed at half the length of the time series. The value of *lambda* controls the degree of smoothing. Tracey & Miller (2012) suggest a value of about 0.6 times the standard deviation of the noise (which is usually unknown but can be estimated by eye).

Hampel

The Hampel filter (Hampel 1974) identifies possible outliers and replaces them with a local median. The length of the sliding window ("points") should be at least *N* = 5. Let *k* = (*N*-1)/2. For input sample x_i the local median is given by

$$m_i = \text{median}(x_{i-k}, x_{i-k+1}, \dots, x_i, \dots, x_{i+k-1}, x_{i+k})$$

The local median absolute deviation (MAD) is

$$MAD_i = \text{median}(x_{i-k} - m_i, \dots, x_{i+k} - m_i)$$

A local standard deviation is estimated as

$$\sigma_i = \frac{1}{\sqrt{2}\text{erf}^{-1}\left(\frac{1}{2}\right)} MAD_i \approx 1.4826\, MAD_i$$

Now, if $|x_i - m_i| > 3\sigma_i$ then x_i is considered an outlier, and is replaced by m_i.

References

Hampel, F.R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69, 382-393.

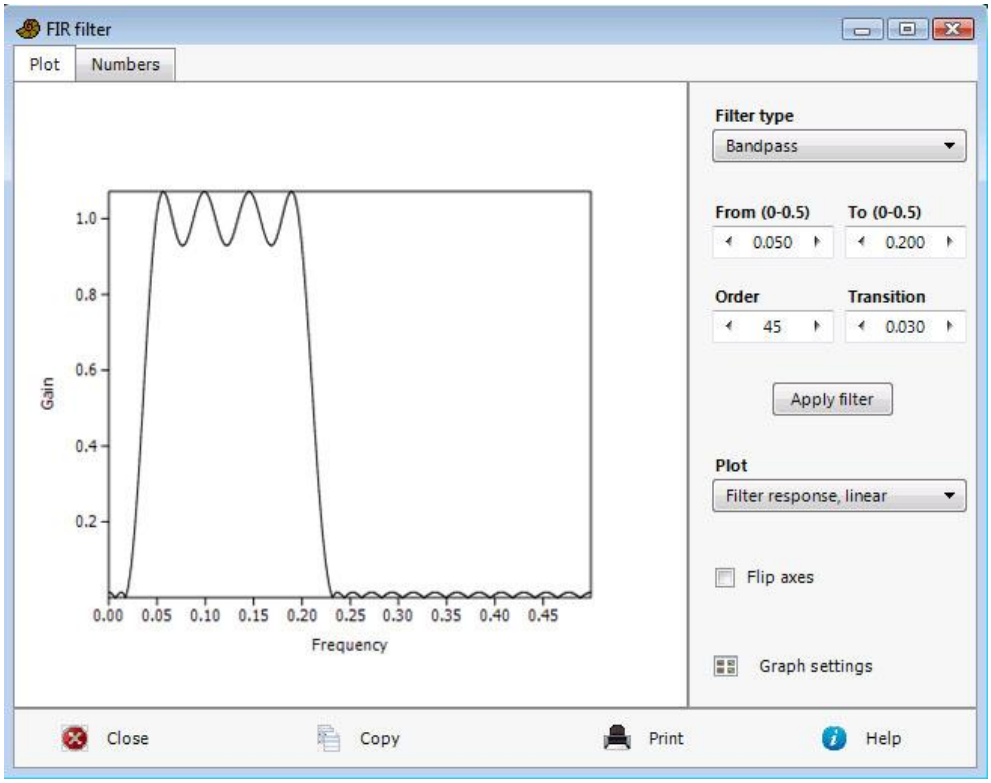
Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 1992. Numerical Recipes in C. 2nd edition. Cambridge University Press.

Savitzky, A. & Golay, M.J.E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36, 1627-1639.

Tracey, B. & Miller, E. 2012. Nonlocal means denoising of ECG signals. *IEEE Transactions on Biomedical Engineering* 59, 2383-2386.

FIR filter

Filtering out certain frequency bands in a time series can be useful to smooth a curve, remove slow variation, or emphasize certain periodicities (e.g. Milankovitch cycles). One column of evenly spaced data is expected. For most applications in data analysis, it is crucial that the filter has linear phase response. Past therefore uses FIR (Finite Impulse Response) filters, which are designed using the Parks-McClellan algorithm. The following filter types are available: Lowpass, highpass, bandpass and bandstop.



Filter parameters

To design an optimal filter takes a little effort. Frequencies are specified in the range 0-0.5, i.e. T_0/T where T_0 is the sampling interval (not specified to the computer) and T is the required period. For example, if your real sampling interval is 1,000 years, a frequency corresponding to a period of 23,000 years is specified as 1,000/23,000=0.043.

After setting the filter type, you should select a transition width (or leave the default of 0.02). Decreasing the transition width will make a sharper filter, at the cost of larger ripple ("waves" in the frequency response).

Note that the values in text fields are not updated until you press Enter. Also, if an invalid combination is entered (e.g. a transition band crossing 0 or 0.5, or upper limit less than lower limit) the program will reset some value to avoid errors. It is therefore required to enter the numbers in an order so that the filter is always valid.

The filter types are as follows:

1. **Lowpass**. The *From* frequency is forced to zero. Frequencies up to the *To* frequency pass the filter. Frequencies from *To*+*Transition* to 0.5 are blocked.
2. **Highpass**. The *To* frequency is forced to 0.5. Frequencies above the *From* frequency pass the filter. Frequencies from 0 to *From*-*Transition* are blocked.
3. **Bandpass**. Frequencies from *From* to *To* pass the filter. Frequencies below *From*-*Transition* and above *To*+*Transition* are blocked.
4. **Bandstop**. Frequencies from *From* to *To* are blocked. Frequencies from 0 to *From*-*Transition* and from *To*+*Transition* to 0.5 pass the filter.

Filter order

The filter order should be large enough to give an acceptably sharp filter with low ripple. However, a filter of length *n* will give less accurate results in the first and last *n*/2 samples of the time series, which puts a practical limit on filter order for short series.

The Parks-McClellan algorithm will not always converge. This gives an obviously incorrect frequency response, and attempting to apply such a filter to the data will give a warning message. Try to change the filter order (usually increase it) to fix the problem.

Missing values are treated using linear interpolation before analysis.

Insolation (solar forcing) model

This module computes solar insolation at any latitude and any time from 100 Ma to the Recent (the results are less accurate before 50 Ma). The calculation can be done for a "true" orbital longitude, "mean" orbital longitude (corresponding to a certain date in the year), averaged over a certain month in each year, or integrated over a whole year.

The implementation in PAST is ported from the code by Laskar et al. (2004), by courtesy of these authors. Please reference Laskar et al. (2004) in any publications.

It is necessary to specify a data file containing orbital parameters. Download the file INSOLN.LA2004.BTL.250.ASC from http://vo.imcce.fr/insola/earth/online/earth/earth.html and put in anywhere on your computer. The first time you run the calculation, PAST will ask for the position of the file.

The amount of data can become excessive for long time spans and short step sizes!

Reference

Laskar, J., P. Robutel, F. Joutel, M. Gastineau, A.C.M. Correia & B. Levrard. 2004. A long-term numerical solution for the insolation quantities of the Earth. *Astronomy & Astrophysics* 428:261-285.

Date/time conversion

Utility to convert dates and/or times to a continuous time unit for analysis. The program expects one or two columns, each containing dates or times. If both are given, then time is added to date to give the final time value.

Dates can be given in the formats Year/Month/Day or Day/Month/Year. Years need all digits (a year given as 11 will mean 11 AD, not 2011). Only Gregorian calendar dates are supported. Leap years are taken into account.

Time can be given as Hours:Minutes or Hours:Minutes:Seconds (seconds can include decimals).

The output units can be years (using the Gregorian mean year of 365.2425 days), days (of 86400 seconds), hours, minutes or seconds.

The starting time (time zero) can be the smallest given time, the beginning of the first day, the beginning of the first year, year 0 (note the "astronomical" convention where the year before year 1 is year 0), or the beginning of the first Julian day (noon, year -4712).

The program operates with simple (UT) time, defined with respect to the Earth's rotation and with a fixed number of seconds (86400) per day.

If your input data consists of space-separated date-time values, such as "2011/12/24 18:00:00.00", then you may have to use the "Import text file" function to read the data such that dates and times are split into separate columns.

The calculation of Julian day (which is used to find number of days between two dates) follows Meeus (1991):

if *month* <= 2 begin *year* := *year* - 1; *month* := *month* + 12; end;

A = floor(*year*/100);

B = 2 – *A* + floor(*A*/4);

JD = floor(365.25(*year* + 4716)) + floor(30.6001(*month*+1)) + *day* + *B* – 1524.5;

Reference

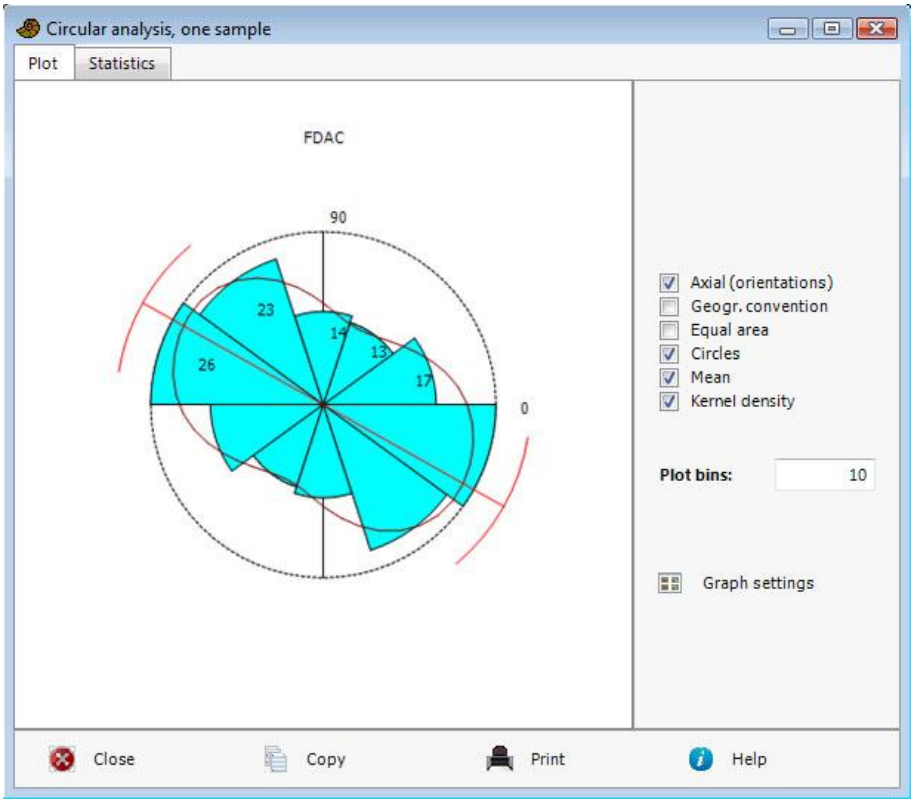
Meeus, J. 1991. *Astronomical algorithms*. Willmann-Bell, Richmond.

Geometrical menu

Circular (one sample)

The module plots a rose diagram (polar histogram) of directions. Used for plotting current-oriented specimens, orientations of trackways, fault lines, etc. Also appropriate for time-of day data (0-24 hours).

One column of directional (0-360) or orientational (0-180) data in degrees is expected. Directional or periodic data in other forms (radians, hours, etc.) must be converted to degrees using e.g. the Evaluate Expression module (Transform menu).



By default, the 'mathematical' angle convention of anticlockwise from east is chosen. If you use the 'geographical' convention of clockwise from north, tick the box.

You can also choose whether to have the abundances proportional to radius in the rose diagram, or proportional to area (equal area).

The "Kernel density" option plots a circular kernel density estimate.

Descriptive statistics

The mean angle takes circularity into account:

$$\bar{\theta} = \tan^{-1} \frac{\sum \sin \theta_i}{\sum \cos \theta_i}$$ (taken to the correct quadrant).

The 95 percent confidence interval on the mean is estimated according to Fisher (1983). It assumes circular normal distribution, and is not accurate for very large variances (confidence interval larger than 45 degrees) or small sample sizes. The bootstrapped 95% confidence interval on the mean uses 5000 bootstrap replicates. The graphic uses the bootstrapped confidence interval.

The concentration parameter κ is estimated by iterative approximation to the solution to the equation

$$\frac{I_1(\kappa)}{I_0(\kappa)} = \overline{R}$$

where I_0 and I_1 are imaginary Bessel functions of orders 0 and 1, estimated according to Press et al. (1992), and *R* defined below (see e.g. Mardia 1972).

Rayleigh's test for uniform distribution

The *R* value (mean resultant length) is given by:

$$\overline{R} = \sqrt{\left(\sum_{i=1}^{n}\cos\theta_i\right)^2 + \left(\sum_{i=1}^{n}\sin\theta_i\right)^2} \Big/ n\,.$$

R is further tested against a random distribution using Rayleigh's test for directional data (Davis 1986). Note that this procedure assumes evenly or unimodally (von Mises) distributed data - the test is not appropriate for e.g. bimodal data. The *p* values are computed using an approximation given by Mardia (1972):

$$K = n\overline{R}^2$$

$$p = e^{-K}\left(1 + \frac{2K - K^2}{4n} - \frac{24K - 132K^2 + 76K^3 - 9K^4}{288n^2}\right)$$

Rao's spacing test for uniform distribution

The Rao's spacing test (Batschelet 1981) for uniform distribution has test statistic

$$U = \frac{1}{2}\sum_{i=1}^{n}\left|T_i - \lambda\right|,$$

where $\lambda = 360^{\circ}/n$. $T_i = \theta_{i+1} - \theta_i$ for *i*<*n*, $T_n = 360^{\circ} - \theta_n + \theta_1$. This test is nonparametric and does not assume e.g. von Mises distribution. The *p* value is estimated by linear interpolation from the probability tables published by Russell & Levitin (1995).

A Chi-square test for uniform distribution is also available, with the number of bins set to 4 if *n*<40, otherwise 8 bins.

The Watson's U^2 goodness-of-fit test for von Mises distribution

Let *f* be the von Mises distribution for estimated parameters of mean angle and concentration:

$$f\left(\theta;\bar{\theta},\kappa\right)=\frac{e^{\kappa\cos\left(\theta-\bar{\theta}\right)}}{2\pi I_0\left(\kappa\right)}.$$

The test statistic (e.g. Lockhart & Stevens 1985) is

$$U^2=\sum\left(z_i-\frac{2i-1}{2n}\right)^2-n\left(\bar{z}-\frac{1}{2}\right)^2+\frac{1}{12n}$$

where

$$z_i=\int_0^{\theta_i} f\left(\theta;\bar{\theta},\kappa\right)d\theta,$$

estimated by numerical integration. Critical values for the test statistic are obtained by linear interpolation into Table 1 of Lockhart & Stevens (1985). They are acceptably accurate for *n*>=20.

Axial data

The 'Orientations' option allows analysis of linear (axial) orientations (0-180 degrees). The Rayleigh and Watson tests are then carried out on doubled angles (this trick is described by Davis 1986); the Chi-square uses four bins from 0-180 degrees; the rose diagram mirrors the histogram around the origin.

References

Batschelet, E. 1981. Circular statistics in biology. Academic Press.

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Fisher, N.I. 1983. Comment on "A Method for Estimating the Standard Deviation of Wind Directions". *Journal of Applied Meteorology* 22:1971.

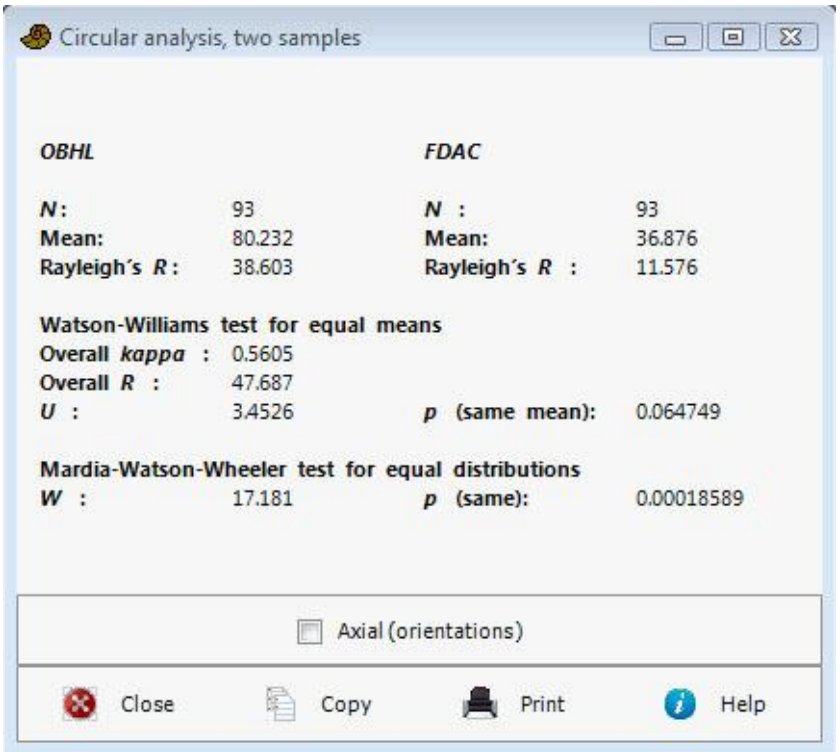
Lockhart, R.A. & M.A. Stephens 1985. Tests of fit for the von Mises distribution. *Biometrika* 72:647-652.

Mardia, K.V. 1972. Statistics of directional data. Academic Press, London.

Russell, G. S. & D.J. Levitin 1995. An expanded table of probability values for Rao's spacing test. *Communications in Statistics: Simulation and Computation* 24:879-888.

Circular (two samples)

The module expects two columns of directional (0-360) or orientational (0-180) data in degrees.



Watson-Williams test

The Watson-Williams test for equal mean angle in two samples is a parametric test, assuming von Mises distribution, but is fairly robust. The concentration parameter κ should be larger than 1.0 for accurate testing. In addition, the test assumes similar angular variances (R values).

