PAST

PAleontological STatistics

Version 2.16

Reference manual

Øyvind Hammer

Natural History Museum

University of Oslo

ohammer@nhm.uio.no

1999-**2012**

Contents

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 ten years, PAST has grown into a comprehensive statistics package that is used not only by paleontologists, but in many fields of life science, earth science, and even engineering and economics.

Further explanations of many of the techniques implemented together with case histories are found in Harper (1999). Also, "Paleontological data analysis" (Hammer & Harper 2005) can be viewed as a companion book to PAST.

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, are found at

http://folk.uio.no/ohammer/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. 2006. Paleontological Data Analysis. Blackwell.

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

The installation of PAST is easy: Just download the file 'Past.exe' and put it anywhere on your harddisk. Double-clicking the file will start the program. 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.

We suggest you make a folder called 'past' anywhere on your harddisk, and put all the files in this folder.

Please note: Problems have been reported for non-standard default font size in Windows - it may be necessary for the user to increase the sizes of windows in order to see all the text and buttons. If this happens, please set the font size to 'Small fonts' in the Screen control panel in Windows.

When you exit PAST, a file called 'pastsetup'will be automatically placed in your personal folder (for example 'My Documents' in Windows 95/98), containing the last used file directories.

The lack of “formal” Windows installation is intentional, and allows installation without administrator privileges.

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. This is only possible when the program is in the 'Edit mode'. To select edit mode, tick the box above the array. When edit mode is off, the array is locked and the data can not be changed. 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 ('?') or the value -1. Unless support for missing data is specifically stated in the documentation for a function, the function *will 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. Note that you can not 'drag out' multiple rows - this will instead move the first row (see below).
- Multiple columns are similarly marked by shift-clicking the additional column labels.
- 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 'dragging out' the area, but this only works when 'Edit mode' is off.

IMPORTANT: Unfortunately, you can not select several columns that are not neighbours. This means that if you want to run an analysis on say the first and the third columns only, you will first have to move the columns together (see next paragraph).

Moving a row or a column

A row or a column (including its label) can 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. Choose 'Rename columns' or 'Rename rows' in the Edit menu. You must select the whole array, or a smaller area as appropriate.

Another way is to select the 'Edit labels' option above the spreadsheet. The first row and column are now editable in the same way as the rest of the cells.

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.

Remember that local blocks of data (not all rows or columns) can only be marked when 'Edit mode' is off.

All modules giving graphic output have a 'Copy graphic' button. This will place the graphical image into the paste buffer for pasting into other programs, such as a drawing program for editing the image. Graphics are copied using the 'Enhanced Metafile Format' in Windows. This allows editing of individual image elements in other programs. When pasting into Coreldraw, you have to choose 'Paste special' in the Edit menu, and then choose 'Enhanced metafile'. Some programs may have idiosyncratic ways of interpreting EMF images - beware of funny things happening.

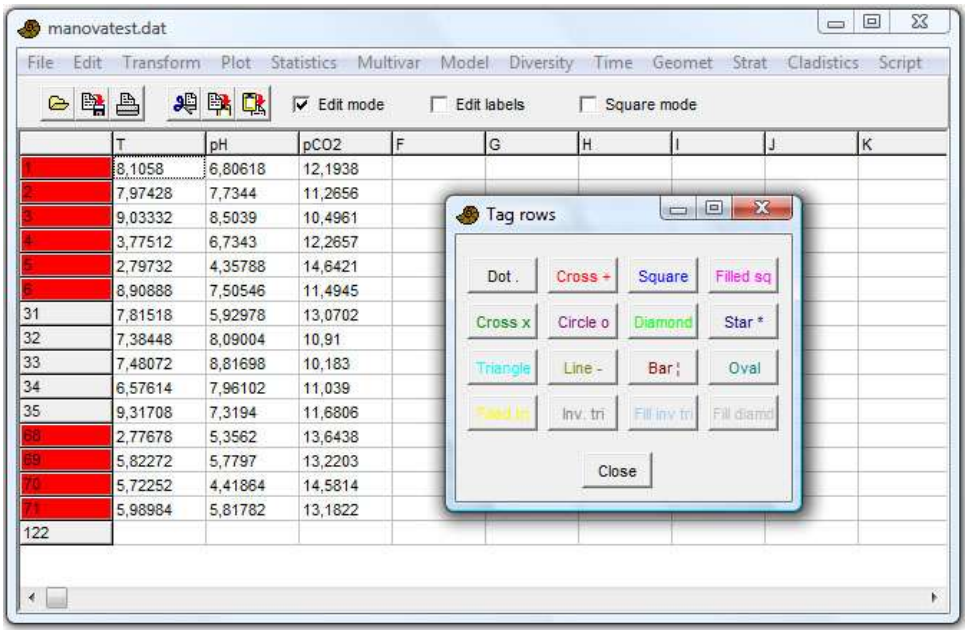
Remove

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.

Grouping (coloring) rows

Selected rows (data points) can be tagged with one of 16 attractive colors using the 'Row color/symbol' option in the Edit menu. Each group is also associated with a symbol (dot, cross, square, diamond, plus, circle, triangle, line, bar, filled square, star, oval, filled triangle, inverted triangle, filled inverted triangle, filled diamond). This is useful for showing different groups of data in plots, and is also required by a number of analysis methods.

Important: For methods that require groupings of rows using colors, rows belonging to one group must be consecutive. If more than 16 groups are required, colors can be re-used. In the example below, three groups have been correctly marked.



	T	pH	pCO2	F	G	H	I	J	K
1	8,1058	6,80618	12,1938						
2	7,97428	7,7344	11,2656						
3	9,03332	8,5039	10,4961						
4	3,77512	6,7343	12,2657						
5	2,79732	4,35788	14,6421						
6	8,90888	7,50546	11,4945						
31	7,81518	5,92978	13,0702						
32	7,38448	8,09004	10,91						
33	7,48072	8,81698	10,183						
34	6,57614	7,96102	11,039						
35	9,31708	7,3194	11,6806						
68	2,77678	5,3562	13,6438						
69	5,82272	5,7797	13,2203						
70	5,72252	4,41864	14,5814						
71	5,98984	5,81782	13,1822						
122									

The 'Numbers to colors' option in the Edit menu allows the numbers 1-16 in one selected column to set corresponding colors (symbols) for the rows.

Selecting datatypes for columns

Selected columns can be tagged with a datatype (continuous/unspecified, ordinal, nominal or binary) using the 'Column data types' option in the Edit menu. This is only required if you wish to use mixed similarity/distance measures.

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 ('?'), and rows or columns with only one non-zero cell (singletons).

Transpose

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

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 colored rows into columns

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

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:

```
.          columnlabel columnlabel columnlabel
rowlabel Data        Data        Data
rowlabel Data        Data        Data
rowlabel Data        Data        Data
```

Empty cells (like the top left cell) are coded with a full stop (.). Cells are separated by white space. If a cell contains space characters, it must be enclosed in double quotes, e.g. “Oxford clay”.

If any rows have been assigned a color other than black, the row labels in the file will start with an underscore, a number from 0 to 15 identifying the color (symbol), and another underscore.

If any columns have been assigned a datatype other than continuous/unspecified, the column labels in the file will similarly start with an underscore, a number from 0-3 identifying the datatype (0=continuous/unspecified, 1=ordinal, 2=nominal, 3=binary), and another underscore.

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

- Excel (only the first worksheet).
- 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.
- 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.

The 'Insert from file' function is useful for concatenating data sets. The loaded file will be inserted into your existing spreadsheet at the selected position (upper left).

Importing data from Excel

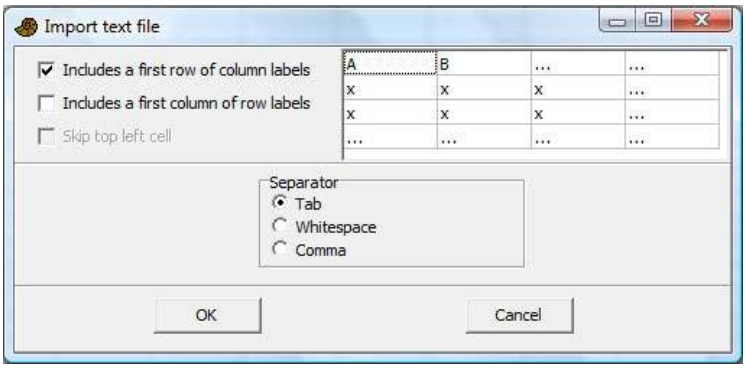
- Copy from Excel and paste into PAST. Note that if you want the first row and column to be copied into the label cells in PAST, you must switch on the "Edit labels" option. Or,
- Open the Excel file from PAST. The "Edit labels" option operates in the same way. Or,
- Make sure that the top left cell in Excel contains a single dot (.) and save as tab-separated text in Excel. The resulting text file can be opened directly 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.

Importing text files

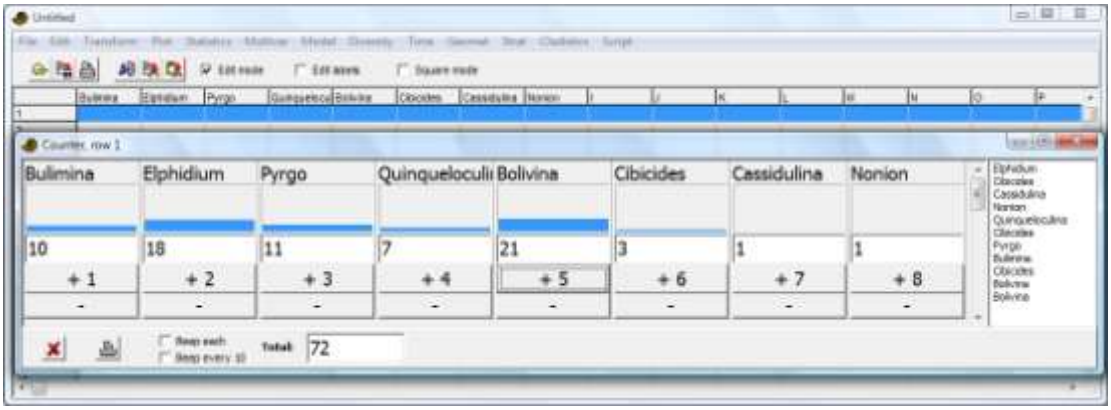
Text files with values separated by white space, tabs or commas can be opened using the 'Import text file' function in the File menu. The spreadsheet in the window illustrates the format of the input file as specified using the check boxes.



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.


Bulimina
Elphidium
Pyrgo
Quinqueloculi
Bolivina
Cibicides
Cassidulina
Nonion
10
18
11
7
21
3
1
1
+ 1
+ 2
+ 3
+ 4
+ 5
+ 6
+ 7
+ 8
72

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.

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.

Subtract mean

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

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:

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

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

Missing values 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.

Missing values supported.

Abundance to presence/absence

All positive (non-zero) numbers are replaced with ones.

Missing values supported.

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.

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.

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

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

Remove size from landmarks

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.

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

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

.

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.

Sort ascending and descending

Sorts the rows in the marked area, based on values in the selected data column.

The "Sort descending" function is useful, for example, to plot taxon abundances against their ranks (this can also be done with the Abundance Model module).

Sort on color

Sorts the rows in the marked area on color.

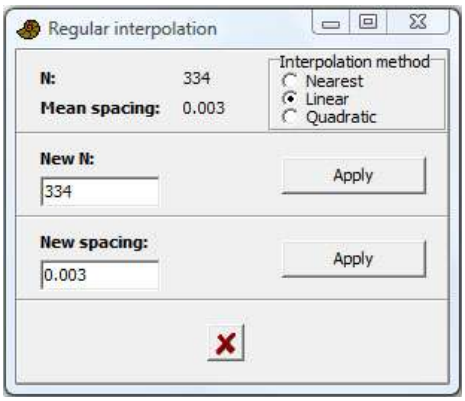
Column difference

Simply subtracts two selected columns, and places the result in the next column.

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. Three interpolation methods are available.



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 inlude any of the operators +, -, *, /, ^ (power), and mod (modulo). Also supported are brackets (), and the functions abs, atan, cos, sin, exp, ln, sqrt, sqr, round 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)
- 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)
- sum (total sum of the current column)

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

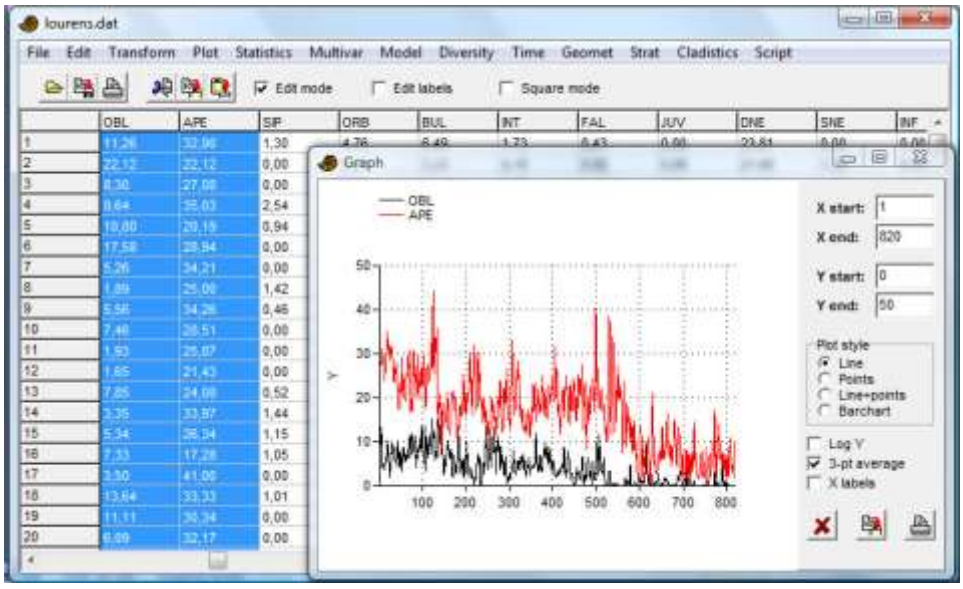
Plot menu

Graph

Plots one or more columns as separate graphs. The *x* coordinates are set automatically to 1,2,3,... There are four plot styles available: Graph (line), points, line with points, and bars. The 'X labels' options sets the *x* axis labels to the appropriate row names.

The “Log Y” option log-transforms the Y values only, to base 10, but with log 0 set to 0.

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 (if necessary, a constant is added to make the minimum log value equal to 0). The curve can also be smoothed using 3-point moving average.

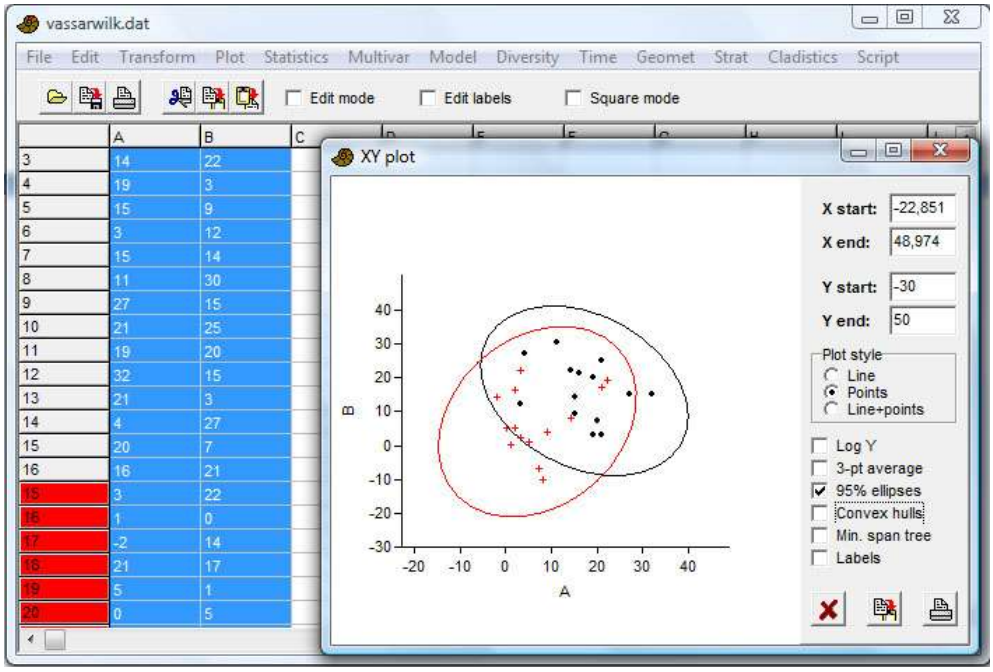
95% concentration ellipses can be plotted in most scatter plots in PAST, such as scores for PCA, CA, DCA, PCO and NMDS. The calculation of these ellipses assumes bivariate normal distribution.

Convex hulls can also be drawn in the scatter plots, in order to show the areas occupied by points of different 'colors'. 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.

Hold the mouse cursor over a point to see its row label.

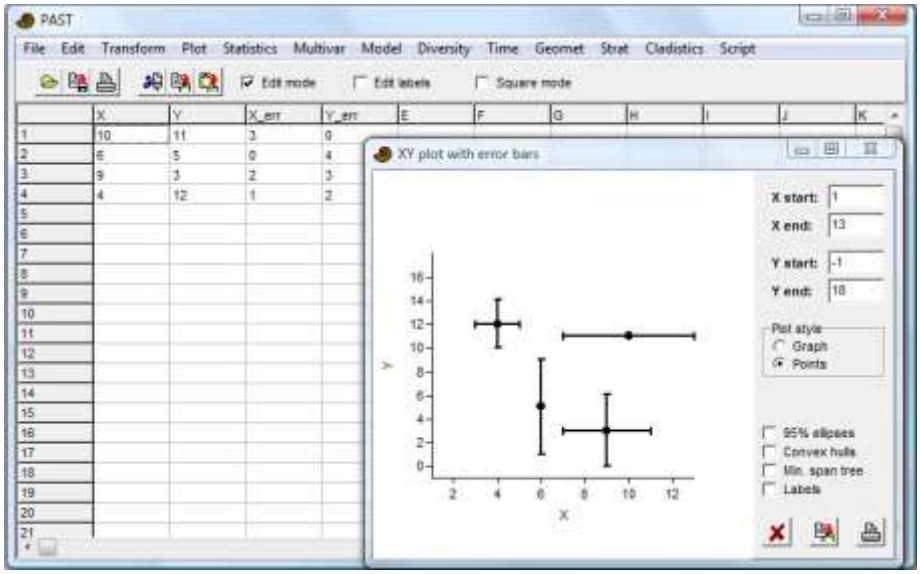
Points with missing values in X and/or Y are disregarded.



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. The number of bins is by default set to an "optimal" number (the zero-stage rule of Wand 1997):

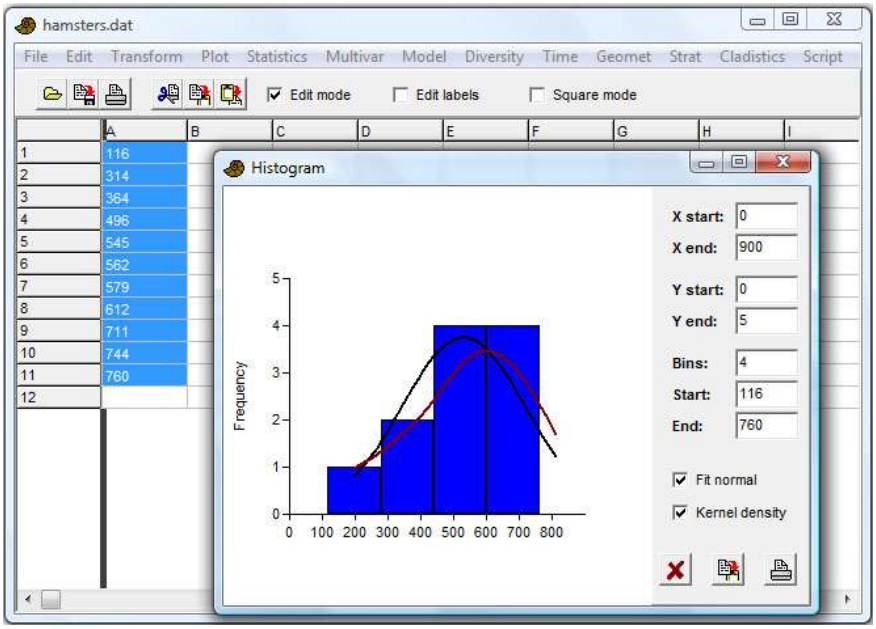
$$h = 3.49 \min(\mathrm{s}, IQ/1.349) n^{-1/3}$$

where *s* is the sample standard deviation and *IQ* the interquartile range.

The number of bins can be changed by the user. 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 or box plot for one or several columns (samples) of univariate data. Missing values are deleted.

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

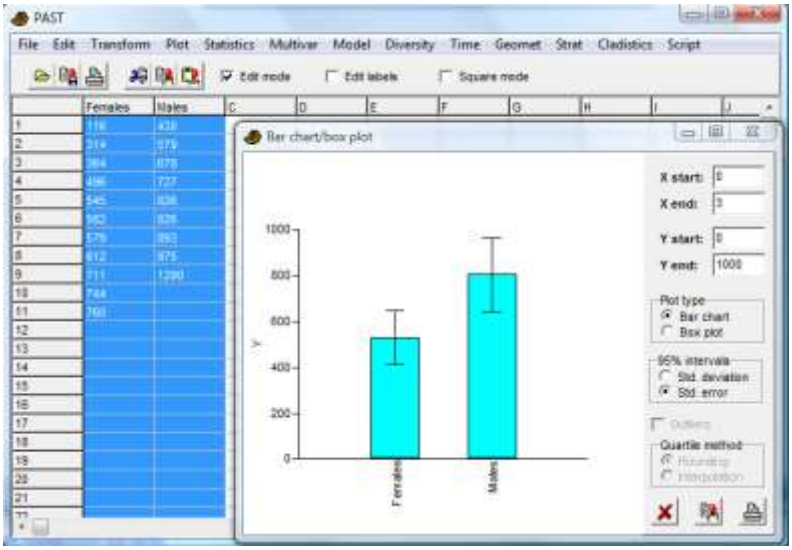
For each sample, the 25-75 percent quartiles are drawn using a box. The median is shown with a horizontal line inside the box. The minimal and maximal values are shown with short horizontal lines ("whiskers").

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.

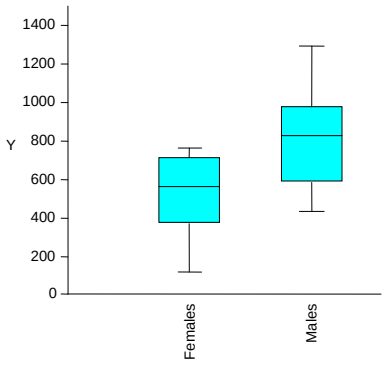
The quartile methods (rounding or interpolation) are described under “Percentiles” below.

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.



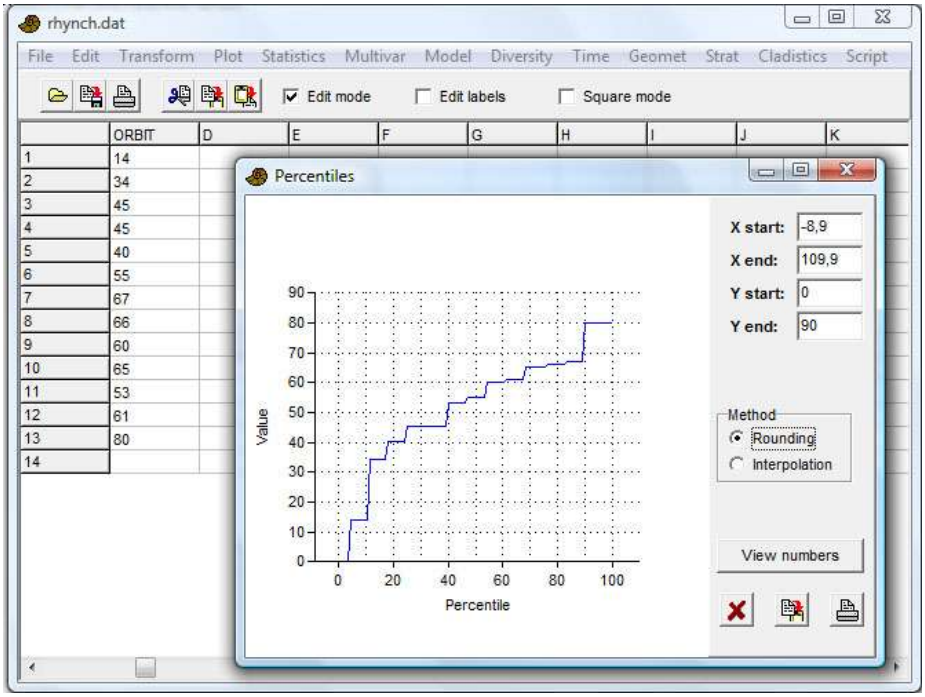
Bar chart



Box plot

Percentiles

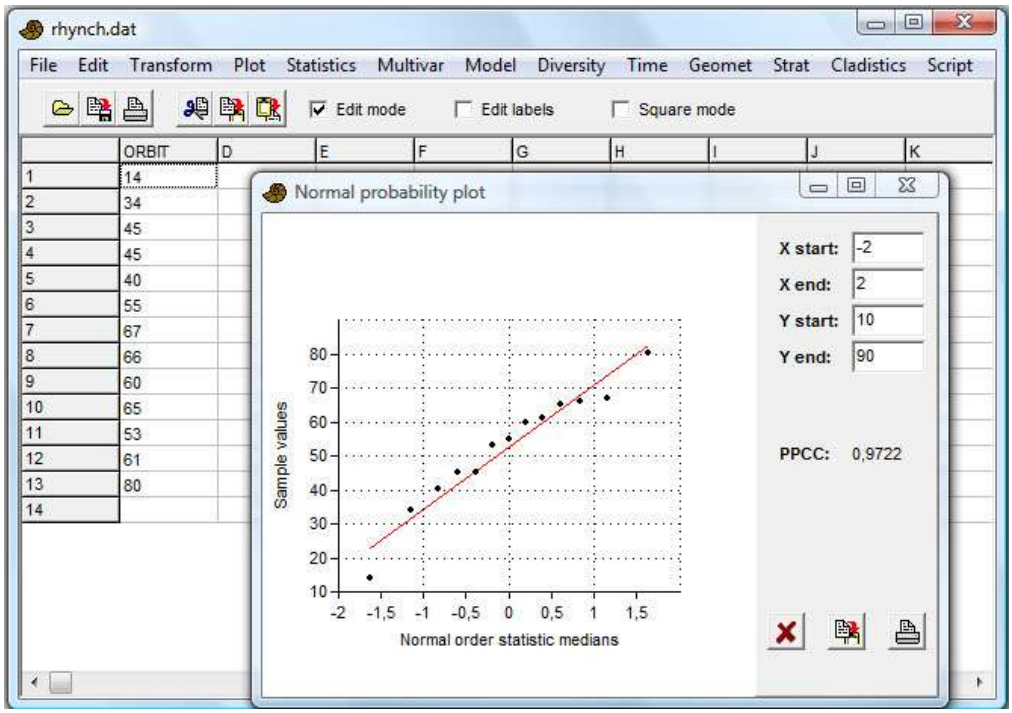
For each percentile *p*, plots the value *y* such that *p* percent of the points are smaller than *y*. Two popular methods are included. For a percentile *p*, the rank is computed according to *k=p(n+1)*/100, and the value that corresponds to that rank taken. In the rounding method, *k* is rounded to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



Missing values are deleted.

Normal probability plot

Plots a normal probability (normal QQ) plot for one column 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.



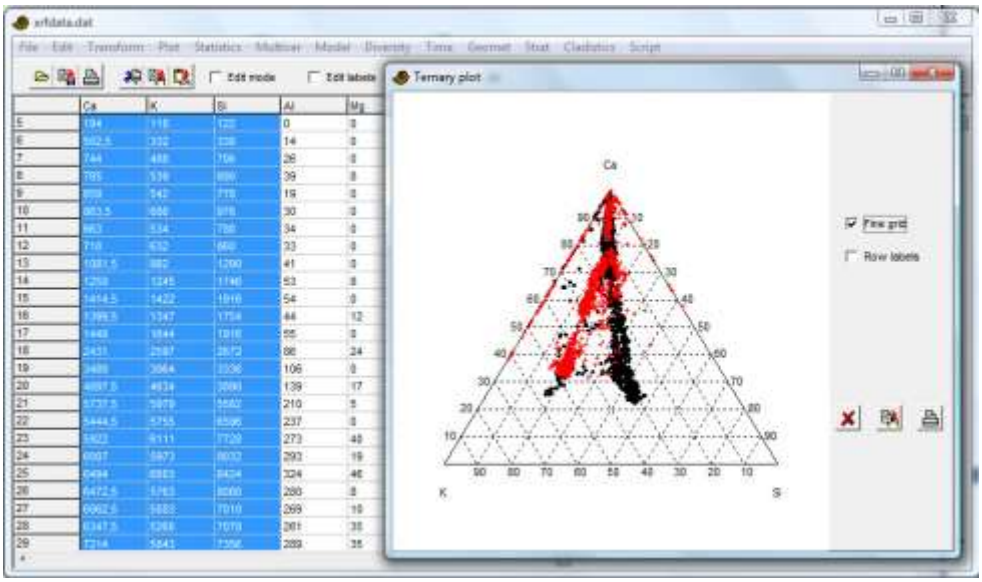
Missing values are deleted.

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

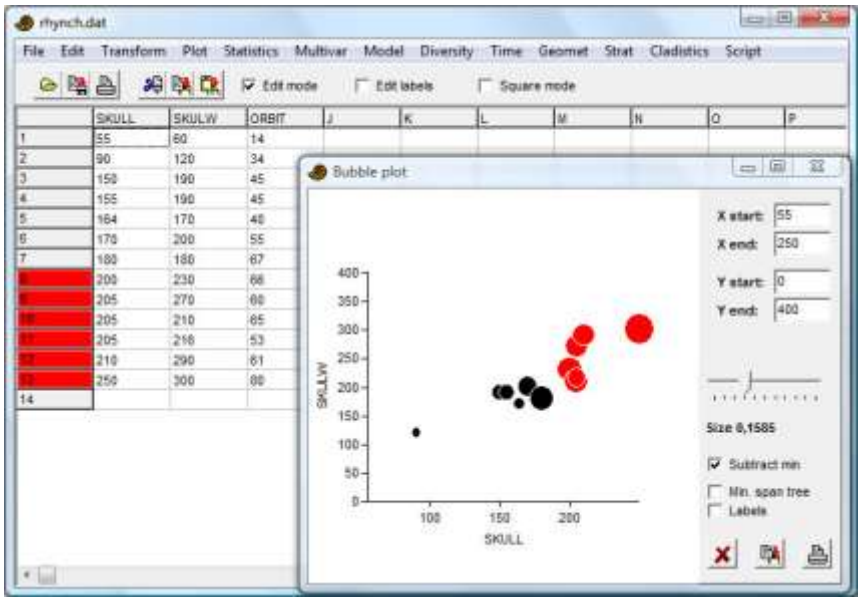
Ternary plot for three columns of data, normally containing proportions of compositions. If a fourth column is included, it will be shown using either a bubble representation or as a color/grayscale map.



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

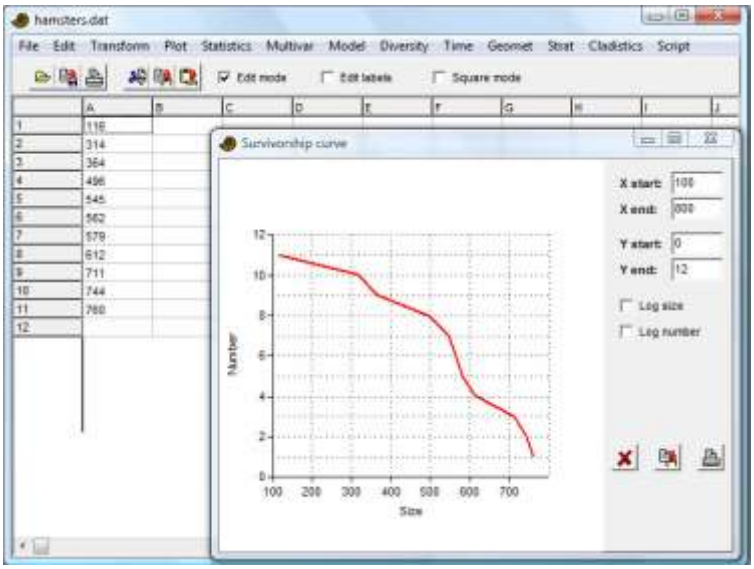


Rows with missing value(s) in any column are deleted.

Survivorship

Survivorship curves for one or more columns of data. The data may consist of age or size values. The plot shows the number of individuals that survived to different ages. Assuming exponential growth (highly questionable!), size may be log-transformed to age. This can be done either in the Transform menu, or directly in the Survivorship dialogue. See also the Survival analysis in the Statistics menu.

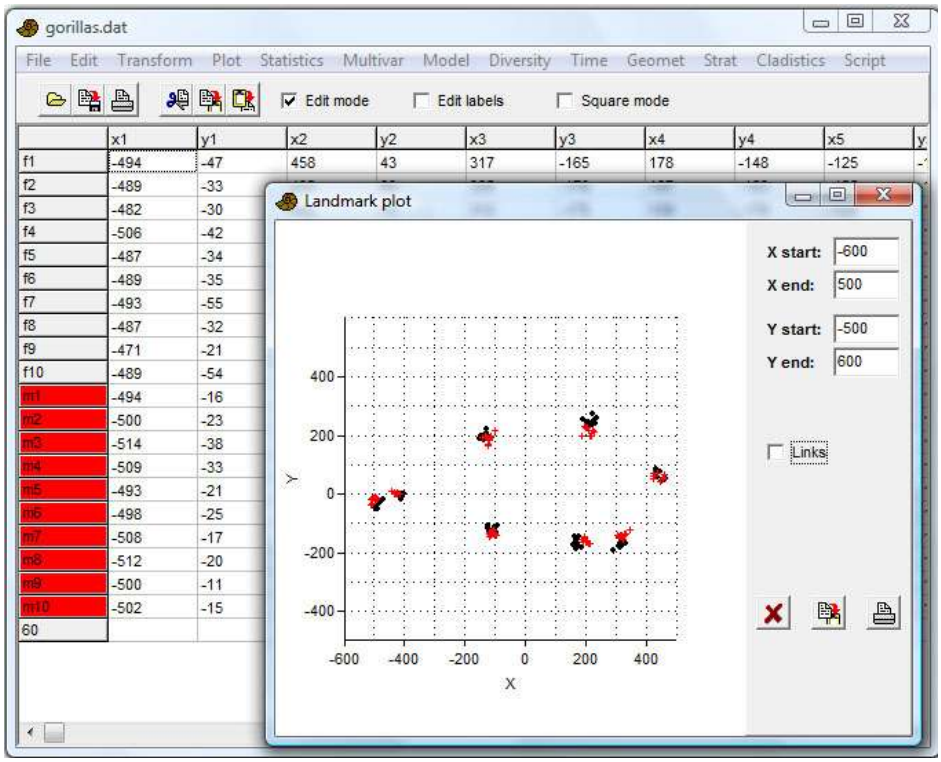
Missing values are deleted.



Landmarks

This function is very similar to the 'XY graph', the only difference being that all XY pairs on each row are plotted with the appropriate row color and symbol. It also forces unit aspect ratio, and is well suited for plotting landmark data. The “links” option plots lines between the landmarks, as specified by the “Landmark linking” option in the Geomet menu.

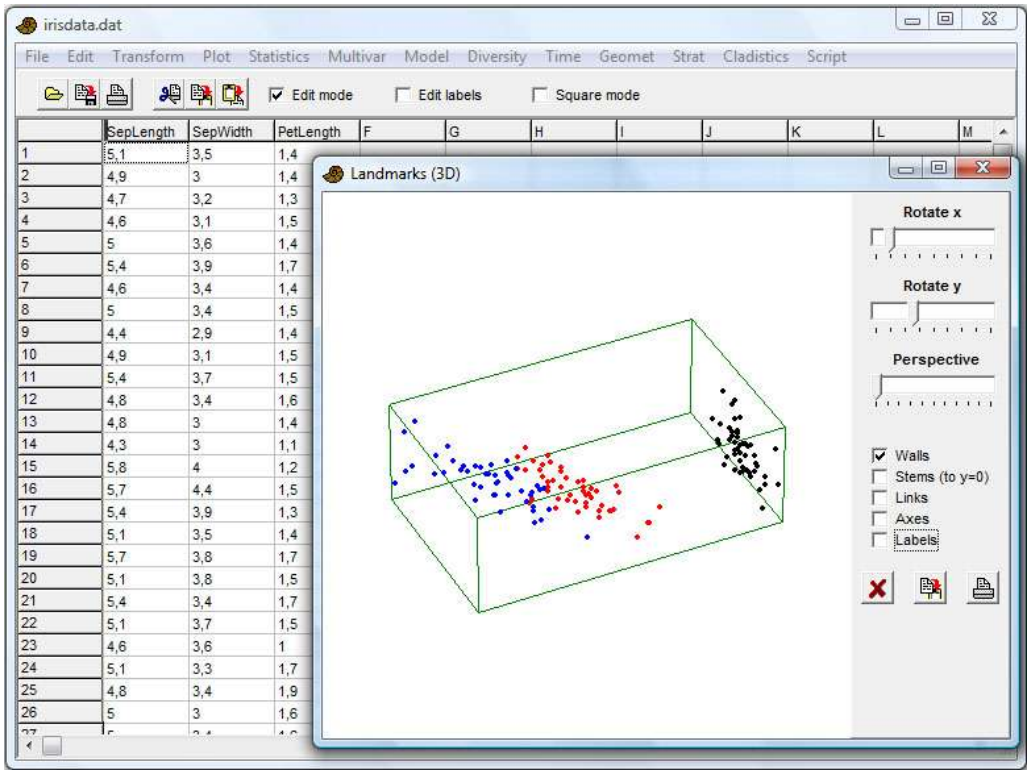
Points with missing values in X and/or Y are disregarded.