The two samples $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ have n_1 and n_2 values. The resultant length R is calculated for each sample and for the combined sample:

$$R_1 = \sqrt{\left(\sum_{i=1}^{n_1} \cos\phi_i\right)^2 + \left(\sum_{i=1}^{n_1} \sin\phi_i\right)^2}$$

$$R_2 = \sqrt{\left(\sum_{i=1}^{n_2} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n_2} \sin\theta_i\right)^2}$$

$$R = \sqrt{\left(\sum_{i=1}^{n_1} \cos\phi_i + \sum_{i=1}^{n_2} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n_1} \sin\phi_i + \sum_{i=1}^{n_2} \sin\theta_i\right)^2} .$$

The test statistic U is computed as

$$U = (n-2)\frac{R_1 + R_2 - R}{n - (R_1 + R_2)}.$$

The significance is computed by first correcting U according to Mardia (1972a):

$$U = \begin{cases} \dfrac{U}{1 - \dfrac{\kappa^2}{8} + \dfrac{1}{n\kappa^2}} & R/n < 0.45 \\ \left(1 + \dfrac{3}{8\kappa}\right)U & R/n < 0.95 \end{cases},$$

where $n = n_1 + n_2$. The p value is then given by the F distribution with 1 and n-2 degrees of freedom. The combined concentration parameter κ is maximum-likelihood, computed as described under "Directions (one sample)" above.

Mardia-Watson-Wheeler test

This non-parametric test for equal distribution is computed according to Mardia (1972b).

$$W = 2\left(\frac{C_1^2 + S_1^2}{n_1} + \frac{C_2^2 + S_2^2}{n_2}\right)$$

where, for the first sample,

$$C_1 = \sum_{i=1}^{n_1} \cos(2\pi r_{1i}/N), \qquad S_1 = \sum_{i=1}^{n_1} \sin(2\pi r_{1i}/N)$$

and similarly for the second sample (N=n_1+n_2). The r_{1i} are the ranks of the values of the first sample within the pooled sample.

For N>14, W is approximately chi-squared with 2 degrees of freedom.

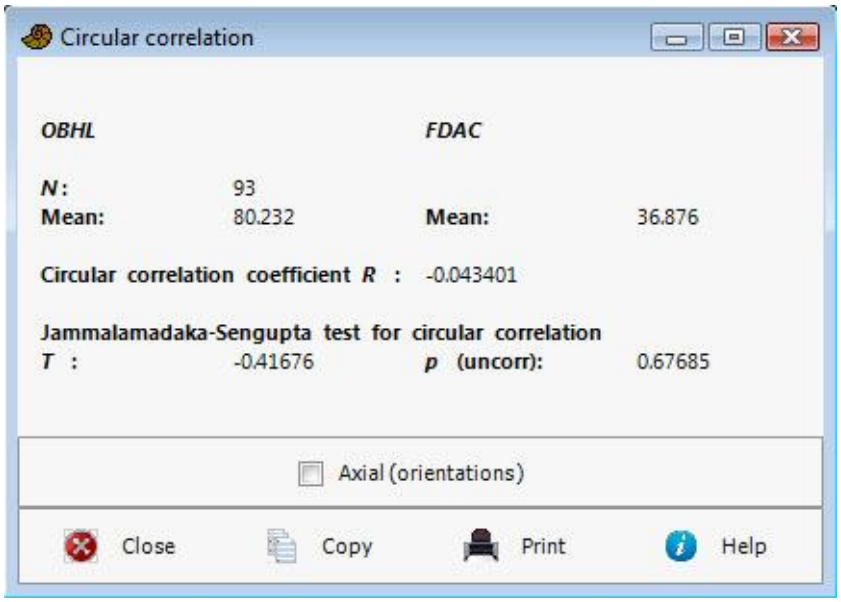
References

Mardia, K.V. 1972a. Statistics of directional data. Academic Press, London.

Mardia, K.V. 1972b. A multi-sample uniform scores test on a circle and its parametric competitor. *Journal of the Royal Statistical Society Series B* 34:102-113.

Circular correlation

Testing for correlation between two directional or orientational variates. Assumes "large" number of observations. Requires two columns of directional (0-360) or orientational (0-180) data in degrees.



This module uses the circular correlation procedure and parametric significance test of Jammalamadaka & Sengupta (2001).

The circular correlation coefficient *r* between vectors of angles **α** and **β** is

$$r = \frac{\sum_{i=1}^{n} \sin(\alpha_i - \bar{\alpha}) \sin(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^{n} \sin^2(\alpha_i - \bar{\alpha}) \sin^2(\beta_i - \bar{\beta})}},$$

where the angular means are calculated as described previously. The test statistic *T* is computed as

$$T = r \sqrt{\frac{\sum_{k=1}^{n} \sin^2(\alpha_k - \bar{\alpha}) \sum_{k=1}^{n} \sin^2(\beta_k - \bar{\beta})}{\sum_{k=1}^{n} \sin^2(\alpha_k - \bar{\alpha}) \sin^2(\beta_k - \bar{\beta})}}.$$

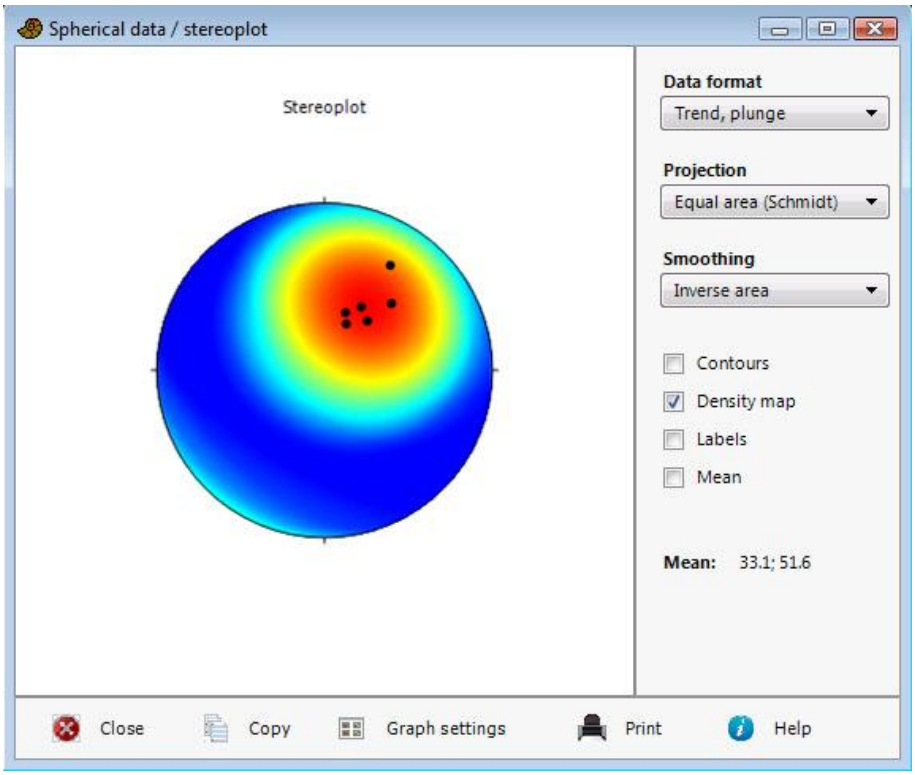
For large *n*, this statistic has asymptotically normal distribution with mean 0 and variance 1 under the null hypothesis of zero correlation, which is the basis for the calculation of *p*.

Reference

Jammalamadaka, S.R. & A. Sengupta. 2001. Topics in circular statistics. World Scientific.

Spherical (one sample)

This module makes stereo plots of axial, spherical data (e.g. strike-dip measurements in structural geology), and performs the Bingham test for uniform distribution.



Three data formats can be used, all using the geographic angle convention (degrees, clockwise from north):

- Trend (azimuth) and plunge (angle down from the horizontal) for axial data
- Dip azimuth and dip angle (down from the horizontal) for planes. The pole (normal vector) of the plane is plotted.
- Strike and dip for planes, using the right-hand rule convention with the dip down to the right from the strike. The pole to the plane is plotted.

Density contouring is based on a modified Kamb method algorithm by Vollmer (1995). Both equal area (Schmidt) and equal angle (Wulff) projections are available. Projections are to the lower hemisphere. Density estimates can use an inverse area, inverse area squared or exponential law, giving progressively higher smoothing.

The Bingham test for uniform distribution of axial data can be used to test for preferred direction (Bingham 1974; Mardia & Jupp 2000, p. 232-233). Past computes the *S* statistic as follows.

The sample scatter matrix is computed as

$$\overline{\boldsymbol{T}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$$

where $\boldsymbol{x}_i$ is the 3-vector of direction cosines for sample *i*. Then,

$$S = \frac{15}{2} n \left(\mathrm{tr}(\overline{\boldsymbol{T}}^2) - \frac{1}{3} \right)$$

An adjusted *S* is computed according to Jupp (2001):

$$S^* = S \left[1 - \frac{1}{n} (B_0 + B_1 S + B_2 S^2) \right]$$

where

$$B_0 = \frac{31}{42}, \qquad B_1 = -\frac{41}{294}, \qquad B_2 = \frac{5}{1323}$$

When *n*<=50 and *S*<=30, the reported *p* value for uniformity is estimated from the *S** value using the chi-squared distribution with 5 degrees of freedom. Otherwise, the *S* value is used (the polynomial equation for *S** turns the wrong way for large *S*). A small *p* (e.g. *p*<0.05) means significantly non-uniform distribution.

References

Bingham, C. 1974. An antipodally symmetric distribution on the sphere. *Annals of Statistics* 2:1201-1225.

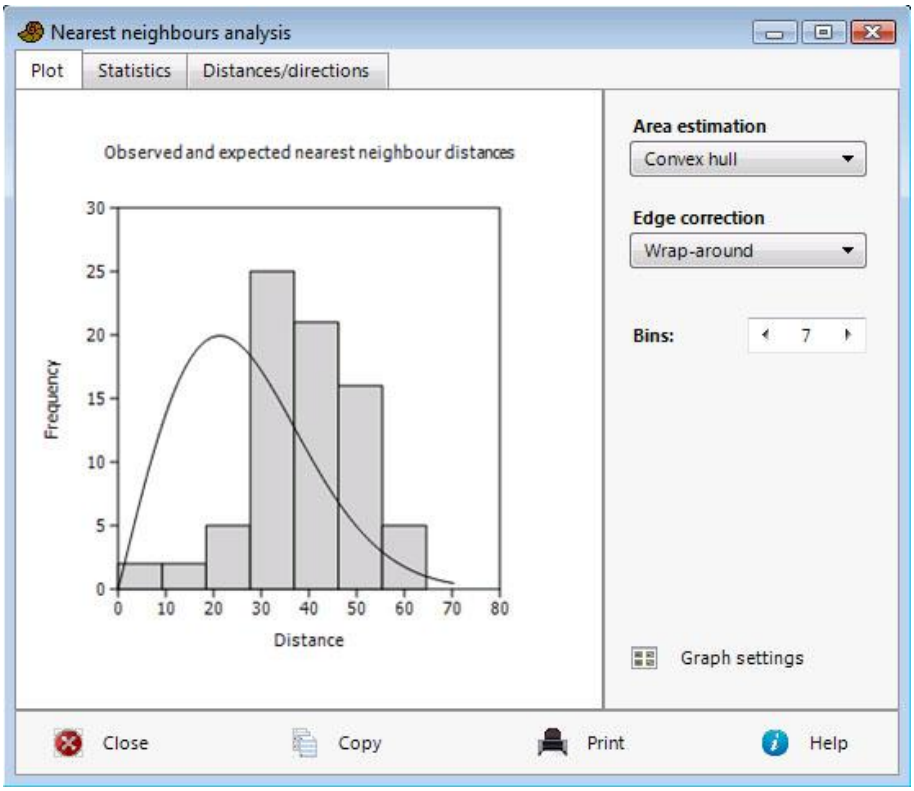
Jupp, P.E. 2001. Modifications of the Rayleigh and Bingham tests for uniformity of directions. *Journal of Multivariate Analysis* 77:1-20.

Mardia, K.V. & Jupp, P.E. 2000. Directional Statistics. John Wiley & Sons.

Vollmer, F.W. 1995. C program for automatic contouring of spherical orientation data using a modified Kamb method. *Computers & Geosciences* 21:31-49.

Point pattern analysis - nearest neighbours

This module tests for clustering or overdispersion of points given as two-dimensional coordinate values. The procedure assumes that elements are small compared to their distances, that the domain is predominantly convex, and *n*>50. Two columns of *x/y* positions are required. Applications of this module include spatial ecology (are in-situ brachiopods clustered), morphology (are trilobite tubercles overdispersed), and geology (distribution of e.g. volcanoes, earthquakes, springs).



The calculation of point distribution statistics using nearest neighbour analysis follows Davis (1986) with modifications. The area is estimated either by the smallest enclosing rectangle or using the convex hull, which is the smallest convex polygon enclosing the points. Both are inappropriate for points in very concave domains. Two different edge effect adjustment methods are available: wrap-around ("torus") and Donnelly's correction. Wrap-around edge detection is only appropriate for rectangular domains.

The null hypothesis is a random Poisson process, giving a modified exponential nearest neighbour distribution (see below) with mean

$$\mu = \frac{\sqrt{A/n}}{2}$$

where *A* is the area and *n* the number of points.

The probability that the distribution is Poisson is presented, together with the *R* value:

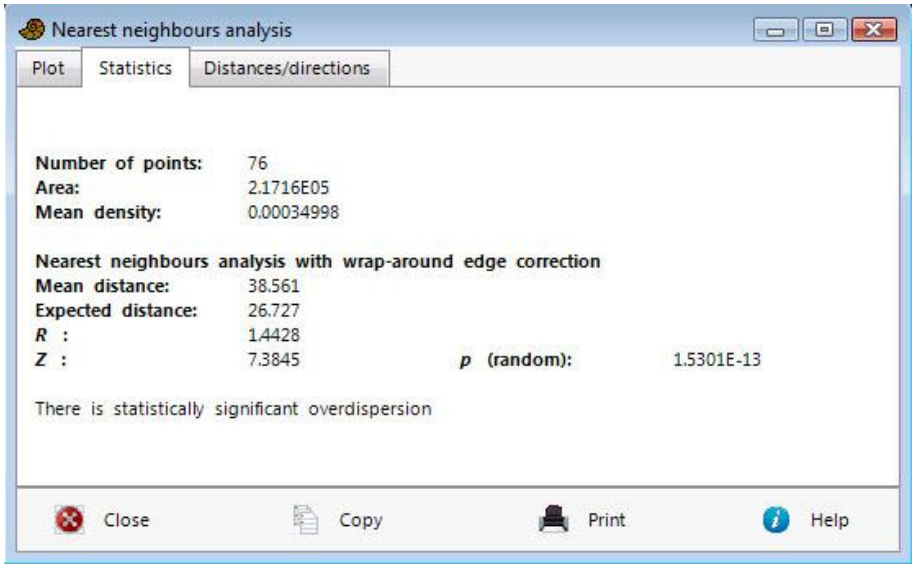
$$R = \frac{\bar{d}}{\mu} = \frac{2\bar{d}}{\sqrt{A/n}}$$

where $\bar{d}$ is the observed mean distance between nearest neighbours. Clustered points give R<1, Poisson patterns give R~1, while overdispersed points give R>1.

The expected (theoretical) distribution under the null hypothesis is plotted as a continuous curve together with the histogram of observed distances. The expected probability density function as a function of distance r is

$$g(r) = 2\rho\pi r \exp\left(-\rho\pi r^2\right)$$

where $\rho = n/A$ is the point density (Clark & Evans 1954).



The orientations (0-180 degrees) and lengths of lines between nearest neighbours, are also included. The orientations can be subjected to directional analysis to test whether the points are organized along lineaments (see Hammer 2009 for more advanced methods).

References

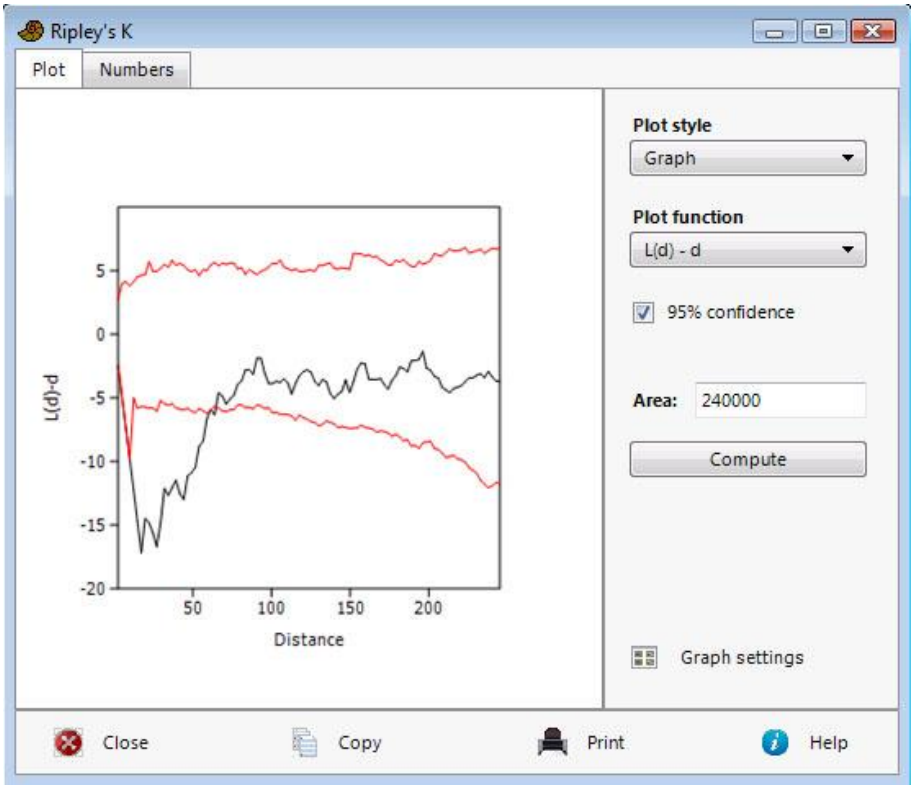
Clark, P.J. & Evans, F.C. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* 35:445-453.

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Hammer, Ø. 2009. New methods for the statistical analysis of point alignments. *Computers & Geosciences* 35:659-666.

Ripley's *K* point pattern analysis

Ripley's *K* (Ripley 1979) is the average point density as a function of distance from every point. It is useful when point pattern characteristics change with scale, e.g. overdispersion over small distances but clustering over large distances. Two columns of *x/y* coordinates in a rectangular domain are expected.



Define the estimated intensity of the point pattern, with *n* points in an area *A*, as $\lambda = n/A$. The distance between points *i* and j is d_{ij} The estimate of Ripley's *K*, as a function of distance, is then computed as

$$K(d) = \frac{1}{\lambda n} \sum_{i=1}^{n} \sum_{j \neq i} I(d_{ij} \leq d),$$

where the indicator function *I* is one if the argument is true, zero otherwise.

The normalization of *K* is such that for complete spatial randomness (CSR), *K*(*d*) is expected to increase as the area of circles, i.e. $K(d) = \pi d^2$. The *L*(*d*) function is a corresponding transformation of *K*(*d*):

$$L(d) = \sqrt{\frac{K(d)}{\pi}}$$

For CSR, *L*(*d*)=*d*, and *L*(*d*)-*d*=0. A 95% confidence interval for CSR is estimated using 1000 Monte Carlo simulations within the bounding rectangle (previous versions used the approximation $1.42\sqrt{A}/n$).

Ripley's edge correction is included, giving weights to counts depending on the proportion of the test circle that is inside the rectangular domain.

The example above shows locations of volcanic pipes. $L(d)$-d is below the 95% confidence interval of CSR, indicating lateral inhibition, up to a distance of ca. 70 m. For larger distances, the curve flattens in the manner expected from CSR.

Area

For the correct calculation of Ripley's K, the area must be known. In the first run, the area is computed using the smallest bounding rectangle, but this can both over- and underestimate the real area. The area can therefore be adjusted by the user. An overestimated area will typically show up as a strong overall linear trend with positive slope for $L(d)$-d.

Fractal dimension

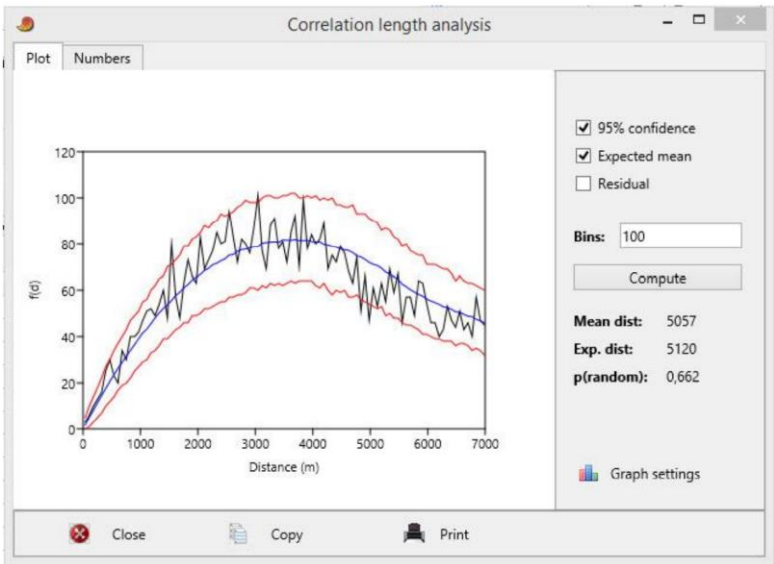
The fractal dimension (if any) can be estimated as the asymptotic linear slope in a log-log plot of $K(d)$. For CSR, the log-log slope should be 2.0. Fractals should have slopes less than 2.

References

Ripley, B.D. 1979. Tests of 'randomness' for spatial point patterns. *Journal of the Royal Statistical Society, ser. B* 41:368-374.

Correlation length analysis

Correlation length analysis (Cartwright & Whitworth 2004; Cartwright et al. 2011) investigates the spatial distribution of a point pattern at different scales, and is an alternative to Ripley's *K*. Two columns of x/y coordinates in a rectangular domain are expected. CLA is simply a histogram of all pairwise distances between points, i.e. a total of *N*(*N*-1)/2 distances (black curve).



The expected curve from a random point pattern (blue curve) and its 95% confidence interval (red curves) are computed from 1000 Monte Carlo simulations of complete spatial randomness (CSR) in a rectangle of the same dimensions as the bounding rectangle of the original data. Thus, distances where the CLA curve from the data (black curve) exceeds the upper red curve, have significantly higher frequencies than expected from a random pattern.

An overall significance test is based on the observed total mean distance compared with the expected mean distance from the Monte Carlo simulations.

The number of bins can be set by the user and should be small to reduce noise, but large enough to capture details.

The "Residual" option flattens the curves on the expected mean (blue curve), i.e. the expected mean is subtracted from the curves at all distances. This can make the figure clearer especially when the confidence interval is narrow.

A comparison between Ripley's *K* and correlation length analysis for a geological data set is given by Cartwright et al. (2011).

References

Cartwright, A. & Whitworth, A.P. 2004. The statistical analysis of star clusters. *Monthly Notices of the Royal Astronomical Society* 348:589-597.

Cartwright, A., Moss, J. & Cartwright, J. 2011. New statistical methods for investigating submarine pockmarks. *Computers & Geosciences* 37:1595-1601.

Minimal spanning tree analysis

Minimal spanning tree analysis (Cartwright & Whitworth 2004; Cartwright et al. 2011) investigates the spatial distribution of a point pattern with focus on small scales, comparable to nearest neighbor analysis but with somewhat different properties. Two columns of x/y coordinates in a rectangular domain are expected. The method is based on a histogram of all the lengths of line segments in the minimal spanning tree (MST). The MST itself can be plotted in the XY graph module (Plot menu).

The expected curve from a random point pattern (blue curve) and its 95% confidence interval (red curves) are computed from 1000 Monte Carlo simulations of complete spatial randomness (CSR) in a rectangle of the same dimensions as the bounding rectangle of the original data. Thus, segment lengths where the histogram from the data (black curve) exceeds the upper red curve, have significantly higher frequencies than expected from a random pattern.

An overall significance test is based on the observed total mean length compared with the expected mean length from the Monte Carlo simulations.

The number of bins can be set by the user and should be small to reduce noise, but large enough to capture details.

The “Residual” option flattens the curves on the expected mean (blue curve), i.e. the expected mean is subtracted from the curves at all distances. This can make the figure clearer especially when the confidence interval is narrow.

A comparison between nearest neighbour and minimal spanning tree analysis for a geological data set is given by Cartwright et al. (2011).

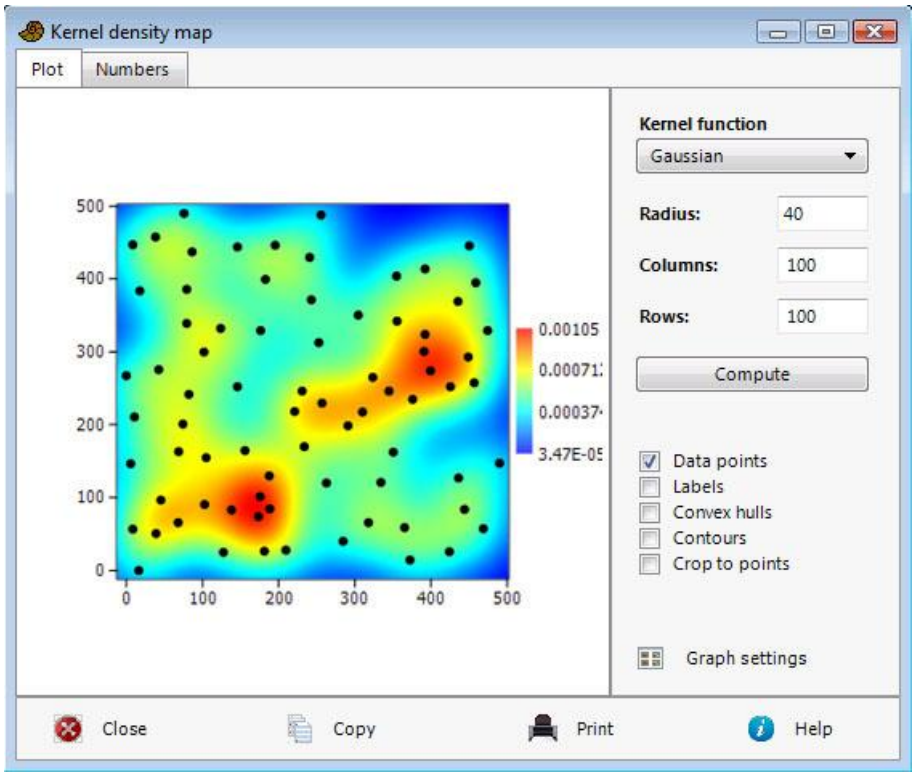
References

Cartwright, A. & Whitworth, A.P. 2004. The statistical analysis of star clusters. *Monthly Notices of the Royal Astronomical Society* 348:589-597

Cartwright, A., Moss, J. & Cartwright, J. 2011. New statistical methods for investigating submarine pockmarks. *Computers & Geosciences* 37:1595-1601.

Kernel density

Makes a smooth map of point density in 2D. Two columns of *x/y* coordinates in a rectangular domain are expected. The user can specify the size of the grid (number of rows and columns). The "Radius" value sets the scale *r* of the kernel. There is currently no automatic selection of "optimal" radius, so this value must be set by the user depending on the scale of interest.



The density estimate is based on one of four kernel functions, with radius parameter *r*. With $d_i = \sqrt{(x-x_i)^2 + (y-y_i)^2}$:

Gaussian (default):

$$\mathrm{f}(x,y) = \frac{1}{\pi r^2} \sum_i \exp\left(-\frac{d_i^2}{2r^2}\right)$$

Paraboloid:

$$\mathrm{f}(x,y) = \frac{3}{2\pi r^2} \sum_i \begin{cases} 1 - \frac{d_i^2}{r^2} & d_i \le r \\ 0 & d_i > r \end{cases}$$

Triangular:

$$\mathrm{f}(x,y) = \frac{2}{\pi r^2} \sum_i \begin{cases} 1 - \frac{d_i}{r} & d_i \le r \\ 0 & d_i > r \end{cases}$$

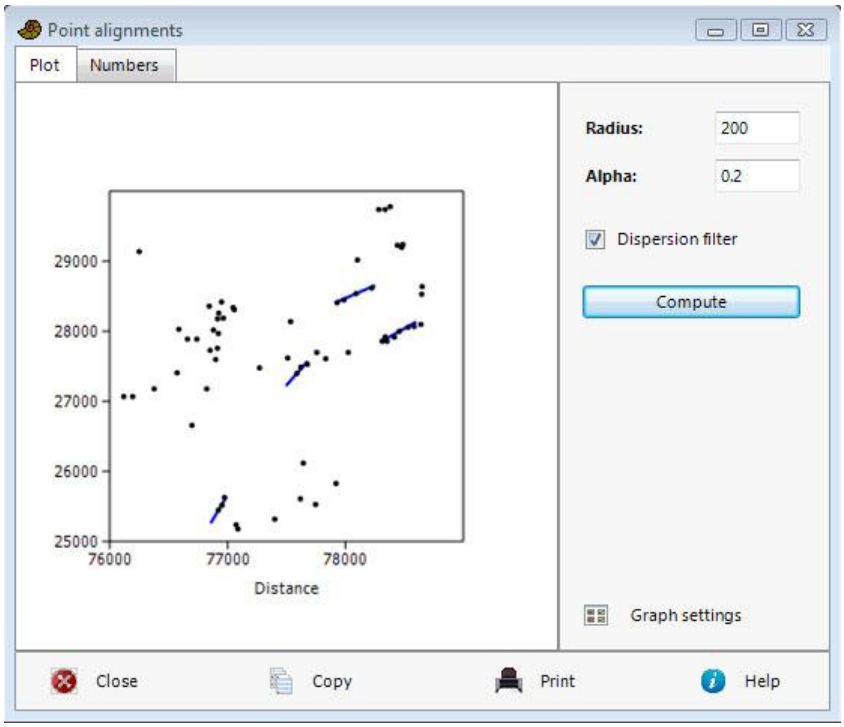
Uniform:

$$\mathrm{f}(x,y) = \frac{1}{\pi r^2} \sum_i \begin{cases} 1 & d_i \le r \\ 0 & d_i > r \end{cases}$$

The scaling gives an estimate of the number of points per area, not a probability density. The gaussian and paraboloid (quadratic) kernels usually perform best. The uniform kernel gives very low smoothness.

Point alignments

Detection of linear alignments in a 2D point pattern, using the continuous sector method (Hammer 2009). Typical applications are in geology and geography, to study the distribution of earthquakes, volcanoes, springs etc. associated with faults and other linear structures.



The *Radius* parameter sets the scale of analysis. In the example above, lineaments of length 1200 m (twice the radius) are detected.

Alpha sets the significance level for the Rayleigh test used by the procedure. Note that this is a pointwise significance, not corrected for the multiple testing of all the points.

The *Dispersion filter* disables alignments with uneven distribution of points along the lineament.

View numbers lists the alignment positions and their orientations, which can be subjected to circular statistics if required (Directions module).

Reference

Hammer, Ø. 2009. New methods for the statistical detection of point alignments. *Computers & Geosciences* 35:659-666.

Quadrat counts

This module provides statistics on the distribution of points in quadrats. The input data consist of a single column of counts of points in equal-sized quadrats (the order is arbitrary). For a random point pattern, the data are expected to follow a Poisson distribution.

The Morisita index (Morisita 1959) is expected to have value *Id*=1 for a random pattern, *Id*<1 for an overdispersed (spaced) pattern, and *Id*>1 (up to *Id*=*n*) for a clustered pattern. It is computed as

$$I_\partial = n\frac{\sum x^2 - \sum x}{(\sum x)^2 - \sum x}$$

where *n* is the number of quadrats, and *x* are the counts. The significance test follows Morisita (1959), with *F* ratio

$$F_0 = \frac{I_\partial(\sum x - 1) + n - \sum x}{n - 1}$$

The degrees of freedom are *n*-1 and ∞. In addition, a Monte Carlo test is carried out with 9999 replicates, each with random distribution of points on quadrats.

The 95% confidence limits (lower and upper) around *Id*=1 (random pattern) are called the uniform and the clumped indices, respectively (Krebs 1999):

$$M_u = \frac{\chi^2_{0.975} - n + \sum x}{\sum x - 1}$$

$$M_c = \frac{\chi^2_{0.025} - n + \sum x}{\sum x - 1}$$

Where $\chi^2_{0.975}$ is the 97.5 percentile point of the chi-squared distribution with *n*-1 degrees of freedom.

The Standardized Morisita Index, *MIS*, was suggested by Smith-Gill (1975). It ranges from -1 to 1, with *MIS*=0 for a random pattern and with 95% confidence limits [-0.5, 0.5]. It is calculated as follows:

$$I_\delta \geq M_c > 1:\ MIS = 0.5 + 0.5\frac{I_\delta - M_c}{n - M_c}$$

$$M_c > I_\delta \geq 1:\ MIS = 0.5\frac{I_\delta - 1}{M_c - 1}$$

$$1 > I_\delta > M_u:\ MIS = -0.5\frac{I_\delta - 1}{M_u - 1}$$

$$1 > M_u > I_\delta:\ MIS = -0.5 + 0.5\frac{I_\delta - M_u}{M_u}$$

References

Krebs, C.J. 1999. Ecological Methodology, 2nd ed. Benjamin Cummings Publishers.

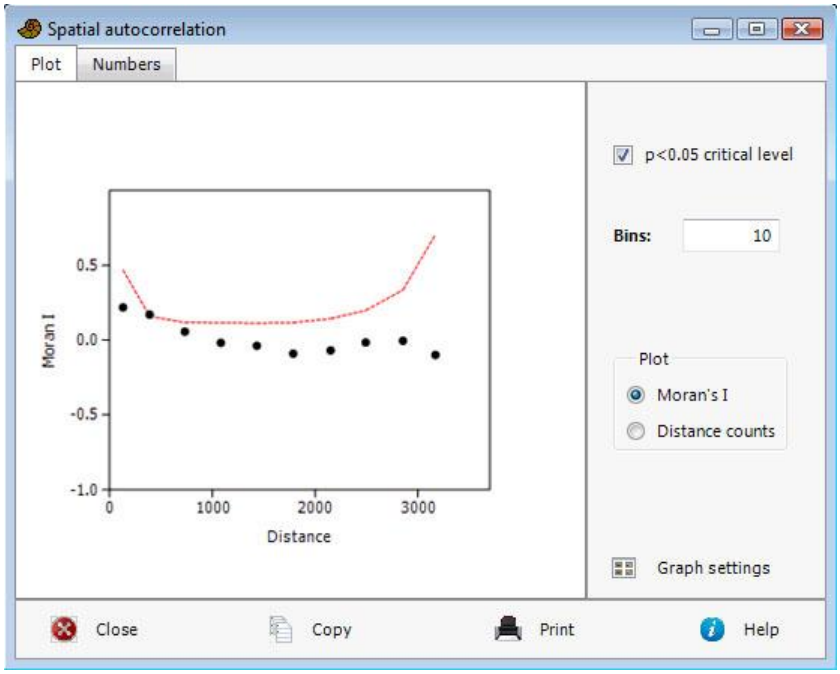
Morisita, M. 1959. Measuring of the dispersion of individuals and analysis of the distributional patterns. *Memoirs of the Faculty of Science, Kyushu University, Series E (biology)* 2:215-235.

Smith-Gill, S. J. 1975. Cytophysiological basis of disruptive pigmentary patterns in the leopard frog, *Rana pipiens*. II. Wild type and mutant cell specific patterns. *Journal of Morphology* 146:35–54.

Spatial autocorrelation (Moran's *I*)

Spatial autocorrelation in Past requires three columns, containing *x* and *y* coordinates and corresponding data values *z* for a number of points. The Moran's *I* correlation statistic is then computed within each of a number of distance classes (bins), ranging from small to large distances.

The one-tailed critical value for $p<0.05$ can be plotted for each bin. Moran's *I* values exceeding the critical value may be considered significant, but Bonferroni or other adjustment for multiple testing should be considered because of the several bins.



The calculation follows Legendre & Legendre (1998). For each distance class *d*, compute

$$I(d)=\frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n}w_{hi}(z_h-\bar{z})(z_i-\bar{z})}{\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2}.$$

Here, *n* is the total number of points, *W* is the number of pairs of points having distances within the distance class, and w_{hi} a weight function such that $w_{hi}=1$ if points *h* and *i* are within the distance class and $w_{hi}=0$ otherwise (Kronecker delta). Note that this equation is incorrect in some publications.