Landmarks 3D

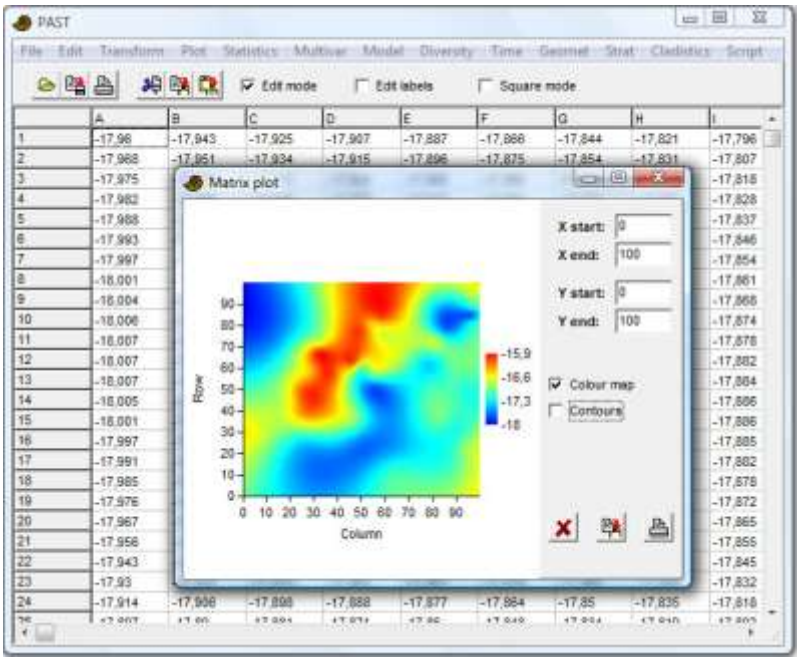
Plotting of points in 3D (XYZ triples). Especially suited for 3D landmark data, but can also be used e.g. for PCA scatter plots along three principal components. The point cloud can be rotated around the *x* and the *y* axes (note: left-handed coordinate system). The 'Perspective' slider is normally not used. The 'Stems' option draws a line from each point down to a bottom plane, which can sometimes enhance 3D information. 'Lines' draws lines between consecutive landmarks within each separate specimen (row). 'Axes' shows the three coordinate axes with the centroid of the points as the origin.

Points with missing values in X, Y or Z are disregarded.



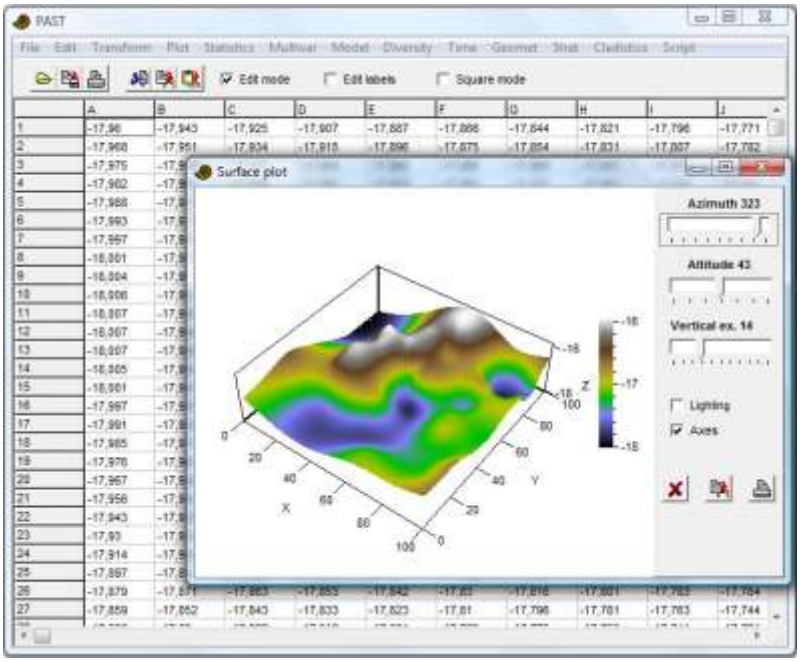
Matrix

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).



Surface

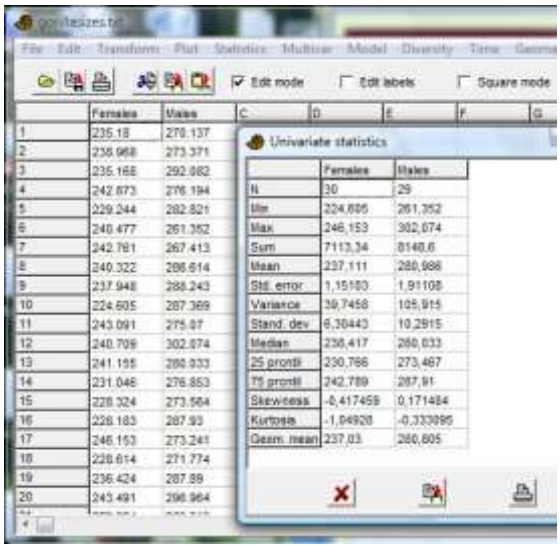
Three-dimensional landscape plot of a data matrix of elevation values. Colors are assigned according to height, or the surface can be gray-shaded using a lighting model with a fixed light source. Vertical exaggeration is adjustable. Missing values are replaced with the average. The data in the example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



Statistics menu

Univariate

This function computes a number of basic descriptive statistics for one or more samples of univariate data. Each sample must have at least 3 values, and occupies one column in the spreadsheet. The columns do not have to contain the same number of values. The example below uses two samples: the sizes in mm of the skulls of 30 female and 29 male gorillas. To run the analysis, the two columns (or the whole spreadsheet) must be selected.



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

Sum: The sum

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).

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 $(x_1 x_2 \cdots x_n)^{1/n}$.

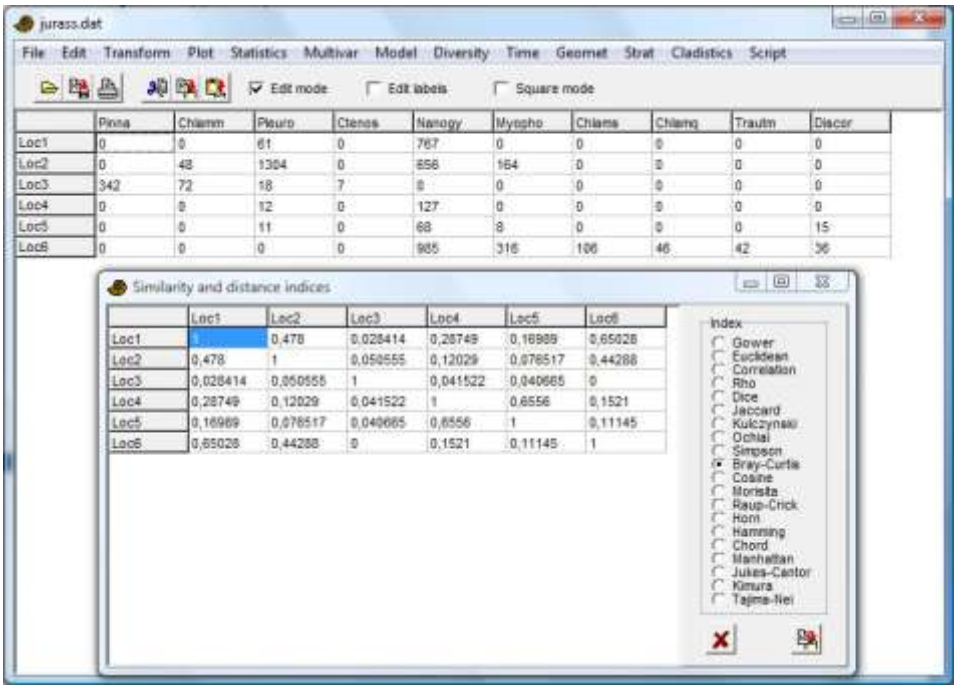
Bootstrapping

Selecting bootstrapping will compute lower and upper limits for 95% confidence intervals, using 9999 bootstrap replicates. Confidence intervals for the min and max values are not given, because bootstrapping is known to not work well for these statistics.

Missing data: Supported by deletion.

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.



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.

Euclidean

Basic Euclidean distance. In early versions of Past, this was normalized for the number of variables (the value is still adjusted for missing data).

$$d_{jk} = \sqrt{\sum_i \left(x_{ji} - x_{ki}\right)^2}.$$

Mahalanobis

A distance measure taking into account the covariance structure of the data. With **S** the variance-covariance matrix:

$$d_{jk} = \sqrt{\left(\mathbf{x_j} - \mathbf{x_k}\right)^T \mathbf{S}^{-1} \left(\mathbf{x_j} - \mathbf{x_k}\right)}.$$

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.

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 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 $\mathrm{M} / \mathrm{N}_{\min}$, where $\mathrm{N}_{\min}$ 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.

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.

$$\lambda_1 = \frac{\sum_i x_{ji}\left(x_{ji} - 1\right)}{\sum_i x_{ji}\left(\sum_i x_{ji} - 1\right)}$$

$$\lambda_2 = \frac{\sum_i x_{ki}\left(x_{ki} - 1\right)}{\sum_i x_{ki}\left(\sum_i x_{ki} - 1\right)}$$

$$d_{jk} = \frac{2\sum_i x_{ji} x_{ki}}{\left(\lambda_1 + \lambda_2\right)\sum_i x_{ji} \sum_i x_{ki}}.$$

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.

Horn

Horn's overlap index for abundance data (Horn 1966).

$$N_j = \sum_i x_{ji}$$

$$N_k = \sum_i x_{ki}$$

$$d_{jk} = \frac{\sum_i \left[\left(x_{ji} + x_{ki}\right)\ln\left(x_{ji} + x_{ki}\right)\right] - \sum_i x_{ji} \ln x_{ji} - \sum_i x_{ki} \ln x_{ki}}{\left(N_j + N_k\right)\ln\left(N_j + N_k\right) - N_j \ln N_j - N_k \ln N_k}.$$

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".

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:

$$d_{jk} = \sum_i \left| x_{ji} - x_{ki} \right|.$$

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.

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.

Vincenty, T. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review* 176:88-93.

Correlation table

A matrix is presented with the correlations between all pairs of columns. 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 (Pearson) and non-parametric (Spearman and Kendall) coefficients and tests are available. Algorithms follow Press et al. (1992) except that the significance of Spearman's coefficient is calculated with an exact test for *n*<=9 (see the section on rank/ordinal correlation below).

Pearson's *r* 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

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

Missing data: Supported by pairwise deletion.

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:

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

Missing data: Supported by column average substitution.

Reference

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

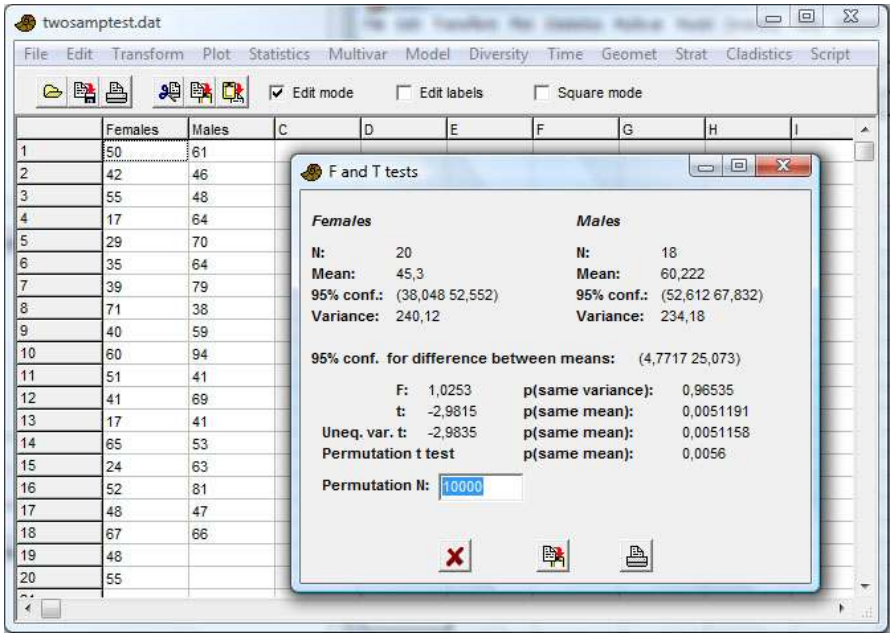
Var-covar

A symmetric matrix is presented with the variances and covariances between all pairs of columns.

Missing data: Supported by pairwise deletion.

F and *t* tests (two samples)

A number of classical, parametric statistics and tests for comparing the means and variances of two univariate samples (in two columns). Normal distribution is assumed.



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. 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, with 9999 replicates.

***F* test**

The *F* test has null hypothesis

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

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.

***t* test**

The *t* test has null hypothesis

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

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 variances are very different, although it can be argued that testing for difference in the means in this case is questionable. 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

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

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. 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.

Missing data: Supported by deletion.

t test (one sample)

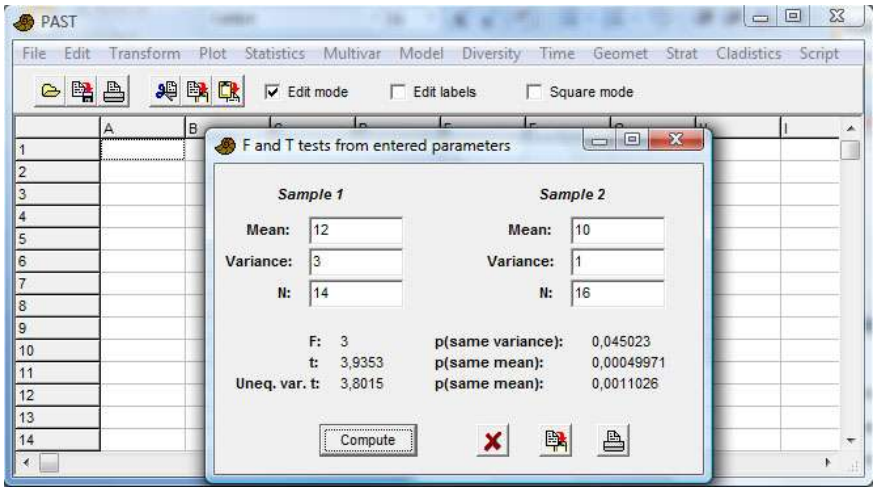
The one-sample *t* test is used to investigate whether the sample is likely to have been taken from a population with a given (theoretical) mean.

The 95% confidence interval for the mean is calculated using the *t* distribution.

Missing data: Supported by deletion.

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.

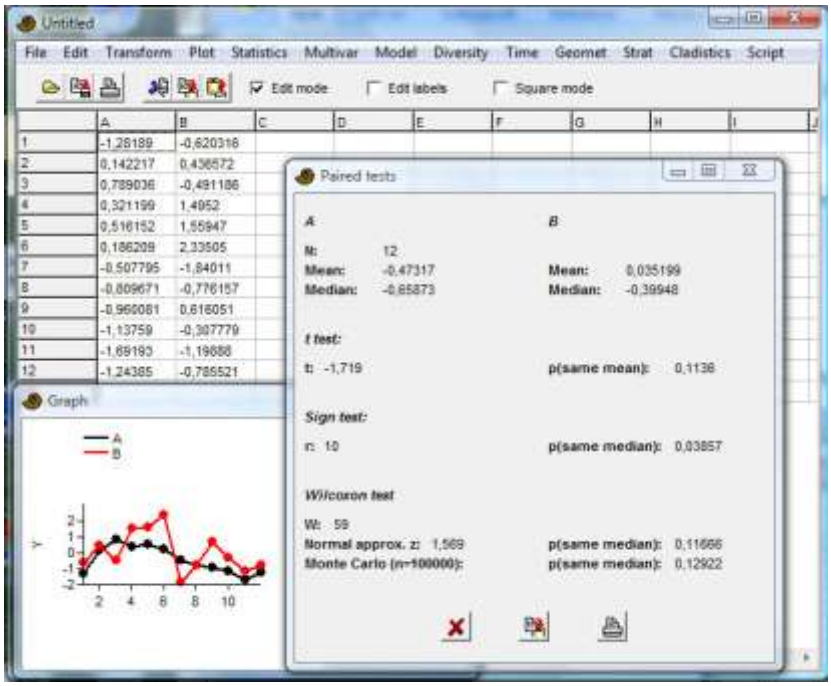


Paired tests (*t*, sign, Wilcoxon)

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 normal one-sample *t* test. 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.

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 will typically have lower power than the other paired tests, but make 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

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

(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*:

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

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

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.

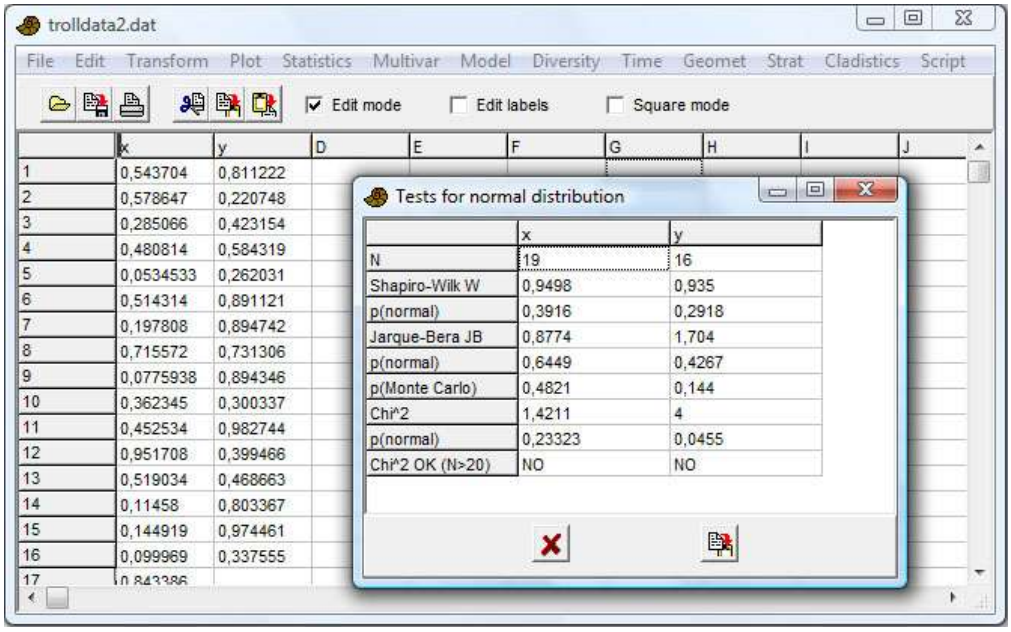
The Monte Carlo significance value is based on 99,999 random reassignments of values to columns, within each pair. This value will be practically identical to the exact *p* value.

For *n*<26, an exact *p* value is computed, by complete enumeration of all possible reassignments (there are 2^n of them, i.e. more than 33 million for *n*=25). This is the preferred *p* value, if available.

Missing data: Supported by deletion of the row.

Normality tests

Four statistical tests for normal distribution of one or several samples of univariate data, given in columns. The data below were generated by a random number generator with uniform distribution.



For all the 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. Of the four given tests, the Shapiro-Wilk and Anderson-Darling are considered to be the more exact, and the two other tests (Jarque-Bera and a chi-square test) are given for reference. 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.

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.

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 above, 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.

Chi-square test

The chi-square test uses an expected normal distribution in four bins, based on the mean and standard deviation estimated from the sample, and constructed to have equal expected frequencies in all bins. The upper limits of all bins, and the observed and expected frequencies, are displayed. A warning message is given if $n<20$, i.e. expected frequency less than 5 in each bin. There is 1 degree of freedom. This test is both theoretically questionable and has low power, and is not recommended. It is included for reference.

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 cumulativedistribution 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

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

Missing data: Supported by deletion.

References

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

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.

Chi^2

The Chi-square test expects two columns with numbers of elements in different bins (compartments). For example, this test can be used to compare two associations (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).

There are two options that you should select or not for correct results. "Sample vs. expected" should be ticked if your second column consists of values from a theoretical distribution (expected values) with zero error bars. If your data are from two counted samples each with error bars, leave this box open. This is *not* a small-sample correction.

"One constraint" should be ticked if your expected values have been normalized in order to fit the total observed number of events, or if two counted samples necessarily have the same totals (for example because they are percentages). This will reduce the number of degrees of freedom by one.

When "one constraint" is selected, a permutation test is available, with 10000 random replicates. For "Sample vs. expected" these replicates are generated by keeping the expected values fixed, while the values in the first column are random with relative probabilities as specified by the expected values, and with constant sum. For two samples, all cells are random but with constant row and column sums.

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

With one constraint, the *Fisher's exact test* is also given (two-tailed). When available, the Fisher's exact test may be far superior to the chi-square. For large tables or large counts, the computation time can be prohibitive and will time out after one minute. 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).

Missing data: Supported by row deletion.

References

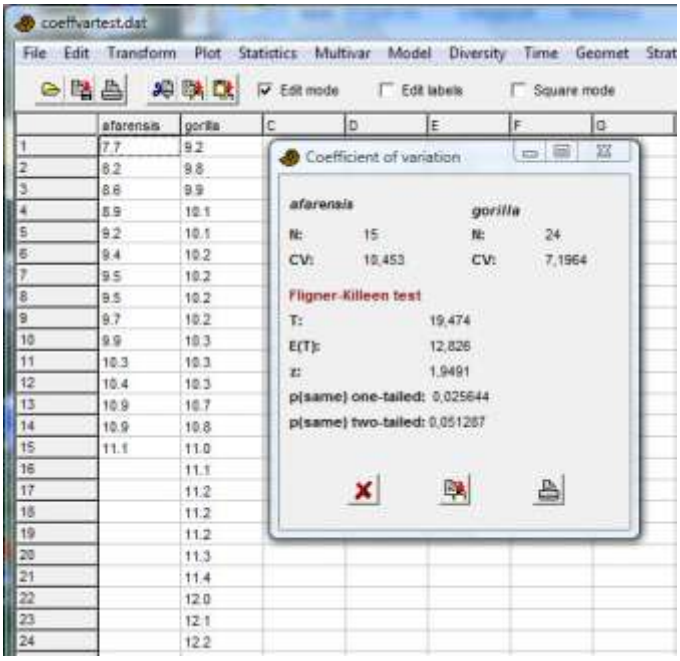
Brown, D. & P. Rothery. 1993. Models in biology: mathematics, statistics and computing. John Wiley & Sons.

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.

Coefficient of variation

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



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, with 9999 replicates.

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.

Missing data: Supported by deletion.

References

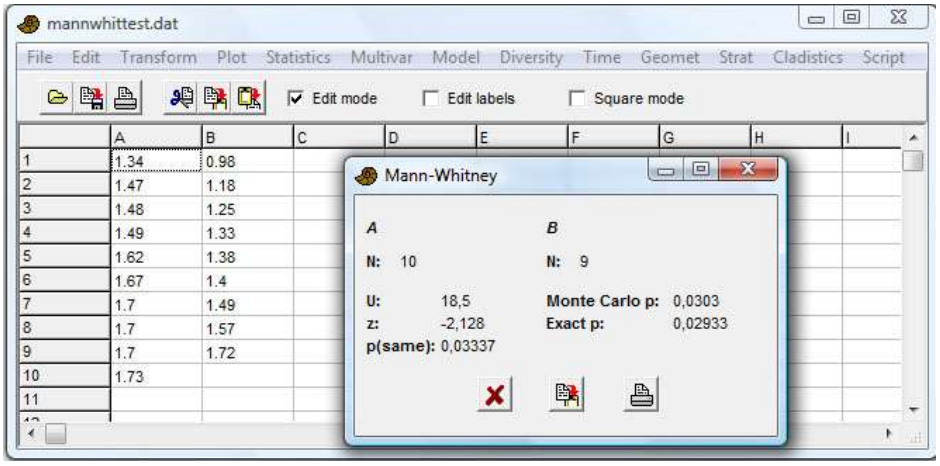
Donnelly, S.M. & Kramer, A. 1999. Testing for multiple species in fossil samples: An evaluation and comparison of tests for equal relative variation. *American Journal of Physical Anthropology* 108:507-529.

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

Mann-Whitney test

The two-tailed (Wilcoxon) Mann-Whitney *U* test can be used to test whether the medians of two independent samples are different. It does not assume normal distribution, but does assume equal-shaped distribution in both groups. The null hypothesis is

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



This test is non-parametric, which means that the distributions can be of any shape.

For each value in sample 1, count 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 $U_1+U_2=n_1n_2$).

In the left column is given 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:

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

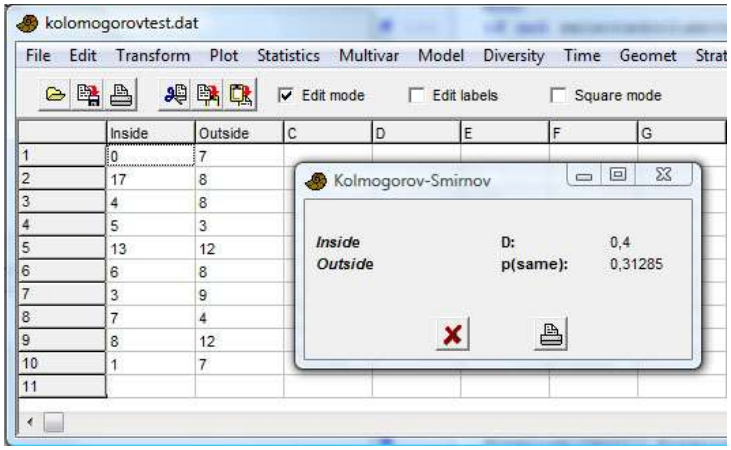
where $n=n_1+n_2$ and f_g is the number of elements in tie *g*.

For $n_1+n_2<=30$ (e.g. 15 values in each group), an exact *p* value is given, based on all possible combinations of group assignment. If available, always use this exact value. For larger samples, the asymptotic approximation is quite accurate. A Monte Carlo value based on 10 000 random assignments is also given – the purpose of this is mainly as a control on the asymptotic value.

Missing data: Supported by deletion.

Kolmogorov-Smirnov

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.



	Inside	Outside	C	D	E	F	G
1	0	7					
2	17	8					
3	4	8					
4	5	3					
5	13	12					
6	6	8					
7	3	9					
8	7	4					
9	8	12					
10	1	7					
11							

In the version of the test provided by Past, both columns must represent samples. You can not 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).

Missing data: Supported by deletion.

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.

Rank/ordinal correlation

These rank-order correlations and tests are used to investigate correlation between two variables, given in two columns.

Spearman's (non-parametric) rank-order correlation coefficient is the linear correlation coefficient (Pearson's *r*) of the ranks. According to 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.

The Monte Carlo permutation test is based on 9999 random replicates.

These statistics are also available through the "Correlation" module, but then without the permutation option.

Missing data: Supported by deletion.

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 $N<100$). For larger N the given asymptotic test (log-ratio test) is accurate.

References

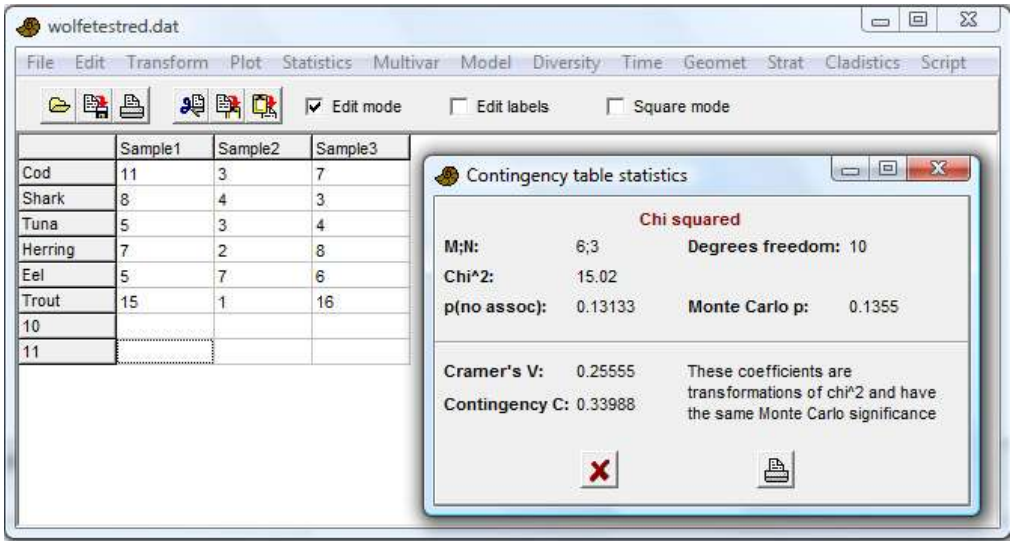
Olsson, U., F. Drasgow & N.J. Dorans. 1982. The polyserial correlation coefficient. *Psychometrika* 47:337-347.

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

Contingency table

A contingency table is input to this routine. Rows represent the different states of one nominal variable, columns represent the states of another nominal variable, and cells contain the counts of occurrences of that specific state (row, column) of the two variables. The significance of association between the two variables (based on chi-squared) is then given, with *p* values from the chi-squared distribution and from a permutation test with 9999 replicates.

For example, rows may represent taxa and columns samples as usual (with specimen counts in the cells). The contingency table analysis then gives information on whether the two variables of taxon and locality are associated. If not, the data matrix is not very informative.



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

Note that for *n*x2 tables, the Fisher's exact test is available in the Chi^2 module.

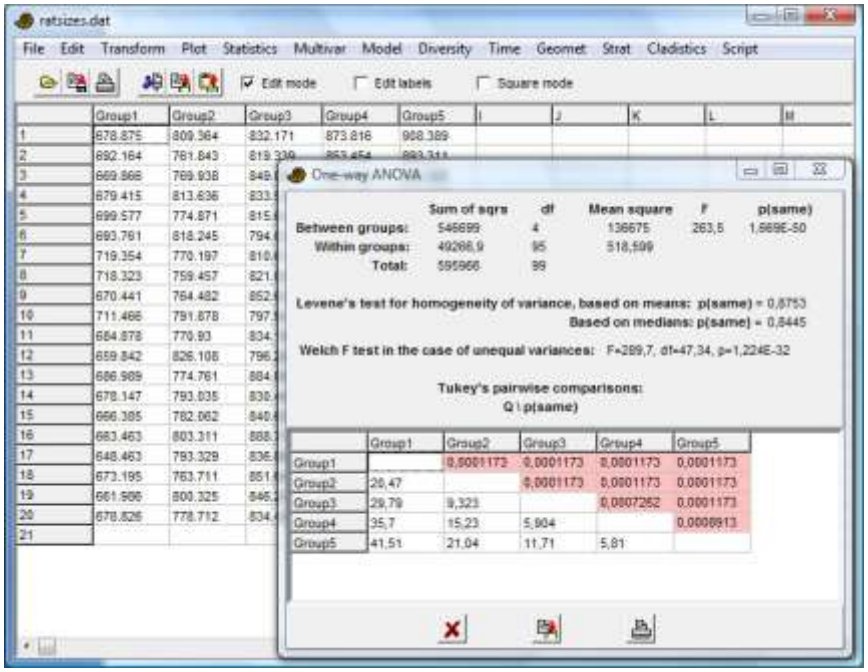
Missing data not supported.

Reference

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

One-way ANOVA

One-way ANOVA (analysis of variance) is a statistical procedure for testing the null hypothesis that several univariate samples (in columns) 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.

Omega squared

The omega squared is a measure of effect size, varying from 0 to 1 (not available for repeated measures ANOVA):

$$\omega^2 = \frac{\mathrm{SS}_{\mathrm{bg}} - \mathrm{df}_{\mathrm{bg}}\mathrm{MS}_{\mathrm{wg}}}{\mathrm{SS}_{\mathrm{total}} + \mathrm{MS}_{\mathrm{wg}}}.$$

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.

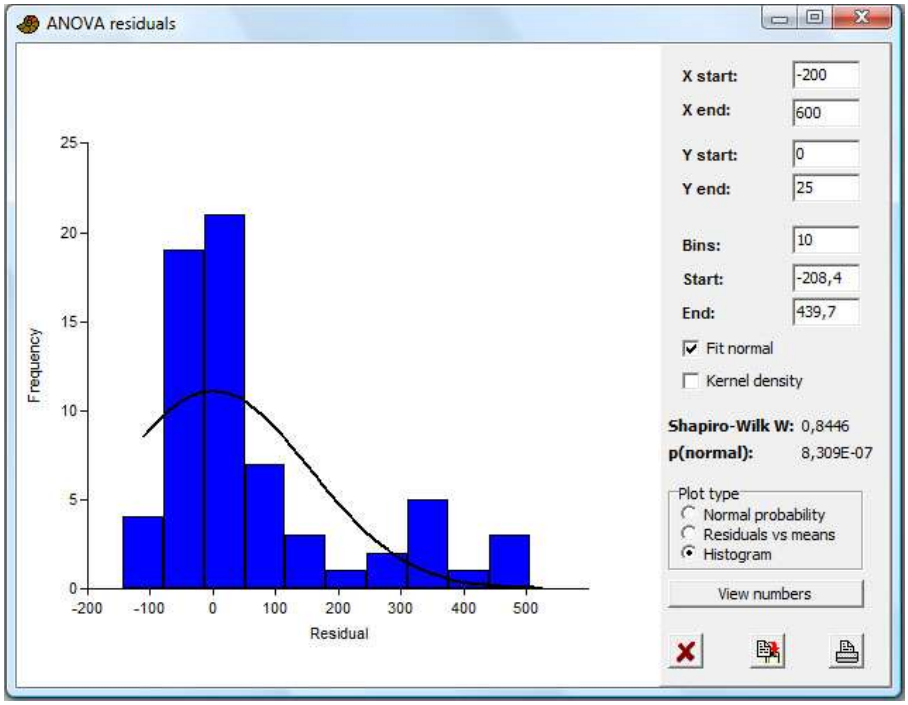
Unequal-variance (Welch) ANOVA

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.

Analysis of residuals

The “Residuals” button opens a window for analysing the 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).



Post-hoc pairwise 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 Tukey's HSD (Honestly Significant Difference) 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. Sample sizes do not have to be equal for the version of Tukey's test used.

Repeated measures (within-subjects) ANOVA

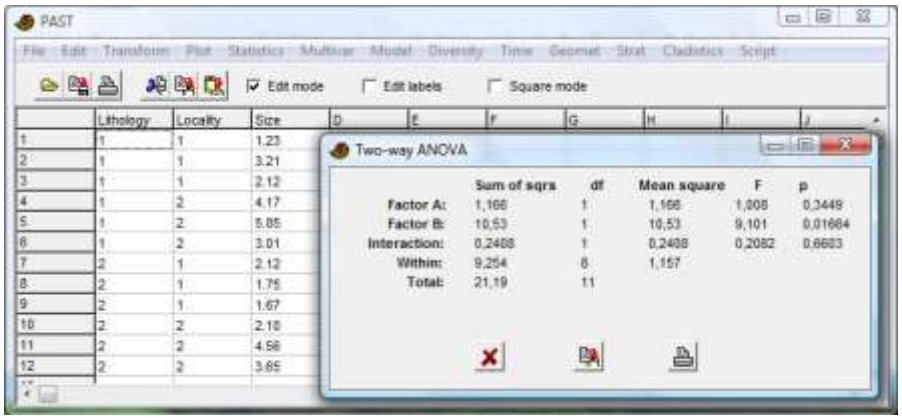
Ticking the “Repeated measures” box selects another type of one-way ANOVA, where the values in each row are observations on the same “subject”. Repeated-measures ANOVA is the extension of the paired t test to several samples. Each column (sample) must contain the same number of values.

Missing values: Supported by deletion, except for repeated measures ANOVA, where missing values are not supported.

Two-way ANOVA

Two-way ANOVA (analysis of variance) is a statistical procedure for testing the null hypotheses that several univariate samples have the same mean across each of the two factors, 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 test assumes a fixed-factor design (the usual case).

Three columns are needed. First, a column with the levels for the first factor (coded as 1, 2, 3 etc.), then a column with the levels for the second factor, and finally the column of the corresponding measured values.



The algorithm uses weighted means for unbalanced designs.

Repeated measures (within-subjects) ANOVA

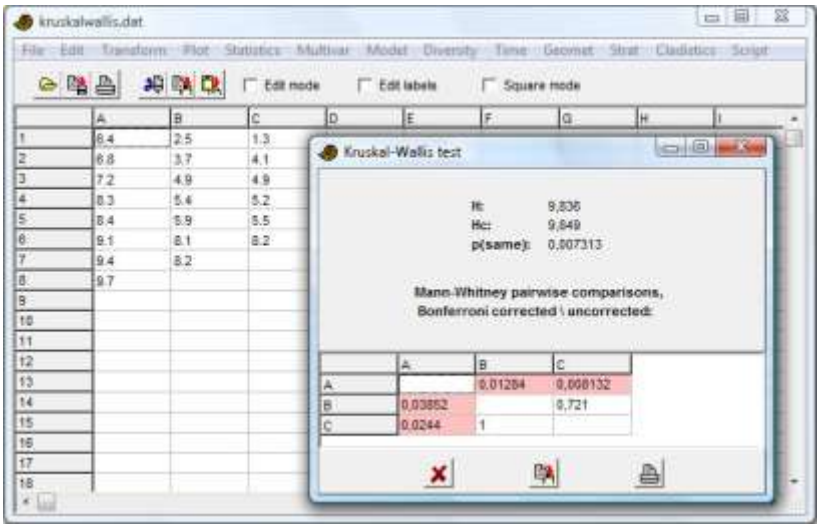
Ticking the “Repeated measures” box selects another type of two-way ANOVA, where each of a number of “subjects” have received several treatments. The data formatting is as above, but it is required that all measurements on the first subject are given in the first rows, then all measurements on the second subject, etc. Each subject must have received all combinations of treatments, and each combination of treatments must be given only once. This means that for e.g. two factors with 2 and 3 levels, each subject must occupy exactly 2x3=6 rows. The program automatically computes the number of subjects from the number of given level combinations and the total number of rows.

Missing values : Rows with missing values are deleted.

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.

Post-hoc pairwise tests

Mann-Whitney pairwise test *p* values are given for all N_p=G(G-1)/2 pairs of groups, in the upper right triangle of the matrix. The lower right triangle gives the corresponding *p* values, but multiplied with N_p as a conservative correction for multiple testing (Bonferroni correction). The values use the asymptotic approximation described under the Mann-Whitney module. If samples are very small, it may be useful to run the exact test available in that module instead.

Missing data: Supported by deletion.

Reference

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

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 groups (treatments) are given in columns, and the cases in rows.

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 are 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):

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

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:

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

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

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.

Missing values not supported.

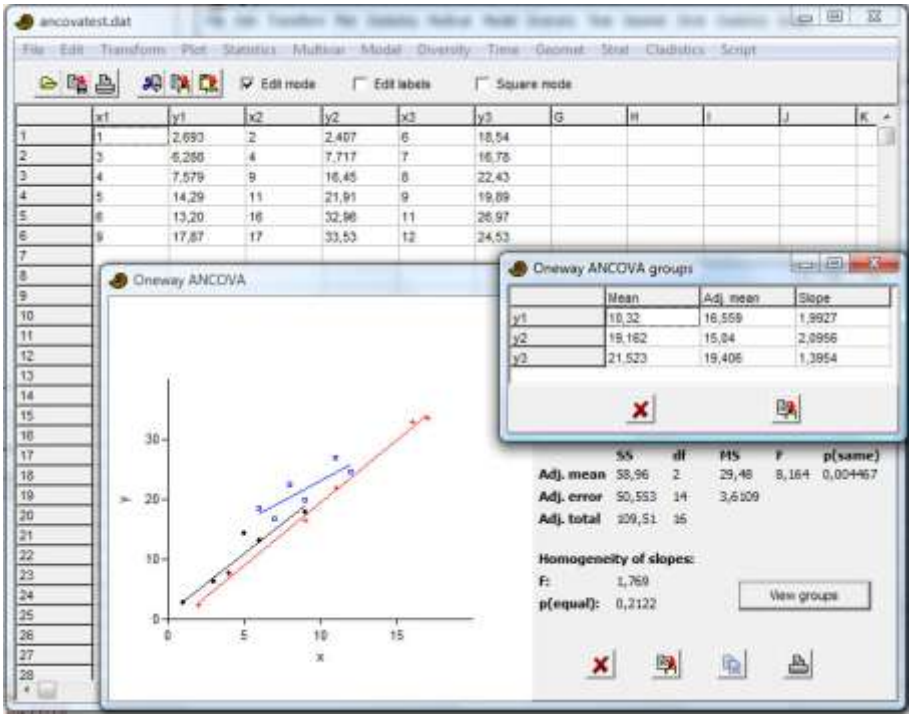
Reference

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

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).



The program 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 $p<0.005$. Equality of slopes can not be rejected ($p=0.21$).

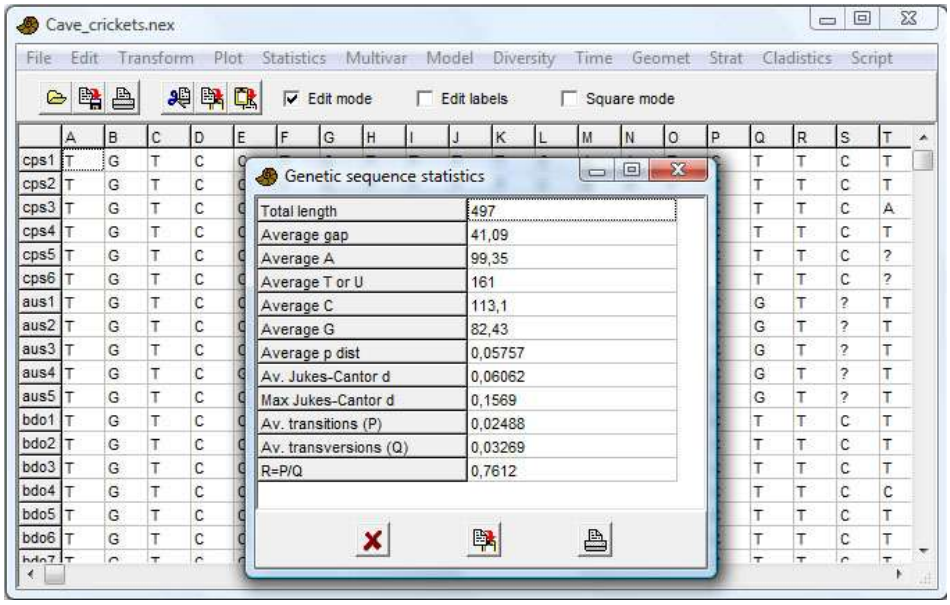
"View groups" gives the summary statistics for each group (mean, adjusted mean and regression slope).

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.

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

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)

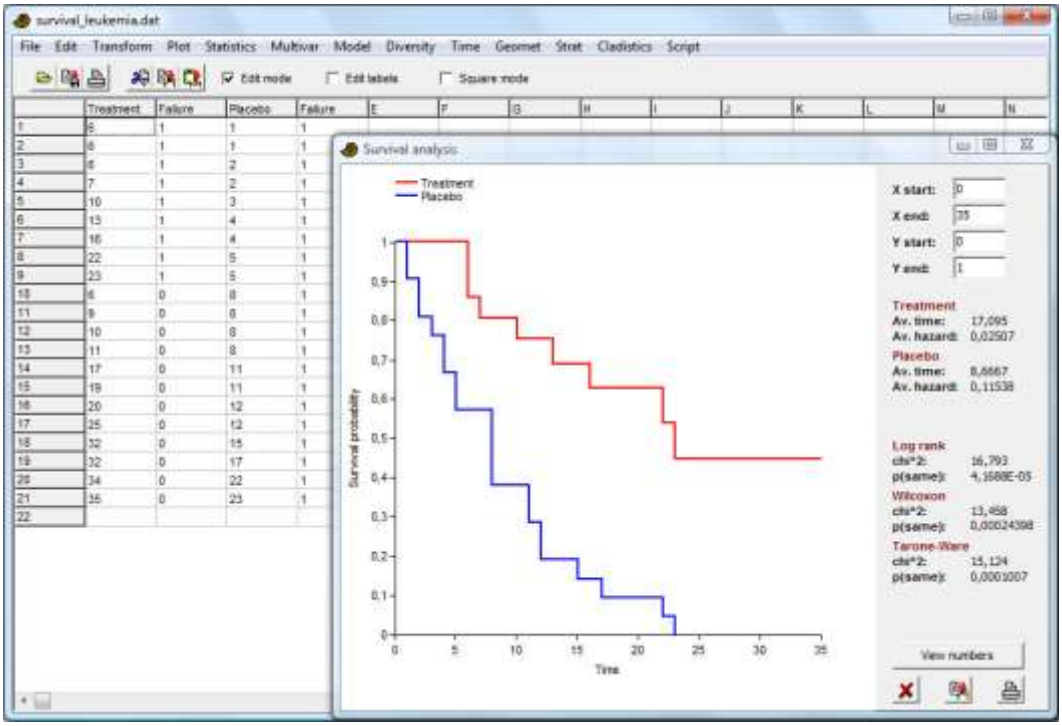
***R*=*P*/*Q*:** The transition/transversion ratio

Missing data: Treated as gaps.

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

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

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.

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"):

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

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

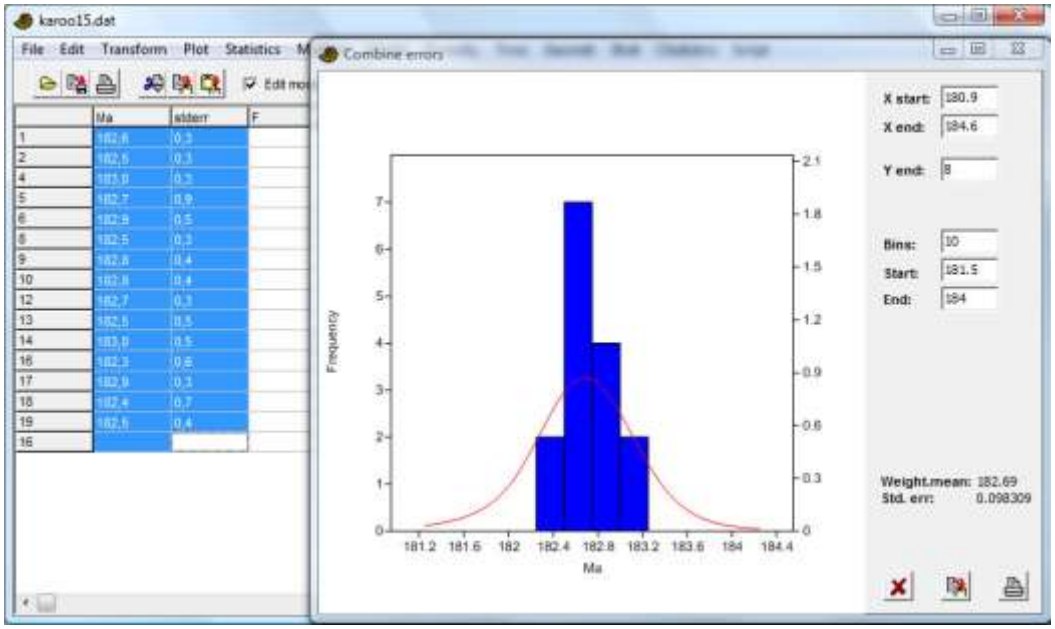
Interval: *OR* / *EF* to *OR* x *EF*

Note there is currently no continuity correction.

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

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.

Multivar menu

Principal components

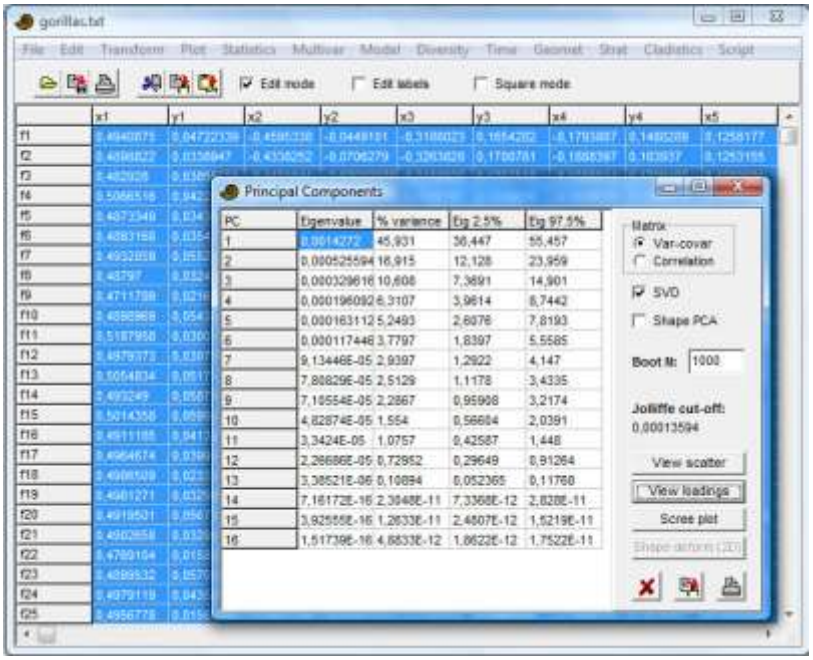
Principal components analysis (PCA) finds hypothetical variables (components) accounting for as much as possible of the variance in your multivariate data (Davis 1986, Harper 1999). These new variables are linear combinations of the original variables. PCA may be used for reduction of the data set to 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). Bruton & Owen (1988) describe a typical morphometrical application of PCA.

The input data is a matrix of multivariate data, with items in rows and variates in columns. There is no separate centering of groups before eigenanalysis – groups are not taken into account.

The PCA routine finds the eigenvalues and eigenvectors of the variance-covariance matrix or the correlation matrix. Use var-covar 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.

Groups: If groups are specified by row color, 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.

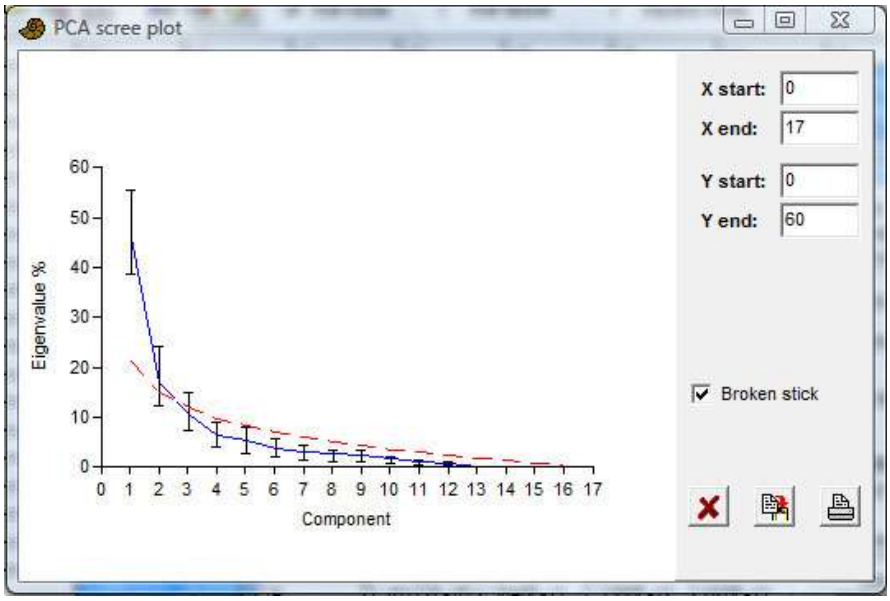
In the example below (landmarks from gorilla skulls), component 1 is strong, explaining 45.9% of variance. The bootstrapped confidence intervals are not shown un less the ‘Boot N’ value is non-zero.



The Jolliffe cut-off value may indicate the number of significant principal components (Jolliffe, 1986). Components with eigenvalues smaller than this value may be considered insignificant, but too much weight should not be put on this criterion.

Row-wise bootstrapping is carried out if a positive number of bootstrap replicates (e.g. 1000) is given in the 'Boot 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.

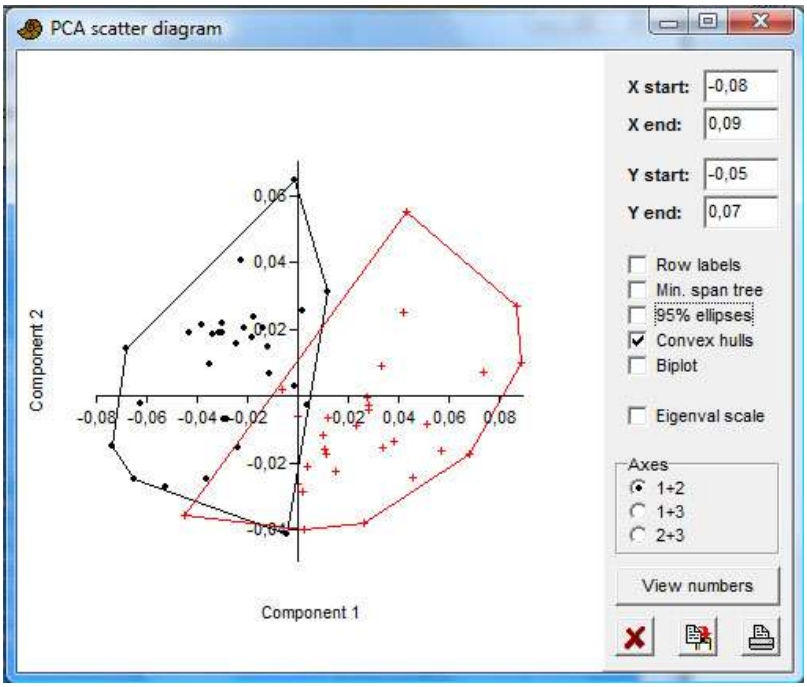
The 'Scree plot' (simple plot of eigenvalues) may also indicate the number of significant components. After this curve starts to flatten out, the components may be regarded as insignificant. 95% confidence intervals are shown if bootstrapping has been carried out. The eigenvalues expected under a random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



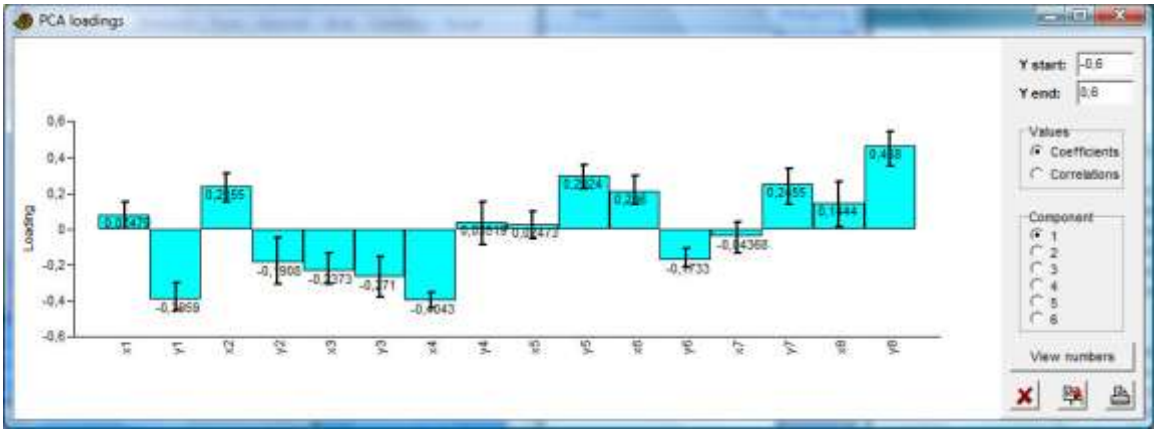
In the gorilla example above, the eigenvalues for the 16 components (blue line) lie above the broken stick values (red dashed line) for the first two components, although the broken stick is inside the 95% confidence interval for the second component.

The 'View scatter' option shows all data points (rows) plotted in the coordinate system given by two of the components. If you have colored (grouped) rows, the groups 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.

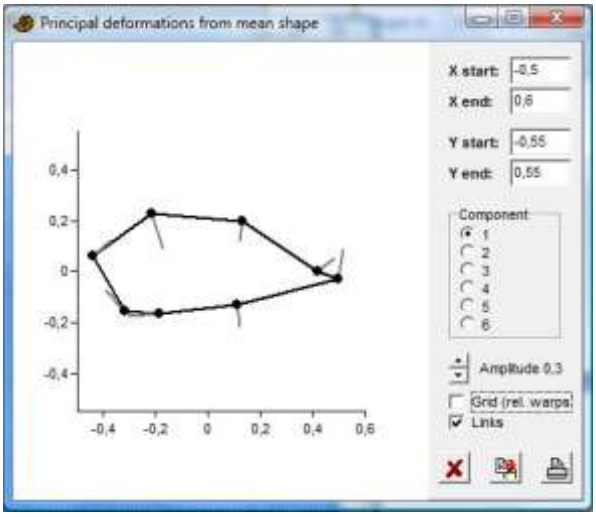


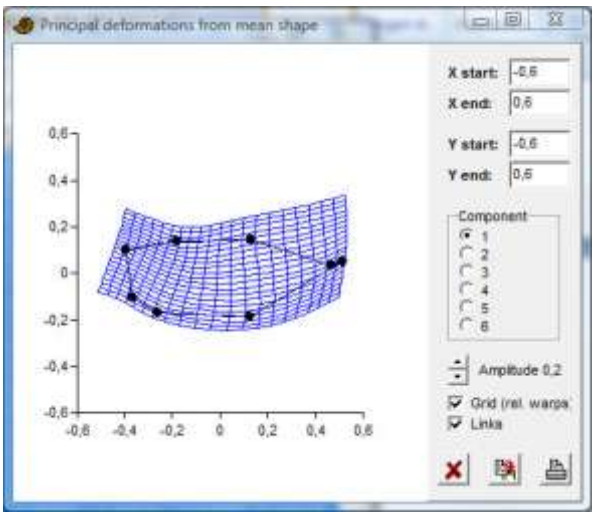
The 'View loadings' option shows to what degree your different original variables (given in the original order along the x axis) enter into the different components (as chosen in the radio button panel). These component loadings are important when you try to interpret the 'meaning' of the components. 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).



The 'SVD' option enforces the superior Singular Value Decomposition algorithm instead of "classical" eigenanalysis. The two algorithms will normally give almost identical results, but axes may be flipped.

The 'Shape deform' option is designed for 2D landmark position data. The default Shape Deformation plot is a ‘lollipop plot’ with the mean shape shown as dots and vectors (lines) pointing in the directions of the axis loadings. The “Grid” option shows the thin-plate spline deformation grids corresponding to the different components. This is in effect a “relative warps” analysis, including the uniform component. For relative warps without the uniform component, see “Relative warps” in the Geometry menu.





Missing data can be handled by one of three 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).
3. *Pairwise deletion*: Pairwise deletion in the var/covar or correlation matrix. Can work when the number of missing values is small. This option will enforce the eigendecomposition method (i.e. not SVD).

References

Bruton, D.L. & A.W. Owen. 1988. The Norwegian Upper Ordovician illaenid trilobites. *Norsk Geologisk Tidsskrift* 68:241-258.

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.

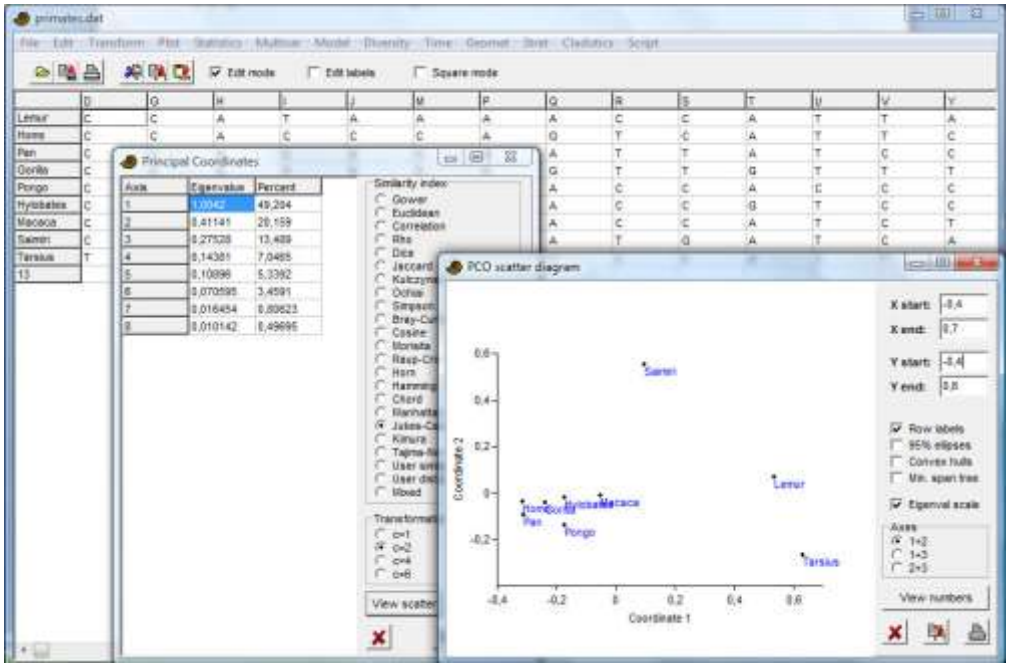
Jackson, D.A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74:2204-2214.

Jolliffe, I.T. 1986. Principal Component Analysis. Springer-Verlag.

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 (PCO) is another ordination method, also known as Metric Multidimensional Scaling. The algorithm is from Davis (1986).



The PCO 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 'View scatter' option allows you to see all your data points (rows) plotted in the coordinate system given by the PCO. 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).

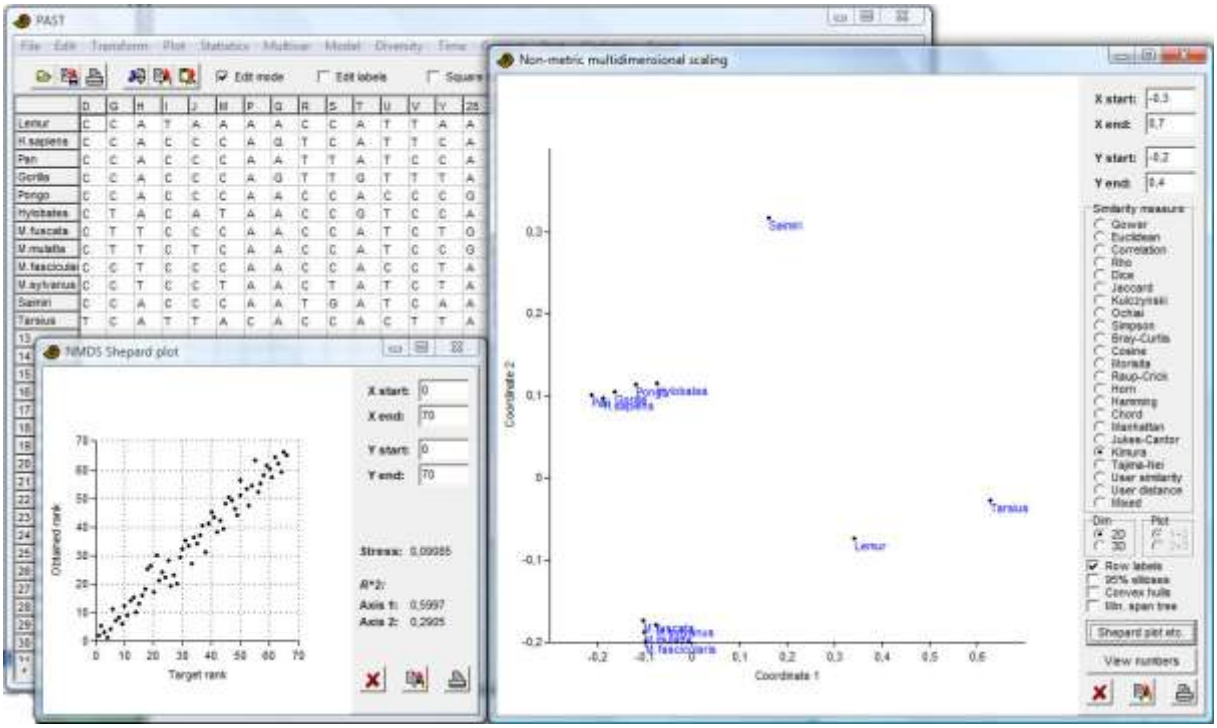
References

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 above. 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 random 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 PCO as the initial condition, but this rarely gives the best solution. 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 & Oono (in press).

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 length of the vectors are arbitrarily scaled to make a readable biplot, so only their directions and relative lengths should be considered.

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).

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.

Correspondence analysis

Correspondence analysis (CA) is yet another ordination method, somewhat similar to PCA but for *counted data*. 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 favour 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 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 'View scatter' option allows you to see all your data points (rows) plotted in the coordinate system given by the CA. If you have colored (grouped) rows, the different groups will be shown using different symbols and colours.

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.

PAST presently uses a symmetric scaling ("Benzecri scaling").

If you have more than two columns in your data set, you can choose to view a scatter plot on the second and third axes.

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).

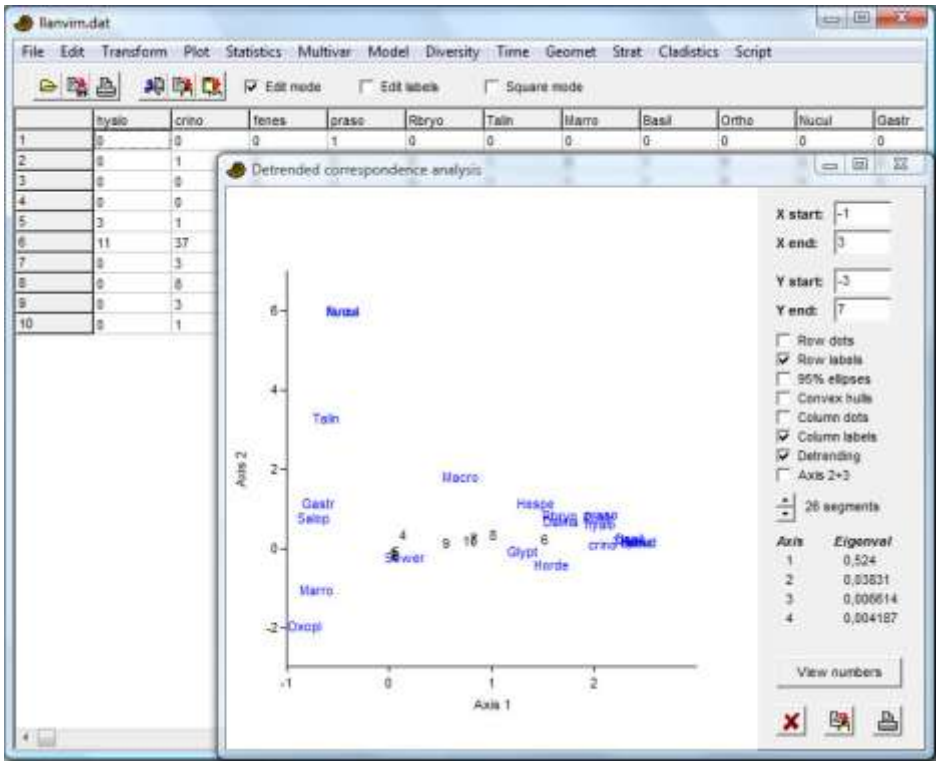
Missing data is supported by column average substitution.

Reference

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.

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 (vice versa prior to v. 1.79). When the 'Detrending' option is switched off, a basic Reciprocal Averaging will be carried out. The result should then be similar to Correspondence Analysis (see above).



Eigenvalues for the first three 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.

Missing data is supported by column average substitution.

References

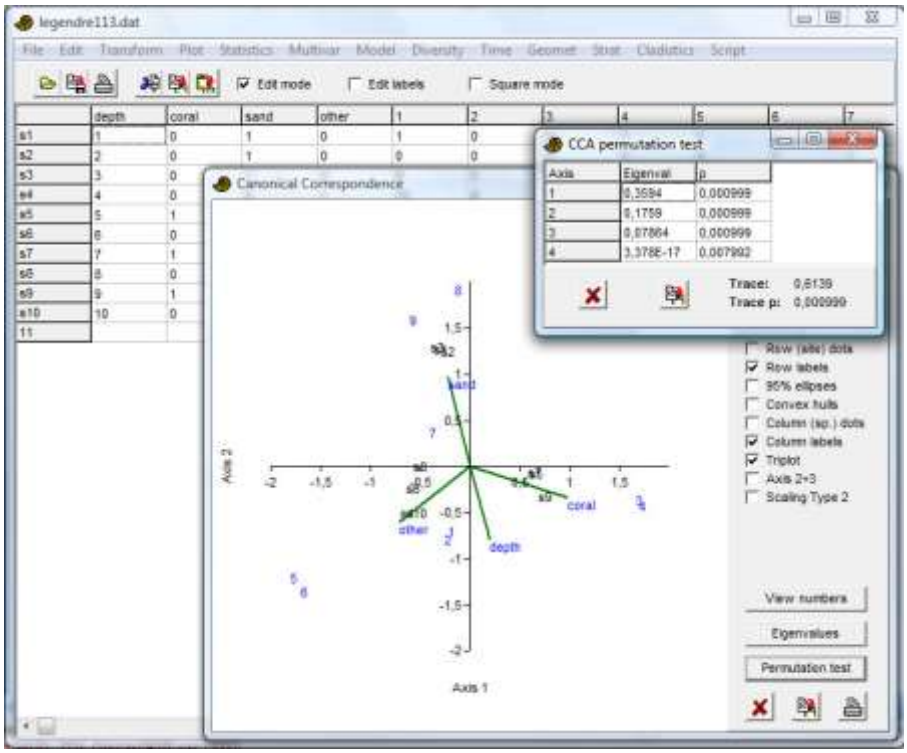
Hill, M.O. & H.G. Gauch Jr. 1980. Detrended Correspondence analysis: an improved ordination technique. *Vegetatio* 42:47-58.

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.

Reference

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

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.

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 $100 \frac{\lambda_i^2}{\sum \lambda_i^2}$.

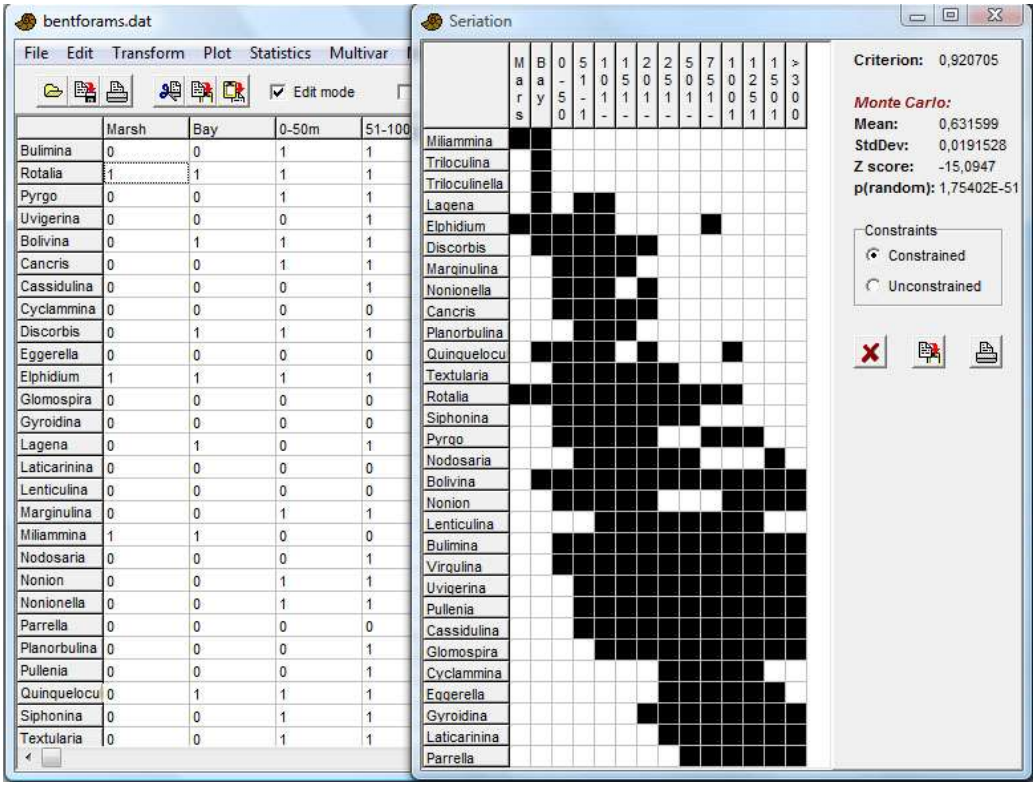
Missing data supported by column average substitution.

Reference

Rohlf, F.J. & M. Corti. 2000. Use of two-block partial least squares to study covariation in shape. *Systematic Biology* 49:740-753.

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.

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 neighbour). 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,

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 in order to informally assess the robustness of the groupings. If a grouping is changed when trying another algorithm, that grouping should perhaps not be trusted.

For Ward's method, a Euclidean distance measure is inherent to the algorithm. For UPGMA and single linkage, the distance matrix can be computed using 20 different indices, as described under the Statistics menu.

Missing data: The cluster analysis algorithm can handle missing data, coded question mark (?). 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.

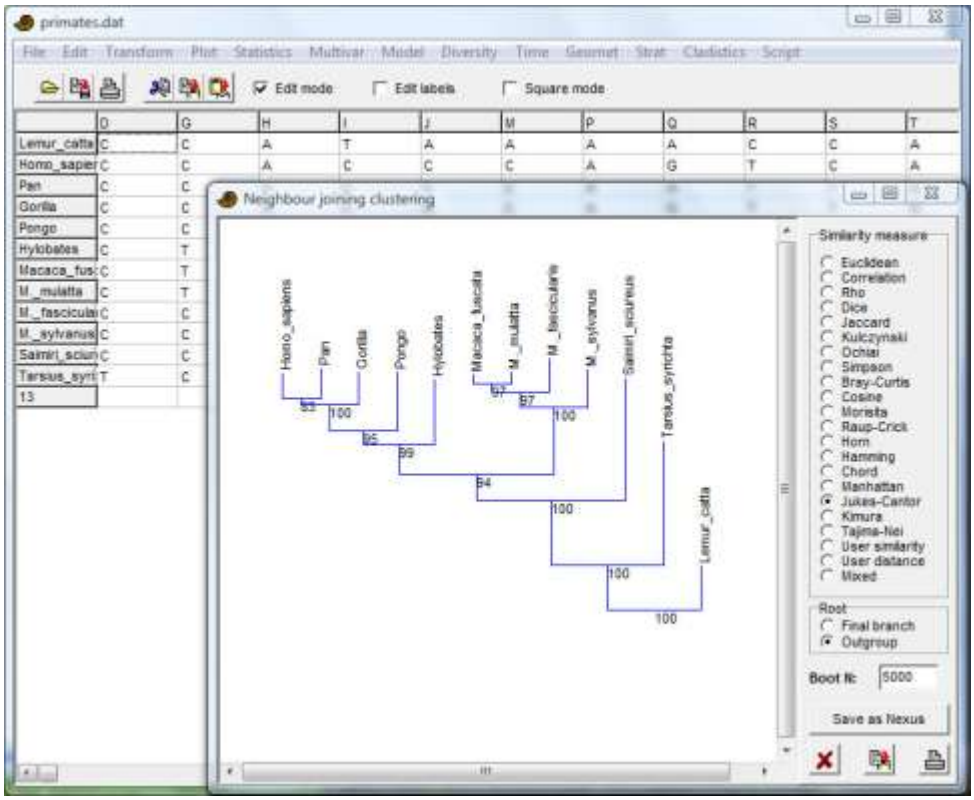
Stratigraphically constrained clustering: This option will allow only adjacent rows or groups of rows to be joined during the agglomerative clustering procedure. May produce strange-looking (but correct) dendrograms.