For the one-tailed critical level $I_{0.05}$, compute

$$S_1 = \frac{1}{2}\sum_{h=1}^{n}\sum_{i=1}^{n}\left(w_{hi} + w_{ih}\right)^2$$

$$S_2 = \sum_{i=1}^{n}\left(w_{i+} + w_{+i}\right)^2$$

$$b_2 = \frac{n\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^4}{\left(\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2\right)^2}$$

$$\mathrm{Var}(I) = \frac{n\left[\left(n^2 - 3n + 3\right)S_1 - nS_2 + 3W^2\right] - b_2\left[\left(n^2 - n\right)S_1 - 2nS_2 + 6W^2\right]}{(n-1)(n-2)(n-3)W^2} - \frac{1}{(n-1)^2}$$

$$I_{0.05} = 1.6452\sqrt{\mathrm{Var}(I)} - k_{0.05}(n-1)^{-1}$$

Here the w_{i+} and w_{+i} are the row and column sums. The correction factor $k_{0.05}$ is set to $\sqrt{10 \cdot 0.05} = 0.707$ if $4\left(n - \sqrt{n}\right) < W \leq 4\left(2n - 3\sqrt{n} + 1\right)$, otherwise $k_{0.05}$=1.

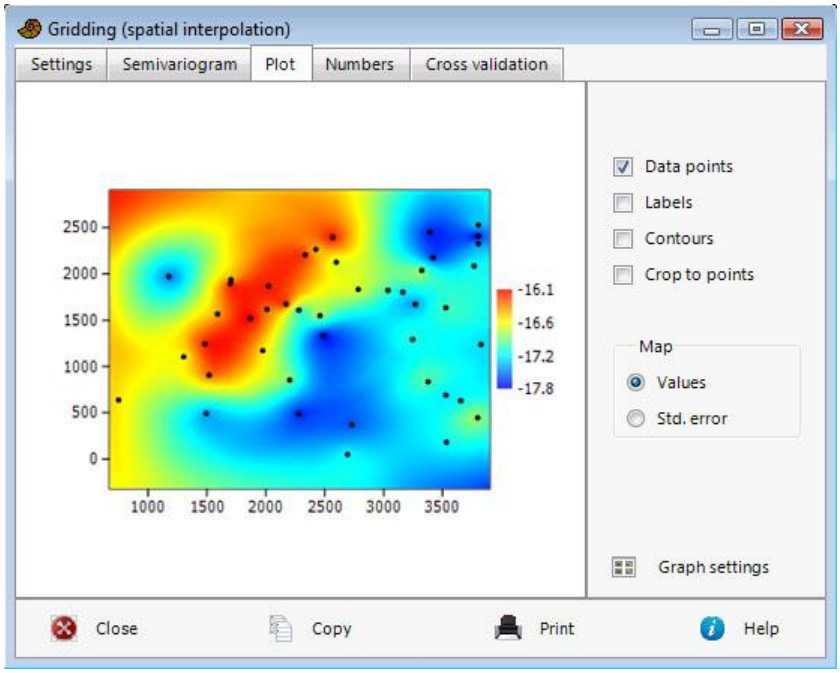
Reference

Legendre, P. & Legendre, L. 1998. Numerical Ecology, 2nd English ed. Elsevier, 853 pp.

Gridding (spatial interpolation)

"Gridding" is the operation of spatial interpolation of scattered 2D data points onto a regular grid. Three columns with position (x,y) and corresponding data values are required.

Gridding produces a map showing a continuous spatial estimate of some variate such as fossil abundance or thickness of a rock unit, based on scattered data points. The user can specify the size of the grid (number of rows and columns). The spatial coverage of the map is set to a square covering the data points. When plotting, this can be reduced to the convex hull of the points.



A least-squares linear surface (trend) is automatically fitted to the data, removed prior to gridding and finally added back in. This is primarily useful for the semivariogram modelling and the kriging method.

Cross validation: This option will remove each data point in turn and re-compute the surface based on the remaining points ("jackknife"). The differences between the original data values and the cross-validated values indicate the prediction accuracy of the surface model. These differences are reported for each point, together with the mean squared error (MSE) over all points.

Four interpolation algorithms are available:

Inverse distance weighting
The value at a grid node is simply the average of the *N* closest data points, as specified by the user (the default is to use all data points). The points are weighted in inverse proportion to distance. This algorithm is fast but will not always give good (smooth) results. A typical artefact is "bull's eyes" around data points. One advantage is that the interpolated values will never exceed the range of the data points. By setting *N*=1, this algorithm reduces to the *nearest-neighbour method*, which sets the value at a grid node to the value of the nearest data point.

Thin-plate spline
Maximally smooth interpolator. Can overshoot in the presence of sharp bends in the surface. This is a radial basis method with radial basis function $\varphi = r \ln r$.

Multiquadric
Radial basis function $\varphi = r$. Popular for terrain modelling.

Kriging
The user is required to specify a model for the semivariogram, by choosing one of four common models and corresponding parameters to fit the empirical semivariances (the residual sum of squares should be as small as possible). The semivariogram is computed within each of a number of bins. Using the histogram option, choose a number of bins so that each bin (except possibly the rightmost ones) contains at least 30 distances.

The *nugget* is a constant added to the model. It implies non-zero variance at zero distance and will therefore allow the surface to not pass exactly through the given data points. The *range* controls the extent of the curve along the distance axis. In the equations below, the normalized distance value *h* represents *distance*/*range*. The *scale* controls the extent of the curve along the variance axis.

Spherical: $$\gamma(h) = \begin{cases} nugget + scale\left(\frac{3h}{2} - \frac{1}{2}h^3\right) & h < 1 \\ nugget + scale & h \geq 1 \end{cases}$$

Exponential: $$\gamma(h) = nugget + scale\left(1 - e^{-h}\right)$$

Gaussian: $$\gamma(h) = nugget + scale\left(1 - e^{-h^2}\right)$$

Cubic: $$\gamma(h) = \begin{cases} nugget + scale\left(7h^2 - 8.75h^3 + 3.5h^5 - 0.75h^7\right) & h < 1 \\ nugget + scale & h \geq 1 \end{cases}$$

The "Optimize all" button will select the model and parameters giving the smallest residual sum of squares in the semivariogram. This may not be what you want: You may wish to use a specific model or to have zero nugget to ensure exact interpolation. You must then set the values manually.

The kriging procedure also provides an estimate of standard errors across the map (this depends on an accurate semivariogram model). Kriging in PAST does not provide for anisotropic semivariance.

Warning: Kriging is slow, do not attempt it for more than ca. 1000 data points on a 100x100 grid.

See e.g. Davis (1986) or de Smith et al. (2009) for more information on gridding.

References

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

de Smith, M.J., M.F. Goodchild & P.A. Longley. 2009. Geospatial Analysis, 3rd ed. Matador.

Multivariate allometry

This module is used for investigating allometry in a multivariate morphometric data set. It expects a multivariate data set with variables (distance measurements) in columns, specimens in rows.

This method for investigating allometry in a multivariate data set is based on Jolicoeur (1963) with extensions by Kowalewski et al. (1997). The data are (automatically) log-transformed and subjected to PCA. The first principal component (PC1) is then regarded as a size axis (this is only valid if the variation accounted for by PC1 is large, say more than 80%). The allometric coefficient for each original variable is estimated by dividing the PC1 loading for that variable by the mean PC1 loading over all variables.

95% confidence intervals for the allometric coefficients are estimated by bootstrapping specimens. 2000 bootstrap replicates are made.

Missing data is supported by column average substitution.

References

Jolicoeur, P. 1963. The multivariate generalization of the allometry equation. *Biometrics* 19:497-499.

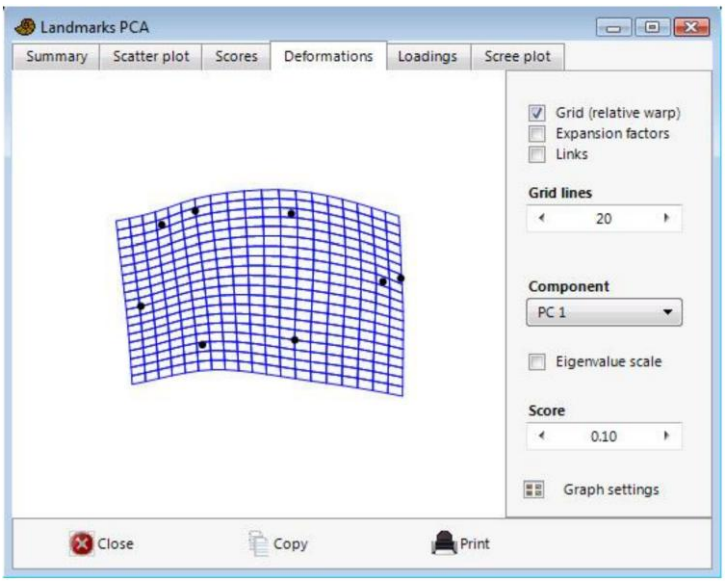
Kowalewski, M., E. Dyreson, J.D. Marcot, J.A. Vargas, K.W. Flessa & D.P. Hallmann. 1997. Phenetic discrimination of biometric simpletons: paleobiological implications of morphospecies in the lingulide brachiopod *Glottidia*. *Paleobiology* 23:444-469.

PCA of 2D landmarks (relative warps)

This module is very similar to the standard PCA module, but with some added functionality for analyzing 2D landmark configurations. The expected data are specimens in rows, alternating *x* and *y* coordinates in columns. Procrustes standardization recommended.

The relative warps (principal components) are ordered according to importance, and the first and second warps are usually the most informative. Note that this module does a straightforward PCA of the landmarks, meaning that the affine component is included in the analysis.

The relative warps are visualized with vectors and/or thin-plate spline transformation grids. When you increase or decrease the score factor away from zero, the original landmark configuration and grid will be progressively deformed according to the selected relative warp. Vectors are drawn from the mean to the deformed (dot) landmark position.



Thin-plate splines for 2D landmarks

This module shows a shape deformation from one landmark configuration to another. The expected data are specimens in rows, alternating *x* and *y* coordinates in columns. Procrustes standardization recommended.

Any shape selected in the "From shape" menu, is taken as a reference, with an associated square grid. The warps from this to all other specimens can be viewed. You can also choose the mean shape as the reference.

The 'Expansion factors' option will display the area expansion (or contraction) factor around each landmark in yellow numbers, indicating the degree of local growth. This is computed using the Jacobian of the warp. Also, the expansions are colour-coded for all grid elements, with green for expansion and purple for contraction.

At each landmark, the principal strains can also be shown, with the major strain in black and minor strain in brown. These vectors indicate directional stretching.

A description of thin-plate spline transformation grids is given by Dryden & Mardia (1998).

Partial warps and scores

From the thin-plate spline window, you can choose to see the partial warps for a particular spline deformation. The first partial warp will represent some long-range (large scale) deformation of the grid, while higher-order warps are normally connected with more local deformations. The affine component of the warp (also known as zeroth warp) represents linear translation, scaling, rotation, and shearing.

When you increase the amplitude factor from zero, the original landmark configuration and a grid will be progressively deformed according to the selected partial warp.

The partial warp scores of all the specimens are given in a table. Each partial warp score has two components (x and y).

Reference

Dryden, I.L. & K.V. Mardia 1998. Statistical Shape Analysis. Wiley.

Linear regression of 2D landmarks

Expects specimens in rows, with a single column of independent data (e.g. size) followed by pairs of columns with Procrustes-fitted landmark positions. Output includes deformation grids and displacement vectors, drawn from the mean to the deformed (dots) landmark positions.

Common Allometric Component analysis for 2D landmarks

Common Allometric Component (CAC) analysis for landmarks was first suggested by Mitteroecker et al. (2004). The principle is simple and logical: Do a linear regression of shape as a function of size (the allometric component) and then a PCA on the residuals (the residual shape components). The required data are one column of sizes, followed by pairs of columns containing Procrustes-fitted x-y coordinates of the landmarks. The data can be obtained from the original landmarks using the "Transform->Landmarks->Procrustes" function, and selecting "Add size column". A final feature of CAC analysis is that if groups are specified, then the data are centered on group means prior to analysis, effectively removing inter-group variation.

X is the *nxm* matrix of (Procrustes-fitted) shape coordinates, centered on group means. **s** is the n-vector with the logarithms of centroid sizes (in Past you can also choose to skip the log-transform of size). The vector of the "common allometric component" is

$$\mathbf{a} = \frac{\mathbf{X}^t\mathbf{s}}{\mathbf{s}^t\mathbf{s}}$$

normalized as $\mathbf{a}'=\mathbf{a}/\sqrt{\mathbf{a}^t\mathbf{a}}$. This common allometric component is projected out to produce a reduced data matrix

$$\mathbf{W} = \mathbf{X}(\mathbf{I} - \boldsymbol{a}'(\boldsymbol{a}')^t)$$

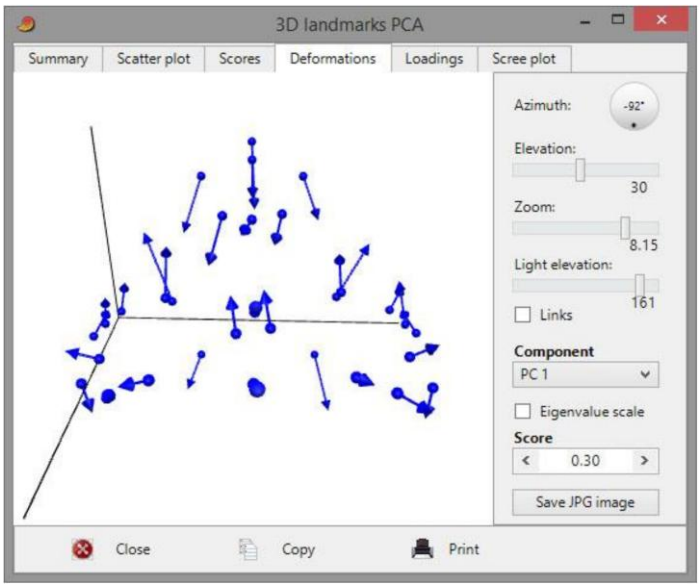
The principal components of **W** constitute the residual shape components. In Past, you can produce scatter plots of scores on the common allometric component and the residual shape components, and landmark deformations along all components can be visualized using vector displacements and thin-plate spline grids, relative to mean shape.

Reference

Mitteroecker, P., Gunz, P., Bernhard, M., Schaefer, K., Bookstein, F.L. 2004. Comparison of cranial ontogenetic trajectories among great apes and humans. *Journal of Human Evolution* 46:679-698.

PCA of 3D landmarks

Specimens in rows, 3D landmarks in triplets of columns (should be Procrustes-fitted first). The module is similar to the standard PCA module, but allows visualization of the principal components as 3D vectors (arrows) away from the mean configuration.



Linear regression of 3D landmarks

Expects specimens in rows, with a single column of independent data (e.g. size) followed by triplets of columns with Procrustes-fitted landmark positions. Output includes 3D plot of the displacement vectors, drawn from the mean to the deformed landmark positions.

Common Allometric Component analysis for 3D landmarks

See above for a description of CAC for 2D landmarks. Expects specimens in rows, with a single column of sizes followed by triplets of columns with Procrustes-fitted landmark positions. In this module, the displacement vectors can be visualized in 3D.

Size from landmarks (2D or 3D) NOT YET IN PAST 4

Digitized *x/y* or *x/y/z* landmark coordinates. Specimens in rows, coordinates with alternating x and y (and z for 3D) values in columns. Must not be Procrustes fitted or normalized for size!

Calculates the centroid size for each specimen (Euclidean norm of the distances from all landmarks to the centroid).

The values in the 'Normalized' column are centroid sizes divided by the square root of the number of landmarks - this might be useful for comparing specimens with different numbers of landmarks.

Normalize size

The 'Normalize size' option in the Transform menu allows you to remove size by dividing all coordinate values by the centroid size for each specimen. For 2D data you may instead use Procrustes coordinates, which are also normalized with respect to size.

See Dryden & Mardia (1998), p. 23-26.

Reference

Dryden, I.L. & K.V. Mardia 1998. Statistical Shape Analysis. Wiley.

Distance from landmarks (2D or 3D) NOT YET IN PAST 4

Digitized *x/y* or *x/y/z* landmark coordinates. Specimens in rows, coordinates with alternating x and y (and z for 3D) values in columns. May or may not be Procrustes fitted or normalized for size.

Calculates the Euclidean distances between two fixed landmarks for one or many specimens. You must choose two landmarks - these are named according to the name of the first column for the landmark (*x* value).

All distances from landmarks (EDMA) NOT YET IN PAST 3

Digitized *x/y* or *x/y/z* landmark coordinates. Specimens in rows, coordinates with alternating x and y (and z for 3D) values in columns. May or may not be Procrustes fitted or normalized for size.

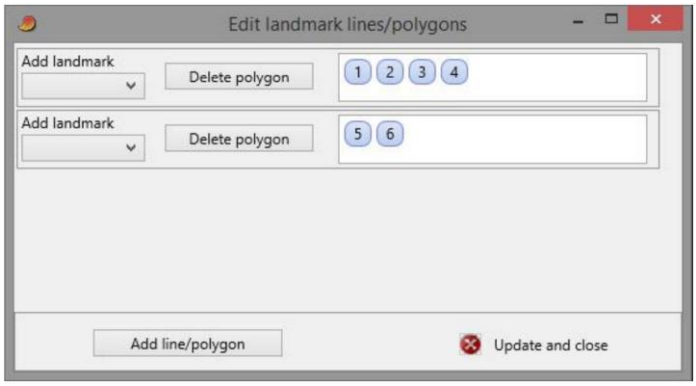
This function will replace the landmark data in the data matrix with a data set consisting of distances between all pairs of landmarks, with one specimen per row. The number of pairs is $N(N-1)/2$ for N landmarks. This transformation will allow multivariate analysis of distance data, which are not sensitive to rotation or translation of the original specimens, so a Procrustes fitting is not mandatory before such analysis. Using distance data also allows log-transformation, and analysis of fit to the allometric equation for pairs of distances.

Missing data is supported by column average substitution.

Edit landmark lines/polygons

This function in the Geometry menu under Landmarks (2D) and Landmarks (3D) allows the selection of landmarks to be linked with lines in the morphometric plots (PCA, thin-plate splines, etc.), to improve readability. The landmarks must be present in the main spreadsheet before links can be defined.

The window contains a list of lines or closed polygons. In the example below, the user has specified one polygon (consisting of four landmarks) and one line (two landmarks). Click on a landmark and press the Delete key to delete it from the list. The data will be written to the Past file when saved.



Elliptic Fourier shape analysis

Requires digitized *x/y* coordinates around outlines. Specimens in rows, coordinates of alternating *x* and *y* values in columns. Elliptic Fourier shape analysis is in several respects superior to simple Fourier shape analysis. One advantage is that the algorithm can handle complicated shapes which may not be expressible as a unique function in polar co-ordinates. Elliptic Fourier shapes is now a standard method of outline analysis. The algorithm used in PAST is described by Ferson et al. (1985).