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.

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 the first row in the data matrix (outgroup).

Missing data supported by pairwise deletion.

References

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. 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.

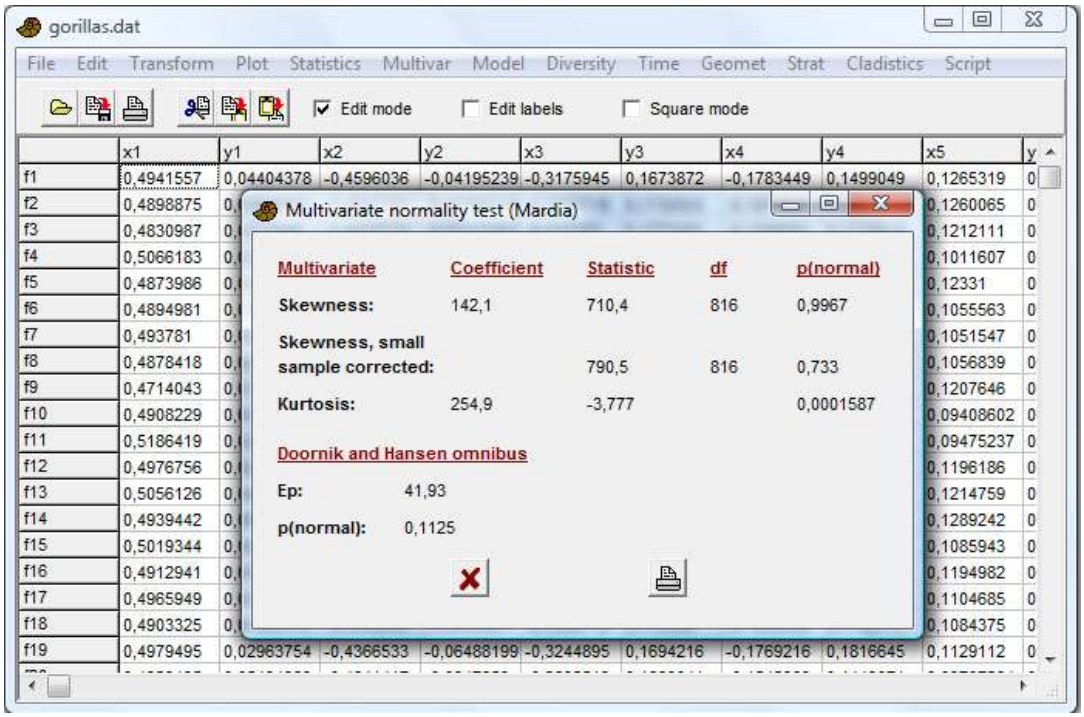
Missing data supported by column average substitution.

Reference

Bow, S.-T. 1984. Pattern recognition. Marcel Dekker, New York.

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.

Discriminant/Hotelling

Given two sets of multivariate data, an axis is constructed which maximizes the difference between the sets (e.g. Davis 1986). The two sets are then plotted along this axis using a histogram. This module expects the rows in the two data sets to be grouped into two sets by coloring the rows, e.g. with black (dots) and red (crosses).

Equality of the means of the two groups is tested by a multivariate analogue to the *t* test, called *Hotelling's T-squared*, and a *p* value for this test is given. Normal distribution of the variables is required, and also that the number of cases is at least two more than the number of variables.

Number of constraints: For correct calculation of the Hotelling's *p* value, 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).

Discriminant analysis can be used for visually confirming or rejecting the hypothesis that two species are morphologically distinct. Using a cutoff point at zero (the midpoint between the means of the discriminant scores of the two groups), a classification into two groups is shown in the "View numbers" option. The percentage of correctly classified items is also given.

Discriminant function: New specimens can be classified according to the discriminant function. Take the inner product between the measurements on the new specimen and the given discriminant function factors, and then subtract the given offset value.

Leave one out (cross-evaluation): An option is available for leaving out one row (specimen) at a time, re-computing the discriminant analysis with the remaining specimens, and classifying the left-out row accordingly (as given by the Score value).

Missing data supported by column average substitution.

Landmark warps

This function should only be used if the discriminant analysis has been carried out on 2D landmark data. It allows the interactive plotting of shape deformations as a function of position along the discriminant axis, either as lollipop plots (vectors away from the mean landmark positions) or as thin-plate spline deformations. TEMPORARILY (?) REMOVED BECAUSE OF LACK OF STABILITY

EFA warps

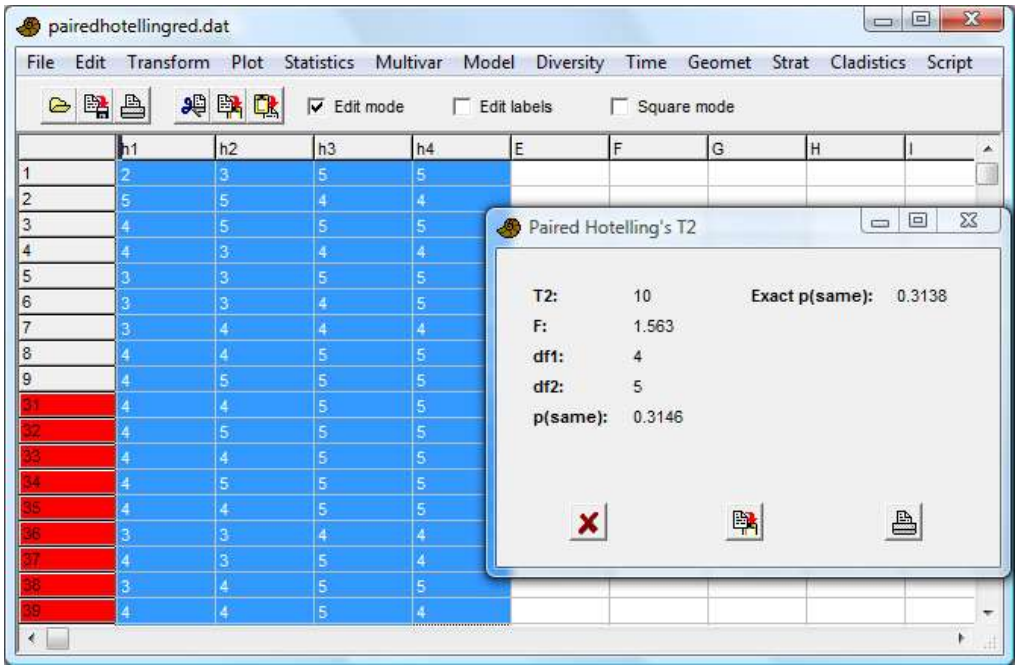
This function should only be used if the discriminant analysis has been carried out on coefficients computed by the Elliptic Fourier Analysis module. It allows the interactive plotting of outlines as a function of position along the discriminant axis. TEMPORARILY (?) REMOVED BECAUSE OF LACK OF STABILITY

Reference

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

Paired Hotelling

The paired Hotelling's test expects two groups of multivariate data, marked with different colours. 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:

$$\mathbf{Y}_i = \mathbf{X}_{1i} - \mathbf{X}_{2i}$$

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_i \mathbf{Y}_i$$

$$\mathbf{S}_y = \frac{1}{n-1}\sum_i (\mathbf{Y}_i - \bar{\mathbf{y}})(\mathbf{Y}_i - \bar{\mathbf{y}})^{\mathbf{T}}$$

$$T^2 = n\bar{\mathbf{y}}^{\mathbf{T}}\mathbf{S}_y^{-1}\bar{\mathbf{y}}$$

$$F = \frac{n-p}{p(n-1)}T^2$$

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.

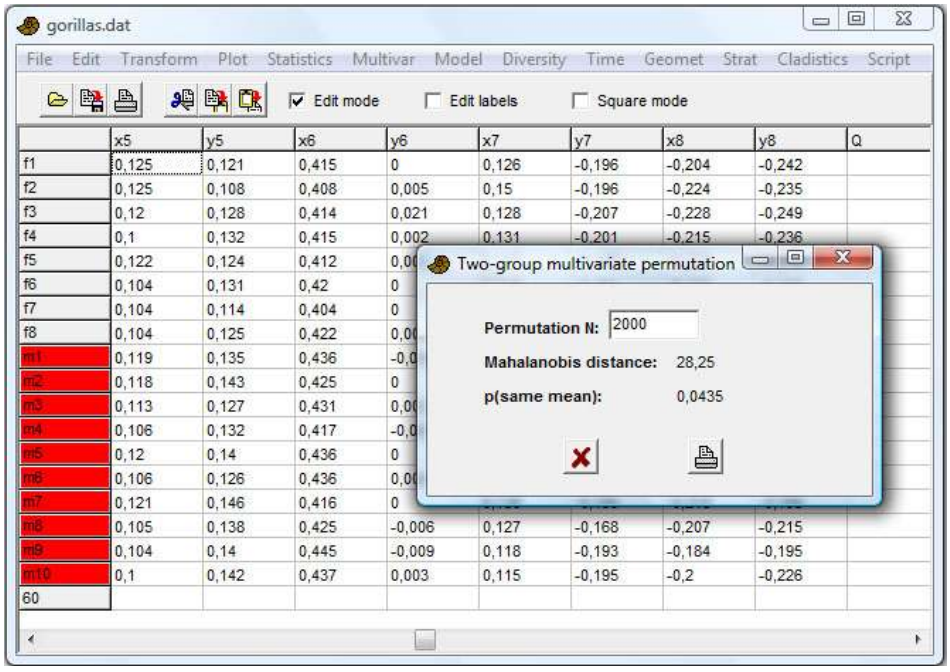
Missing data supported by column average substitution.

Two-group permutation

This module expects the rows in the two data sets to be grouped into two sets by coloring the rows, e.g. with black (dots) and red (crosses).

Equality of the means of the two groups is tested using permutation with 2000 replicates (can be changed by the user), and the Mahalanobis squared distance measure. The permutation test is an alternative to Hotelling's test when the assumptions of multivariate normal distributions and equal covariance matrices do not hold.

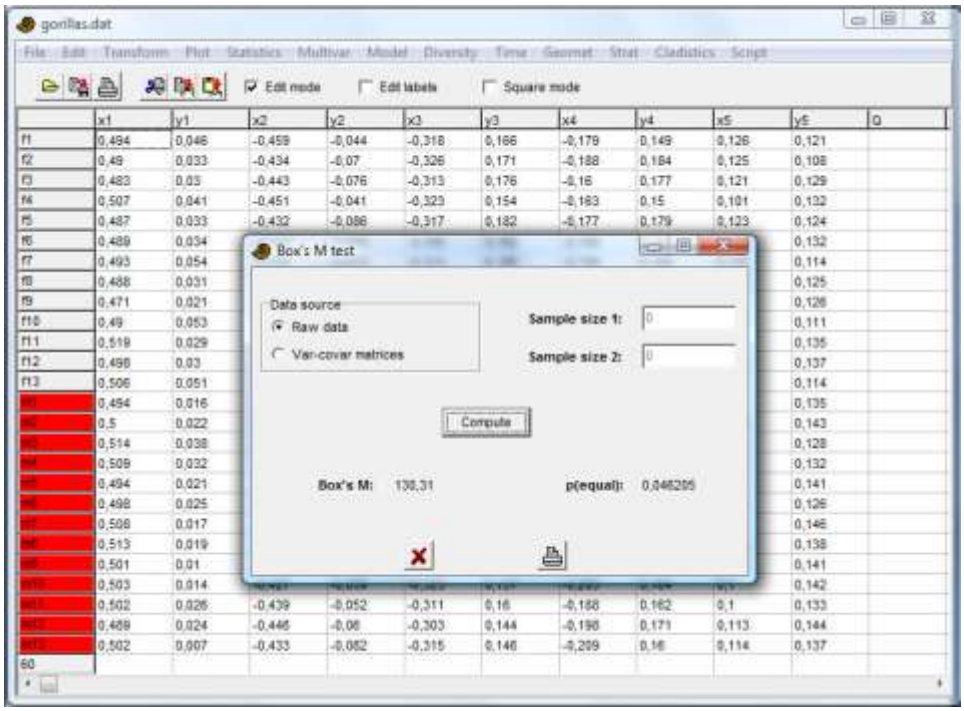
Missing data supported by column average substitution.



	x5	y5	x6	y6	x7	y7	x8	y8	Q
f1	0,125	0,121	0,415	0	0,126	-0,196	-0,204	-0,242	
f2	0,125	0,108	0,408	0,005	0,15	-0,196	-0,224	-0,235	
f3	0,12	0,128	0,414	0,021	0,128	-0,207	-0,228	-0,249	
f4	0,1	0,132	0,415	0,002	0,131	-0,201	-0,215	-0,236	
f5	0,122	0,124	0,412	0,00					
f6	0,104	0,131	0,42	0					
f7	0,104	0,114	0,404	0					
f8	0,104	0,125	0,422	0,00					
m1	0,119	0,135	0,436	-0,0					
m2	0,118	0,143	0,425	0					
m3	0,113	0,127	0,431	0,00					
m4	0,106	0,132	0,417	-0,0					
m5	0,12	0,14	0,436	0					
m6	0,106	0,126	0,436	0,00					
m7	0,121	0,146	0,416	0					
m8	0,105	0,138	0,425	-0,006	0,127	-0,168	-0,207	-0,215	
m9	0,104	0,14	0,445	-0,009	0,118	-0,193	-0,184	-0,195	
m10	0,1	0,142	0,437	0,003	0,115	-0,195	-0,2	-0,226	
60									

Box's *M*

Test for the equivalence of the covariance matrices for two multivariate samples marked with different colors. 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, $\mathbf{S}$ 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.

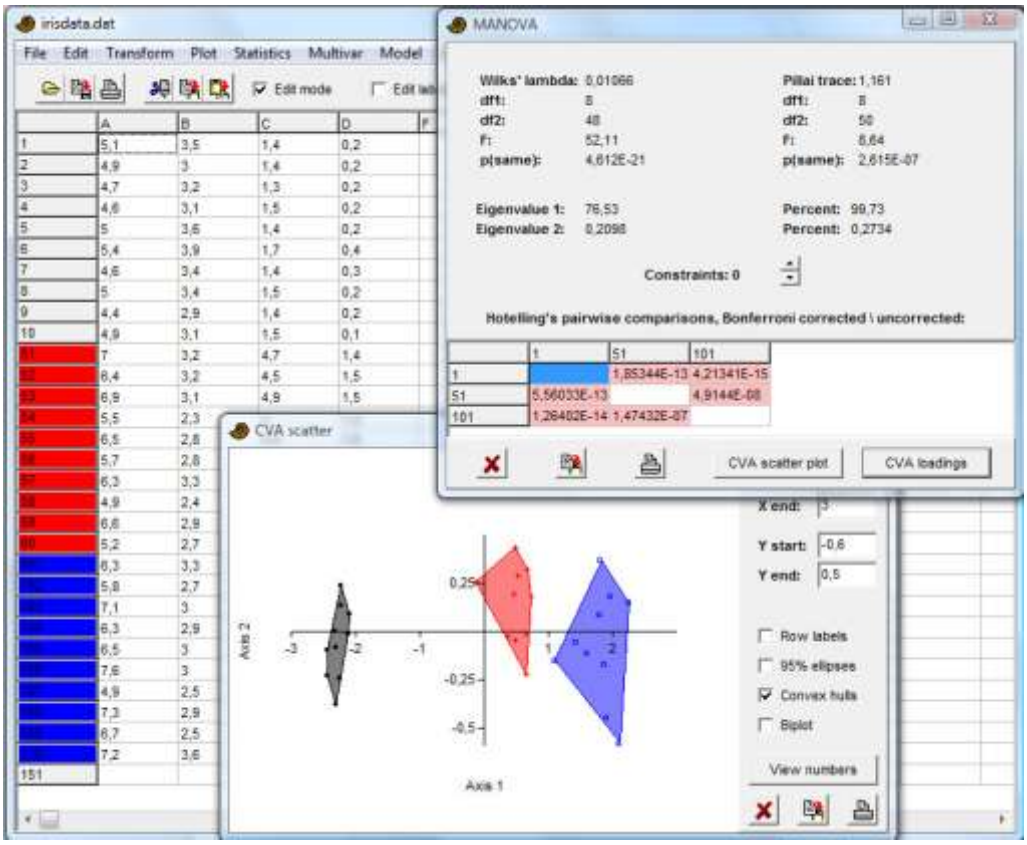
Missing data supported by column average substitution.

Reference

Rencher, A.C. 2002. Methods of multivariate analysis, 2nd ed. Wiley.

MANOVA/CVA

One-way MANOVA (Multivariate ANalysis Of VAriance) is the multivariate version of the univariate ANOVA, testing whether several samples have the same mean. If you have only two samples, you would perhaps rather use the two-sample Hotelling's T^2 test.



Two statistics are provided: Wilk's lambda with it's 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 quite simple, by pairwise Hotelling's tests. In the post-hoc table, groups are named according to the row label of the first item in the group. 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 differ from the values reported in the "Two-group permutation" and the "Discriminant" modules, which pool only the covariance matrices from the two groups being compared.

Missing data supported by column average substitution.

Canonical Variates Analysis

An option under MANOVA, CVA produces a scatter plot of specimens along the two first canonical axes, producing maximal and second to maximal separation between all groups (multigroup discriminant analysis). The axes are linear combinations of the original variables as in PCA, and eigenvalues indicate amount of variation explained by these axes.

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.

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.

Landmark warps

This function should only be used if the CVA analysis has been carried out on 2D landmark data. It allows the interactive plotting of shape deformations as a function of position along a discriminant axis, either as lollipop plots (vectors away from the mean landmark positions) or as thin-plate spline deformations.

EFA warps

This function should only be used if the CVA analysis has been carried out on coefficients computed by the Elliptic Fourier Analysis module. It allows the interactive plotting of outlines as a function of position along a discriminant axis.

CVA 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 $\mathbf{W}_{cov}$:

$$\mathbf{W} = \mathbf{B}_2'\mathbf{B}_2$$

$$\mathbf{W}_{cov} = \frac{1}{n-g}\mathbf{W}.$$

e and **U** are the eigenvalues and eigenvectors of **W**; $\mathbf{e}_c$ and $\mathbf{U}_c$ are the eigenvalues and eigenvectors of $\mathbf{W}_{cov}$. Then,

$$\mathbf{Z}'\mathbf{Z} = \mathrm{diag}(1/\mathbf{e})\mathbf{U}'\mathbf{X}'\mathbf{X}\mathbf{U}\,\mathrm{diag}(1/\mathbf{e}).$$

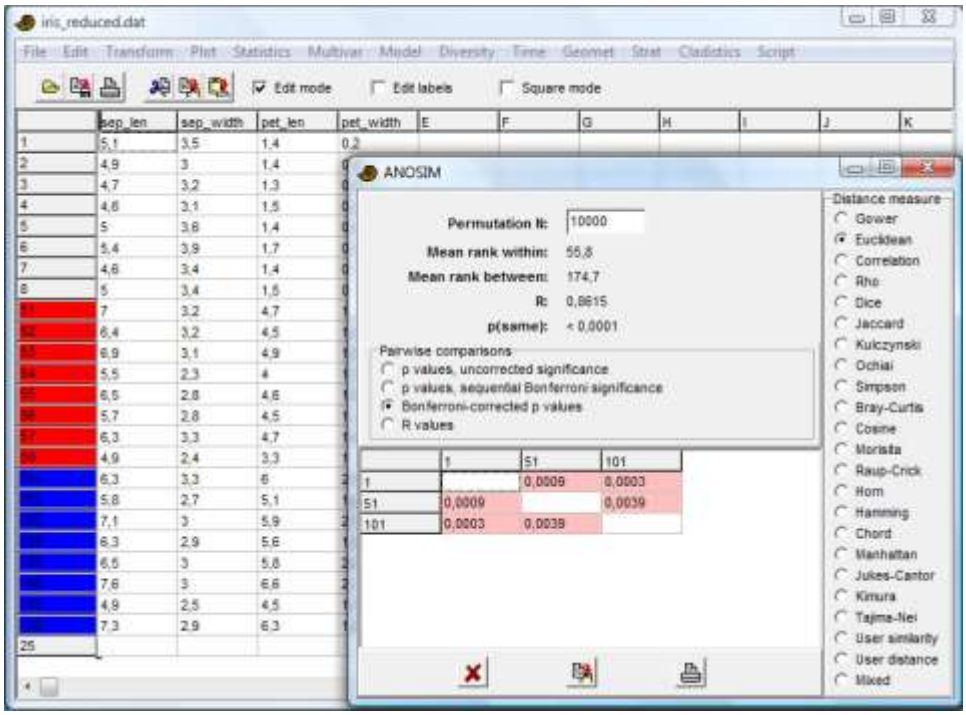
a and **A** are the eigenvalues and eigenvectors of **Z'Z**. We take only the first *g*-1 eigenvectors (columns of **A**), as the rest will be zero. The canonical variates are now

$$\mathbf{C} = \mathbf{U}\,\mathrm{diag}(1/\mathbf{e}_c)\mathbf{A}.$$

The CVA scores are then **BC**. The visualizations of shape derformations are shown along vectors $\mathbf{W}_{cov}\mathbf{C}$.

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 using row colors 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.

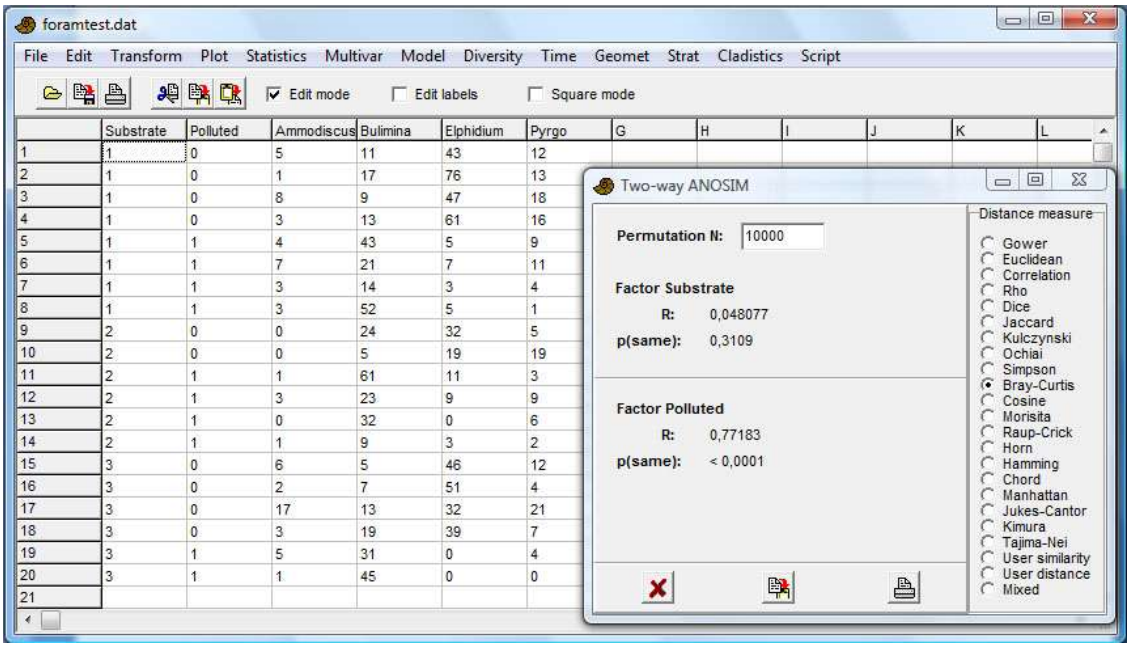
Missing data supported by pairwise deletion (not for the Raup-Crick, Rho and user-defined indices).

Reference

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18:117-143.

Two-way ANOSIM

The two-way ANOSIM in PAST uses the crossed design (Clarke 1993). For more information see one-way ANOSIM, but note that groups (levels) are not coded with colors but with integer numbers in the first two columns.



	Substrate	Polluted	Ammodiscus	Bulimina	Elphidium	Pyrgo
1	1	0	5	11	43	12
2	1	0	1	17	76	13
3	1	0	8	9	47	18
4	1	0	3	13	61	16
5	1	1	4	43	5	9
6	1	1	7	21	7	11
7	1	1	3	14	3	4
8	1	1	3	52	5	1
9	2	0	0	24	32	5
10	2	0	0	5	19	19
11	2	1	1	61	11	3
12	2	1	3	23	9	9
13	2	1	0	32	0	6
14	2	1	1	9	3	2
15	3	0	6	5	46	12
16	3	0	2	7	51	4
17	3	0	17	13	32	21
18	3	0	3	19	39	7
19	3	1	5	31	0	4
20	3	1	1	45	0	0
21						

In the example above, the foraminiferan fauna is significantly different in the polluted and non-polluted samples, while not significantly different between substrates.

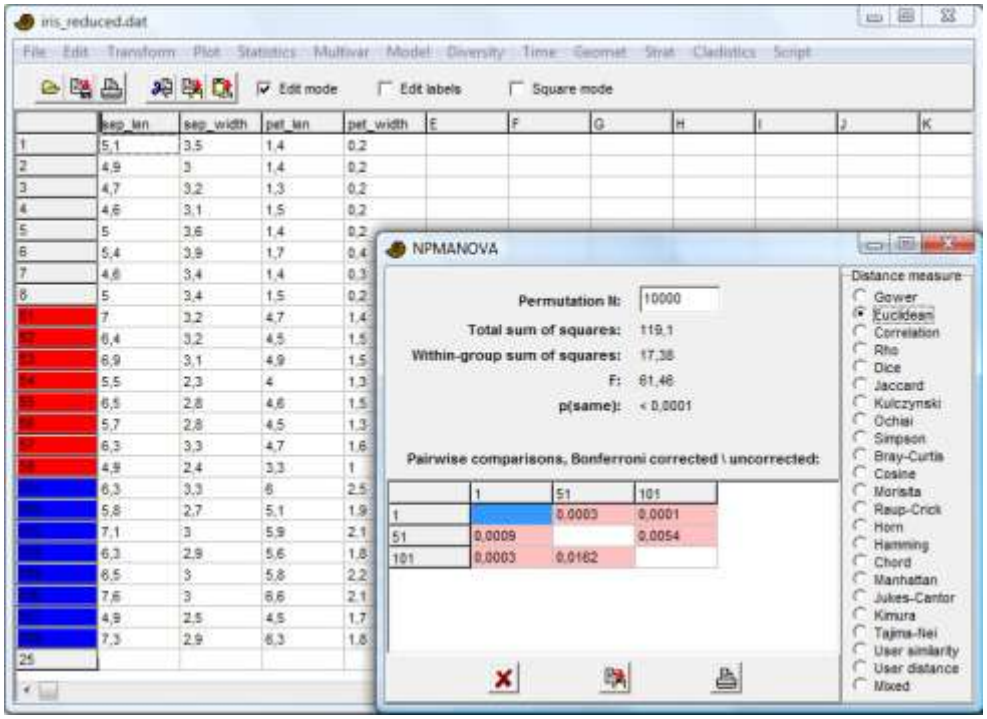
Reference

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18:117-143.

One-way NPMANOVA

NPMANOVA (Non-Parametric MANOVA, also known as PERMANOVA) is a non-parametric test of significant difference between two or more groups, based on any distance measure (Anderson 2001). NPMANOVA 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 using row colors as usual.



NPMANOVA calculates an *F* value in analogy with ANOVA. In fact, for univariate data sets and the Euclidean distance measure, NPMANOVA 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).

Pairwise NPMANOVAs 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 NPMANOVA

The two-way NPMANOVA (Anderson, 2001) in PAST uses the crossed design. The design must be balanced, i.e. each combination of levels must contain the same number of rows. For more information see one-way NPMANOVA, but note that groups (levels) are not coded with colors but with integer numbers in the first two columns (as for two-way ANOSIM).

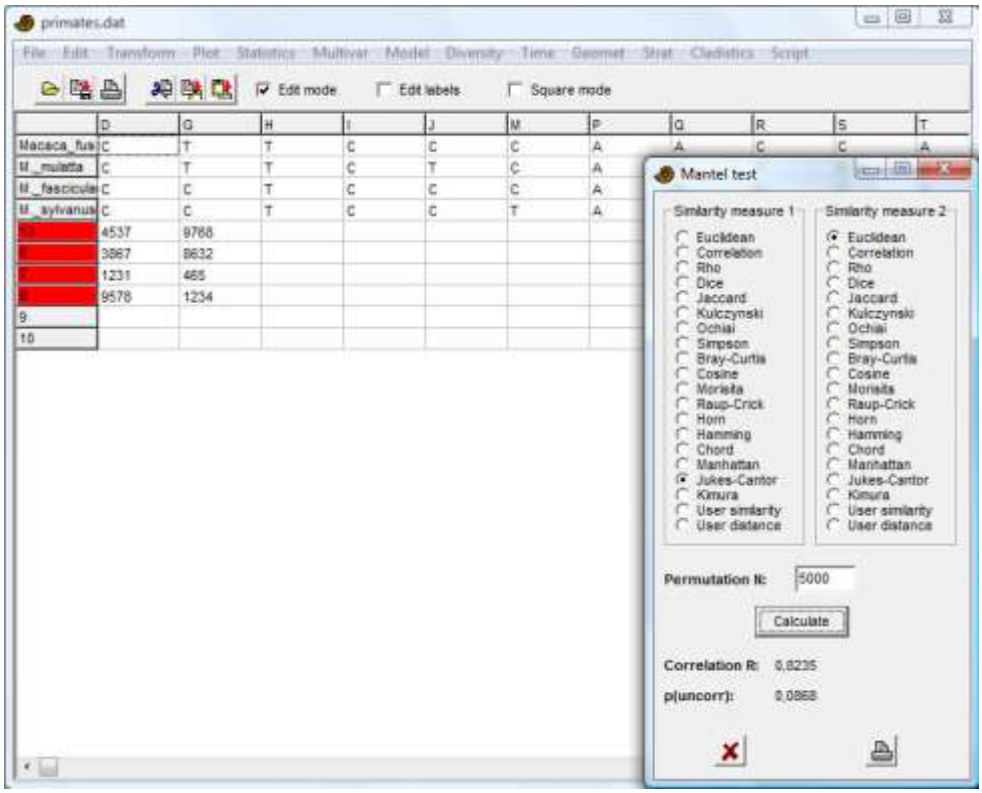
Reference

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32-46.

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 marked with two different colors. 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.

In the example below, the first matrix consists of sequence data from four species of *Macaca*, while the second matrix contains their geographical coordinates. The user has selected the Jukes-Cantor measure for the first matrix, and Euclidean distance for the second. The two data sets seem to be correlated (R=0.82), but significance at p<0.05 is not achieved.



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 needed to correlate the lower triangles). It ranges from -1 to +1. The permutation test compares the original R to R computed in e.g. 5000 random permutations. The reported p value is one-tailed.

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.

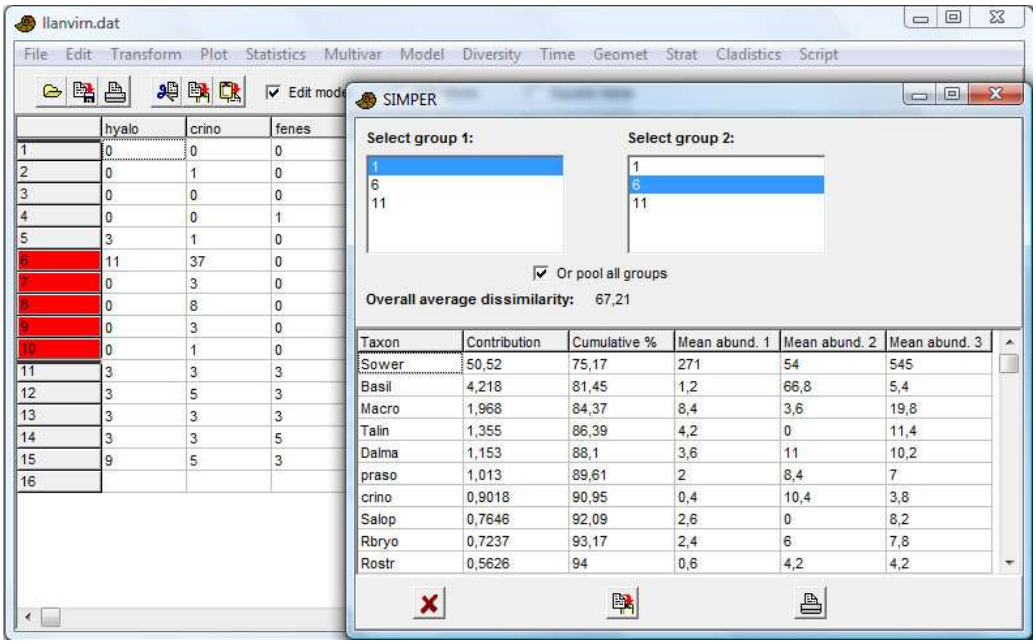
Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209-220.

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.



Samples go in rows, grouped with colors, and taxa in columns. In this example, all three groups (of five samples each) are compared. 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 three 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.

Calibration from CABFAC

This module will reconstruct a (single) environmental parameter from taxa-in-samples abundance data. The program will also ask for a CABFAC transfer function file, as previously made by CABFAC factor analysis. The set of taxa (columns) must be identical in the spreadsheet and the transfer function file.

Calibration from optima

The first three rows can be generated from known (Recent) abundance and environmental data by the "Species packing" option in the Model menu. The third row (peak abundance) is not used, and the second row (tolerance) is used only when the "Equal tolerances" box is not ticked.

The algorithm is weighted averaging, optionally with tolerance weighting, according to ter Braak & van Dam (1989).

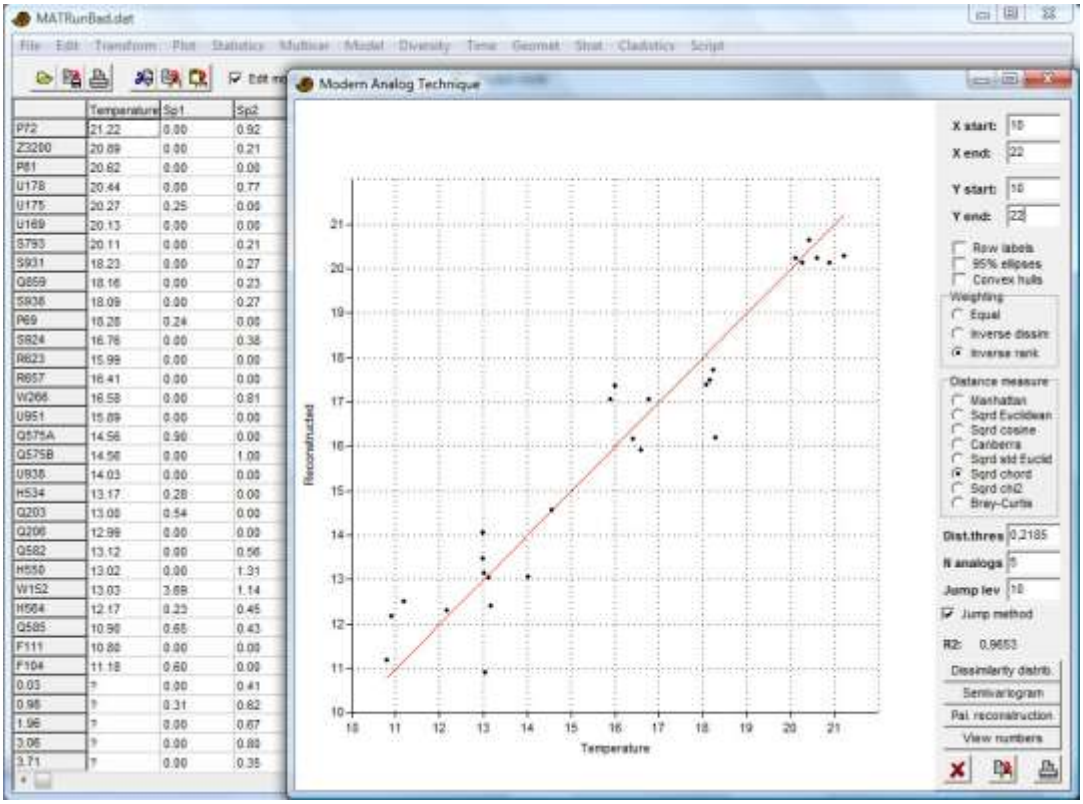
Reference

ter Braak, C.J.F & H. van Dam. 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* 178:209-223.

Modern Analog Technique

The Modern Analog Technique 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). In the example below, the last five visible rows contain the first core samples.



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.

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.

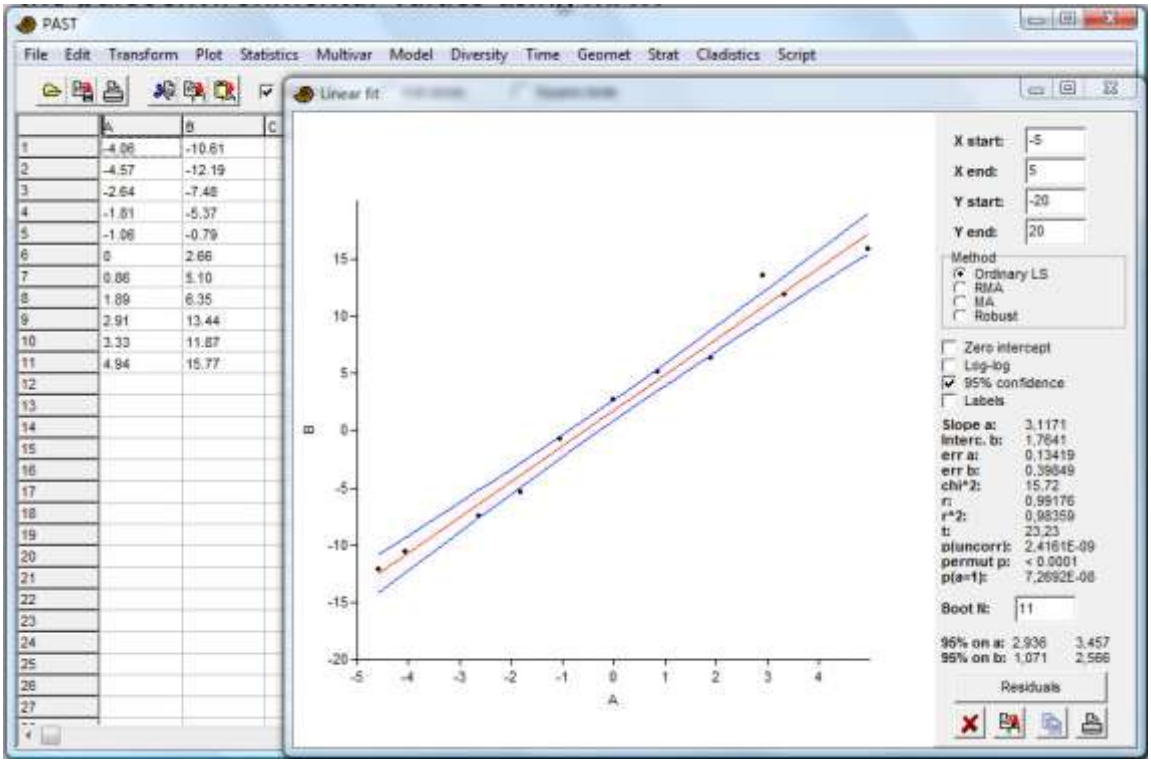
Pal. reconstruction

Reconstruction of the paleoenvironmental values using MAT.

Model menu

Linear

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. There are four different algorithms available: Ordinary Least Squares (OLS), Reduced Major Axis (RMA), Major Axis (MA), and Robust. 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 and standard error estimation is according to Warton *et al*. (2006), *non* Davis (1986)!



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).

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, a Chi-square correlation value (not for RMA/MA), 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 (2000 replicates). The number of random points selected for each replicate should normally be kept as *N*, but may be reduced for special applications.

The permutation test on correlation (r^2) uses 10,000 replicates.

Confidence band

In OLS regression (not RMA/MA/Robust), a 95 percent "Working-Hotelling" confidence band for the fitted line (not for the data points!) 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

$$\mathrm{CI} = ax \pm t_{0.05/2,n-1}\sqrt{SE_{reg}^2\frac{x^2}{\sum x_i^2}}\ .$$

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 four 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.

RMA equations

Slope $a = \text{sign}(r)\sqrt{\frac{\sum(y-\bar{y})^2}{\sum(x-\bar{x})^2}}$.

Standard error on $a = \text{abs}(a)\sqrt{\frac{1-r^2}{n-2}}$.

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

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

Farebrother, R.W. 1980. Pan's procedure for the tail probabilities of the Durbin-Watson statistic. *Applied Statistics* 29:224–227.

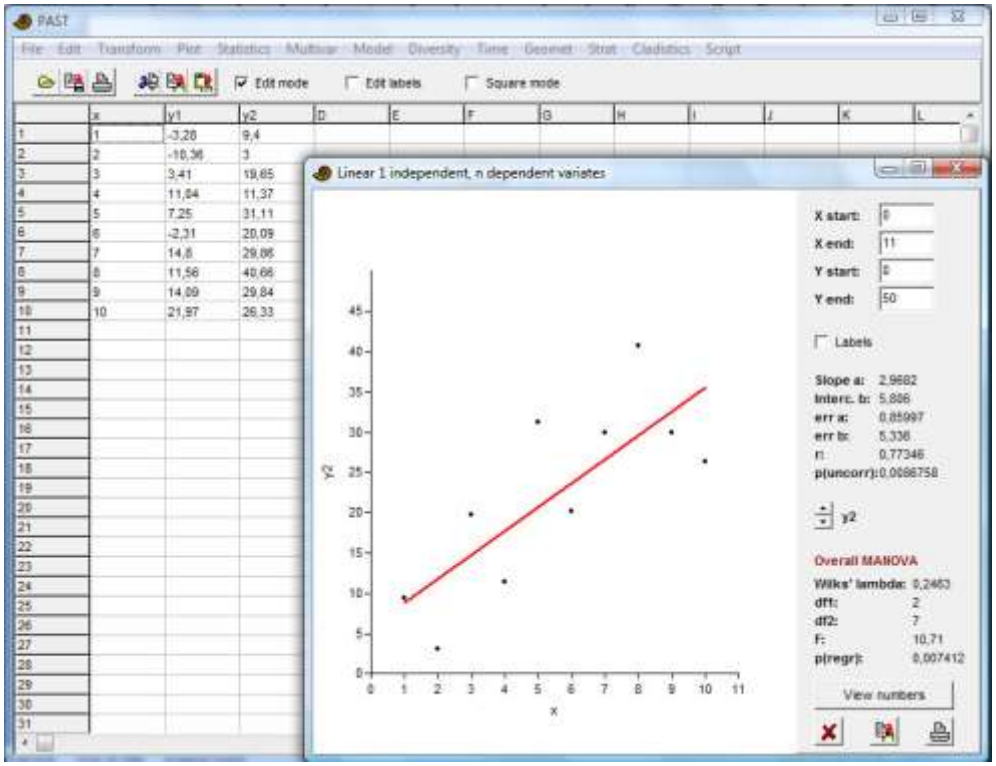
Rousseeuw, P.J. & van Driessen, K. 1999. Computing LTS regression for large data sets. Institute of Mathematical Statistics Bulletin.

Warton, D.I., Wright, I.J., Falster, D.S. & Westoby, M. 2006. Bivariate line-fitting methods for allometry. *Biological Review* 81:259-291.

Linear, one independent, n dependent (multivariate regression)

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).

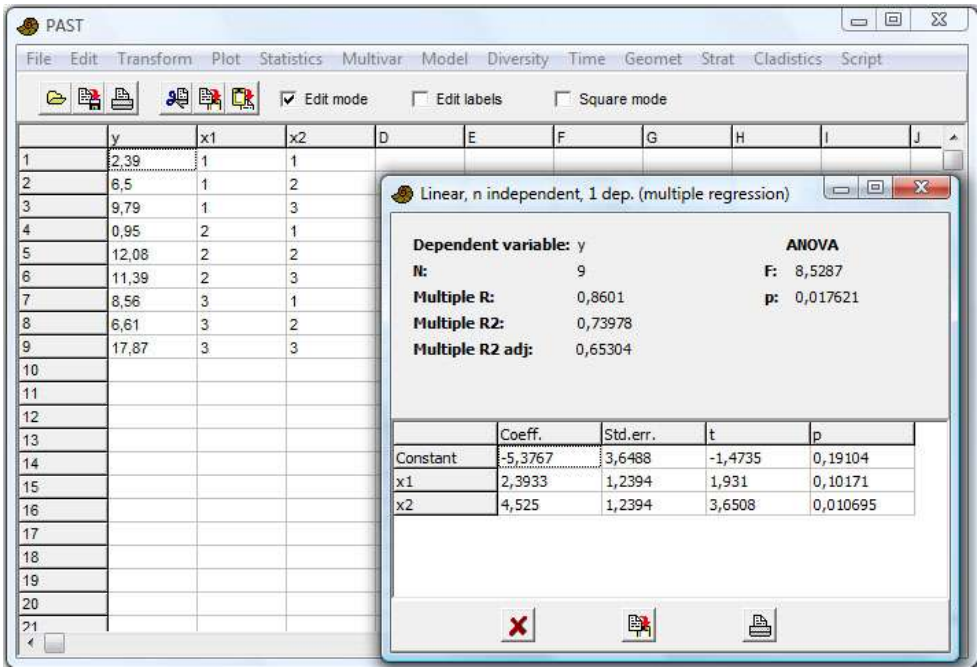
Missing data supported by column average substitution.

Landmark warps and EFA deformation

If the regression has been carried out with Procrustes-fitted landmarks or Elliptic Fourier coefficients as the dependent variables, this window allows the visualization of shapes as a function of the independent variable.

Linear, n independent, one dependent (multiple regression)

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,

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1},$$

$$F = \frac{SSR/k}{SSE/(n-k-1)}.$$

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, n independent, n dependent (multivariate multiple regression)

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:

$$m = n - q - 1 - \frac{1}{2}(p - q + 1)$$

$$\tau = \begin{cases} \sqrt{\dfrac{p^2 q^2 - 4}{p^2 + q^2 - 5}} & \text{if } p^2 + q^2 - 5 > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$F = \frac{1 - \Lambda^{1/\tau}}{\Lambda^{1/\tau}} \cdot \frac{m\tau + 1 - pq/2}{pq}$$

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.

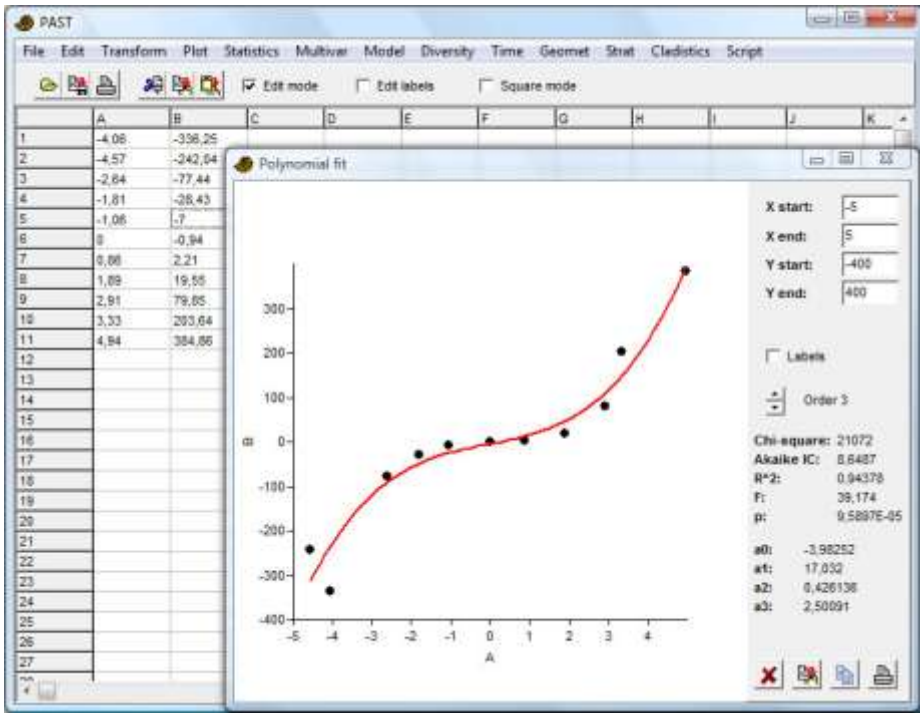
Regression coefficients and statistics

The complete set of coefficients and their significances for all combinations of independent and dependent variables.

Missing data supported by column average substitution.

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.

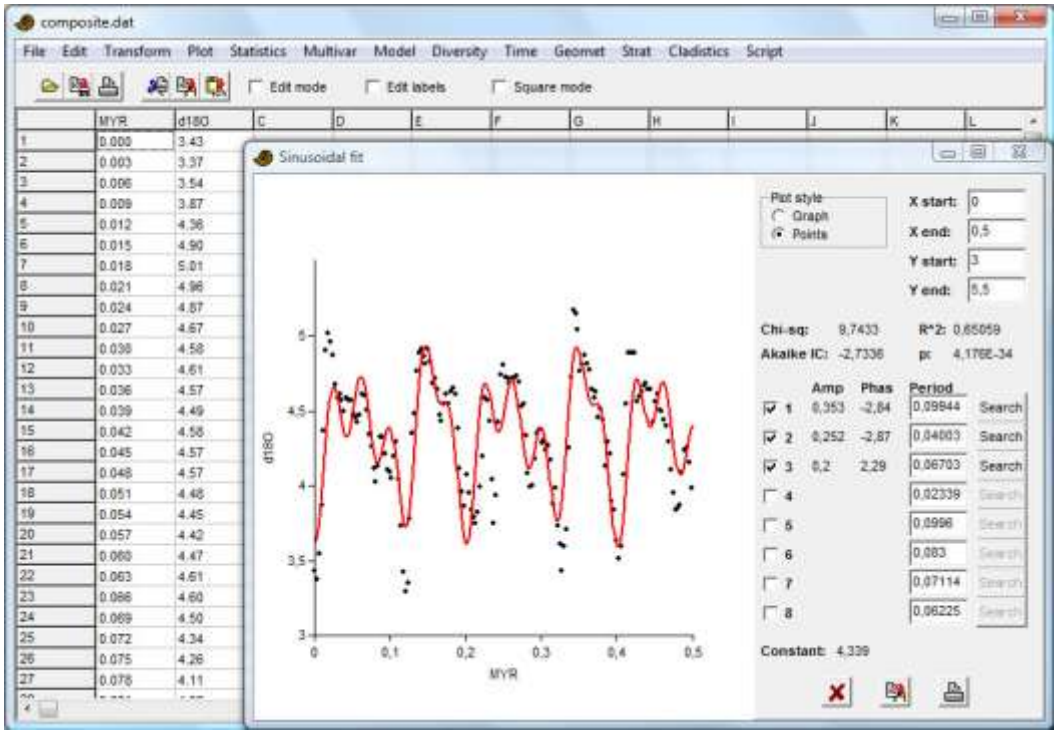
R^2 is the coefficient of determination, or proportion of variance explained by the model. Finally, a *p* value, based on an *F* test, gives the significance of the fit.

Reference

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

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 (Press et al. 1992). 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 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.

A "search" function for each sinusoid will optimize the frequency of that sinusoid (over the full meaningful range from one period to the Nyquist frequency), holding all other selected sinusoid frequencies constant. The algorithm is slow but very robust and almost guaranteed to find the global optimum.

For a "blind" spectral analysis, finding all parameters of an optimal number of sinusoids, follow this procedure: Start with only the first sinusoid selected. Click "search" to optimize period, amplitude and phase. This will find the strongest sinusoid in the data. Note the AIC. Add (select) the second sinusoid, and click its search button to optimize all parameters of both sinusoids except the period of the first sinusoid. This will find the second strongest sinusoid. Continue until the AIC no longer decreases.

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.

There are also options to enforce a pure sine or cosine series, i.e. with fixed phases.

Reference

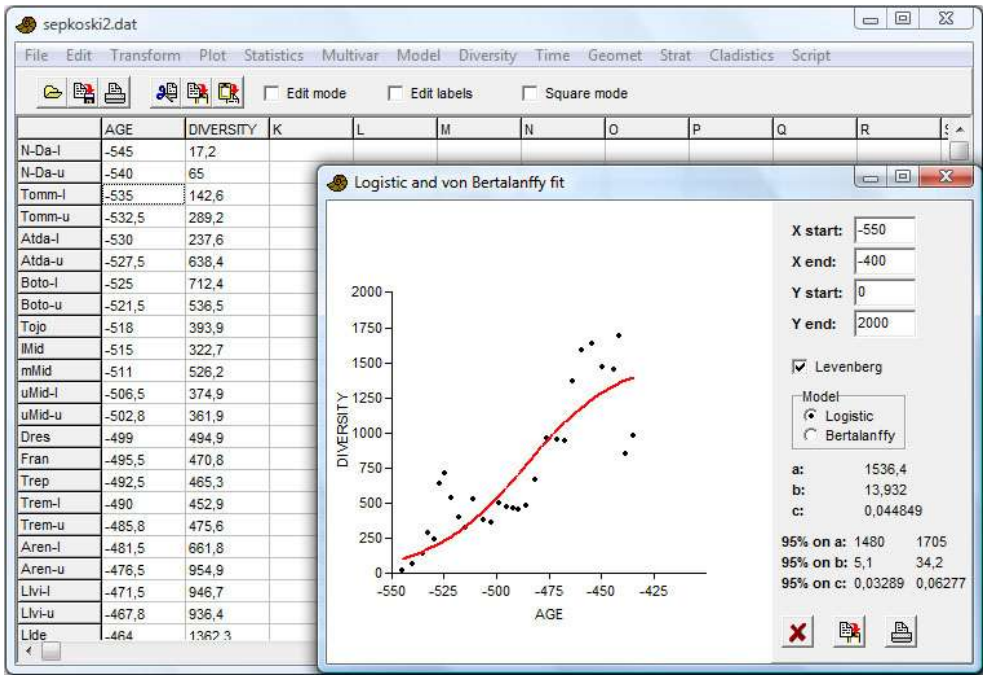
Press, W.H., S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. 1992. Numerical Recipes in C. Cambridge University Press.

Logistic/Bertalanffy/Michaelis-Menten/Gompertz

Attempts to fit two columns of x-y data to one of three “saturation” models.

The *logistic* equation is $y=a/(1+be^{-cx})$. The value of *a* is first estimated by the maximal value of *y*. The values of *b* and *c* are then estimated using a straight-line fit to a linearized model.

The fit can be improved by using the estimated values as an initial guess for a Levenberg-Marquardt optimization (Press et al. 1992). Due to numerical instability this can fail with an error message, especially during bootstrapping and for the Gompertz curve.



The 95% confidence intervals are based on 2000 bootstrap replicates.

The *von Bertalanffy* option uses the same algorithm as above, but fits to the equation $y=a(1-be^{-cx})$. This equation is used for modelling growth of multi-celled animals (in units of length or width, not volume).

The *Michaelis-Menten* option fits to the equation $y=ax/(b+x)$. The algorithm uses maximum-likelihood estimators for the so-called Eadie-Hofstee transformation (Raaijmakers 1987; Colwell & Coddington 1994). This estimate usually improves when using the Levenberg optimization.

The *Gompertz* option fits to the equation $y=a*\exp(b*\exp(cx))$. The initial estimate is computed using regression on a linearized model .

The logistic equation can model growth with saturation, and was used by Sepkoski (1984) to describe the proposed stabilization of marine diversity in the late Palaeozoic. The logistic and the von Bertalanffy growth models are described by Brown & Rothery (1993). 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).

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.

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.

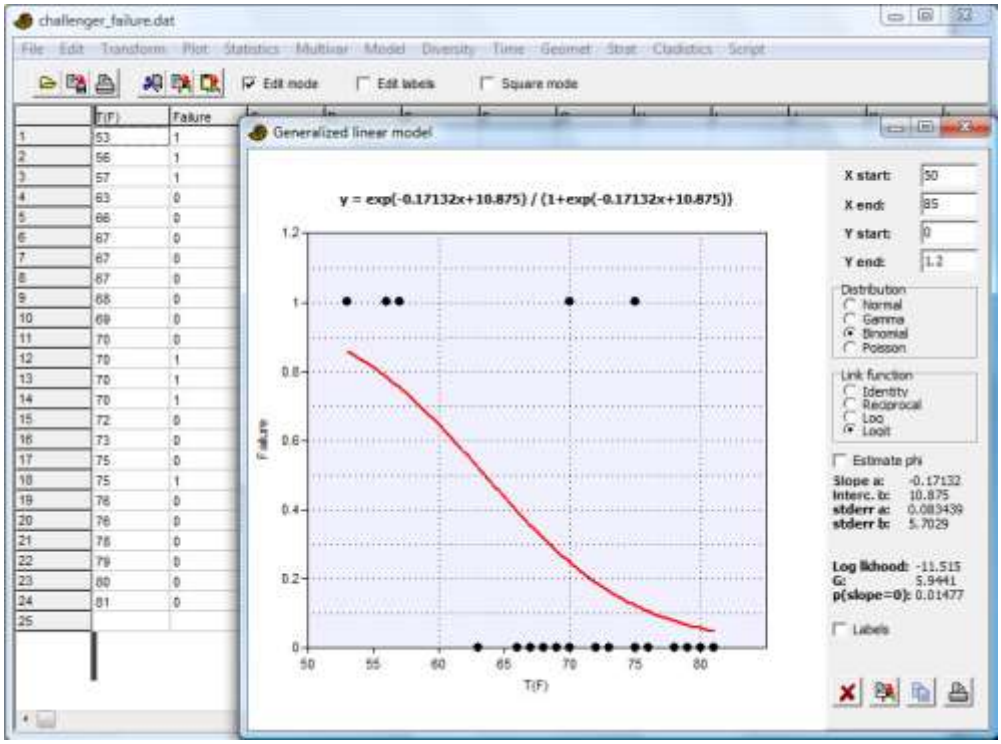
Press, W.H., S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. 1992. Numerical Recipes in C. Cambridge University Press.

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.

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).



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 $y=\exp(ax+b)$.

Binomial (Bernoulli) distribution and the logit link: Logistic regression for a binary response variable (see figure above).

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 $\varphi = 1$, unless the "Estimate phi" option is selected, in which case it is estimated using Pearson's chi-square. Typically, φ is assumed to be 1 for the Poisson and binomial 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: $D = 2\sum_i \left[-\ln \frac{y_i}{\mu_i} + \frac{y_i - \mu_i}{\mu_i} \right]$

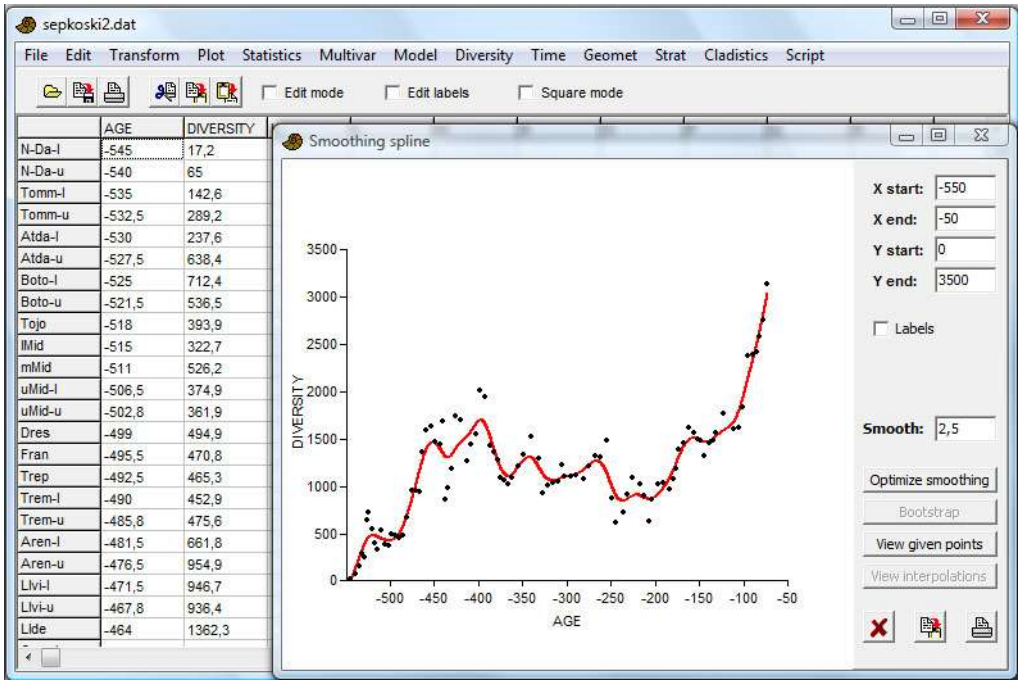
Bernoulli: $D = 2\sum_i \left[y_i \ln \frac{y_i}{\mu_i} + (1 - y_i) \ln \frac{1 - y_i}{1 - \mu_i} \right]$ (the first term defined as zero if y_i=0)

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.

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.



	AGE	DIVERSITY
N-Da-l	-545	17,2
N-Da-u	-540	65
Tomm-l	-535	142,6
Tomm-u	-532,5	289,2
Atda-l	-530	237,6
Atda-u	-527,5	638,4
Boto-l	-525	712,4
Boto-u	-521,5	536,5
Tojo	-518	393,9
lMid	-515	322,7
mMid	-511	526,2
uMid-l	-506,5	374,9
uMid-u	-502,8	361,9
Dres	-499	494,9
Fran	-495,5	470,8
Trep	-492,5	465,3
Trem-l	-490	452,9
Trem-u	-485,8	475,6
Aren-l	-481,5	661,8
Aren-u	-476,5	954,9
Llvi-l	-471,5	946,7
Llvi-u	-467,8	936,4
Llde	-464	1362,3

An optional third columns 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 crossvalidation 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.

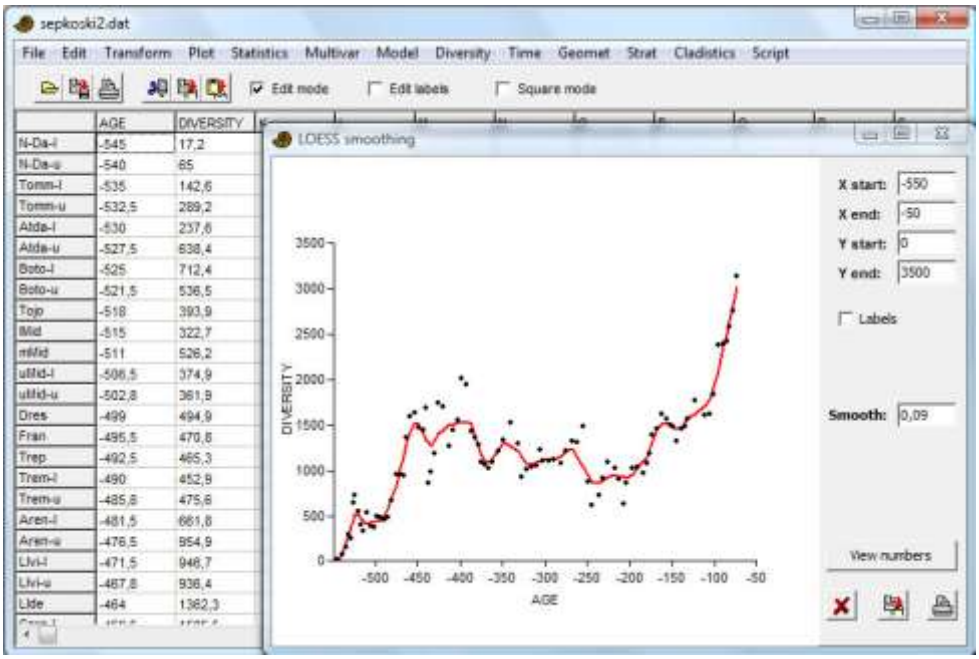
Reference

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.



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

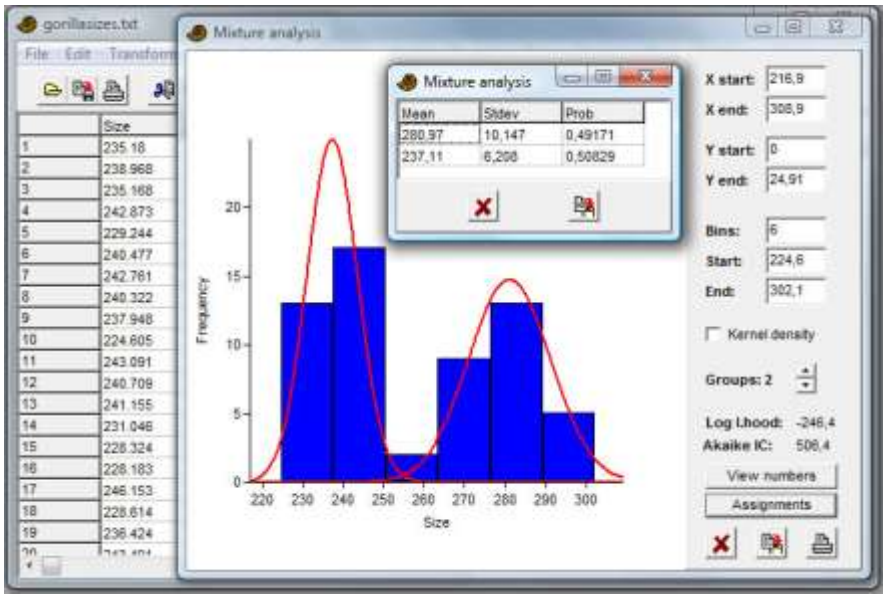
Cleveland, W.S. 1979. Robust locally weighted fitting and smoothing scatterplots. *Journal of the American Statistical Association* 74:829-836.

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 univariate sample. The program can also estimate mean and proportion of exponential and Poisson 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 male and female gorillas have been pooled in one sample. The means, standard deviations and proportions of the two original samples have been almost perfectly recovered (see “Univariate” above).



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

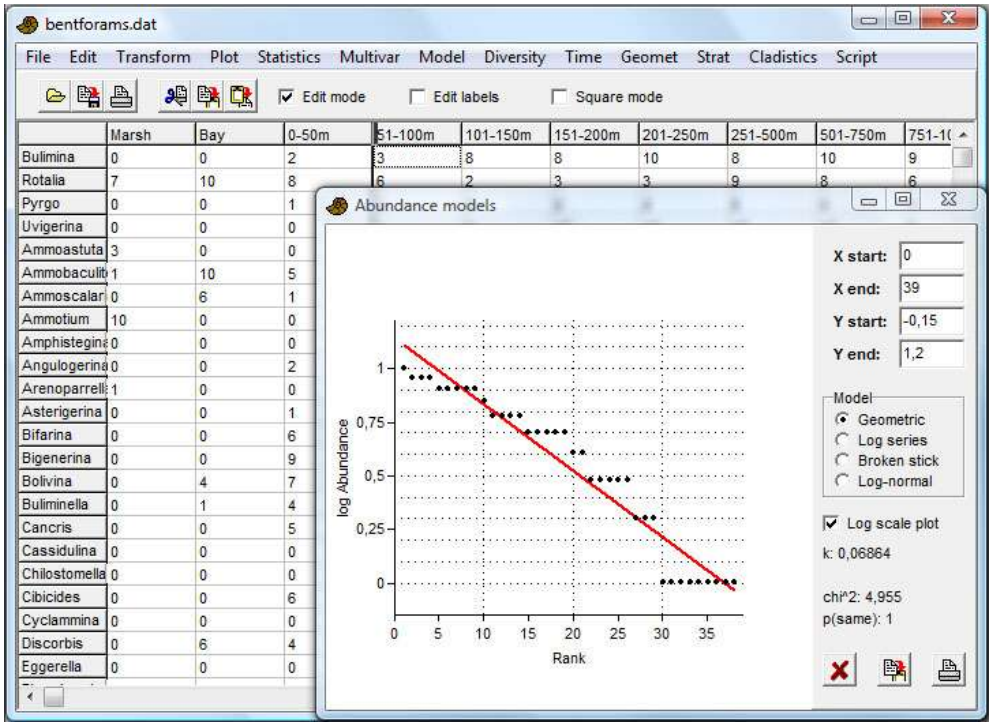
Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716-723.

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. 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 ith 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):

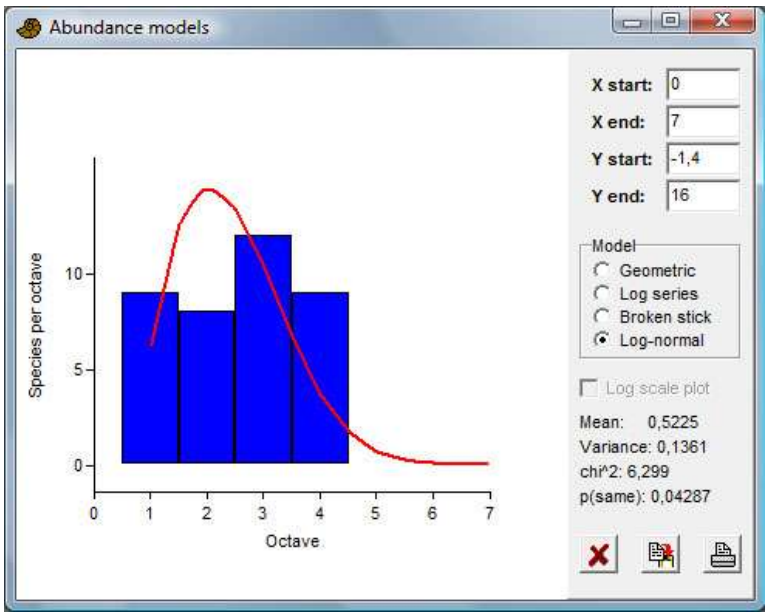
$$S_n = \frac{\alpha x^n}{n}$$

- 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.

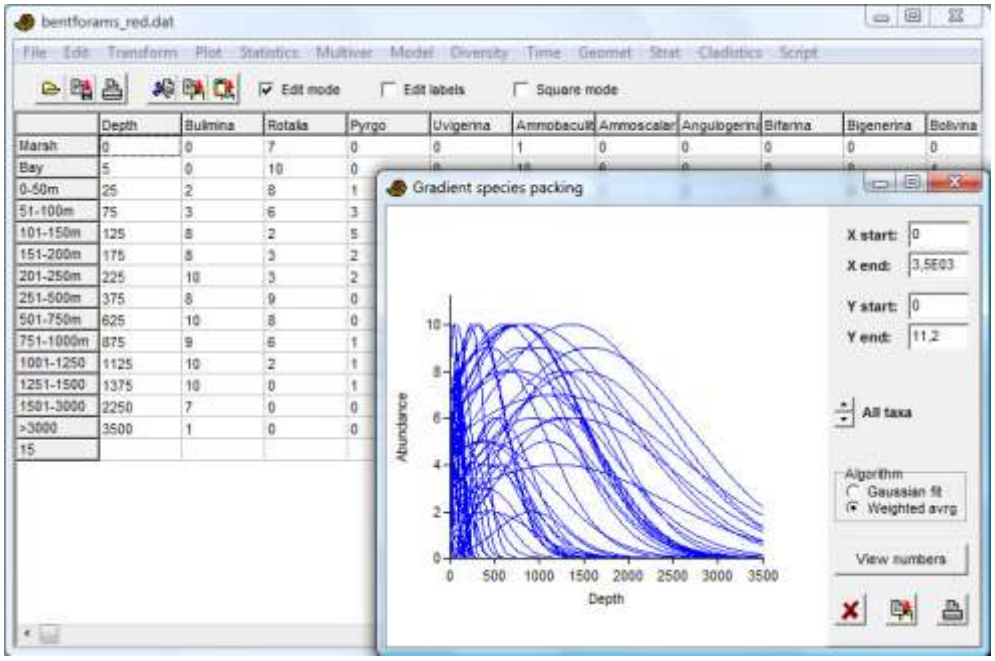
MacArthur, R.H. 1957. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences, USA* 43:293-295.

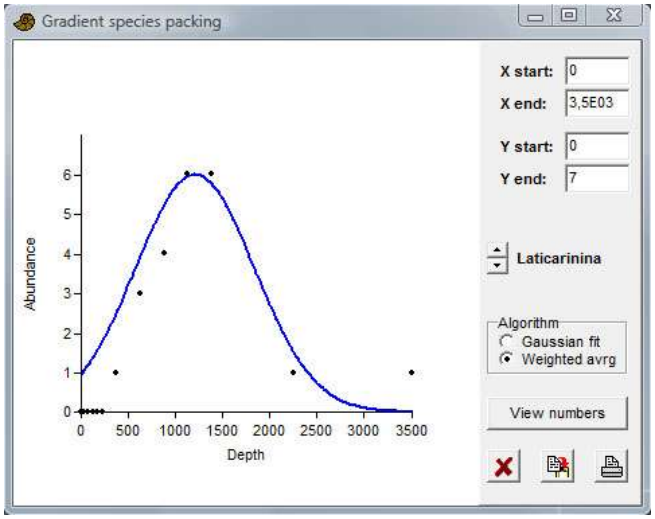
Species packing (Gaussian)

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.

One column of environmental measurements in samples (e.g. temperature), and one or more columns of abundance data (taxa in columns).

The algorithm is based on weighted averaging according to ter Braak & van Dam (1989).



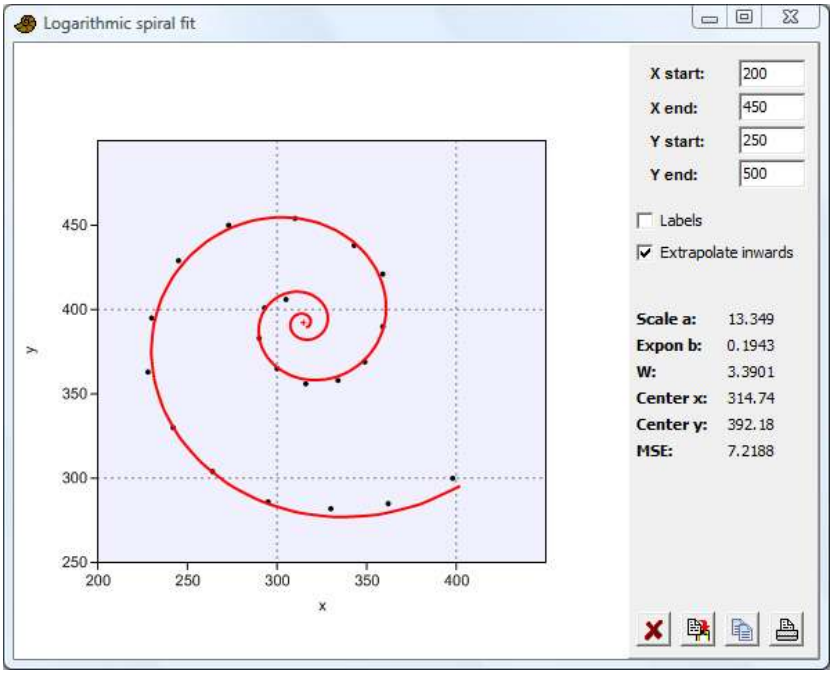


Reference

ter Braak, C.J.F & H. van Dam. 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* 178:209-223.