EFA coefficients

Cosine and sine components of *x* and *y* increments along the outline for the first 30 harmonics are given, but only the first *N*/2 harmonics should be used, where *N* is the number of digitized points. Size and positional translation are normalized away, and do not enter in the coefficients. The size (before normalization) is given in the first column. The optional standardization for rotation or starting point, following Ferson et al., sometimes flips shapes around . This should be checked with the 'Shape view' (see below) – it may be necessary to remove such specimens.

The coefficients can be copied to the main spreadsheet for further analysis such as PCA and discriminant analysis.

The 'Shape view' window allows graphical viewing of the elliptic Fourier shape approximation(s).

EFA PCA

Principal Components Analysis of the EFA coefficients of the given outlines, with visualization of the principal components as EFA deformations. For more details on PCA in Past, see the description of PCA.

Reference

Ferson, S.F., F.J. Rohlf & R.K. Koehn. 1985. Measuring shape variation of two-dimensional outlines. *Systematic Zoology* 34:59-68.

Hangle Fourier shape analysis

Requires digitized *x/y* coordinates around outlines. Specimens in rows, coordinates of alternating *x* and *y* values in columns.

The "Hangle" method for analysing closed outlines, proposed by Haines & Crampton (2000) is a competitor to Elliptic Fourier Analysis. Hangle has certain advantages over EFA, the most important being that fewer coefficients are needed to capture the outline to a given precision. This is of importance for statistical testing (e.g. MANOVA) and discriminant analysis. The implementation in Past is based on the Hangle/Hmatch/Htree/Hshape package of Haines & Crampton (thanks to the authors for providing the source code).

The output consists of 46 Fourier coefficients, which are the cos and sin coefficients of the first 24 harmonics (modes), starting on harmonic number 2. Copy these numbers back to a Past spreadshhet for further multivariate shape analysis.

Starting point normalization

Usually leave at 'Match all', either with the 'Hmatch' or (perhaps preferably) the 'Htree' method to align all the outlines. Alternatively, select 2.-4. harmonic, which will phase shift each outline according to the selected mode (see Haines & Crampton 2000).

Smoothing

Increasing the smoothing parameter can reduce high-frequency noise, at the cost of dampening potentially informative high-frequency shape information.

Shape view

Use this function to inspect the shapes reconstructed from the Fourier coefficients. Check that the matching routine has not rotated any shape incorrectly. Also, use this function to select the minimum number of modes necessary for capturing the shape. In the example above, the number of modes has been set to 14, which captures 99.88% of the total integrated power (amplitude squared) of the selected shape. The number of modes is shown by the red line in the power spectrum – make sure that the main features of the spectrum are to the left of this line for all the shapes.

Note: PCA visualization and regression (as for EFA) has not yet been implemented for Hangle.

Reference

Haines, A.J. & J.S. Crampton. 2000. Improvements to the method of Fourier shape analysis as applied in morphometric studies. *Palaeontology* 43:765-783

Eigenshapes

Eigenshape analysis (Lohmann, 1983; Rohlf, 1986; MacLeod, 1999) can, somewhat imprecisely, be thought of as PCA of the raw outlines, without going through a transformation stage such as Fourier analysis. Past carries out the following steps:

1. Produce an equally spaced set of points by interpolation between the original points. The number of interpolated points along the outline is optimized automatically.

2. Going along the outline from a fixed point (homologous on all outlines), calculate the *tangent angle* from one point to the next. For *m* interpolated points on a closed outline there are *m* tangent angles, constituting a vector $\boldsymbol{\varphi}$ describing the shape.

Given a set of *n* equally spaced (*x*, *y*)-coordinates around a contour, the tangent angles are calculated in the following way (modified after Zahn and Roskies, 1972). First, compute the angular orientations of the tangent to the curve:

$$\varphi_i = \tan^{-1}\frac{y_i - y_{i-1}}{x_i - x_{i-1}}$$

For a closed contour of *n* points, we use

$$\varphi_1 = \tan^{-1}\frac{y_1 - y_n}{x_1 - x_n}$$

This vector is then normalized by subtracting the angular orientations that would be observed for a circle:

$$\varphi_i^* = \varphi_i - \varphi_1 - \frac{2\pi(i-1)}{n}$$

3. The shape vectors for the *n* shapes are subjected to PCA, giving principal components that are referred to as *eigenshapes*. The eigenshapes are themselves tangent angle vectors, given in decreasing order of amount of shape variation they explain. The first (most important) eigenshapes define a low-dimensional space into which the original specimens can be projected.

References

Lohmann, G.P. 1983. Eigenshape analysis of microfossils: a general morphometric method for describing changes in shape. *Mathematical Geology* 15: 659–672.

MacLeod, N. 1999. Generalizing and extending the eigenshape method of shape space visualization and analysis. *Paleobiology* 25: 107–138.

Rohlf, F.J. 1986. Relationships among eigenshape analysis, Fourier analysis and analysis of coordinates. *Mathematical Geology* 18: 845–854.

Zahn, C.T., Roskies, R.Z. 1972. Fourier descriptors for plane closed curves. *IEEE Transactions, Computers* C-21: 269–281.

Coordinate transformation

Conversion between geographical coordinates in different grids and datums. The number of input columns depends on the data type, as described below.

Decimal degrees (WGS84)

Two columns: Latitude and longitude, in decimal degrees (60.5 is 60 degrees, 30 minutes). Negative values for south of equator and west of Greenwich. Referenced to the WGS84 datum.

Deg/decimal mins (WGS84)

Four columns: Latitude degrees, deci mal minutes (40.5 is 40 minutes, 30 seconds), longitude degrees, decimal minutes. Referenced to the WGS84 datum.

Deg/min/sec (WGS84)

Six columns: Latitude degrees, minutes, seconds, longitude degrees, minutes, seconds. Referenced to the WGS84 datum.

UTM-ED50 (Intl 1924)

Three columns: Easting (meters), northing (meters), and zone. Use negative zone numbers for the southern hemisphere. The handling of UTM zones takes into account the special cases of Svalbard and western Norway. Referenced to the ED50 European datum at Potsdam.

UTM-WGS84 (WGS84)

Three columns: Easting (meters), northing (meters), and zone. Referenced to the WGS84 datum.

UTM-NAD27 (Clarke 1866)

Three columns: Easting (meters), northing (meters), and zone. Referenced to the NAD27 datum. Conversion to/from this format is slightly inaccurate (5-6 meters).

UTM-NAD83 (GRS80)

Three columns: Easting (meters), northing (meters), and zone. Referenced to the NAD83 datum (practically identical to WGS84).

Sweden (RT90)

Two columns: Easting (meters) and northing (meters).

The transformations are based on code generously provided by I. Scollar.

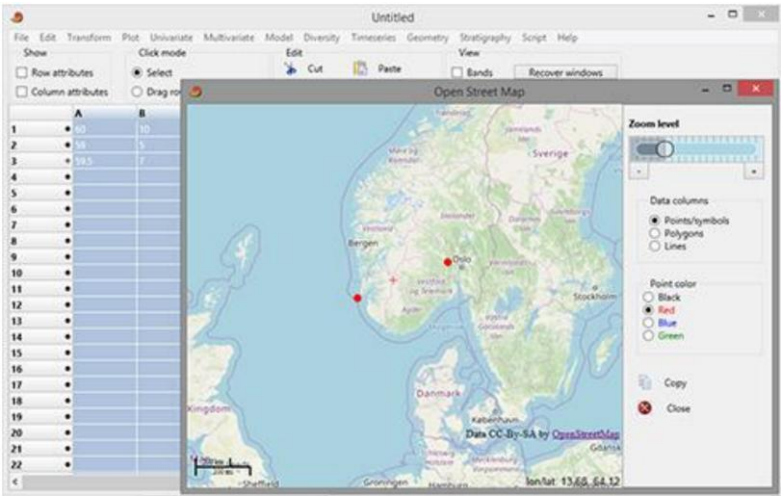
Open Street Map

Takes two columns of latitudes and longitudes in decimal degrees (WGS84) and shows an Open Street Map window with graphic elements at the given coordinates in one of these ways:

- Points with symbols and colours taken from the Past spreadsheet
- Filled polygons. The colour is taken from the row colour of the first point. Use an additional group column to specify multiple polygons.
- Multi-segment lines, with colour and groups as for filled polygons
- Bubbles, with the radius in km of each bubble given in the third data column.
- Pie charts, with multiple data columns starting in the third column.

An optional third column can be used to specify the sizes of bubbles (radius in km) or thicknesses of lines (pixels).

This module requires Internet connection.



Please note: Use of these maps requires that you include the appropriate copyright notice. For OpenStreetMap, this is:

"© OpenStreetMap contributors (ODbL)"

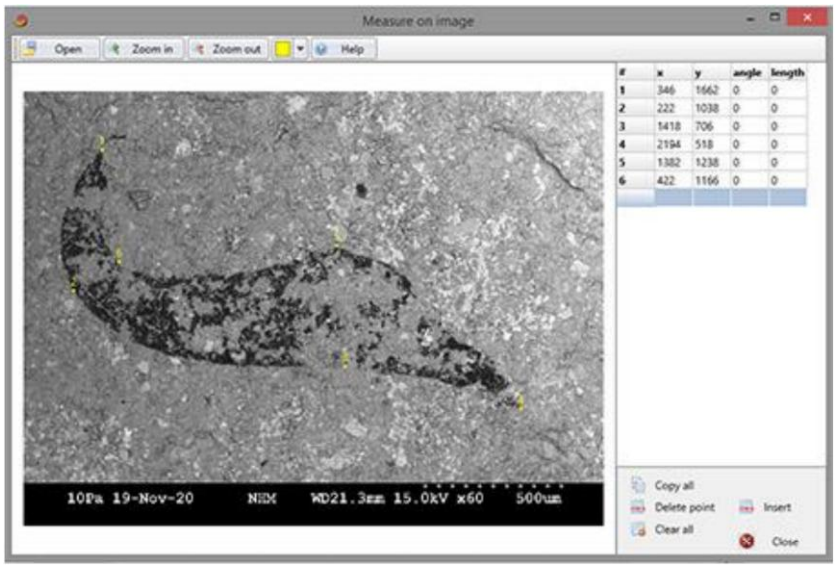
For the open topomap:

"Map data: © OpenStreetMap contributors, SRTM. Map design: © OpenTopoMap (CC-BY-SA)"

Measure on image

A simple tool to digitize point co-ordinates, distances and directions on images. Click “Open” to open an image. A list of measurements is given on the right. If you start this tool with selected data in the main spreadsheet, these data will be pre-loaded into the measurement list. The measurement list can be copy-pasted back to the Past spreadsheet with the “Copy all” button.

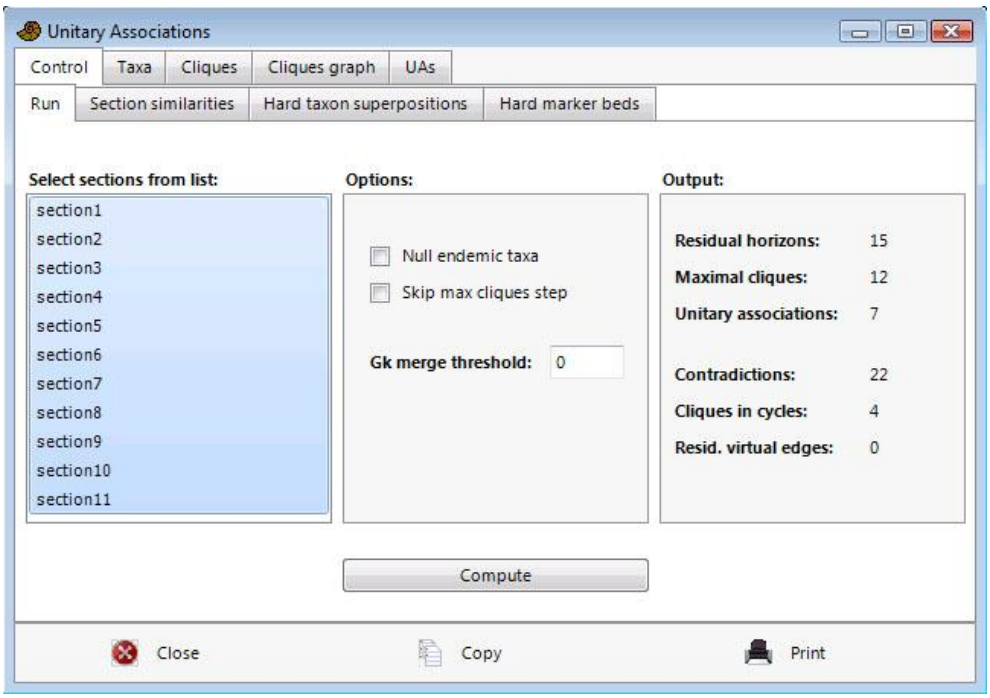
New measurements will enter in the selected row in the list. You can move a point by clicking on the row in the list, then clicking on the new position in the image. You can also delete or insert points in the list.



Stratigraphy menu

Unitary Associations

Unitary Associations analysis (Guex 1991; Guex et al. 2016) is a method for biostratigraphical correlation (see Angiolini & Bucher 1999 for a typical application). The data input consists of a presence/absence matrix with samples in rows and taxa in columns. Samples belonging to the same section (locality) must be assigned to the same group, and ordered stratigraphically within each section such that the lowermost sample enters in the lowest row.



Overview of the method

The method of Unitary Associations is logical, but rather complicated, consisting of a number of steps. For details, see Guex (1991). The implementation in PAST includes most of the features found in the original program, called BioGraph (Savary & Guex 1999), and thanks to a fruitful co-operation with Jean Guex it also includes a number of additional options and improvements.

The basic idea is to generate a number of assemblage zones (similar to 'Oppel zones') which are optimal in the sense that they give maximal stratigraphic resolution with a minimum of superpositional contradictions. One example of such a contradiction would be a section containing a species A above a species B, while assemblage 1 (containing species A) is placed below assemblage 2 (containing species B). PAST carries out the following steps:

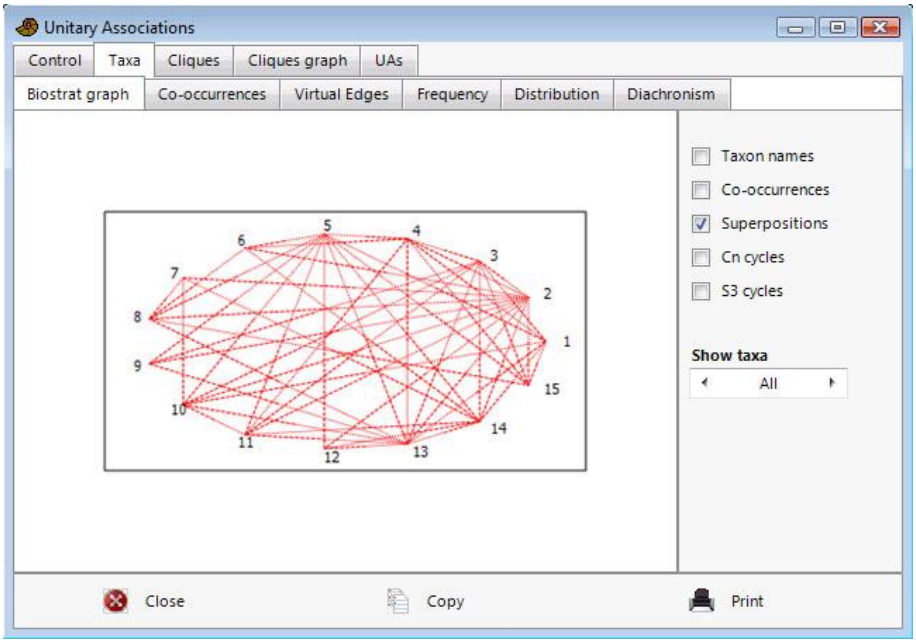
1. Residual maximal horizons

The method makes the range-through assumption, meaning that taxa are considered to have been present at all levels between the first and last appearance in any section. Then, any samples with a set of taxa that is contained in another sample are discarded. The remaining samples are called *residual maximal horizons*. The idea behind this throwing away of data is that the absent taxa in the discarded samples may simply not have been found even though they originally existed. Absences are therefore not as informative as presences.

2. Superposition and co-occurrence of taxa

Next, all pairs (A,B) of taxa are inspected for their superpositional relationships: A below B, B below A, A together with B, or unknown. If A occurs below B in one locality and B below A in another, they are considered to be co-occurring although they have never actually been found together.

The superpositions and co-occurrences of taxa can be viewed in the *biostratigraphic graph*. In this graph, taxa are coded as numbers. Co-occurrences between pairs of taxa are shown as solid blue lines. Superpositions are shown as arrow lines, pointing from the below-occurring taxon to the above-occurring taxon.



Some taxa may occur in so-called *forbidden sub-graphs*, which indicate inconsistencies in their superpositional relationships. Two of the several types of such sub-graphs can be plotted in PAST: C_n *cycles*, which are superpositional cycles (A->B->C->A), and S_3 *circuits*, which are inconsistencies of the type 'A co-occurring with B, C above A, and C below B'. Interpretations of such forbidden sub-graphs are suggested by Guex (1991).

3. Maximal cliques

Maximal cliques are groups of co-occurring taxa not contained in any larger group of co-occurring taxa. The maximal cliques are candidates for the status of unitary associations, but will be further processed below. In PAST, maximal cliques receive a number and are also named after a maximal horizon in the original data set which is identical to, or contained in (marked with asterisk), the maximal clique.

4. Superposition of maximal cliques

The superpositional relationships between maximal cliques are decided by inspecting the superpositional relationships between their constituent taxa, as computed in step 2. Contradictions (some taxa in clique A occur below some taxa in clique B, and vice versa) are resolved by a 'majority vote'. The contradictions between cliques can be viewed in PAST.

The superpositions and co-occurrences of cliques can be viewed in the *maximal clique graph*. In this graph, cliques are coded as numbers. Co-occurrences between pairs of cliques are shown as solid blue lines. Superpositions are shown as red arrow lines, pointing from the below-occurring clique to the above-occurring clique. Also, cycles between maximal cliques (see below) can be viewed as green lines.