Logarithmic spiral

Fits a set of points in the plane to a logarithmic 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 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.

Diversity menu

Diversity indices

These statistics apply to association data, where number of individuals are tabulated in rows (taxa) and possibly several columns (associations). The available statistics are as follows, for each association:

- 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*.
- 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.

  $H = -\sum_i \frac{n_i}{n} \ln \frac{n_i}{n}$
- 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: $(S-1) / \ln(n)$
- Equitability. Shannon diversity divided by the logarithm of number of taxa. This measures the evenness with which individuals are divided among the taxa present.
- 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*.

Many of these indices are explained in Harper (1999).

Approximate confidence intervals for all these indices can be computed with a bootstrap procedure. 1000 random samples are produced (200 prior to version 0.87b), each with the same total number of individuals as in each original sample. The random samples are taken from the total, pooled data set (all columns). For each individual in the random sample, the taxon is chosen with probabilities according to the original, pooled 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.

Since these confidence intervals are all computed with respect to the pooled data set, they do not represent confidence intervals for the individual samples. They are mainly useful for identifying samples where the given diversity index falls outside the confidence interval. Bootstrapped comparison of diversity indices in two samples is provided in the Compare diversities module.

Reference

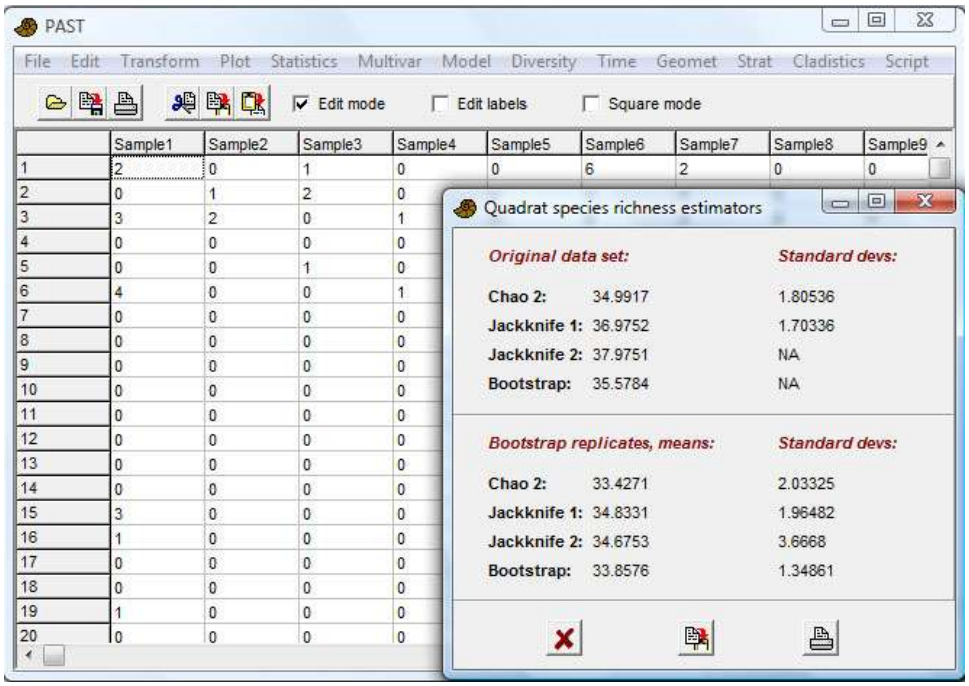
Harper, D.A.T. (ed.). 1999. Numerical Palaeobiology. John Wiley & Sons.

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

$$\hat{\text{var}}\left(\hat{S}_{Chao2}\right) = \left(\frac{m-1}{m}\right)\frac{Q_1(Q_1-1)}{2(Q_2+1)} + \left(\frac{m-1}{m}\right)^2\frac{Q_1(2Q_1-1)^2}{4(Q_2+1)^2} + \left(\frac{m-1}{m}\right)^2\frac{Q_1^2Q_2(Q_1-1)^2}{4(Q_2+1)^4}.$$

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):

$$\hat{S}_{jack1} = S_{obs} + \left(\frac{m-1}{m}\right)Q_1.$$

$$\hat{\text{var}}\left(\hat{S}_{jack1}\right) = \left(\frac{m-1}{m}\right)\left(\sum_{j=0}^{S} j^2 f_j - \frac{{Q_1}^2}{m}\right),$$

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 distinctness

One or more columns, each containing counts of individuals of different taxa down the rows. In addition, the leftmost row(s) must contain names of genera/families etc. (see below).

Taxonomic diversity and taxonomic distinctness as defined by Clarke & Warwick (1998), including confidence intervals computed from 200 random replicates taken from the pooled data set (all columns). 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, fixed column), genus names in column 1, family in column 2 etc. (of course you can substitute for other taxonomic levels as long as they are in ascending order). Species counts follow in the columns thereafter. The program will ask for the number of columns containing taxonomic information above the species level.

For presence-absence data, taxonomic diversity and distinctness will be valid but equal to each other.

Taxonomic distinctness 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:

$$\Delta^* = \frac{\sum\sum_{i<j} w_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j}.$$

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 your *largest* 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 *Ni* 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 variances) are given by the program. In the graphical plot, these standard errors are converted to 95 percent confidence intervals.

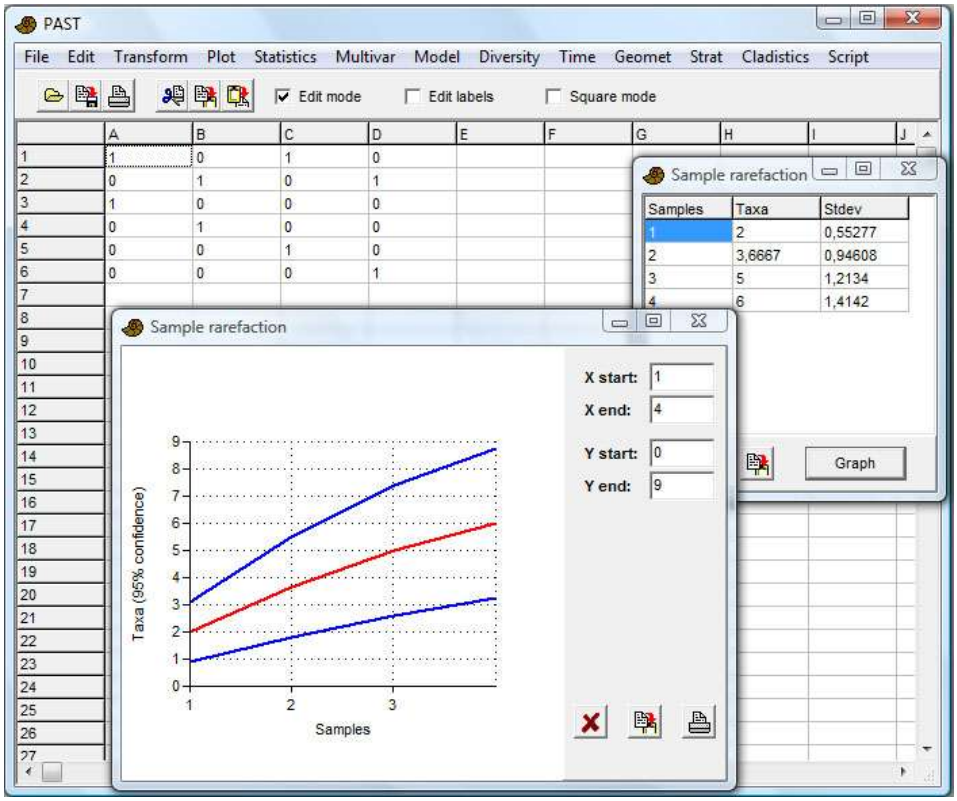
References

Adrain, J.M., S.R. Westrop & D.E. Chatterton. 2000. Silurian trilobite alpha diversity and the end-Ordovician mass extinction. *Paleobiology* 26:625-646.

Krebs, C.J. 1989. Ecological Methodology. Harper & Row, New York.

Sample rarefaction (Mao 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 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

$$\tilde{\tau}(h) = S_{obs} - \sum_{j=1}^{H} \alpha_{jh} s_j \quad .$$

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,

$$\tilde{S} = S_{obs} + \frac{(H-1)s_1(s_1-1)}{2H(s_2+1)}.$$

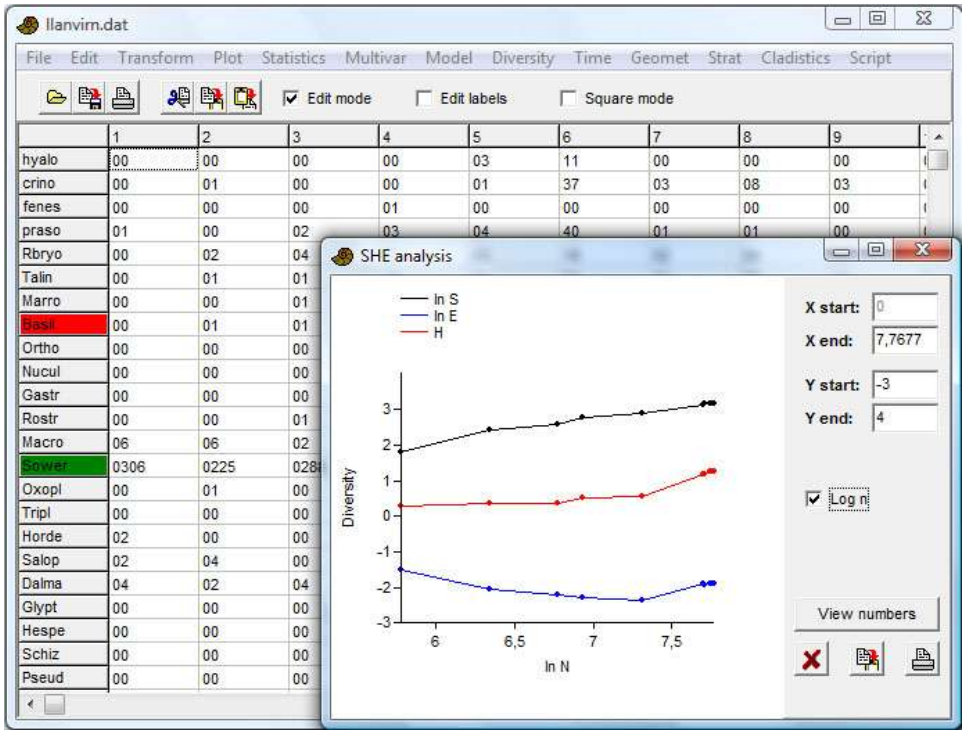
For modeling and extrapolating the curve using the Michaelis-Menten equation, use the Copy Data button, paste to a new Past spreadsheet, and use the fitting module in the Model menu.

Reference

Colwell, R.K., C.X. Mao & J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* 85:2717-2727.

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.

Compare diversities

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 two different randomisation procedures as follows.

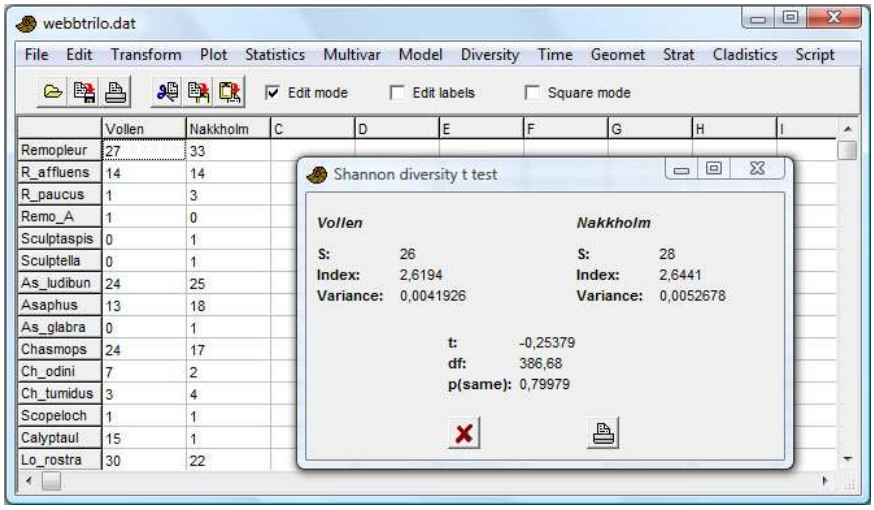
Bootstrapping
The two samples A and B are pooled. 1000 random pairs of samples (A_i,B_i) are then taken from this pool, with the same numbers of individuals as in the original two samples. For each replicate pair, the diversity indices div(A_i) and div(B_i) are computed. The number of times |div(A_i)-div(B_i)| exceeds or equals |div(A)-div(B)| indicates the probability that the observed difference could have occurred by random sampling from one parent population as estimated by the pooled sample.

A small probability value *p(same)* then indicates a significant difference in diversity index between the two samples.

Permutation
1000 random matrices with two columns (samples) are generated, each with the same row and column totals as in the original data matrix. The *p* value is computed as for the bootstrap test.

Diversity *t* test

Comparison of the Shannon diversities in two samples, using a *t* test described by e.g. Hutcheson (1970), Poole (1974), Magurran (1988). This is an alternative to the randomization test available in the *Compare diversities* module. Requires two columns of abundance data with taxa down the rows.



	Vollen	Nakkholm
Remopleur	27	33
R_affluens	14	14
R_paucus	1	3
Remo_A	1	0
Sculptaspis	0	1
Sculptella	0	1
As_ludibun	24	25
Asaphus	13	18
As_glabra	0	1
Chasmops	24	17
Ch_odini	7	2
Ch_tumidus	3	4
Scopeloch	1	1
Calyptaul	15	1
Lo_rostra	30	22

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

$$\mathrm{Var}\ H' = \frac{\sum p_i (\ln p_i)^2 - \left[\sum (p_i \ln p_i)\right]^2}{N} + \frac{S-1}{2N^2}.$$

The *t* test statistic is given by

$$t = \frac{H_1' - H_2'}{\sqrt{\mathrm{Var}\ H_1' + \mathrm{Var}\ H_2'}}.$$

The degrees of freedom for the *t* test is

$$df = \frac{\left(\mathrm{Var}\ H_1' + \mathrm{Var}\ H_2'\right)^2}{\frac{\left(\mathrm{Var}\ H_1'\right)^2}{N_1} + \frac{\left(\mathrm{Var}\ H_2'\right)^2}{N_2}}.$$

References

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

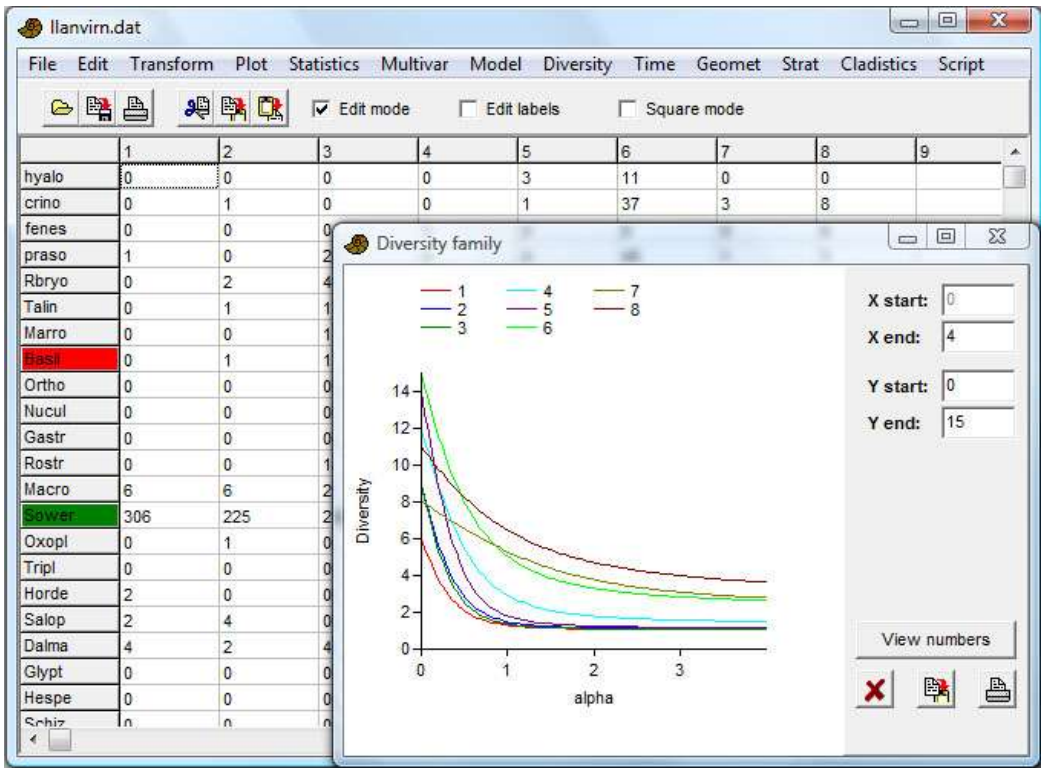
This module requires one or more columns of abundance data with taxa down the rows. The main purpose is to compare diversities in several samples.

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.



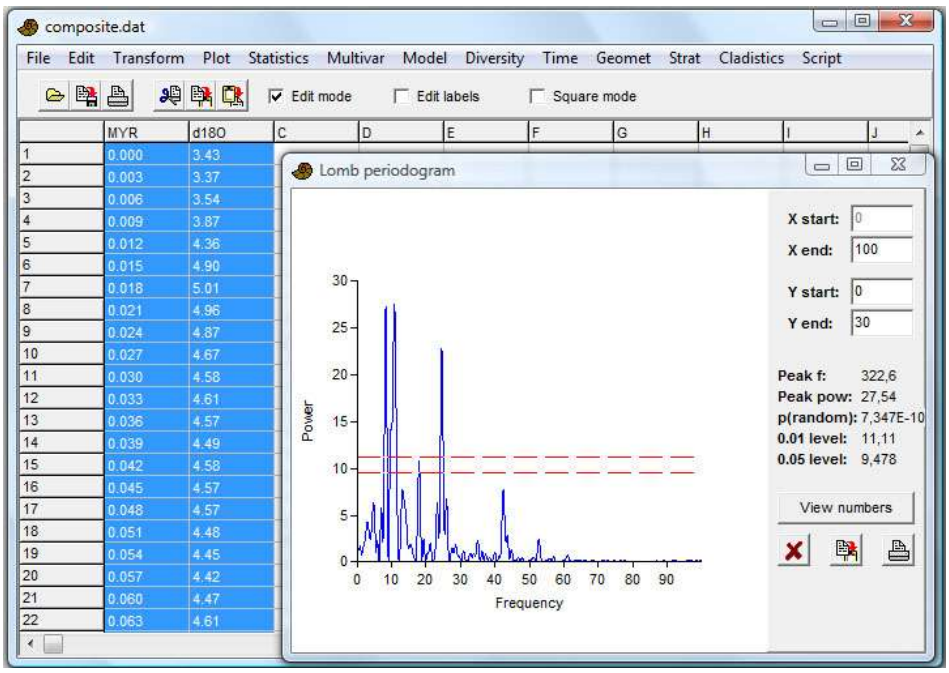
Reference

Tothmeresz, B. 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science* 6:283-290.

Time series menu

Spectral analysis

Since palaeontological data are often unevenly sampled, Fourier-based methods can be difficult to use. PAST therefore includes the Lomb periodogram algorithm 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, an 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.

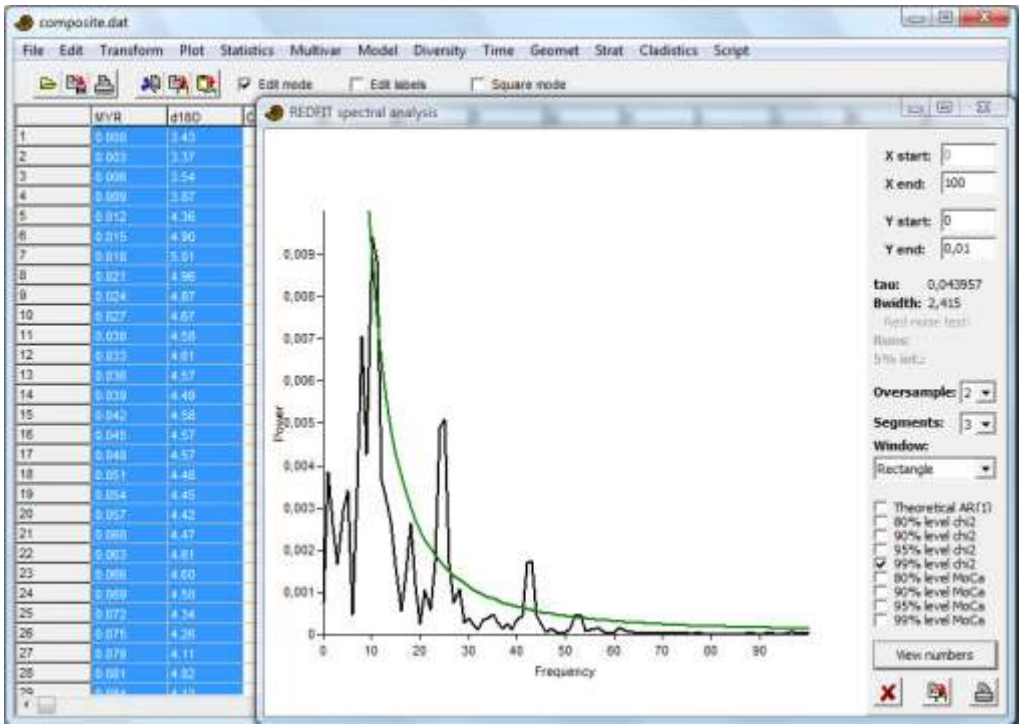
Reference

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).

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) in order to reduce spectral leakage. In the multitaper method, several different (orthogonal) window functions are applied, and the results combined. The resulting spectrum has low leakage, low variance, and retains information contained in the beginning and end of the time series. In addition, statistical testing can take advantage of the multiple spectral estimates. One possible disadvantage is reduced spectral resolution.

The multitaper method requires evenly spaced data, given in one column.

The implementation in Past is based on the code of Lees & Park (1995). The multitaper spectrum can be compared with a simple periodogram (FFT with a 10% cosine window) and a smoothed periodogram. The number of tapers (NWIN) can be set to 3, 4 or 5, for different tradeoffs between variance reduction and resolution. The “time-bandwidth product” *p* is fixed at 3.0.

The *F* test for significance of periodicity follows Lees & Park (1995). The 0.05 and 0.01 significance levels are shown as horizontal lines, based on 2 and 2*NWIN-2 degrees of freedom.

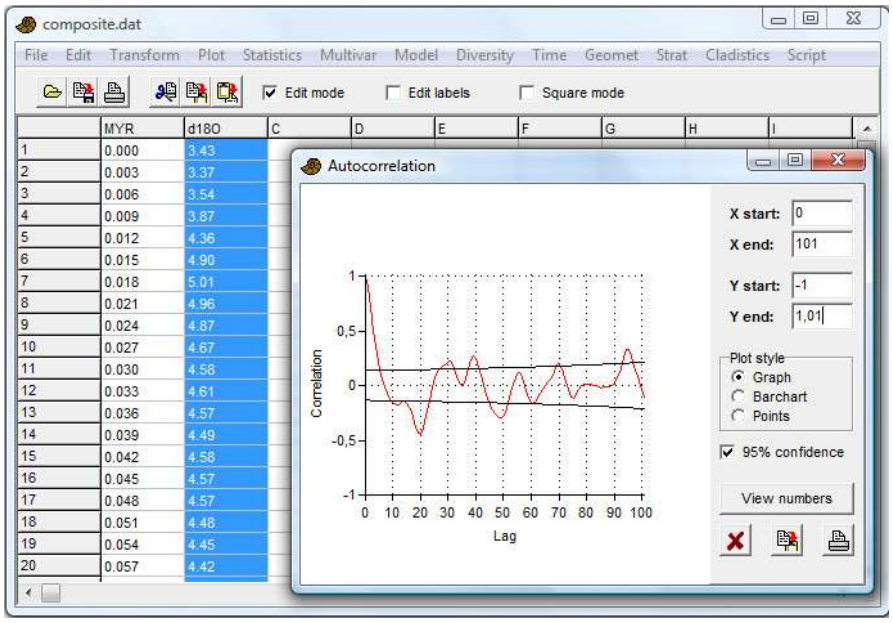
The data are zero-padded to the second lowest power of 2 above the length of the sequence. This is required to reproduce the test results given by Lees & Park (1995).

Reference

Lees, J.M. & J. Park. 1995. Multiple-taper spectral analysis: a stand-alone C-subroutine. *Computers & Geosciences* 21:199-236.

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).

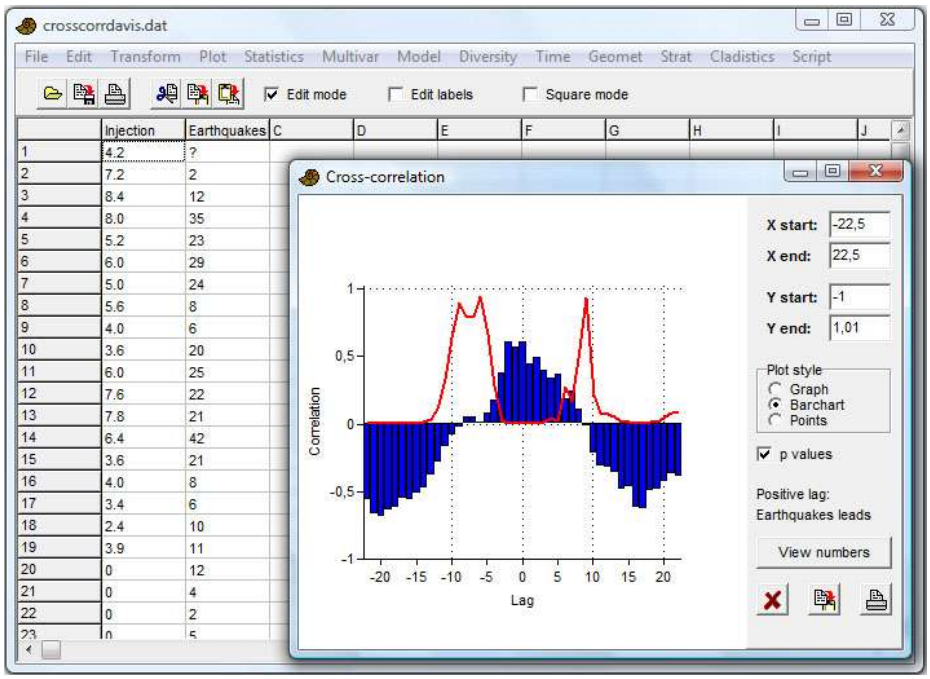
Missing data supported.

Reference

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

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

$$r_m = \frac{\sum (x_i - \bar{x}) \sum (y_{i-m} - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_{i-m} - \bar{y})}} .$$

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.

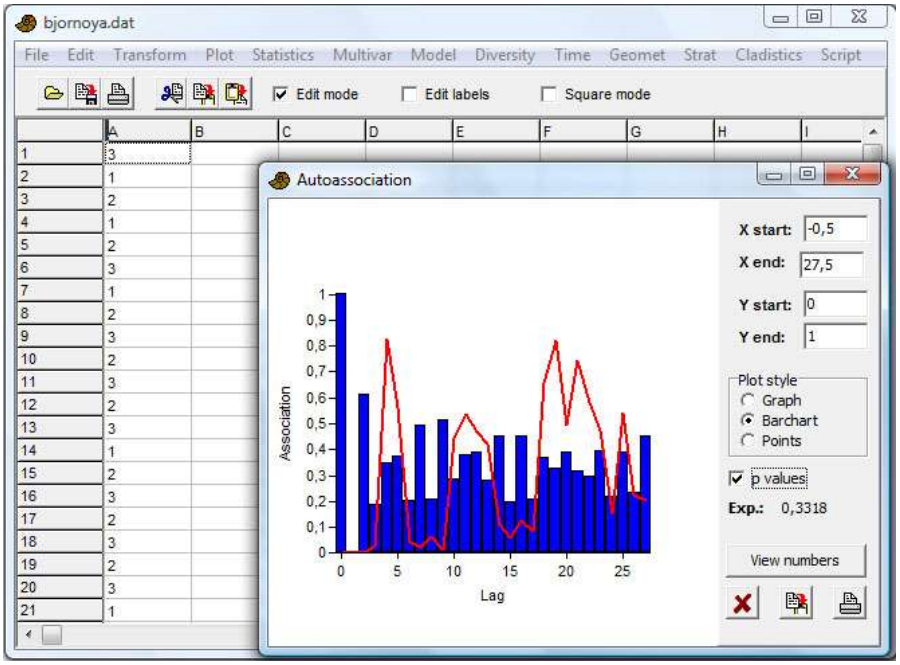
Missing data supported.

Reference

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

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 (0.3318 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 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. This test will take a long time to run for $n>30$, and the option is not available for $n>40$.

Missing data supported.

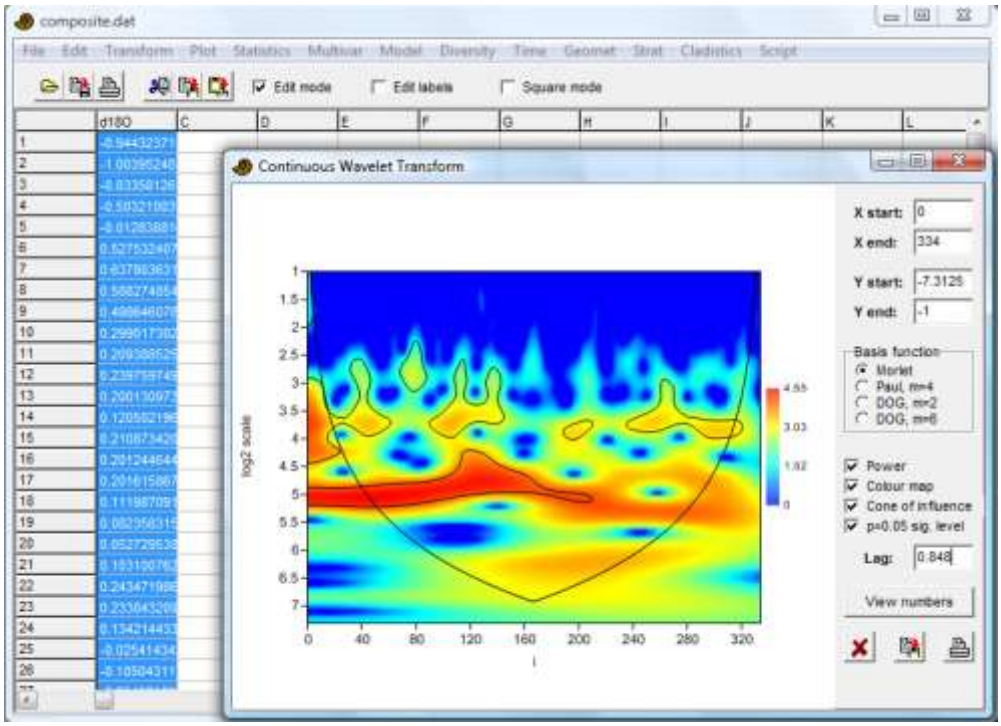
Reference

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

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.

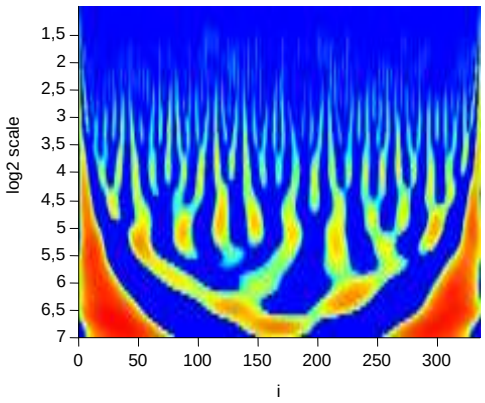
The example above is based on a foram oxygen isotope record from 1 Ma to Recent, with an even spacing of 0.003 Ma (3 ka). A band can be seen at a scale of about 2^5=32 samples, or about 100 ka. A weaker band around $2^{3.7}$=13 samples corresponds to a scale of about 40 ka. These are orbital periodicities. 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 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, note the MA values are given with negative sign).

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 numb ers' window, each row shows one scale, with sample number (position) along the columns.

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).

Reference

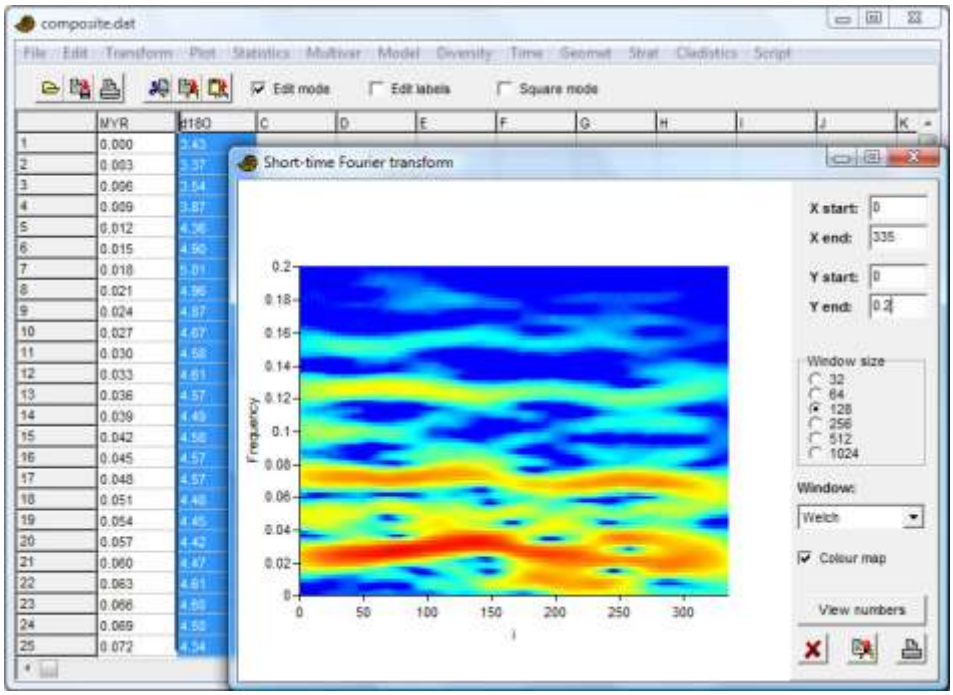
Prokoph, A., A.D. Fowler & R.T. Patterson. 2000. Evidence for periodicity and nonlinearity in a high-resolution fossil record of long-term evolution. *Geology* 28:867-870.

Torrence, C. & G.P. Compo. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* 79:61-78.

Short-time 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.

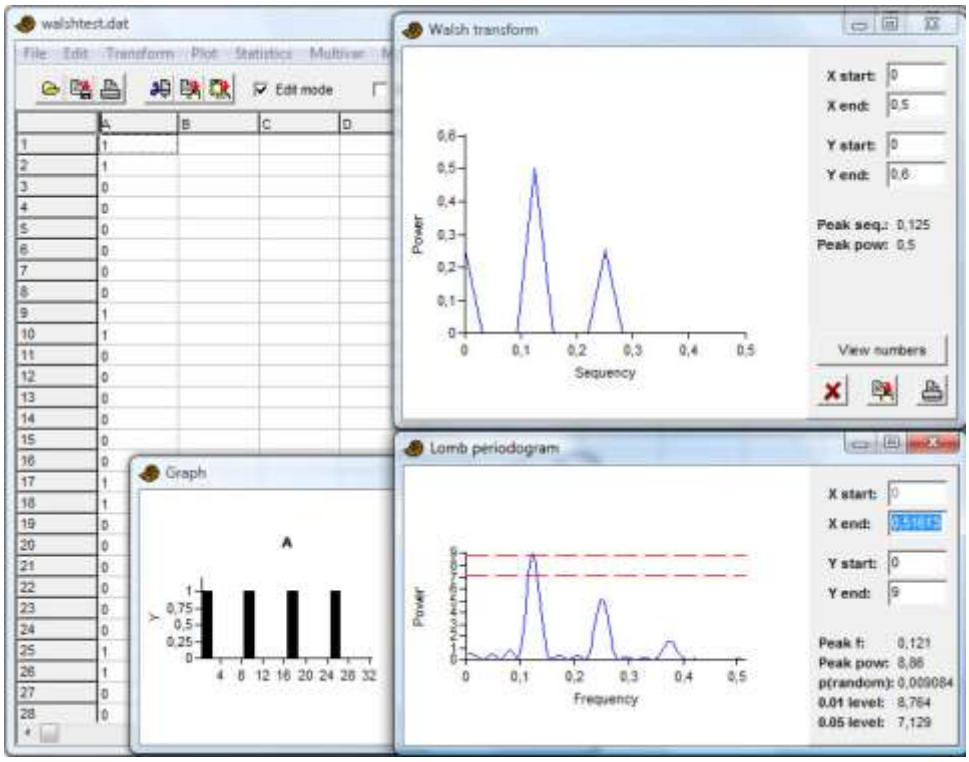
The Short-time Fourier Transform (STFT) can be compared with wavelet analysis, but with a linear frequency scale and with constant time resolution independent of frequency.



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.

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.

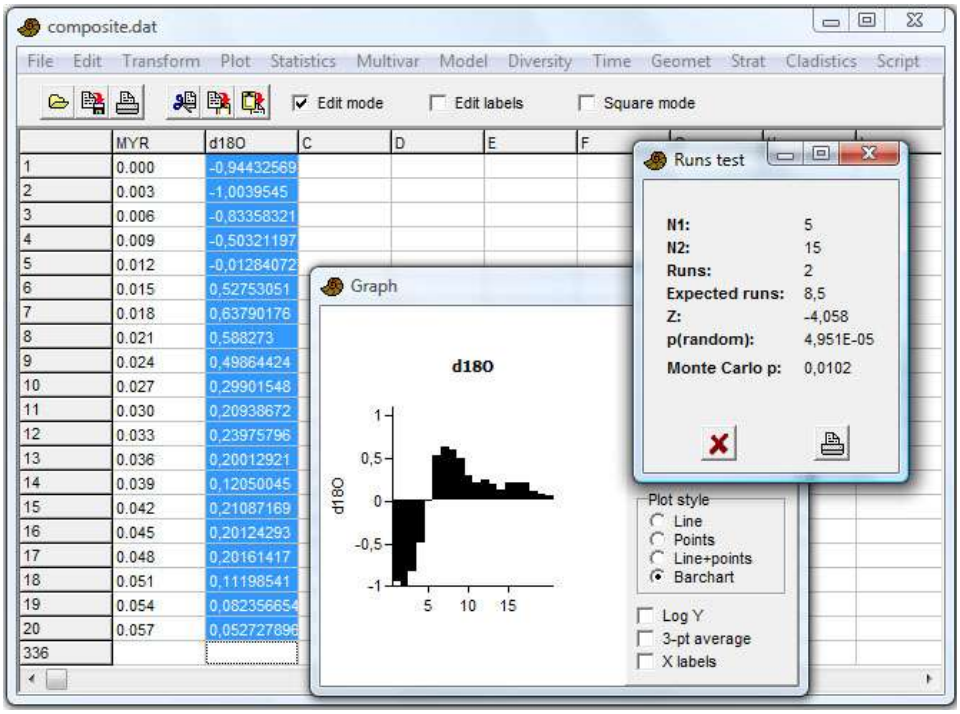
In the example above, compare the Walsh periodogram (top) to the Lomb periodogram (bottom). The data set has 0.125 periods per sample. Both analyses show harmonics.

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.

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

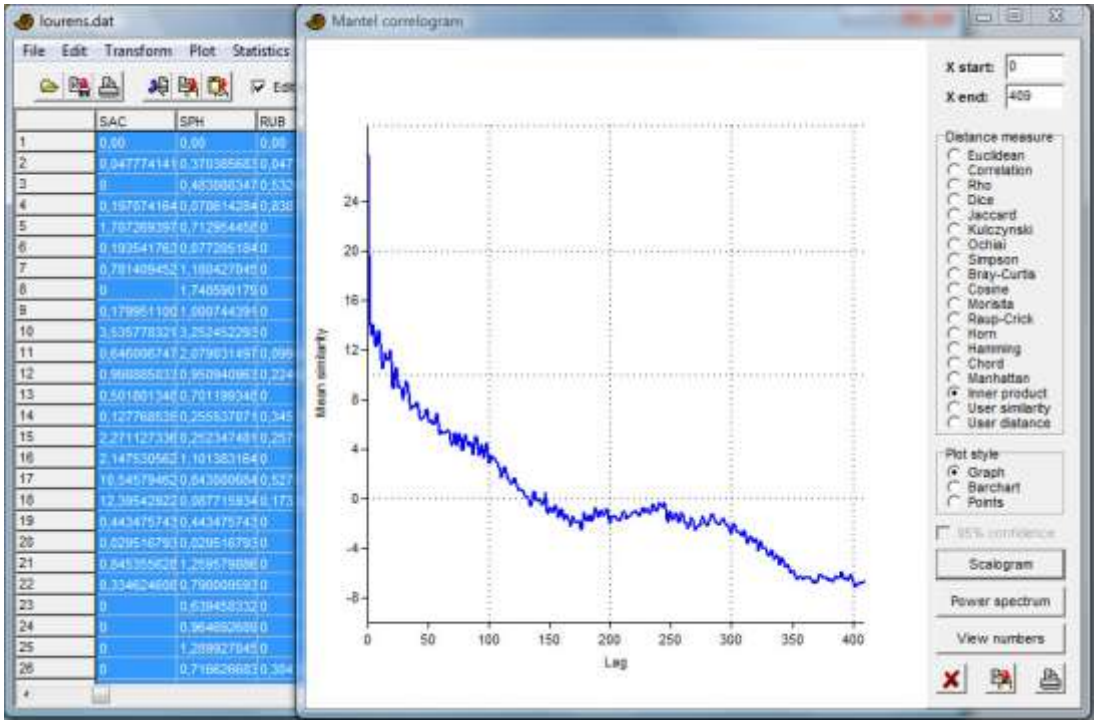
$$z = \frac{R - E(R)}{\sqrt{Var(R)}}.$$

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.

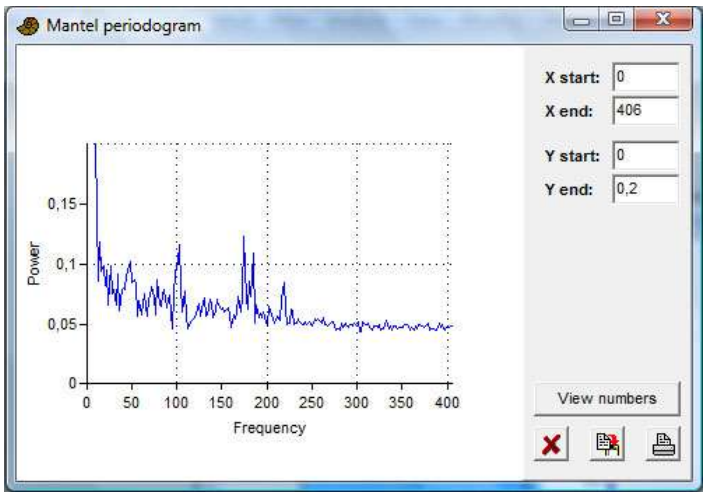
Mantel correlogram (and periodogram)

This module expects several rows of multivariate data, one row for each sample. Samples are assumed to be evenly spaced in time.

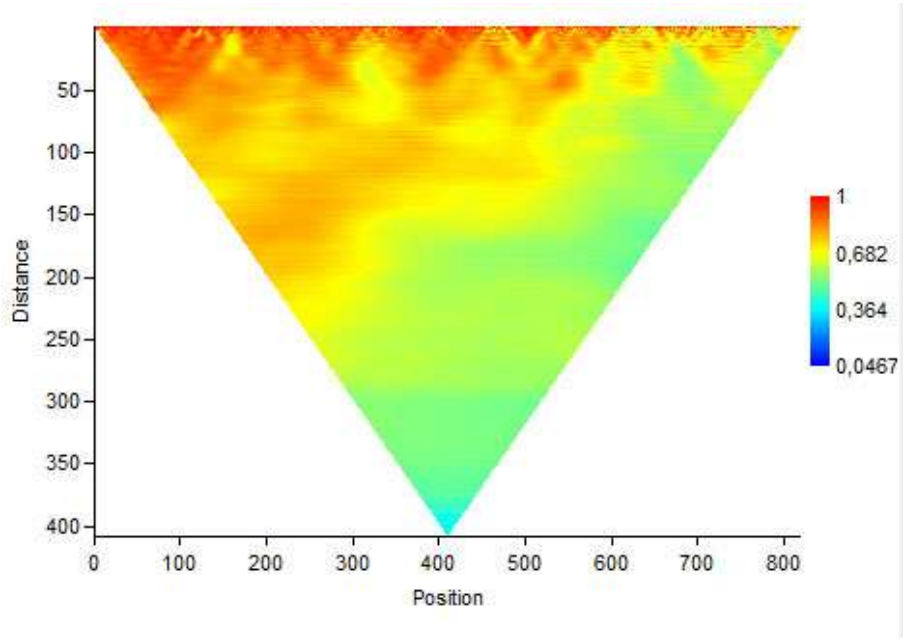
The Mantel correlogram (e.g. Legendre & Legendre 1998) is a multivariate extension to autocorrelation, based on any similarity or distance measure. The Mantel correlogram in PAST shows the average similarity between the time series and a time lagged copy, for different lags.



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.

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. Stationary time series, except for a single intervention. 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.

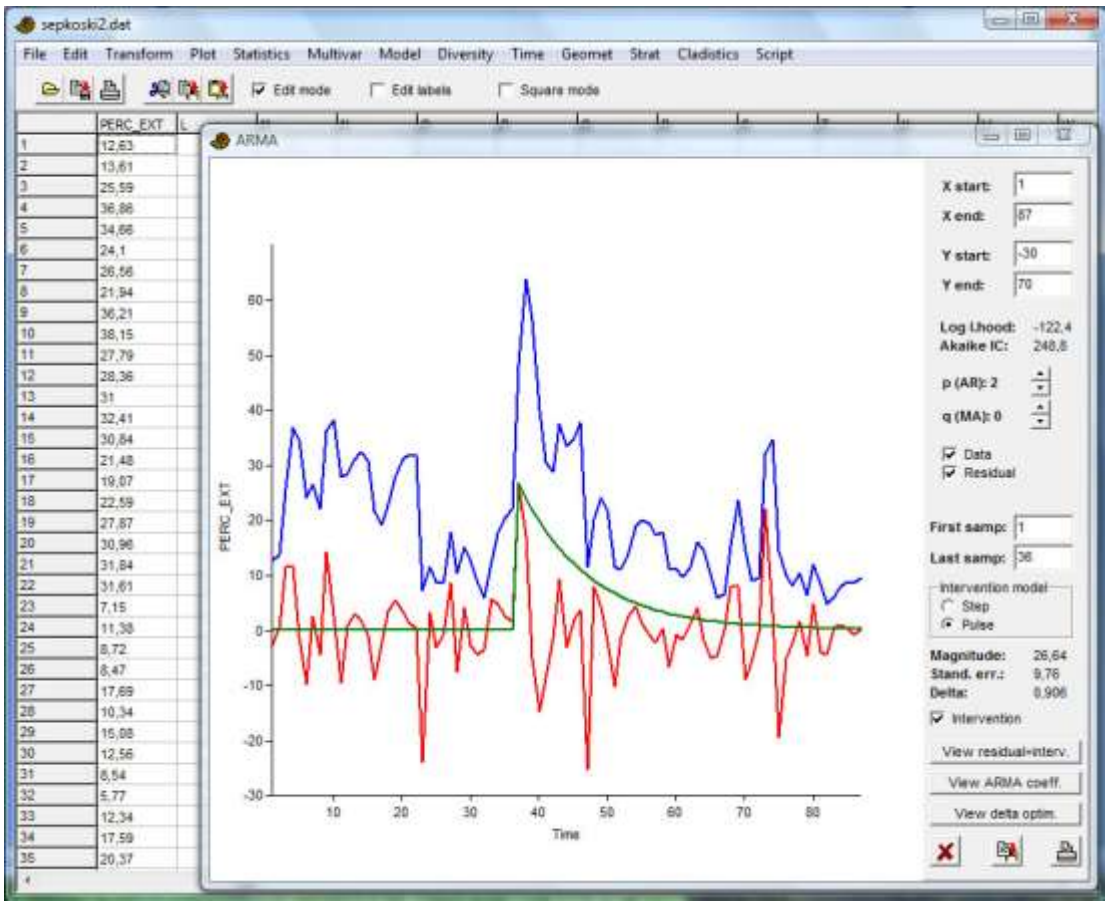
Intervention analysis proceeds as follows. First, carry out ARMA analysis on only the samples preceding the intervention, by typing the last pre-intervention sample number in the "last samp" box. It is also possible to run the ARMA analysis only on the samples following the intervention, by typing the 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.

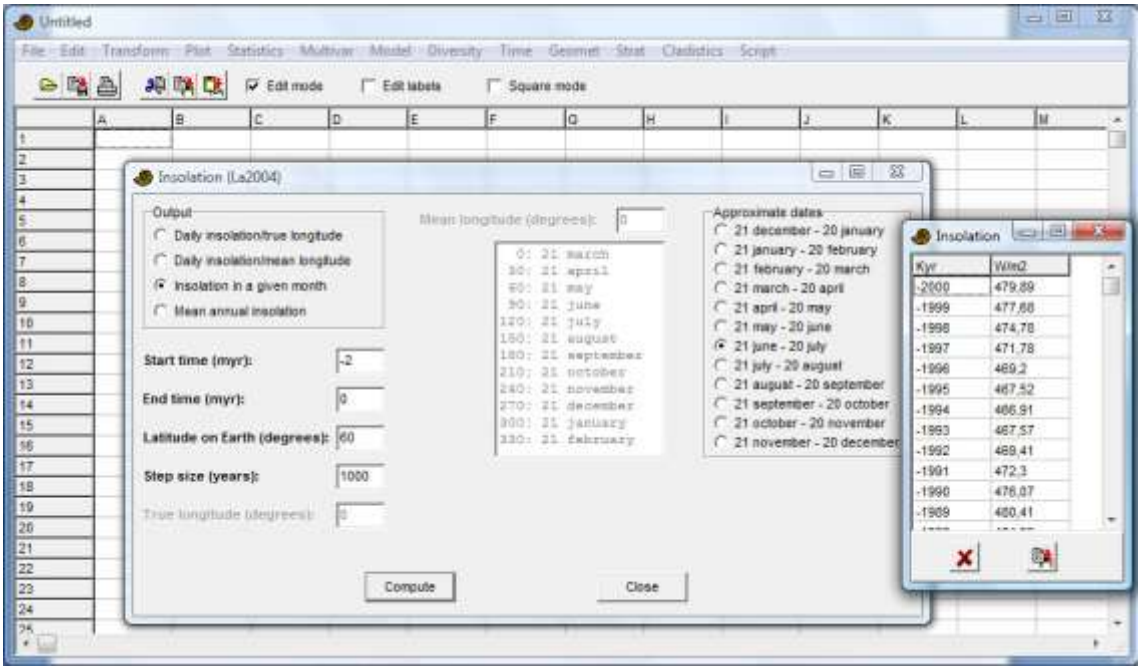
Insolation (solar forcing) model

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.100.ASC from http://www.imcce.fr/Equipes/ASD/insola/earth/La2004 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!

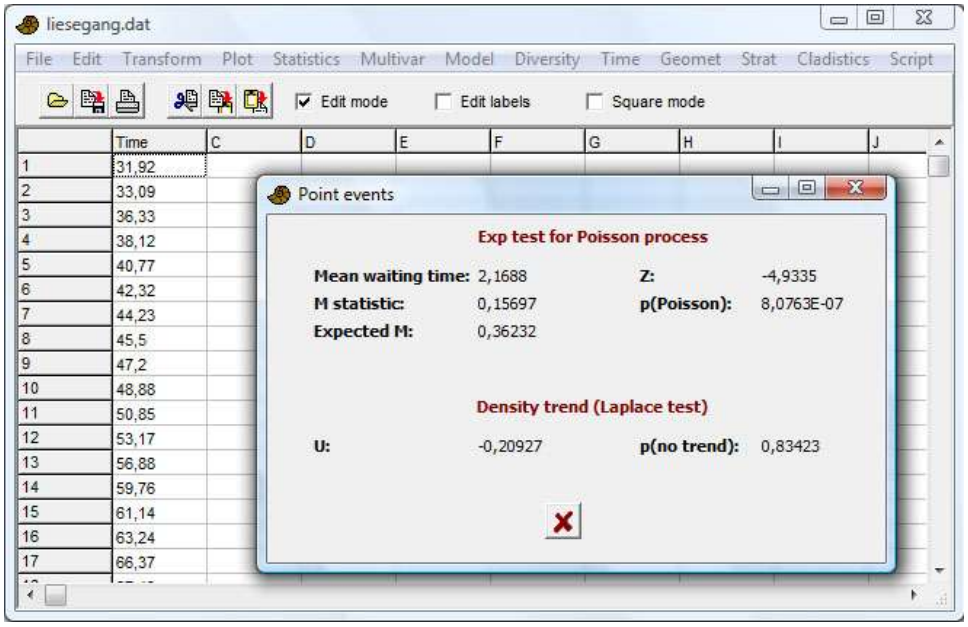


Reference

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.

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.



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.

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 now $\bar{t}$ 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 above are violated.

References

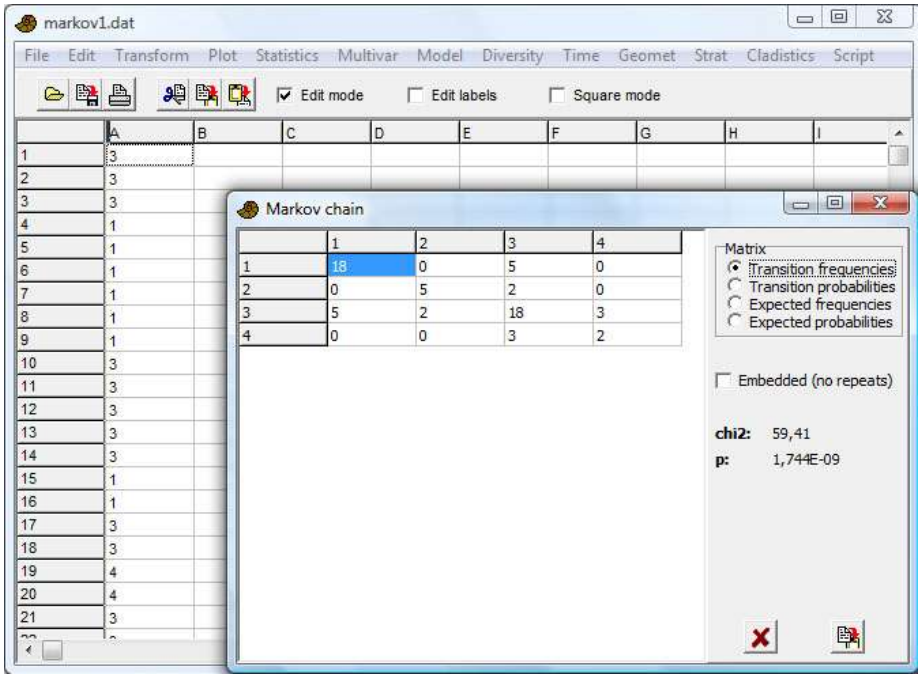
Cox, D. R. & P. A. W. Lewis. 1978. *The Statistical Analysis of Series of Events*. Chapman and Hall, London.

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.

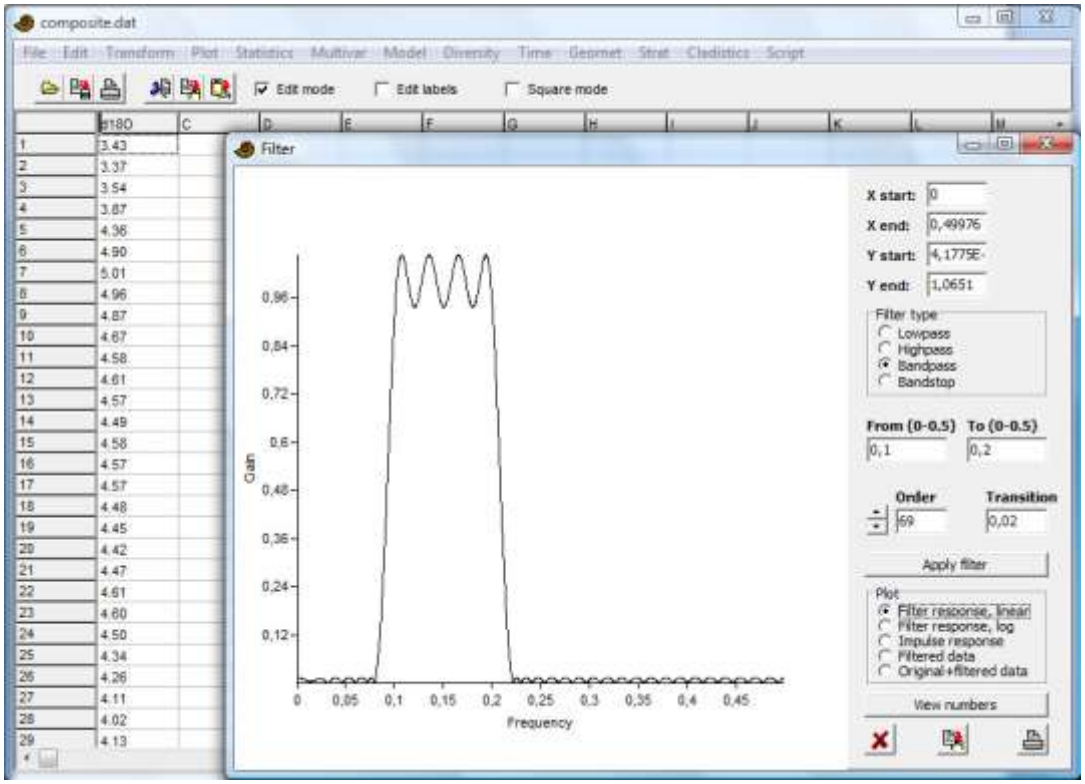
The algorithms, including an iterative algorithm for embedded Markov chains, are according to Davis (1986).

Reference

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

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 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 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.

Simple smoothers

A set of simple smoothers for a single column of evenly spaced data.

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 probably the best overall method in the module.

Moving median

Similar to moving average but takes the median instead of the mean. This method is more robust to outliers.

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.

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

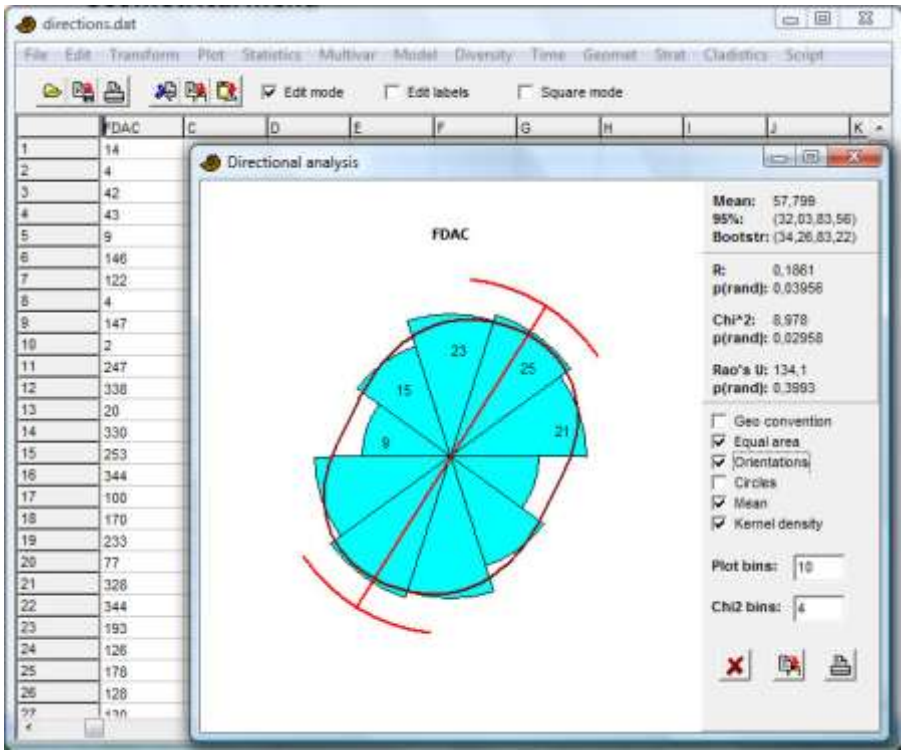
Meeus, J. 1991. *Astronomical algorithms*. Willmann-Bell, Richmond.

Geometrical menu

Directions (one sample)

The module plots a rose diagram (polar histogram) of directions. Used for plotting current-oriented specimens, orientations of trackways, orientations of morphological features (e.g. terrace lines), etc.

One column of directional (0-360) or orientational (0-180) data in degrees is expected. Directional or periodic data in other forms (radians, 0-24 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 a user-defined number of bins (default 4).

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.

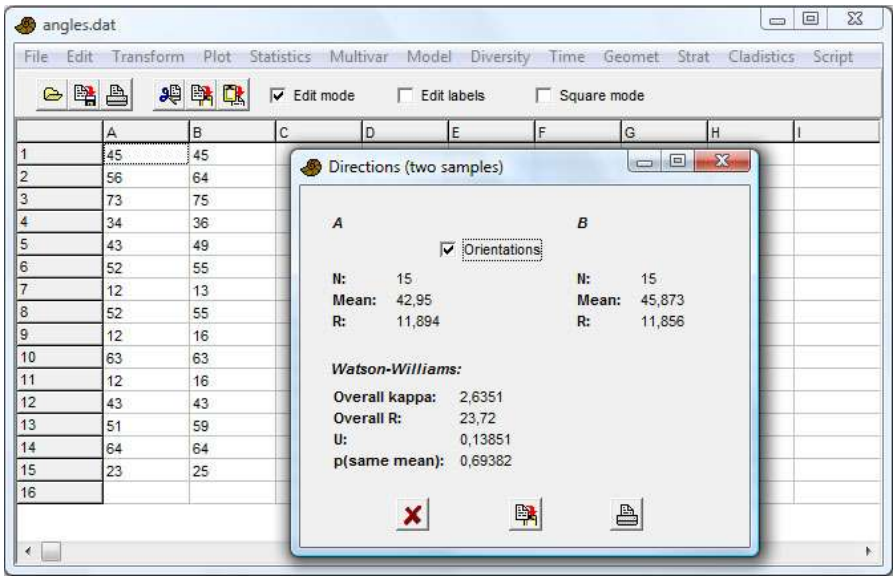
Russell, G. S. & D.J. Levitin 1995. An expanded table of probability values for Rao's spacing test. *Communications in Statistics: Simulation and Computation* 24:879-888.

Directions (two samples)

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 module expects two columns of directional (0-360) or orientational (0-180) data in degrees.

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 **ϕ** and **θ** have n_1 and n_2 values. Rayleigh's spread *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 correlations

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

$$r = \frac{\sum_{i=1}^{n} \sin(\alpha_i - \overline{\alpha}) \sin(\beta_i - \overline{\beta})}{\sqrt{\sum_{i=1}^{n} \sin^2(\alpha_i - \overline{\alpha}) \sin^2(\beta_i - \overline{\beta})}},$$

where the angular means are calculated as described previously. The test statistic *T* is computed as

$$T = r \sqrt{\frac{\sum_{k=1}^{n} \sin^2(\alpha_k - \overline{\alpha}) \sum_{k=1}^{n} \sin^2(\beta_k - \overline{\beta})}{\sum_{k=1}^{n} \sin^2(\alpha_k - \overline{\alpha}) \sin^2(\beta_k - \overline{\beta})}}.$$

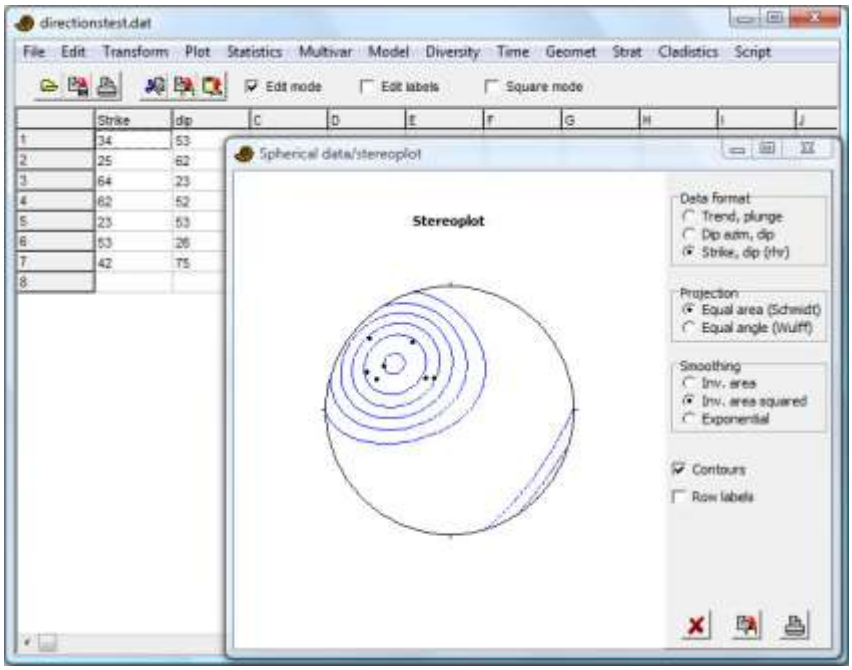
For large *n*, this statistic has asymptotically normal distribution with mean 0 and variance 1 under the null hypothesis of zero correlation, which is the basis for the calculation of *p*.

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). Spherical statistics may be added in future versions.



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.

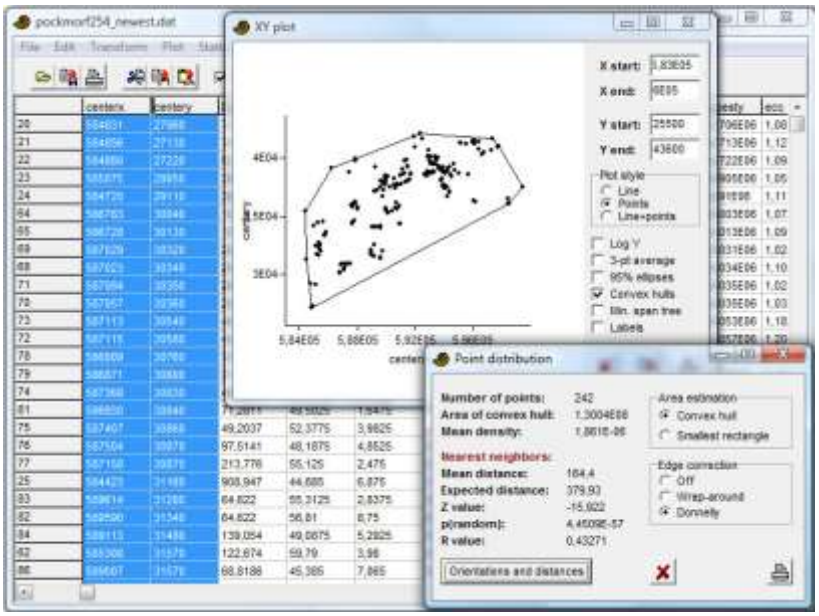
Density contouring is based on a modified Kamb method algorithm by Vollmer (1995). Both equal area (Schmidt) and equal angle (Wulff) projections are available. Projections are to the lower hemisphere. Density estimates can use an inverse area, inverse area squared or exponential law, giving progressively higher smoothing.

Reference

Vollmer, F.W. 1995. C program for automatic contouring of spherical orientation data using a modified Kamb method. *Computers & Geosciences* 21:31-49.

Nearest neighbour point pattern analysis

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 organised along lineaments (see Hammer 2009 for more advanced methods).

References

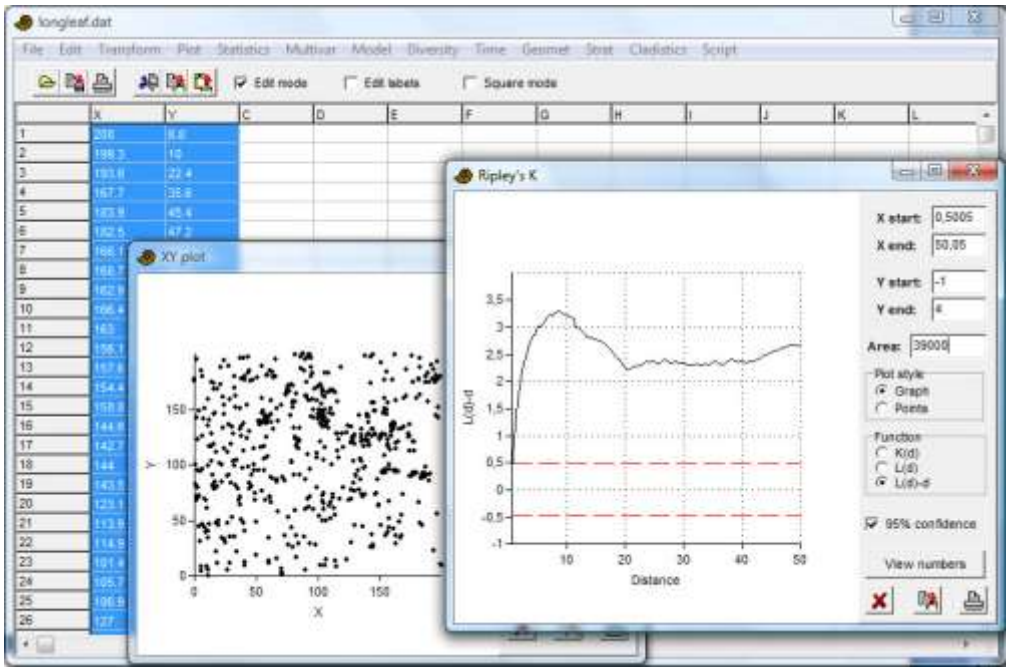
Clark, P.J. & Evans, F.C. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* 35:445-453.

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 trees in a forest. *L*(*d*)-*d* is above the 95% confidence interval of CSR, indicating clustering. In addition, the spatial interactions seem most prominent at a scale of around 10 m, above which 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

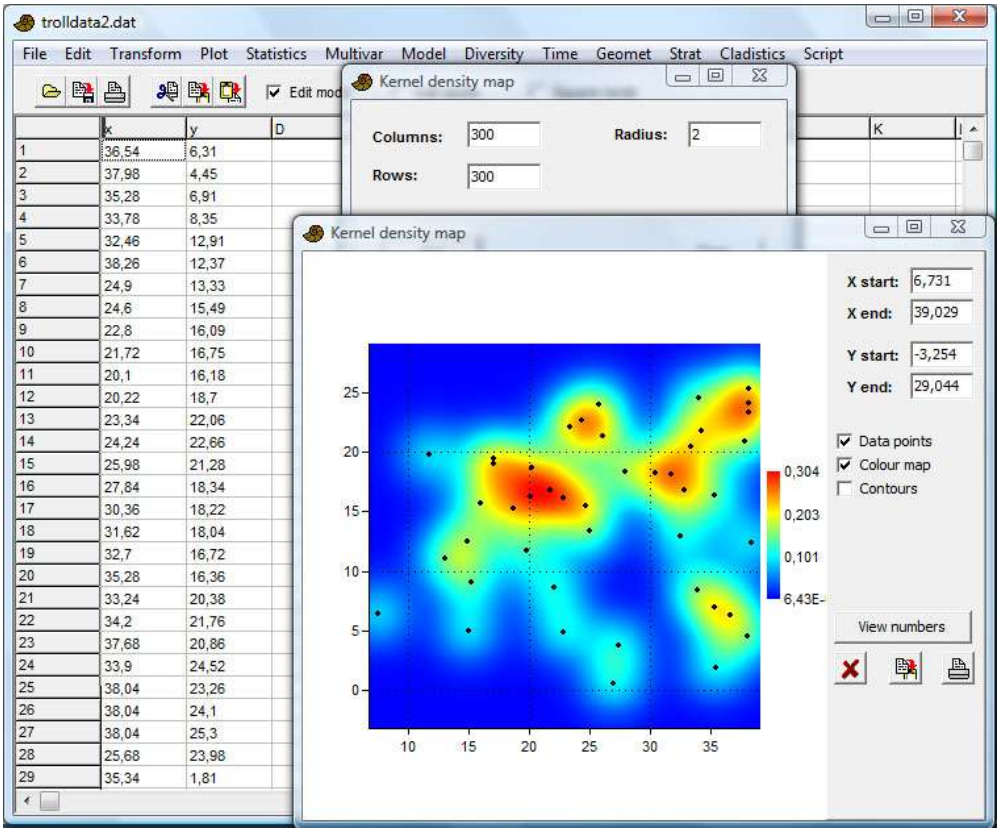
The fractal dimension (if any) can be estimated as the asymptotic linear slope in a log-log plot of *K*(*d*). For CSR, the log-log slope should be 2.0. Fractals should have slopes less than 2.

References

Ripley, B.D. 1979. Tests of 'randomness' for spatial point patterns. *Journal of the Royal Statistical Society, ser. B* 41:368-374.