5. Resolving cycles

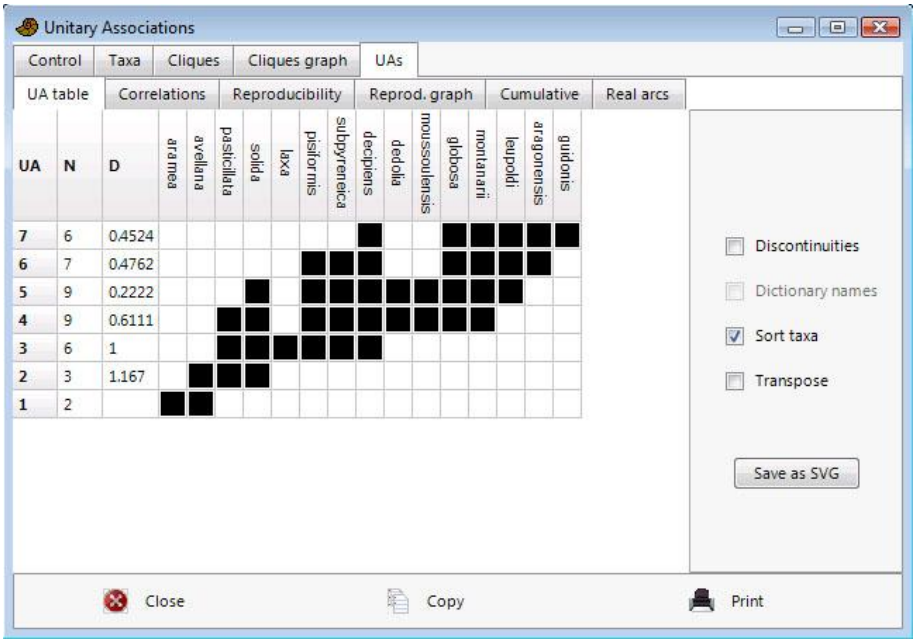
It will sometimes be the case that maximal cliques are now ordered in cycles: A is below B, which is below C, which is below A again. This is clearly contradictory. The 'weakest link' (superpositional relationship supported by fewest taxa) in such cycles is destroyed.

6. Reduction to unique path

At this stage, we should ideally have a single path (chain) of superpositional relationships between maximal cliques, from bottom to top. This is however often not the case, for example if A and B are below C, which is below D, or if we have isolated paths without any relationships (A below B and C below D). To produce a single path, it is necessary to merge cliques according to special rules.

7. Post-processing of maximal cliques

Finally, a number of minor manipulations are carried out to 'polish' the result: Generation of the 'consecutive ones' property, reinsertion of residual virtual co-occurrences and superpositions, and compaction to remove any generated non-maximal cliques. For details on these procedures, see Guex (1991). At last, we now have the Unitary Associations, which can be viewed in PAST.



The unitary associations have associated with them an index of similarity from one UA to the next, called D:

$D_i=|UA_i-UA_{i-1}| / |UA_i| +|UA_{i-1}-UA_i| / |UA_{i-1}|$

8. Correlation using the Unitary Associations

The original samples are now correlated using the unitary associations. A sample may contain taxa which uniquely places it in a unitary association, or it may lack key taxa which could differentiate between two or more unitary associations, in which case only a range can be given. These correlations can be viewed in PAST.

9. Reproducibility matrix

Some unitary associations may be identified in only one or a few sections, in which case one may consider to merge unitary associations to improve the geographical reproducibility (see below). The reproducibility matrix should be inspected to identify such unitary associations. A UA which is uniquely identified in a section is shown as a black square, while ranges of UAs (as given in the correlation list) are shown in gray.

10. Reproducibility graph and suggested UA merges (biozonation)

The reproducibility graph (Gk' in Guex 1991) shows the superpositions of unitary associations that are actually observed in the sections. PAST will internally reduce this graph to a unique maximal path (Guex 1991, section 5.6.3), and in the process of doing so it may merge some UAs. These mergers are

shown as red lines in the reproducibility graph. The sequence of single and merged UAs can be viewed as a suggested biozonation.

Special functionality

The implementation of the Unitary Associations method in PAST includes a number of options and functions which have not yet been described in the literature. For questions about these, please contact us.

References

Angiolini, L., Bucher, H. 1999. Taxonomy and quantitative biochronology of Guadalupian brachiopods from the Khuff Formation, Southeastern Oman. *Geobios* 32:665-699.

Guex, J. 1991. *Biochronological Correlations*. Springer Verlag.

Guex, J., Galster, F., Hammer, Ø. 2016. *Discrete Biochronological Time Scales*. Springer.

Savary, J., Guex, J. 1999. Discrete Biochronological Scales and Unitary Associations: Description of the BioGraph Computer Program. *Meomoires de Geologie (Lausanne)* 34.

Ranking-Scaling

Ranking-Scaling (Agterberg & Gradstein 1999) is a method for quantitative biostratigraphy based on *events* in a number of wells or sections. The data input consists of wells in rows with one well per row, and events (e.g. FADs and/or LADs) in columns. The values in the matrix are depths of each event in each well, increasing upwards (you may want to use negative values to achieve this). Absences are coded as zero. If only the order of events is known, this can be coded as increasing whole numbers (ranks, with possible ties for co-occurring events) within each well.

The implementation of ranking-scaling in PAST is not comprehensive, and advanced users are referred to the RASC and CASC programs of Agterberg and Gradstein.

Overview of the method

The method of Ranking-Scaling proceeds in two steps:

1. Ranking

The first step of Ranking-Scaling is to produce a single, comprehensive stratigraphic ordering of events, even if the data contains contradictions (event A over B in one well, but B over A in another), or longer cycles (A over B over C over A). This is done by 'majority vote', counting the number of times each event occurs above, below or together with all others. Technically, this is achieved by Presorting followed by the Modified Hay Method (Agterberg & Gradstein 1999).

2. Scaling

The biostratigraphic analysis may end with ranking, but additional insight may be gained by estimating stratigraphic distances between the consecutive events. This is done by counting the number of observed superpositional relationships (A above or below B) between each pair (A,B) of consecutive events. A low number of contradictions implies long distance.

Some computed distances may turn out to be negative, indicating that the ordering given by the ranking step was not optimal. If this happens, the events are re-ordered and the distances re-computed in order to ensure only positive inter-event distances.

RASC in PAST

Parameters

- Well threshold: The minimum number of wells in which an event must occur in order to be included in the analysis
- Pair threshold: The minimum number of times a relationship between events A and B must be observed in order for the pair (A,B) to be included in the ranking step
- Scaling threshold: Pair threshold for the scaling step
- Tolerance: Used in the ranking step (see Agterberg & Gradstein)

Ranking

The ordering of events after the ranking step is given, with the first event at the bottom of the list. The "Range" column indicates uncertainty in the position.

Scaling

The ordering of the events after the scaling step is given, with the first event at the bottom of the list. For an explanation of all the columns, see Agterberg & Gradstein (1999).

Event distribution

A plot showing the number of events in each well, with the wells ordered according to number of events.

Scattergrams

For each well, the depth of each event in the well is plotted against the optimum sequence (after scaling). Ideally, the events should plot in an ascending sequence.

Dendrogram

Plot of the distances between events in the scaled sequence, including a dendrogram which may aid in zonation.

Variance analysis

For each event, this function plots the deviations from the line of correlation (see above) across all the wells. This gives a graphical representation of the biostratigraphic quality of each event.

Reference

Agterberg, F.P. & F.M. Gradstein. 1999. The RASC method for Ranking and Scaling of Biostratigraphic Events. In: Proceedings Conference 75th Birthday C.W. Drooger, Utrecht, November 1997. *Earth Science Review* 46(1-4):1-25.

Constrained optimization (CONOP)

Requires a table of depths/levels, with wells/sections in rows and event pairs in columns: FADs in odd columns and LADs in even columns. Missing events coded with zeros.

PAST includes a simple version of Constrained Optimization (Kemple et al. 1989). Both FAD and LAD of each taxon must be specified in alternate columns. Using so-called Simulated Annealing, the program searches for a global (composite) sequence of events that implies a minimal total amount of range extension (penalty) in the individual wells/sections. The parameters for the optimization procedure include an initial annealing temperature, the number of cooling steps, the cooling ratio (percentage lower than 100), and the number of trials per step. For explanation and recommendations, see Kemple et al. (1989).

Output windows include the optimization history with the temperature and penalty as function of cooling step, the global composite solution and the implied ranges in each individual section.

The implementation of CONOP in PAST is based on a FORTRAN optimization core provided by Sadler and Kemple.

Reference

Kemple, W.G., P.M. Sadler & D.J. Strauss. 1989. A prototype constrained optimization solution to the time correlation problem. In Agterberg, F.P. & G.F. Bonham-Carter (eds), Statistical Applications in the Earth Sciences. *Geological Survey of Canada Paper* 89-9:417-425.

Range confidence intervals

Estimation of confidence intervals for first or last appearances or total stratigraphic range, for one taxon.

Assuming a random (Poisson) distribution of fossiliferous horizons, and given the first occurrence datum (level), last occurrence datum, and total number of horizons where the taxon is found, we can calculate confidence intervals for the stratigraphic range of one taxon (Strauss & Sadler 1989, Marshall 1990).

No data are needed in the spreadsheet. The program will ask for the number of horizons where the taxon is found, and levels or dates for the first and last appearances. If necessary, use negative values to ensure that the last appearance datum has a higher numerical value than the first appearance datum. 80%, 95% and 99% confidence intervals are calculated for the FAD considered in isolation, the LAD considered in isolation, and the total range.

The value α is the length of the confidence interval divided by the length of the observed range.

For the single endpoint case:

$$\alpha = (1 - C_1)^{-1/(H-1)} - 1$$

where C_1 is the confidence level and H the number of fossiliferous horizons.

For the joint endpoint (total range) case, α is found by iterative solution of the equation

$$C_2 = 1 - 2(1 + \alpha)^{-(H-1)} + (1 + 2\alpha)^{-(H-1)}$$

The assumption of random distribution will of course not hold in many real situations.

Solow's method

Past also includes a method due to Solow (2003) that does not assume a uniform (stationary) distribution of finds. It only uses the occurrence datums of the lowest, second lowest, highest and second highest finds.

References

Marshall, C.R. 1990. Confidence intervals on stratigraphic ranges. *Paleobiology* 16:1-10.

Solow, A.R. 2003. Estimation of stratigraphic ranges when fossil finds are not randomly distributed. *Paleobiology* 29:181-185.

Strauss, D. & P.M. Sadler. 1989. Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. *Mathematical Geology* 21:411-427.

Distribution-free range confidence intervals

Estimation of confidence intervals for first or last appearances. Assumes there is no correlation between stratigraphic position and gap size. Section should be continuously sampled. Expects one column per taxon, with levels or dates of all horizons where the taxon is found. This method (Marshall 1994) does not assume random distribution of fossiliferous horizons. It does require that the levels or dates of all horizons containing the taxon are given. The program outputs upper and lower bounds on the lengths of the confidence intervals, using a 95 percent confidence probability, for confidence levels of 50, 80 and 95 percent. Values which cannot be calculated are marked with an asterisk (see Marshall 1994).

Reference

Marshall, C.R. 1994. Confidence intervals on stratigraphic ranges: partial relaxation of the assumption of randomly distributed fossil horizons. *Paleobiology* 20:459-469.

Sequence slotting

Requires two groups (marked with a group column) of univariate or multivariate data, the rows in each group arranged in sequence. Sequence slotting will merge the two sequences into one, in a way that minimizes the total path length from point to point, according to the selected distance measure. A typical application is the correlation of geophysical or geochemical data from two wells or sections. The algorithm is based on Clark (1985).

An optional third group with up to three columns specifies constraints on the sequence. Each row in this group specifies one constraint using three numbers: a constraint type code followed by up to two integer parameters A and B, referring to row numbers within each original sequence.

Code	A	B	Function
1	s	t	A_s and B_t are near-neighbors
2	s	t	A_s and B_t are adjacent
3	s	t	A_s precedes B_t
4	s	t	A_s follows B_t
5	s	t	Subsequence A_s , B_t , A_{s+1}
6	s	t	Subsequence B_t , A_s , B_{t+1}
7			Sequence starts with A_1
8			Sequence starts with B_1
9			Sequence ends with A_m
10			Sequence ends with B_n
11	s		No B between A_s and A_{s+1}
12		t	No A between B_t and B_{t+1}

Reference

Clark, R.M. 1985. A FORTRAN program for constrained sequence-slotting based on minimum combined path length. *Computers & Geosciences* 11: 605-617.

Diversity and rates

This module computes curves for diversity (species richness), origination and extinction from a presence-absence (binary) matrix with the oldest sample or interval in the last (lowest) row, and the youngest in the first row.

The many options reflect the considerable literature on such curves – for a short overview see Hammer and Harper (2024, pp. 192-196).

Richness

Species richness is estimated simply by counting the number of presences in each row. The *Range-through* option will fill in any absences between the first and last appearance of each species. The *Endpoint adjust* option will count any species beginning or ending in the row as 0.5 unit, while species confined to the row (singletons) will count 1/3 unit. This option should be used if rows represent time intervals instead of point samples (Foote 2000; Hammer & Harper 2024).

Origination and extinction – method "Simple"

Counts metric: The simple number of first appearances in the row (originations) and the number of last appearances (extinctions). These numbers are calculated with or without range-through, as selected by the Range-through option.

Ratios metric: Simple origination/extinction counts divided by the number of presences in the row.

Rates metric: Calculated according to Foote (2000) as follows.

Origination rate: $R_o = \ln \frac{N_{bt}+N_t}{N_{bt}}$ where N_t is the number of taxa present both in the focus row and the row above, and N_{bt} is the number of taxa present both in the focus row and in the rows below and above.

Extinction rate: $R_e = \ln \frac{N_{bt}+N_b}{N_{bt}}$ where N_b is the number of taxa present both in the focus row and the row below.

Origination and extinction, method "Three-timer"

The data are not subjected to range-through.

Counts metric: Not available

Ratios metric: Calculated according to Alroy (2008) as follows.

Per-interval sampling probability $s = \frac{t_3}{t_3+p}$ where t_3 is number of taxa present both in the focus row and in the rows below and above ("three-timers"), and p is the number of taxa present below and above but not in the focus row ("part-timers").

Origination ratio: $Ra_o = s\frac{t_2^U}{t_3}$ where $t_2{}^U$ is the number of species present in the focus row and the row above ("upper two-timers").

Extinction ratio: $Ra_e = s\frac{t_2^L}{t_3}$ where $t_2{}^L$ is the number of species present in the focus row and the row below ("lower two-timers").

Rates metric: Logarithms of the ratios.

Origination and extinction, method "Gap-filler"

Counts metric: Not available

Ratios metric: Calculated according to Alroy (2014) as

Origination ratio: $Ra_o = \frac{t_2^U+p}{t_3+p+g}$ where *g* is the number of species present in the row above and two rows below ("gap fillers"). It does not matter if it was found in the focus row.

Extinction ratio: $Ra_e = \frac{t_2^L+p}{t_3+p+g}$ where *g* is the number of species present in the row below and two rows above.

Rates metric: Logarithms of the ratios.

References

Alroy, J. 2008. Dynamics of origination and extinction in the marine fossil record. *Proceedings of the National Academy of Sciences USA* 105:11,536–11,542.

Alroy, J. 2014. Accurate and precise estimates of origination and extinction rates. *Paleobiology* 40: 374–397.

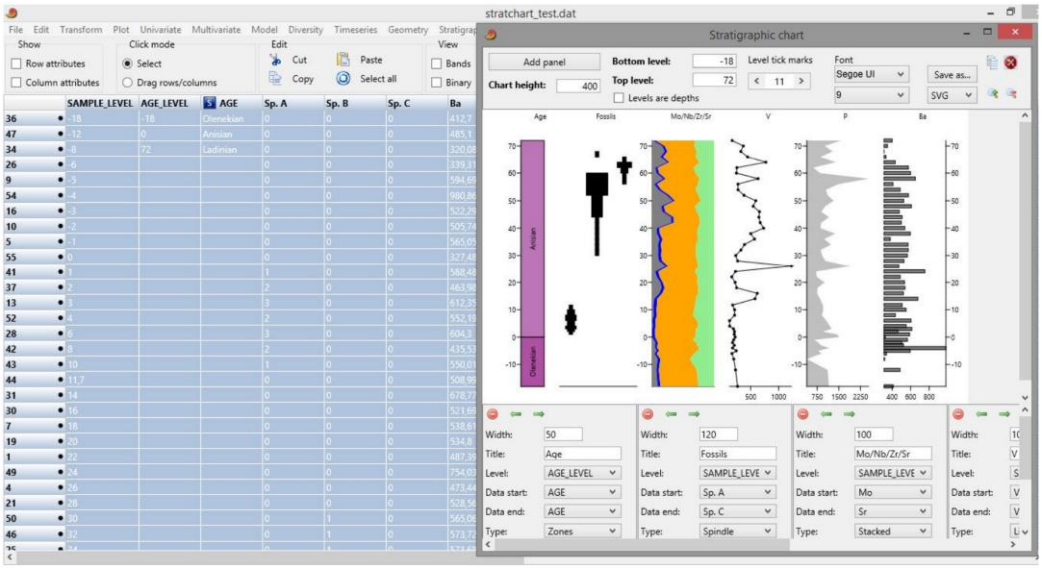
Foote, M. 2000. Origination and extinction components of taxonomic diversity: general problems. *Paleobiology* 26:74–102.

Hammer, Ø. & Harper, D.A.T. 2024. Paleontological Data Analysis, 2nd ed. Elsevier.

Stratigraphic chart

This flexible plotting module can produce well logs and logs of stratigraphic sections. It provides for multiple plots, called panels, of different types, making it possible to combine e.g. biozones, lithology, geochemical and geophysical logs and pollen diagram in one figure. The plotting setup is automatically saved in your Past file together with the data in the spreadsheet.

Note: This module is under construction; expect bugs and lack of features. Especially, you should not delete, add or rearrange columns in the spreadsheet while working with a stratigraphic chart, this will cause unpredictable behaviour.



The spreadsheet should contain one or more columns of stratigraphic levels (typically meter levels), and several columns of data collected at the given levels. When opening the module you will see a blank page. Use the "Add panel" button to add plots. The Zoom buttons are useful for navigating in large charts.

Global settings

Above the plot are found the settings affecting all the panels. The *Chart height* is the height of the chart in pixels. The *Bottom* and *Top* levels are the vertical limits of the chart. The *Levels are depths* option controls the vertical orientation – select this if your levels increase downwards in the section or core.

Panel settings

These are the settings affecting each panel. The *Width* is the width of the panel in pixels. *Level* sets the column containing the vertical levels for each data point. *Data start* and *Data end* set the range of columns for the data in this panel. Often, each panel will show only one data series, and then Data start and Data end should be identical.

Type sets the plot type (Line, Dots, Line+dots, Silhouette, Bars, Stacked, Spindle, Zones, Lithology).

The spindle diagram is used for fossil occurrences (abundances or presence-absence).

The Zones plot displays a stratigraphic succession of intervals such as biozones, periods or stages (see the leftmost panel in the example above). The levels represent the basal levels of each interval. An extra data point is required for the top of the last interval. The data should be a single column of type 'String', with the names of the intervals. Chronostratigraphic units such as 'Precambrian', 'Ordovician', 'Upper Ordovician', 'Pliensbachian' are recognized by the program and will be plotted using the colors specified by the International Commission on Stratigraphy.

The Lithology plot displays a column with lithology patterns. Levels are given as for the Zones plot (above). The data should consist of a column of type 'String', containing lithology codes (not case sensitive). A second data column may be given with a width in percent (0-100).

Sst	Sandstone
Cst	Coarse sandstone
Cgl	Conglomerate
Slt	Siltstone
Sha	Shale
Lst	Limestone
Sls	Silty limestone
Ign	Igneous rock
Mis	Missing/covered (cross)

Radiocarbon calibration

This module expects one or more rows of radiocarbon dates, with two columns for uncalibrated age (BP 1950) and its laboratory-reported standard error (one-sigma). Three calibration curves are included: IntCal20 (Reimer et al. 2020) for atmospheric (terrestrial) samples, SHcal20 (Hogg et al. 2020) for atmospheric southern hemisphere samples, and MarineCal20 (Heaton et al. 2020) for marine samples. Please include a reference to Reimer et al. (2020), Hogg et al. (2020) or Heaton et al. (2020) when reporting calibration results (as well as Past of course!).

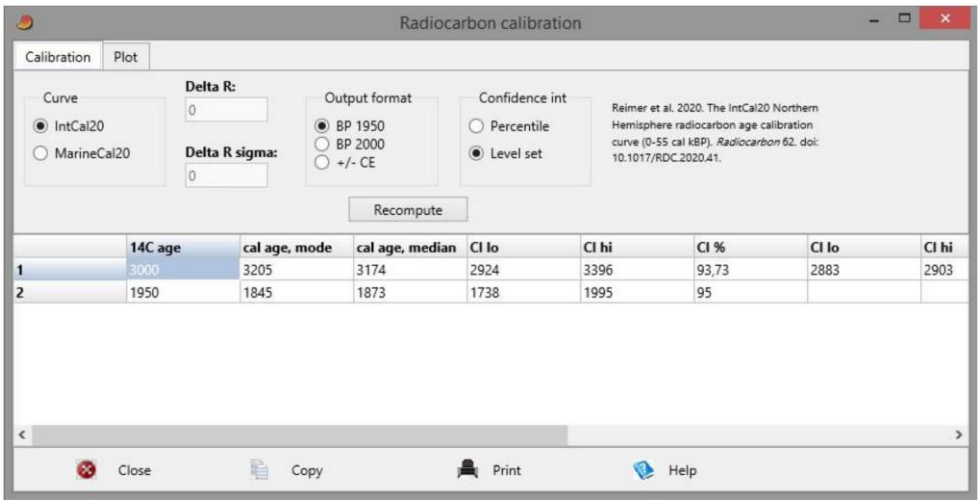
For the marine calibration, a globally averaged reservoir age (around 400 years, but varying with time) is included in the calibration. However, a local reservoir age correction (delta R) and its one-sigma error should be entered. These values can be looked up e.g. in the Marine Reservoir Correction Database at http://calib.org/marine/

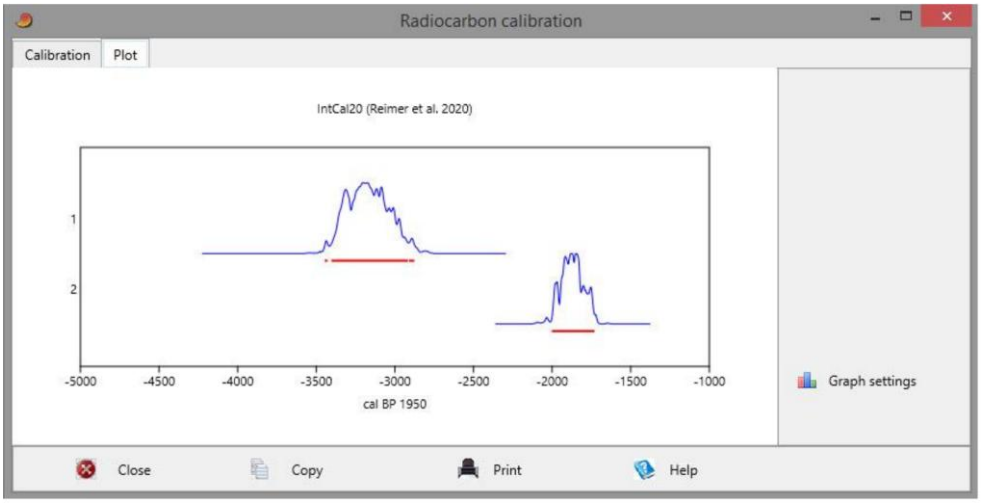
The output ages can be reported in cal BP 1950 (years before 1950), BP 2000, or CE (Common Era). The CE values are reported with negative or positive sign and with year zero included (ISO 8601).

Very young and old samples can be somewhat inaccurately calibrated because of the limited range of the calibration curve. Check the calibration plot to see if the probability curve is truncated.

Plotted probability curves are normalized to a peak value of 1.

Calibrated ages should always be reported with confidence intervals (and ideally the complete probability curve), but the median and mode (age of peak probability) are also given by Past as convenient short-hands. The median is more commonly used, but Michczyński (2007) found the mode to perform better in simulation studies.





Confidence intervals

For consistency with other modules in Past, the confidence intervals on calibrated ages are given with respect to the 95% range, not 95.4% as in some other calibration software. This can give slightly shorter confidence intervals in Past.

The "Percentile" option calculates a single confidence interval from the 2.5 to the 97.5 percentile points. The "Level set" option calculates a probability threshold such that the sum of all probabilities larger than the threshold is 0.95. All the corresponding t values are included in the CI. Because the probability curve can be multimodal, this can give rise to several disjoint segments in the CI. Past reports the three of these segments containing the largest areas (reported in %), which is usually sufficient to reach 95%.

Both these options produce valid 95% confidence intervals. The "Level set" option gives a more detailed view, and is used in most other calibration software, but the "Percentile" option is more convenient and gives a sufficiently informative CI for most purposes (personal opinion!). If only a single segment is sufficient for the level set CI, this will normally be very similar to the percentile CI.

The confidence interval is shown as red lines in the plot.

Computational details

The pointwise calculation of the probability curve $p(t)$ as a function of calibrated time follows Equation (2) in Bronk Ramsey (2001):

$$p(t) = \frac{\exp\left[-0.5\left(r_m - r(t)\right)^2 / \left(\sigma_m^2 + \sigma^2(t)\right)\right]}{\sqrt{\sigma_m^2 + \sigma^2(t)}}$$

where the calibration curve is $r(t)$ with standard error $\sigma(t)$, and the given radiocarbon date is $r_m \pm \sigma_m$. This equation is evaluated with a time step of 1 year.

Values for $r(t)$ and $\sigma(t)$ are linearly interpolated from the tables, which is reasonably accurate given the fine spacing of the IntCal20 and MarineCal20 curves.

The given error in delta R is included in σ_m (added in quadrature) before calculation of $p(t)$.

References

Bronk Ramsey, C. 2001. Development of the radiocarbon calibration program. *Radiocarbon* 43:355-363

Heaton TJ, Köhler P, Butzin M, Bard E, Reimer RW, Austin WEN, Bronk Ramsey C, Grootes PM, Hughen KA, Kromer B, Reimer PJ, Adkins J, Burke A, Cook MS, Olsen J, Skinner LC. 2020. Marine20-the marine radiocarbon age calibration curve (0-55,000 cal BP). *Radiocarbon* 62. doi: 10.1017/RDC.2020.68.

Hogg AG, Heaton TJ, Hua Q, Palmer JG, Turney CSM, Southon J, Bayliss A, Blackwell PG, Boswijk G, Bronk Ramsey C, Pearson C, Petchey F, Reimer P, Reimer R, Wacker L. 2020. SHCal20 Southern Hemisphere calibration, 0-55,000 years cal BP. *Radiocarbon* 62. doi: 10.1017/RDC.2020.59.

Michczyński, A. 2007. Is it possible to find a good point estimate of a calibrated radiocarbon date? *Radiocarbon* 49:393-401.

Reimer P, Austin WEN, Bard E, Bayliss A, Blackwell PG, Bronk Ramsey C, Butzin M, Cheng H, Edwards RL, Friedrich M, Grootes PM, Guilderson TP, Hajdas I, Heaton TJ, Hogg AG, Hughen KA, Kromer B, Manning SW, Muscheler R, Palmer JG, Pearson C, van der Plicht J, Reimer RW, Richards DA, Scott EM, Southon JR, Turney CSM, Wacker L, Adolphi F, Büntgen U, Capano M, Fahrni S, Fogtmann-Schulz A, Friedrich R, Köhler P, Kudsk S, Miyake F, Olsen J, Reinig F, Sakamoto M, Sookdeo A, Talamo S. 2020. The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0-55 cal kBP). *Radiocarbon* 62. doi: 10.1017/RDC.2020.41.

Grain size analysis

Classical plotting and characterization of grain size distributions.

The input data should consist of, first, one column of sieve sizes, in mm, from smallest to largest. The smallest size can be 0, which is interpreted as phi = 10, or 0.001 mm (1 µm). This can be used for unanalyzed sediment in the bottom pan. The largest sieve size should be larger than the largest particle in the sample and will contain no sediment. This is used for defining an upper size limit.

The second column contains sample weights for each size fraction, in any unit, or percent. The last value is assumed to be 0 and does not need to be entered. Additional columns can contain sample weights for additional samples. The example below defines two samples, with the sieve sizes in the first column:

0	4.0	5.2
0.020	7.1	5.3
0.063	9.3	7.1
0.125	9.4	8.0
0.25	11.3	9.1
0.5	8.1	4.3
1	4.2	2.0
2	2.1	1.5
4		

Past follows recommendations and equations of Blott and Pye (2001). Grain sizes are reported in the logarithmic phi scale (Krumbein scale), where $\varphi = 0$ corresponds to grain diameter 1 mm, more positive values are finer, and more negative coarser:

$$\varphi = -\log_2 \frac{D}{1\ mm}$$

where D is in mm. The denominator makes the φ scale dimensionless.

Gravel is defined as $\varphi < -1$ ($D > 2$ mm); sand is $-1 < \varphi < 4$ (0.0625 mm > $D > 2$ mm); mud is $\varphi > 4$ ($D <$ 0.0625 mm).

The sediment name (textural group) is defined on the proportions of gravel, sand and mud, following Folk (1954) and with linear interpolation when sieve sizes do not correspond to the definition boundaries. The refined classification that also includes proportions of silt and clay within the mud class is not included in Past.

Mean grain size is computed both with the method-of-moments (MOM) and the Folk & Ward (1957) "graphical" method. MOM is potentially more accurate when the sediment analysis is complete and finely binned, but the F&W method is more robust and therefore recommended for less complete analyses. Results are given both in φ units and in mm.

Mean grain size in φ units, method of moments:

$$\bar{x}_\varphi = \frac{\sum f m_\varphi}{\sum f}$$

where f are the given sieve sample weights and m_φ are the midpoints of the sieve size classes.

Mean grain size in φ units, Folk & Ward:

$$M_Z = \frac{\varphi_{16} + \varphi_{50} + \varphi_{84}}{3}$$

where the φ are the 16, 50 and 84 percentile points of the cumulative grain size distribution.

Sorting is defined using standard deviations, as follows.

Sorting using φ units, method of moments:

$$\sigma_\varphi = \frac{\sum f\left(m_\varphi - \bar{x}_\varphi\right)^2}{\sum f}$$

Sorting using φ units, Folk & Ward:

$$\sigma_I = \frac{\varphi_{84} - \varphi_{16}}{4} + \frac{\varphi_{95} - \varphi_5}{6.6}$$

Standardized names of degree of sorting follow Folk & Ward (1957), from "very well sorted" (σ < 0.35) to "extremely poorly sorted" (σ > 4.00)

Skewness is a somewhat less commonly used parameter. Because skewness is very sensitive to incomplete sieve analysis, only the more robust Folk & Ward method is calculated:

$$Sk_I = \frac{\varphi_{16} + \varphi_{84} - 2\varphi_{50}}{2(\varphi_{84} - \varphi_{16})} + \frac{\varphi_5 + \varphi_{95} - 2\varphi_{50}}{2(\varphi_{95} - \varphi_5)}$$

Standardized names of skewness follow Folk & Ward (1957), from "very coarse skewed" (Sk_I < -0.3) to "very fine skewed" (Sk_I > 0.3).

References

Blott, S.J., Pye, K. 2001. GRADISTAT: A grain size distribution and statistics package for the analysis of unconsolidated sediments. *Earth Surface Processes and Landforms* 26:1237–1248.

Folk, R.L. 1954. The distinction between grain size and mineral composition in sedimentary-rock nomenclature. *Journal of Geology*, 62:344–359.

Folk, R.L., Ward, W.C. 1957. Brazos River bar: a study in the significance of grain size parameters. *Journal of Sedimentary Petrology* 27:3–26.

Scripting

Past includes a fairly rich scripting language, which allows you to program your own modules taking advantage of the Past user interface, graphics package and mathematical and statistical routines. The scripter is already fully functional, but it will be substantially extended in future versions.

Language structure

Scripts are written in Pascal-style syntax. The fundamental language elements are:

begin .. end constructor
procedure and function declarations
if .. then .. else constructor
for .. to .. do .. step constructor
while .. do constructor
repeat .. until constructor
try .. except and try .. finally blocks
case statements
array constructors (x:=[1, 2, 3];)
^ , * , / , and , + , - , or , <> , >=, <= , = , > , < , div , mod , xor , shl , shr operators
access to object properties and methods (ObjectName.SubObject.Property)

Script structure

A script is made of two major blocks: a) procedure and function declarations and b) main block. Both are optional, but at least one should be present. There is no need for the main block to be inside begin..end. It could be a single statement. Some examples:

SCRIPT 1:

```
procedure DoSomething;
begin
  CallSomethingElse;
end;

begin
  DoSomething;
end;
```

SCRIPT 2:

```
begin
  CallSomethingElse;
end;
```

SCRIPT 3:

```
function MyFunction;
begin
  result:='Ok!';
end;
```

SCRIPT 4:

```
CallSomethingElse;
```

Statements should be terminated by the ";" character. Begin..end blocks are used to group statements.

Identifiers

Identifier names in script (variable names, function and procedure names, etc.) should begin with a character (a..z or A..Z), or '_', and can be followed by alphanumeric chars or the '_' char. They cannot contain any other characters or spaces.

Valid identifiers:

VarName
_Some
V1A2
_____Some____

Invalid identifiers:

2Var
My Name
Some-more
This,is,not,valid

Assign statements

Assign statements (assigning a value or expression result to a variable or object property) are built using ":=". Examples:

```
MyVar := 2;

Button.Caption := 'This ' + 'is ok.';
```

Character strings

Strings (sequence of characters) are declared using single quote (') characters. Double quotes (") are not used. You can also use #nn to declare a character inside a string. There is no need to use the '+' operator to add a character to a string. Some examples:

```
A := 'This is a text';
Str := 'Text '+'concat';
B := 'String with CR and LF char at the end'#13#10;
C := 'String with '#33#34' characters in the middle';
```

Comments

Comments are defined by // chars or (* *) or { } blocks. With the // char, the comment will finish at the end of line.

```
//This is a comment before ShowMessage
ShowMessage('Ok');
(* This is another comment *)
ShowMessage('More ok!');

{ And this is a comment
with two lines }

ShowMessage('End of okays');
```

Variables

There is no need to declare variable types. Thus, you declare a variable just using the var directive and its name. Also, it is optional to declare variables at all. Variables and their types are implicitly declared at first usage. Examples:

SCRIPT 1:

```
procedure Msg;
var S;
begin
  S:='Hello world!';
  ShowMessage(S);
end;
```

SCRIPT 2:

```
var A;
begin
```

```
  A:=0;
  A:=A+1;
end;
```

SCRIPT 3:

```
var S: string;
begin
  S:='Hello World!';
  ShowMessage(S);
end;
```

Var declarations are not strictly necessary in any of scripts above.

Indexing

Strings, arrays and array properties can be indexed using "[" and "]" chars. For example, if Str is a string variable, the expression Str[3] returns the third character in the string denoted by Str, while Str[I + 1] returns the character immediately after the one indexed by I. More examples:

```
MyChar:=MyStr[2];

MyStr[1]:='A';

MyArray[1,2]:=1530;

Lines.Strings[2]:='Some text';
```

Arrays

To construct an array, use "[" and "]" chars. You can construct multi-index arrays, nesting array constructors. You can then access arrays using indexes. A variable is an array if it was assigned using an array constructor or if it was created using the array or vector procedures. Some examples:

```
NewArray := [ 2,4,6,8 ];
Num:=NewArray[1]; //Num receives "4"
MultiArray := [['green','red','blue'] , ['apple','orange','lemon']];
Str:=MultiArray[0,2]; //Str receives 'blue'
MultiArray[1,1]:='new orange';
V:=vector(100);
A:=array(100,100);
```

Arrays defined using the array constructors can contain elements of any type, but arrays defined by the vector and array procedures are of type Double.

Arrays constructed using the array constructors are indexed from 0. Also, arrays defined using the vector and array procedures are indexed from 0, but the 0 element is often not used and these arrays contain n+1 elements, indexed from 0 to n.

If statements

There are two forms of if statement: if...then and if...then...else. If the if expression is true, the statement (or block) is executed. If there is an else part and the expression is false, the statement (or block) after else is executed. Examples:

```
if J <> 0 then Result := I/J;
if J = 0 then Exit else Result := I/J;
if J <> 0 then
begin
  Result := I/J;
  Count := Count + 1;
end else
  Done := True;
```

while statements

A while statement is used to repeat a statement or a block, while a control condition (expression) is evaluated as true. The control condition is evaluated before the statement. Hence, if the control condition is false at first iteration, the statement sequence is never executed. The while statement executes its constituent statement (or block) repeatedly, testing the expression before each iteration. As long as expression returns True, execution continues. Examples:

```
while Data[I] <> X do I := I + 1;

while I > 0 do
begin
  if Odd(I) then Z := Z * X;
  I := I div 2;
  X := Sqr(X);
end;

while not Eof(InputFile) do
begin
  Readln(InputFile, Line);
  Process(Line);
end;
```

repeat statements

The syntax of a repeat statement is

```
repeat statement_1; ...; statement_n; until expression
```

where *expression* returns a Boolean value. The repeat statement executes its sequence of constituent statements continually, testing the expression after each iteration. When expression returns True, the repeat statement terminates. The sequence is always executed at least once because *expression* is not evaluated until after the first iteration. Examples:

```
repeat
  K := I mod J;
  I := J;
  J := K;
until J = 0;

repeat
  Write('Enter a value (0..9): ');
  Readln(I);
until (I >= 0) and (I <= 9);
```

for statements

For statements have the following syntax:

```
for counter := initialValue to finalValue do statement
```

The For statement sets counter to initialValue, repeats execution of the statement (or block) and increments the value of counter until counter reaches finalValue. Examples:

SCRIPT 1:

```
for c:=1 to 10 do
  a:=a+c;
```

SCRIPT 2:

```
for i:=a to b do
begin
  j:=i^2;
  sum:=sum+j;
end;
```

case statements

Case statements have the following syntax:

```
case selectorExpression of
  caseexpr1: statement1;
  ...
  caseexprn: statementn;
else
  elsestatement;
end
```

If selectorExpression matches the result of one of caseexprn expressions, the respective statement (or block) will be executed. Otherwise, elsestatement will be executed.The Else part of the case statement is optional. A Case statement doesn't need to use only ordinal values. You can use expressions of any type in both the selector expression and the case expression. Example:

```
case uppercase(Fruit) of
  'lime': ShowMessage('green');
  'orange': ShowMessage('orange');
  'apple': ShowMessage('red');
else
  ShowMessage('black');
end;
```

function and procedure declaration

Declaration of functions and procedures are similar to Pascal, with the difference that you don't specify variable types. To return function values, use the implicitly declared *result* variable.

Parameters by reference can also be used, with the restriction mentioned: no need to specify variable types. Some examples:

```
procedure HelloWord;
begin
  ShowMessage('Hello world!');
end;

procedure UpcaseMessage(Msg);
begin
  ShowMessage(Uppercase(Msg));
end;

function TodayAsString;
begin
  result:=DateToStr(Date);
end;

function Max(A,B);
begin
  if A>B then
    result:=A
  else
    result:=B;
end;
```

```
procedure SwapValues(var A, B);
Var Temp;
begin
  Temp:=A;
  A:=B;
  B:=Temp;
end;
```

The output window

When you run a script, an output window will be opened automatically. It contains three tabs: Text, graphic and table.

The text tab

The text tab contains a window to which the script can write output. The text can be copy-pasted to other programs by the user. The following procedures are available:

cleartext	Clears the text window
textout(s)	Writes a line to the text window. Handles numerical, string, vector and array types.

The graphic tab

A resizable graphic canvas with the usual Past functionality such as a graph preferences window with export to vector (SVG or PDF) or bitmap formats. The window will automatically scale to its contents, so you do not need to consider the scale of coordinates. For efficiency, no graphics will appear until the redraw procedure is called.

Colors must be given as one of the following constants: black, red, blue, green, purple, yellow, gray, brown.

redraw	Redraw the graphic window with automatic axis ranges
setaxes(x1, x2, y1, y2)	Redraw the graphic with the given axis ranges
cleargraphic	Clears the graphic window
savegraphic(filename)	Depending on the file extension, will save the graphic in one of the following formats: svg, pdf, jpg, tif, gif, png, bmp
drawpoints(x, y, color)	Draws one point (if x and y are single numbers) or several (if x and y are vectors). Color is a single integer (see above).
drawsymbols(x, y, color, symbol)	Draws one symbol (if x and y are single numbers) or several (if x and y are vectors). Color and symbol are single integers, see 'spreadsheet_symbols' for symbol coding.

drawline(x1, y1, x2, y2, color)	A line from (x1, y1) to (x2, y2)
drawpolyline(x, y, color)	A polyline with lists of x and y coordinates in vectors x and y
drawrectangle(x1, y1, x2, y2, color)	A rectangle with the given corners
drawellipse(x, y, major, minor, angle, color)	An ellipse with center (x, y), given major and minor axes and with the major axis at the given angle (radians) to the x axis.
drawtext(x, y, string)	Draws text at position (x, y) – may reposition to reduce overlap
drawmatrix(A, interpolate)	Draws the matrix A. Set interpolate to false or true to select drawing mode. Does an automatic redraw.
drawhistogram(V, nbins, color, kde)	A histogram of vector V, with the given number of bins. If kde=true, a kernel density estimate is also drawn.
drawbars(V, color)	A bar chart of vector V.
drawboxplot(V, x, outliers)	A box plot of vector V at the given x position. Outliers is true or false.
drawconvexhull(Vx, Vy, color)	The convex hull of the points in vectors Vx and Vy.
drawrose(V, n, equalarea, kde)	A rose plot of angles in V (degrees), with n bins. Equalarea and kde (kernel density estimate) are true or false.

The table tab

A table (spreadsheet) window with copy-paste function.

tablesize(rows, columns: integer)	Set the number of rows and columns in the table
tableout(row, col, value)	Write value to a particular cell in the table, indexing starting at 0.

Accessing the main Past spreadsheet and menus

clickmenu(name: string)	Executes an item in the Past menu. Specify the name as given in the menu, e.g. clickmenu('save as'). Parts of the name in brackets are not included: Use 'species packing', not 'species packing (Gaussian)'.
spreadsheet_array	Returns an array containing the selected area in the Past spreadsheet. Group columns are not included
spreadsheet_column(n: integer)	Returns a vector with the numbers in column n in the Past spreadsheet.
spreadsheet_groups(n: integer)	Returns a vector with group numbers corresponding to the rows in spreadsheet. For n=1, the first group column is returned, for n=2 the second group column (if any), etc.
spreadsheet_rowlabels	Returns a string vector with the row labels in the selected area.
spreadsheet_columnlabels	Returns a string vector with the column labels in the selected area.
spreadsheet_symbols	Returns a vector with numbers (0-15) identifying the symbols corresponding to the rows in spreadsheetarray. 0=dot, 1=+, 2=square, 3=X, 4=triangle, 5=O, 6=diamond, 7=-, 8=l, 9=fillsquare, 10=*, 11=oval, 12=filltriangle 13=invtriangle, 14=fillinvtriangle, 15=filldiamond
spreadsheet_set(row, col, s)	Sets the contents of the cell at (row, col) in the Past spreadsheet to s (number or string). Indexing starts from 0 (label cells).

Array and vector operations

array(m, n: integer) Allocates and returns a Float (Double) array with m rows and n columns, indexing starting at 1.
vector(n: integer) Allocates and returns a vector (one-dimensional array) with n elements, indexing starting at 1
column(A, n) Returns column n in array A, as a vector
row(A, m) Return row n in array A, as a vector
ncols(A) Returns the number of columns in array A
nrows(A) Returns the number of rows in array A
inv(A) Inverse of square matrix A
arrmult(A, B) Returns array multiplication A*B
mean(V) Mean of vector V
variance(V) Variance of vector V
median(V) Median of vector V
skew(V) Skew of vector V
kurtosis(V) Kurtosis of vector V
svd(A) Singular Value Decomposition of A, returning V augmented by an extra column containing D.
cov(A) Returns variance-covariance matrix of A.
eig(A) Returns the eigenvectors of A, augmented by an extra column with the eigenvalues.
pearsonr(X, Y) Returns the Pearson correlation (r) between vectors X and Y
spearmanrs(X, Y) Returns the Spearman rank-order correlation (rs) between vectors X and Y
linfit(X, Y) Ordinary least-squares regression of vectors X, Y. Returns a 4-vector with slope, intercept, standard error of slope, standard error of intercept.

Some common mathematical functions are available for arrays and vectors, for efficiency. The function is applied to each element of the array or vector, returning an array or vector.

arrAbs(A) Absolute value
arrCos(A) Cosine (radians)
arrExp(A) e^x
arrLn(A) Natural logarithm (base *e*)
arrSin(A) Sine (radians)
arrSqrt(A) Square root
arrTan(A) Tangent (radians)

Scalar math functions

abs(x) Absolute value
arctan(x) Inverse tangent (radians)
arctan2(y, x) Inverse tangent of y/x extended to correct quadrant.
cos(x) Cosine (radians)
cumulnorm(x, s) Cumulative normal distribution function, mean=0, stdev=s
exp(x) e^x
frac(x) Fractional part of *x*
fresnel(x) Returns a 2-vector with the S and C Fresnel integrals to x.
ln(x) Natural logarithm (base *e*)
normal(m, s) Normally distributed random number, mean=m, stdev=s.
gamma(k) Gamma distributed random number, shape=k, scale=1.
invnorm(x) Inverse of the cumulative normal distribution, mean=0, stdev=1, $0<x<1$

invchi2(x, df)	Inverse of the cumulative chi-squared distribution with df degrees of freedom
odd(x)	True if *x* is odd
random	Random number, uniform distribution, 0 <= *x* < 1
round(x)	Rounds to the nearest integer
sin(x)	Sine (radians)
sqr(x)	Square (*x***x*)
sqrt(x)	Square root
studentp(t, df)	Two-tailed *p* value from Student's *t* distribution with *df* degrees of freedom
tan(x)	Tangent (radians)
trunc(x)	Rounds down

File I/O

`function` **`Append`**`(var F: File): Integer;`

Prepares an existing file for adding text to its end. F is a text file variable and must be associated with an external file, using AssignFile. If the external file does not exist, an error occurs. If F is already open, it is closed, then reopened. The current file position is set at the end of the file.

`function` **`AssignFile`**`(var F: File; FileName: String): Integer;`

Associates the name of an external file with a file variable. After calling AssignFile, F is associated with the external file until F is closed. All further operations on the file variable F operate on the external file named by FileName.

`procedure` **`ChDir`**`(S: string);`

Changes the current directory to the path specified by S.

`procedure` **`CloseFile`**`(var F: File);`

Terminates the association between a file variable and an external disk file. F is a file variable opened using Reset, Rewrite, or Append. The external file associated with F is completely updated and then closed, freeing the file handle for reuse.

`function` **`Eof`**`(var F: File): Boolean;`

Tests whether the file position is at the end of a file.

`function` **`FilePos`**`(var F: File): Integer;`

Use on an open file to determine the current position. If the current position is at the beginning, FilePos returns 0. Otherwise, FilePos returns the byte offset from the beginning of the file.

`function` **`FileSize`**`(var F: File): Integer;`

Returns the number of records in a file.

```
function ReadLn(var F: File): String;
```

Reads a line of text and then skips to the next line of the file.

```
procedure Reset(var F: File);
```

Opens the existing external file with the name assigned to F. An error results if no existing external file of the given name exists or if the file cannot be opened. If F is already open, it is first closed and then reopened. The current file position is set to the beginning of the file.

```
procedure Rewrite(var F: File);
```

Creates a new external file with the name assigned to F. F is associated with an external file using AssignFile. If a file with the same name already exists, it is deleted and a new empty file is created in its place. If F is already open, it is first closed and then re-created. The current file position is set to the beginning of the empty file.

```
procedure WriteLn(var F: File; S: string);
```

Writes to a text file and adds an end-of-line marker.

String operations

```
function Chr(X: Byte): Char;
```

Returns the character for a specified ASCII value.

```
function CompareStr(S1, S2: string): Integer;
```

Compares S1 to S2, with case sensitivity. The return value is less than 0 if S1 is less than S2, 0 if S1 equals S2, or greater than 0 if S1 is greater than S2.

```
function CompareText(S1, S2: string): Integer;
```

Compares S1 to S2, without case sensitivity. The return value is less than 0 if S1 is less than S2, 0 if S1 equals S2, or greater than 0 if S1 is greater than S2.

```
function Copy(S: string; Index: Integer; Count: Integer): string;
```

Returns a substring of a string S. Index and Count are integer-type expressions. Copy returns a substring or subarray containing Count characters or elements starting at S[Index].

```
function FloatToStr(Value: Double): string;
```

Converts the floating-point value given by Value to its string representation. The conversion uses general number format with 15 significant digits.

```
procedure Insert(Source: string; var Dest: string; Index: Integer);
```

Inserts a substring into a string, from a specified position. If Index is less than 1, it is set to 1. If it is past the end of Dest, it is set to the length of Dest, turning the operation into an append.

```
function IntToStr(Value: Integer): string;
```

Converts an integer to a string that contains its decimal representation.

```
function Length(S: string): Integer;
```

Returns the number of characters in a string.

```
function LowerCase(S: string): string;
```

Returns a string with the same text as in S, but with all letters converted to lowercase.

```
function Pos(SubStr, Str: string): Integer;
```

Returns an index of the first occurrence of Substr in Str. Returns zero if Substr is not found.

```
function StrToFloat(S: string): Double;
```

Converts a string to a floating-point value (leading and trailing blanks are ignored).

```
function StrToInt(S: string): Integer;
```

Converts a string that represents an integer into a number.

```
function StrToIntDef(S: string; Default: Integer): Integer;
```

Converts the string S, which represents an integer, into a number. If S does not represent a valid number, StrToIntDef returns Default.

```
function Trim(S: string): string;
```

Trims leading and trailing spaces and control characters from a string.

```
function TrimRight(S: string): string;
```

Trims trailing spaces and control characters from a string.

```
function UpperCase(S: string): string;
```

Returns a copy of a string in uppercase.

Other functions

```
procedure ShowMessage(S: string);
```

Shows a message box and waits for user to click Ok.

```
function InputQuery(Caption, Prompt: string; var Value: string):
Boolean;
```

Displays an input dialog box that lets the user enter a value. Caption is the caption of the dialog box. Prompt is the text that prompts the user to enter input. Value is the value that appears in the edit box when the dialog first appears and returns the value that the user enters. InputQuery returns true if the user chooses OK, false if the user chooses Cancel or presses Esc.

```
procedure sleep(ms: integer);
```

Suspends the execution of the script for the given number of milliseconds.

Calling dll functions (Windows only)

Past allows importing and calling external DLL functions, by declaration of script routines, indicating library name and, optionally, the calling convention, in addition to the function signature. External libraries are loaded by Past on demand, before function calls, if not already loaded (dynamically or statically). To load and unload libraries explicitly, functions LoadLibrary and FreeLibrary from unit Windows can be used.

Syntax

```
function functionName(arguments): resultType; [callingConvention];
external 'libName.dll' [name ExternalFunctionName];
```

For example, the following declaration:

```
function MyFunction(arg: integer): integer;
external 'CustomLib.dll';
```

imports a function called `MyFunction` from `CustomLib.dll`. Default calling convention, if not specified, is register. Past also allows declaring a different calling convention (stdcall, register, pascal, cdecl or safecall) and to use a different name for the DLL function, like the following declaration:

```
function MessageBox(hwnd: pointer; text, caption: string; msgtype:
integer): integer; stdcall;

external 'User32.dll' name 'MessageBoxA';
```

that imports 'MessageBoxA' function from User32.dll (Windows API library), named 'MessageBox' to be used in the script.

The declaration above can be used for functions and procedures (routines without result value).

Supported types

Past supports the following basic data types on arguments and result of external functions:

Integer
Boolean
Char
Extended
String
Pointer
PChar
Object
Class
WideChar
PWideChar
AnsiString
Currency
Variant

Interface
WideString
Longint
Cardinal
Longword
Single
Byte
Shortint
Word
Smallint
Double
Real
DateTime
TObject descendants (class must be registered in scripter with DefineClass)

Other types (records, arrays, etc.) are not supported yet. Arguments of above types can be passed by reference, by adding var in the param declaration of the function.

Libraries and classes

Forms and components

For user input more complex than provided by the InputQuery function (see above), you can build your own forms (windows) with the following components: Labels, edits (where the user can enter text and numbers), buttons and check boxes.

Form

A form is defined with a variable of type TForm, and created using "TForm.CreateNew(nil, 0)". Some useful properties of the TForm class are:

Caption	The text at the top of the form
Width	Width in pixels
Height	Height in pixels

The method showModal displays the form, returning mrOk or mrCancel depending on buttons clicked in the form (see Button below).

Label

A label, of type TLabel, created with Tlabel.create(parentform) shows a simple text. Properties of the TLabel class include

Parent	The parent form, must be specified here in addition to in TLabel.create
Text	The text of the caption
Position.x	x position, in pixels, relative to the parent form
Position.y	y position, in pixels
Width	Width in pixels

Edit

A box, of type TEdit, created with TEdit.create(parentform), where the user can enter text or numbers. Properties of the TEdit class include

Parent	The parent form, must be specified here in addition to in TEdit.create
Text	The text of the caption
Position.x	x position, in pixels, relative to the parent form
Position.y	y position, in pixels
Width	Width in pixels

Button

A button, of type TButton, created with TButton.create(parentform). Properties of the TButton class include

Parent	The parent form, must be specified here in addition to in TButton.create
Text	The text of the caption
Position.x	x position, in pixels, relative to the parent form
Position.y	y position, in pixels
Width	Width in pixels

Height	Height in pixels
modalResult	Can be set to mrOk or mrCancel. When the button is clicked, this value is returned by the form's showModal method.

Check box

A check box, of type TCheckBox, created with TCheckBox.create(parentform). Properties of the TCheckBox class include

Parent	The parent form, must be specified here in addition to in TCheckBox.create
Text	The text of the caption
Position.x	x position, in pixels, relative to the parent form
Position.y	y position, in pixels
Width	Width in pixels
isChecked	Boolean (True or False). Read only, don't set to True from code, it won't work

Example

The following script shows a form with a label, an edit control and an OK button. When the user clicks the button, the text of the edit control is written to the text window.

```
var
  fm: TForm;
  lb: TLabel;
  ed: TEdit;
  bt: TButton;

begin
  fm : = TForm.CreateNew(nil, 0);
  fm.Caption := 'A brand new form!';
  fm.Width := 300;
  fm.Height := 150;

  lb := TLabel.Create(fm);
  lb.Parent := fm;
  lb.Position.X := 10;
  lb.Position.Y := 10;
  lb.Text := 'Your name:';

  ed := TEdit.Create(fm);
  ed.Parent := fm;
  ed.Position.X := lb.Position.X;
  ed.Position.Y := lb.Position.Y + lb.Height + 10;

  bt := TButton.Create(fm);
  bt.Parent := fm;
  bt.Position.X := ed.Position.X;
  bt.Position.Y := ed.Position.Y + ed.Height + 10;
  bt.Text := 'Ok';
  bt.Default := True;
  bt.ModalResult := mrOk;

  fm.ActiveControl := ed; // Sets the focus to the edit control
```

```
  if fm.ShowModal = mrOk then
    textout('Hello '+ed.Text+'!');
  fm.Free;
end;
```