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): $$f(x,y) = \frac{1}{\pi r^2} \sum_i \exp\left(-\frac{d_i^2}{2r^2}\right)$$

Paraboloid: $$f(x,y) = \frac{3}{2\pi r^2} \sum_i \begin{cases} 1 - \frac{d_i^2}{r^2} & d_i \le r \\ 0 & d_i > r \end{cases}$$

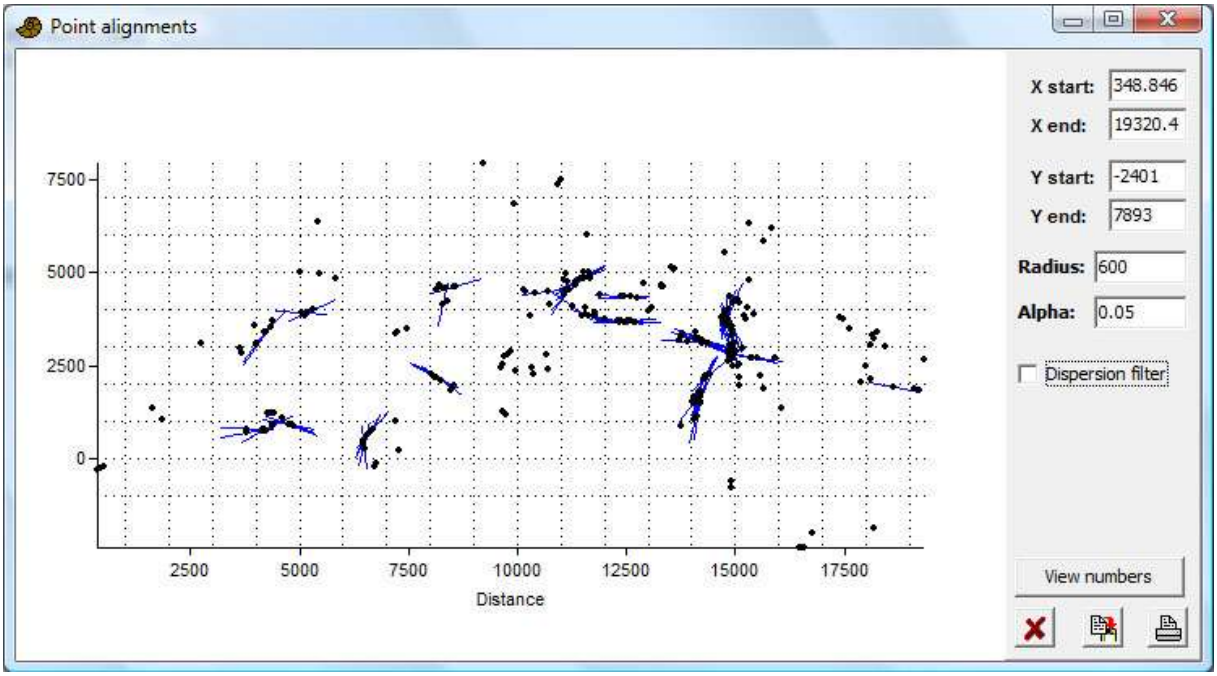
Triangular: $$\mathrm{f}(x,y)=\frac{2}{\pi r^2}\sum_i \begin{cases} 1-\dfrac{d_i}{r} & d_i \le r \\ 0 & d_i > r \end{cases}$$

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.

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

$$S_1 = \frac{1}{2}\sum_{h=1}^{n}\sum_{i=1}^{n}(w_{hi} + w_{ih})^2$$

$$S_2 = \sum_{i=1}^{n}(w_{i+} + w_{+i})^2$$

$$b_2 = \frac{n\sum_{i=1}^{n}(z_i - \bar{z})^4}{\left(\sum_{i=1}^{n}(z_i - \bar{z})^2\right)^2}$$

$$\mathrm{Var}(I) = \frac{n\left[(n^2 - 3n + 3)S_1 - nS_2 + 3W^2\right] - b_2\left[(n^2 - n)S_1 - 2nS_2 + 6W^2\right]}{(n-1)(n-2)(n-3)W^2} - \frac{1}{(n-1)^2}$$

$$I_{0.05} = 1.6452\sqrt{\mathrm{Var}(I)} - k_{0.05}(n-1)^{-1}$$

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 $4(n - \sqrt{n}) < W \le 4(2n - 3\sqrt{n} + 1)$, otherwise $k_{0.05}=1$.

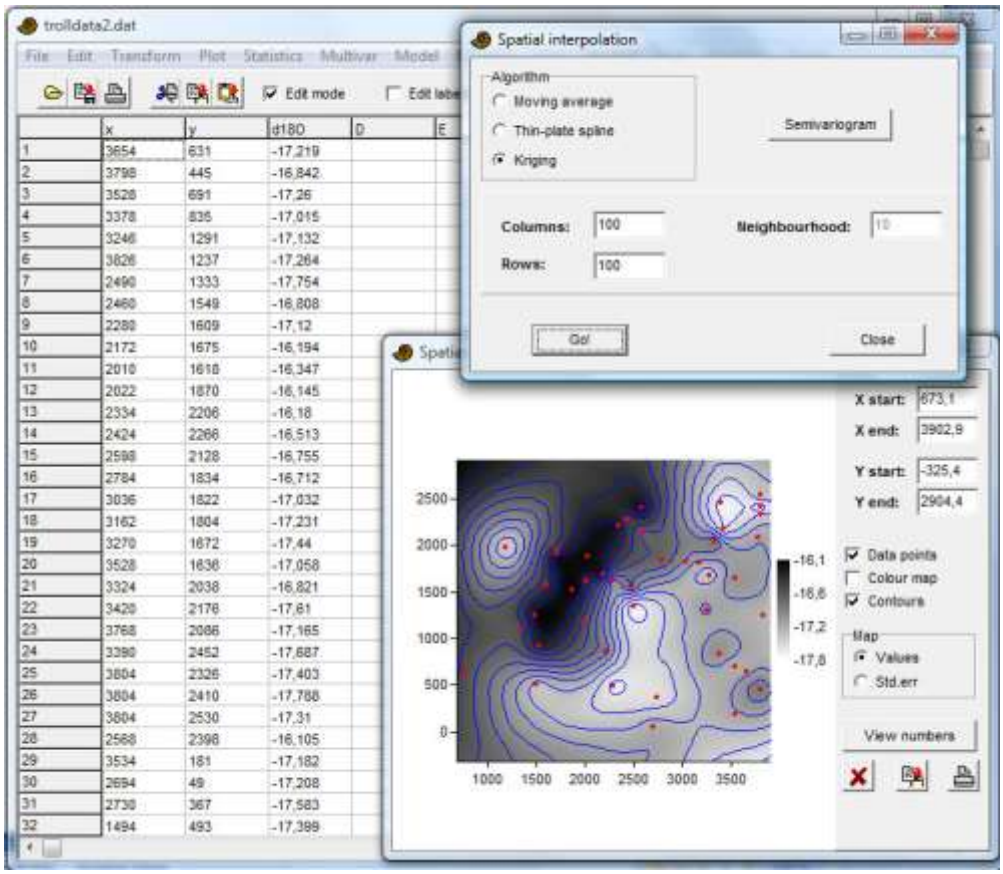
Reference

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 allows the production of 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 generated automatically as 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* parameter is a constant added to the model. It implies a 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: $$\gamma(h) = nugget + scale\left(1 - e^{-h}\right)$$

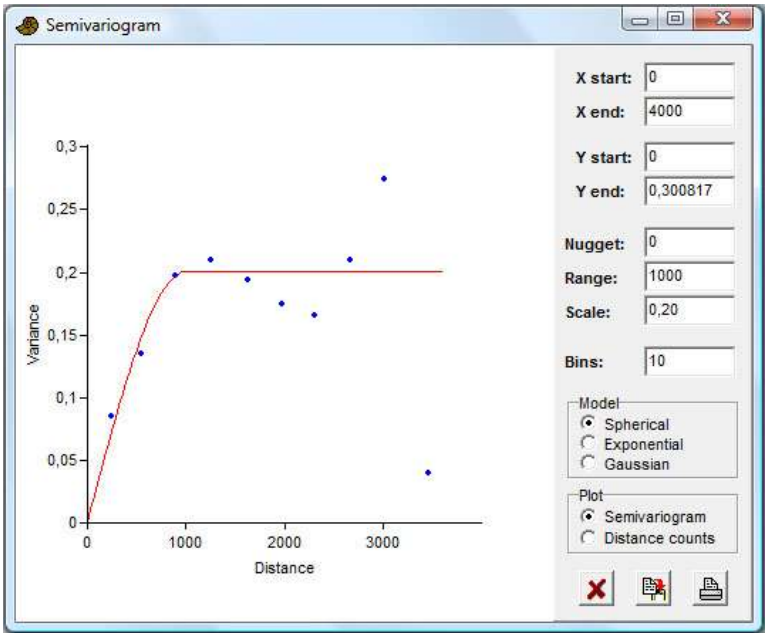
Gaussian: $$\gamma(h) = nugget + scale\left(1 - e^{-h^2}\right)$$

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: For example you may wish to use a specific model or to have zero nugget in order to ensure exact interpolation. This will require setting the values manually.

The kriging procedure also provides an estimate of standard errors across the map (this depends on the semivariogram model being accurate). 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.

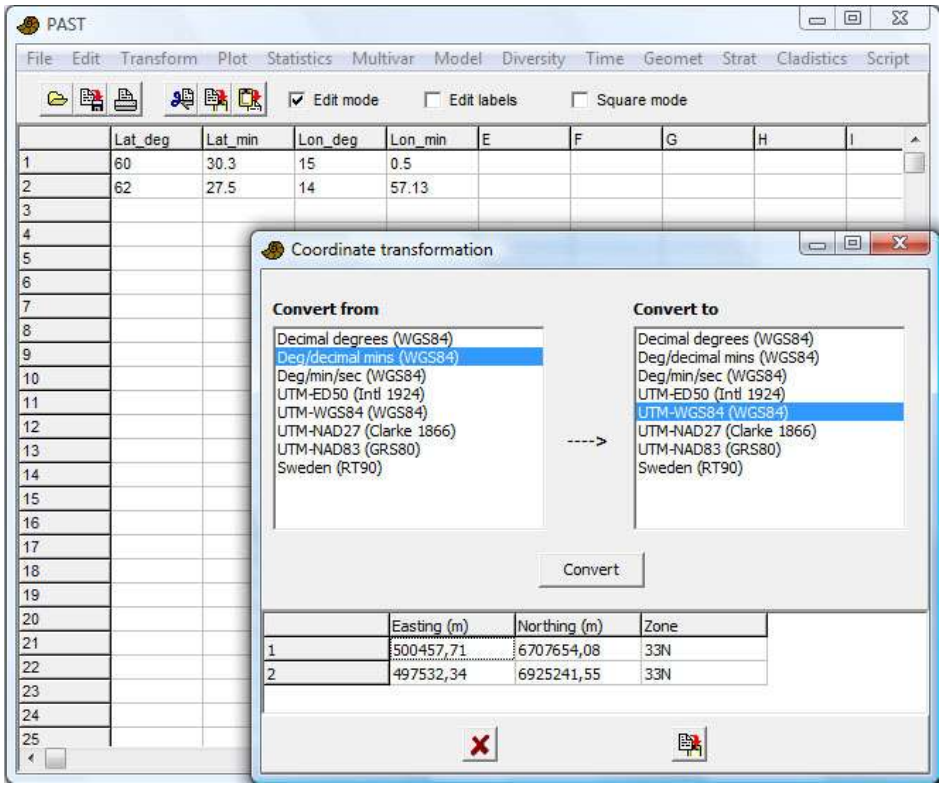
References

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.

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).

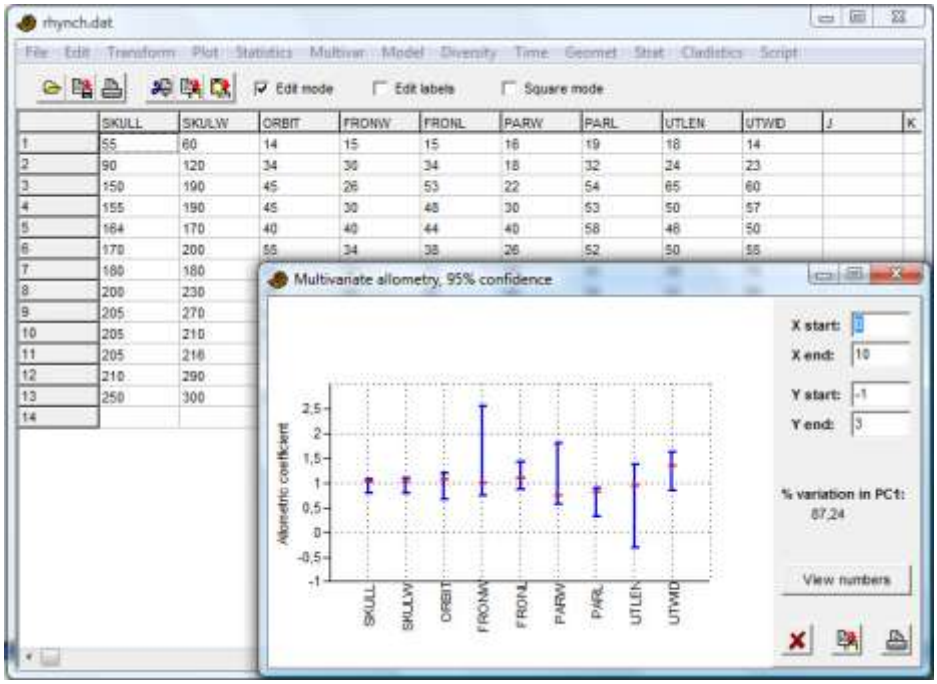
Sweden (RT90)

Two columns: Easting (meters) and northing (meters).

The transformations are based on code generously provided by I. Scollar.

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.

Fourier shape (2D)

Analysis of fossil outline shape (2D). Shape expressible in polar coordinates, sufficient number of digitized points to capture features. Digitized *x/y* coordinates around an outline. Specimens in rows, coordinates of alternating *x* and *y* values in columns (see Procrustes fitting in the Transform menu).

Accepts X-Y coordinates digitized around an outline. More than one shape (row) can be simultaneously analyzed. Points do not need to be totally evenly spaced. The shape must be expressible as a unique function in polar co-ordinates, that is, any straight line radiating from the centre of the shape must cross the outline only once.

The algorithm follows Davis (1986). The origin for the polar coordinate system is found by numerical approximation to the centroid. 128 points are then produced at equal angular increments around the outline, through linear interpolation. The centroid is then re-computed, and the radii normalized (size is thus removed from the analysis). The cosine and sine components are given for the first twenty harmonics, but note that only *N*/2 harmonics are 'valid', where *N* is the number of digitized points. The coefficients can be copied to the main spreadsheet for further analysis (e.g. by PCA).

The 'Shape view' window allows graphical viewing of the Fourier shape approximation(s).

Reference

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

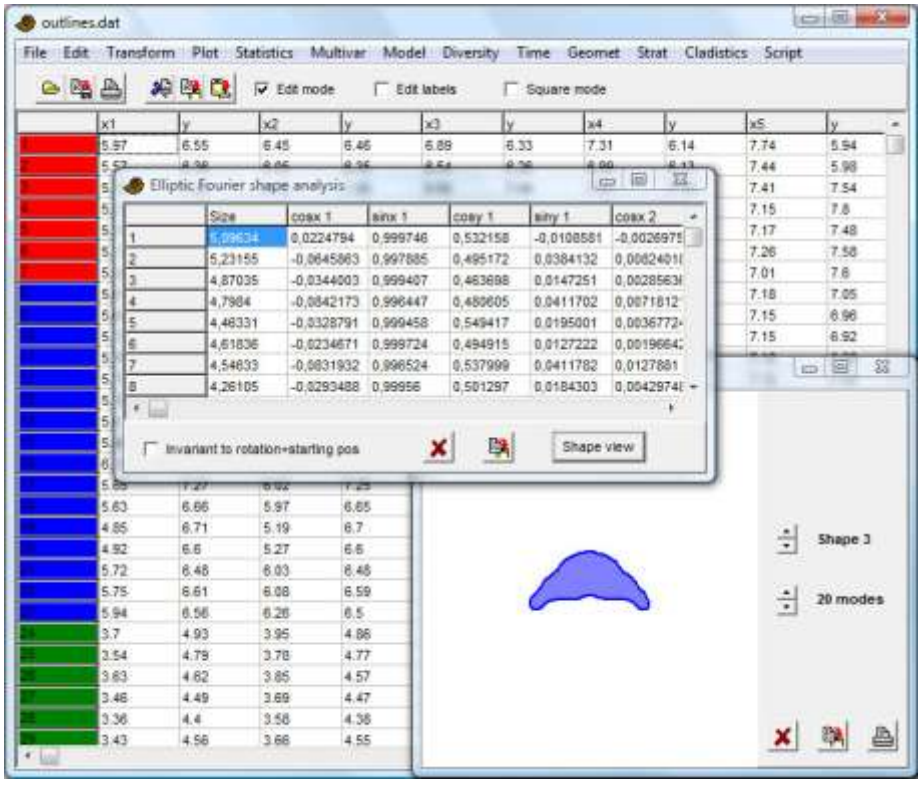
Elliptic Fourier shape analysis

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).

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 PCA and linear regression (1 independent, *n* dependent) modules contain functions for displaying the outline shapes corresponding to given PCA scores or values for the independent variable.

The 'Shape view' window allows graphical viewing of the elliptic Fourier shape approximation(s).



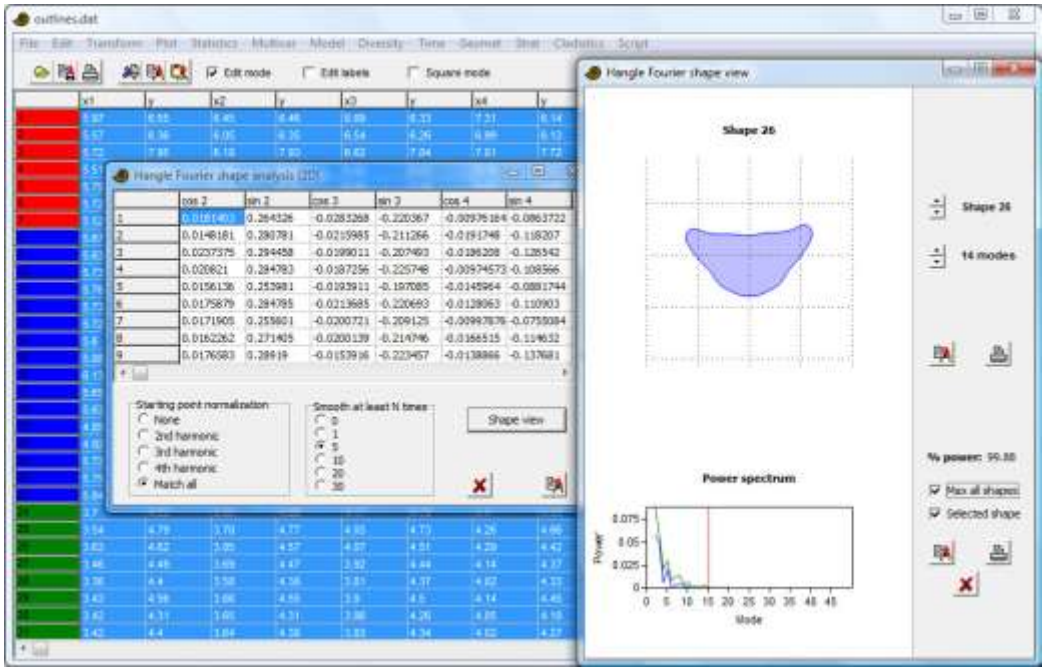
Reference

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: Shape reconstruction in PCA, regression and CVA (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

Eigenshape analysis

Digitized *x/y* coordinates around an outline. Specimens in rows, coordinates of alternating *x* and *y* values in columns (see Procrustes fitting in the Transform menu).

Eigenshapes are principal components of outlines. The scatter plot of outlines in principal component space can be shown, and linear combinations of the eigenshapes themselves can be visualized.

The implementation in PAST is partly based on MacLeod (1999). It finds the optimal number of equally spaced points around the outline using an iterative search, so the original points need not be equally spaced. The eigenanalysis is based on the covariance matrix of the non-normalized turning angle increments around the outlines. The algorithm does not assume a closed curve, and the endpoints are therefore not constrained to coincide in the reconstructed shapes. Landmark-registered eigenshape analysis is not included. All outlines must start at the 'same' point.

Reference

MacLeod, N. 1999. Generalizing and extending the eigenshape method of shape space visualization and analysis. *Paleobiology* 25:107-138.

Thin-plate splines and warps

Digitized *x/y* landmark coordinates. Specimens in rows, coordinates of alternating *x* and *y* values in columns. Procrustes standardization recommended.

The first specimen (first row) 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

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 will normally be connected with more local deformations. The affine component of the warp (also known as zeroth warp) represents linear translation, scaling, rotation and shearing. In the present version of PAST you can not view the principal warps.

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.

Partial warp scores

From the thin-plate spline window, you can also choose to see the partial warp scores of all the specimens. Each partial warp score has two components (*x* and *y*), and the scores are therefore presented in scatter plots.

Reference

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

Relative warps

Ordination of a set of shapes. Digitized *x/y* landmark coordinates. Specimens in rows, coordinates of alternating *x* and *y* values in columns. Procrustes standardization recommended.

The relative warps can be viewed as the principal components of the set of thin-plate transformations from the mean shape to each of the shapes under study. It provides an alternative to direct PCA of the landmarks (see Shape PCA above).

The parameter *alpha* can be set to one of three values:

- *alpha=-1* emphasizes small-scale variation.
- *alpha=0* is PCA of the landmarks directly, and is equivalent to Shape PCA (see above) *but without including the affine (uniform) component*.
- *alpha=1* emphasizes large-scale variation.

The relative warps are ordered according to importance, and the first and second warps are usually the most informative. Note that the percentage values of the eigenvalues are relative to the total non-affine part of the transformation - the affine part is not included (see Shape PCA for relative warps with the affine component included).

The relative warps are visualized with thin-plate spline transformation grids. When you increase or decrease the amplitude factor away from zero, the original landmark configuration and grid will be progressively deformed according to the selected relative warp.

The relative warp scores of pairs of consecutive relative warps can shown in scatter plots, and all scores can be shown in a numerical matrix.

The algorithm for computing the relative warps is taken from Dryden & Mardia (1998).

Reference

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

Size from landmarks (2D or 3D)

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)

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)

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.

Landmark linking

This function in the Geomet menu allows the selection of any pairs of landmarks to be linked with lines in the morphometric plots (thin-plate splines, partial and relative warps, etc.), to improve readability. The landmarks must be present in the main spreadsheet before links can be defined.

Pairs of landmarks are selected or deselected by clicking in the symmetric matrix. The set of links can also be saved in a text file. Note that there is little error checking in this module.

Strat menu

Unitary Associations

Unitary Associations analysis (Guex 1991) 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 the same color, and ordered stratigraphically within each section such that the lowermost sample enters in the lowest row. Colors can be re-used in data sets with large numbers of sections .

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 options and improvements not found in the present version of that program.

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 (and BioGraph) 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 dashed red lines, with long dashes from the above-occurring taxon and short dashes from the below-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 dashed red lines, with long dashes from the above-occurring clique and short dashes from the below-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. & H. Bucher. 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.

Savary, J. & J. Guex. 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.

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.

CONOP

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.

Appearance Event Ordination

Appearance Event Ordination (Alroy 1994, 2000) is a method for biostratigraphical seriation and correlation. The data input is in the same format as for Unitary Associations, consisting of a presence/absence matrix with samples in rows and taxa in columns. Samples belonging to the same section (locality) must be assigned the same color, and ordered stratigraphically within each section such that the lowermost sample enters in the lowest row. Colors can be re-used in data sets with large numbers of sections.

The implementation in PAST is based on code provided by John Alroy. It includes Maximum Likelihood AEO (Alroy 2000).

References

Alroy, J. 1994. Appearance event ordination: a new biochronologic method. *Paleobiology* 20:191-207.

Alroy, J. 2000. New methods for quantifying macroevolutionary patterns and processes. *Paleobiology* 26:707-733.

Diversity curve

Abundance or presence/absence matrix with samples in rows (lowest sample at bottom) and taxa in columns

Found in the 'Strat' menu, this simple tool allows plotting of diversity curves from occurrence data in a stratigraphical column. Note that samples should be in stratigraphical order, with the uppermost (youngest) sample in the uppermost row. Data are subjected to the range-through assumption (absences between first and last appearance are treated as presences). Originations and extinctions are in absolute numbers, not percentages.

The 'Endpoint correction' option counts a first or last occurrence (FAD or LAD) in a sample as 0.5 instead of 1 in that sample. Both FAD and LAD in the sample (singleton) counts as 0.33. See Hammer & Harper (2006).

Reference

Hammer, Ø. & Harper, D.A.T. 2006. Paleontological Data Analysis. Blackwell.

Range confidence intervals

Estimation of confidence intervals for first or last appearances or total range, for one taxon. Assumes random distribution of fossiliferous horizons through the stratigraphic column or through time. Section should be continuously sampled.

Assuming a random (Poisson) distribution of fossiliferous horizons, confidence intervals for the stratigraphic range of one taxon can be calculated given the first occurrence datum (level), last occurrence datum, and total number of horizons where the taxon is found (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:

$$\alpha = \left(1 - C_1\right)^{-1/(H-1)} - 1,$$

where C_1 is the confidence level and H is the number of fossiliferous horizons.

For the joint endpoint (total range) case, α is found by iterative solution of the equation

$$C_2 = 1 - 2\left(1+\alpha\right)^{-(H-1)} + \left(1+2\alpha\right)^{-(H-1)}.$$

Be aware that the assumption of random distribution will not hold in many real situations.

References

Marshall, C.R. 1990. Confidence intervals on stratigraphic ranges. *Paleobiology* 16:1-10.

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 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 requires 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 can not be calculated are marked with an asterisk (see Marshall 1994).

Reference

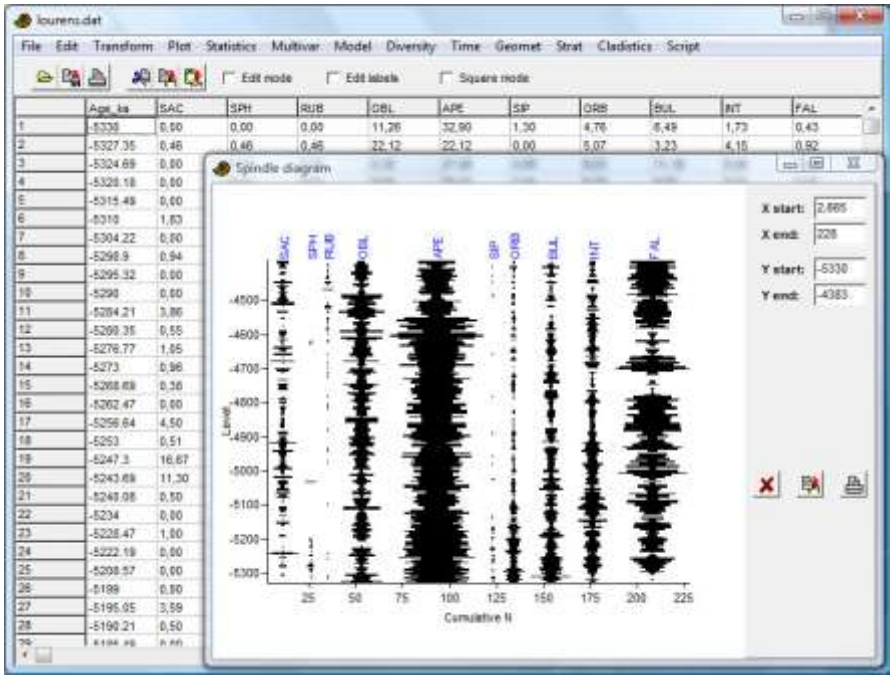
Marshall, C.R. 1994. Confidence intervals on stratigraphic ranges: partial relaxation of the assumption of randomly distributed fossil horizons. *Paleobiology* 20:459-469.

Spindle diagram

Standard plot type used in paleontology to illustrate abundance of fossil taxa through a stratigraphic section or core. Samples are entered in rows, taxa in columns. The program will ask whether the first column contains stratigraphic levels (e.g. meters or years).

If stratigraphic levels are given, each box will be drawn from the given level to the level in the next row. Therefore a last, extra “dummy” row should be given, with a final stratigraphic level but zero for all the taxa. If levels are given in depths or ages, negative numbers should be used in order to ensure correct orientation of the figure.

If stratigraphic levels are not given, all boxes will be given equal height. The topmost sample should be entered in the first row.



Filter events

text

Cladistics

Parsimony analysis

Warning: the cladistics package in PAST is fully operational, but lacking in comprehensive functionality. The heuristic algorithms seem not to perform quite as well as in some other programs (this is being looked into). The PAST cladistics package is adequate for education and initial data exploration, but for more 'serious' work we recommend a specialized program such as PAUP.

Semi-objective analysis of relationships between taxa from morphological or genetic evidence.

Character matrix with taxa in rows, outgroup in first row. For calculation of stratigraphic congruence indices, first and last appearance datums must be given in the first two columns.

Algorithms are from Kitching et al. (1998).

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 (?) or the value -1. 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.

There are four algorithms available for finding short trees:

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 12 taxa.

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. For 12 taxa, 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). For more than 8 taxa, the histogram is based on a subset of tree lengths and may not be accurate.

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 permutated and another heuristic search attempted.

Please 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 optimization criteria

Three different optimization criteria are availiable:

Wagner

Characters are reversible and ordered, meaning that 0->2 costs more than 0->1, but has the same cost as 2->0.

Fitch

Characters are reversible and unordered, meaning that all changes have equal cost. This is the criterion with fewest assumptions, and is therefore generally preferable.

Dollo

Characters are ordered, but acquisition of a character state (from lower to higher value) can happen only once. All homoplasy is accounted for by secondary reversals. Hence, 0->1 can only happen once, normally relatively close to the root of the tree, but 1->0 can happen any number of times further up in the tree. (This definition has been debated on the PAST mailing list, especially whether Dollo characters need to be ordered).

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' buttons. 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 amount of synapomorphy on the tree, and varies from 0 to 1.

Please note that in the present version, the retention index is only correctly calculated when using Fitch optimisation.

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.

Stratigraphic congruence indices

For calculation of stratigraphic congruence indices, the first two columns in the data matrix must contain the first and last appearance datums, respectively, for each taxon. These datums must be

given such that the youngest age (or highest stratigraphic level) has the highest numerical value. You may need to use negative values to achieve this (e.g. 400 million years before present is coded as -400.0). The box "FADs/LADs in first columns" in the Parsimony dialogue must be ticked.

The Stratigraphic Congruence Index (SCI) of Huelsenbeck (1994) is defined as the proportion of stratigraphically consistent nodes on the cladogram, and varies from 0 to 1. A node is stratigraphically consistent when the oldest first occurrence above the node is the same age or younger than the first occurrence of its sister taxon (node).

The Relative Completeness Index (RCI) of Benton & Storrs (1994) is defined as (1-MIG/SRL)x100%, where MIG (Minimum Implied Gap) is the sum of the durations of ghost ranges and SRL is the sum of the durations of observed ranges. The RCI can become negative, but will normally vary from 0 to 100.

The Gap Excess Ratio (GER) of Wills (1999) is defined as $1-(MIG-G_{min})/(G_{max}-G_{min})$ where G_{min} is the minimum possible sum of ghost ranges on any tree (that is, the sum of distances between successive FADs), and G_{max} is the maximum (that is, the sum of distances from first FAD to all other FADs).

These indices are further subjected to a permutation test, where all dates are randomly redistributed on the different taxa 1000 times. The proportion of permutations where the recalculated index exceeds the original index is given. If small (e.g. $p<0.05$), this indicates a statistically significant departure from the null hypothesis of no congruency between cladogram and stratigraphy (in other words, you have significant congruency). The permutation probabilities of RCI and GER are equal for any given set of permutations, because they are based on the same value for MIG.

References

Benton, M.J. & G.W. Storrs. 1994. Testing the quality of the fossil record: paleontological knowledge is improving. *Geology* 22:111-114.

Farris, J.S. 1989. The retention index and the rescaled consistency index. *Cladistics* 5:417-419.

Huelsenbeck, J.P. 1994. Comparing the stratigraphic record to estimates of phylogeny. *Paleobiology* 20:470-483.

Kitching, I.J., P.L. Forey, C.J. Humphries & D.M. Williams. 1998. Cladistics. Oxford University Press.

Wills, M.A. 1999. The gap excess ratio, randomization tests, and the goodness of fit of trees to stratigraphy. *Systematic Biology* 48:559-580.

Scripting

PAST includes a simple scripting (programming) language that allows you to make your own algorithms within the PAST framework. The language is a simplified Pascal, but with full matrix support and a library of mathematical, statistical and user-interface functions.

Important: As of the present version, this new scripting capability is rudimentary and not entirely stable. It will be rapidly improved in coming versions!

LANGUAGE STRUCTURE

Variables
Variable names can be of any length. There is no declaration of variables. All variables have global scope. The assignment operator is :=.

Comments
Upon encountering the comment character '#' in the source, the rest of the line is skipped.

begin ... end
Blocks are marked with begin ... end. The entire script must be contained within a begin .... end pair.

if ... else
Example:

```
if fad>0 begin
  lad:=fad+1;
  fad:=0;
end else lad:=0;
```

for ... to
Example:

```
for i:=1 to 10 a[i]:=0;
```

while
Example:

```
while a[i]=0 i:=i+1;
```

procedure
Procedures have no arguments, return values or local variables. All communication with the procedure must go through global variables. A procedure call must contain a dummy argument. The procedure definition must be inside the outermost begin ... end pair. Example:

```
begin
  procedure writehello
  begin
    message("Hello");
  end;
```

```
  writehello(0);
end;
```

Types

Four types are available: Double-precision numbers, vectors, arrays and strings. Types are implicitly defined at first usage. "Sensible" type conversions are performed automatically.

Examples:

```
a:=14.3;
b:="Hi!";
c:=array(5,10); c[3,7]:=a;
d:=vector(10); d[7]:=17;
```

Operators

Operator	Supported types
+	double+double, vector+vector, array+array, array+double, vector+double
-	double-double, vector-vector, array-array, array-double, vector-double
*	double*double, array*array (array multiplication), array*double, vector*double
/	double/double, array/double, vector/double
^	double^double (power)
%	double%double (integer modulo)
&	string&string, string&double, double&string (concatenation)
=	double=double, string=string
>	double>double, string>string
<	double<string<>
>=	double>=double, string>=string
<=	double<=double, string<=string
<>	double<>double, string<>string
And	double and double
Or	double or double

MATHEMATICAL FUNCTIONS

Most of these functions support double, vector and array types.

Function	Comments
abs(x)	Absolute value
cos(x)	Cosine
sin(x)	Sine

tan(x)	Tangent
exp(x)	Exponential, e^x
log(x)	Natural logarithm
sqrt(x)	Square root
int(x)	Round down to integer
rnd(x)	Random, uniformly distributed number in the range [0..x)
fprob(f, df1, df2)	p value for F distribution
tprob(t, df)	p value for Student's t distribution
zprob(z)	p value for standardized normal distribution (z test)
chi2prob(chi2, df)	p value for chi^2 distribution
anova(col1,col2)	F statistic for one-way ANOVA, columns col1-col2 from spreadsheet

Array and vector operations

Function	Comments
nrows(x)	Number of rows in vector or array x
ncols(x)	Number of columns in array x
array(m,n)	Allocate an array of size m,n
vector(m)	Allocate a vector of length m
row(x,m)	Row m in array x, as a vector
column(x,n)	Column n in array x, as a vector
inv(x)	Inverse of double, or square array
mean(x)	Mean value of vector or array
eigen(x)	For a square NxN matrix x, returns a Nx(N+1) matrix with N sorted eigenvectors in the first N columns, and the eigenvalues in the last column.
cov(x)	Variance-covariance matrix of array x
sort2(x1, x2)	Sort the vectors x1 and x2, on x1. Returns an array with the sorted vectors in the two columns.

Distance matrices

Note that all these functions return symmetric distance matrices. For similarity indices such as Jaccard, the complement 1-x is returned. A value -1 in the input matrix is treated as missing value.

Function	Comments
eucliddist(x)	Symmetric Euclidean distance matrix of array x
chorddist(x)	Chord distance
cosdist(x)	Complement of cosine similarity
dicedist(x)	Complement of Dice similarity
jaccarddist(x)	Complement of Jaccard similarity
morisitadist(x)	Complement of Morisita similarity

horndist(x)	Complement of Horn similarity
manhattandist(x)	Manhattan distance
corrdist(x)	Complement of correlation similarity
rhodist(x)	Complement of non-parametric correlation
raupcrickdist(x)	Complement of Raup-Crick similarity
hammingdist(x1, x2)	Hamming distance
simpsondist(x1, x2)	Complement of Simpson similarity

USER INTERFACE FUNCTIONS AND PROCEDURES

Functions:

Function	Comments
spreadsheetarray(0|1)	Returns the selected array of numbers in the spreadsheet. Call with argument 1 for replacing missing values with column averages.
spreadsheetcolumn(n)	Returns column n from the spreadsheet, as a vector
spreadsheetsymbols(0)	Returns a vector of integer values for symbols (colors) in the selected array.
selectedrowlabel(m)	Returns a string containing the row label of row m in the selected array.
selectedrowlabel(n)	Returns a string containing the column label of column n in the selected array.

Procedures:

message(x)	Displays a message dialogue showing x (string or double)
setspreadsheet(m, n, x)	Set the contents of cell m,n in spreadsheet to x (string or double)
opennumwindow(m, n)	Opens a numerical output (table) window with m rows and n columns.
setnumwindow(m, n, x)	Set the contents of cell m,n in the numerical output window to x (string or double)

Graphics

Only one graphics window is available at any one time. Before displaying graphics, the window must be opened using the procedure openwindow. Coordinates can be in any unit, as the window is automatically scaled to accomodate its contents.

Colors are defined as 24-bit integers (R, G, B has 8 bits each), or using the pre-defined constants colred, colblack, colblue, colyellow, colgreen, colpurple.

Procedures:

openwindow	Opens the graphics window
drawline(x1, y1, x2, y2)	Draws a blue line.
drawstring(x1, y1, string,	Draws a text string.

color)	
drawvectorpoints(x, color)	Draws a set of points contained in the vector x, with indices on the horizontal axis.
drawxypoints(x, y, color)	Draws a set of points with x,y coordinates contained in the vectors x and y.
drawxysymbols(x, y, symbols)	Draws a set of points with x,y coordinates contained in the vectors x and y. The symbols vector contains integer values for symbols (colors).
drawhistogram(x, nbins, color)	Draws a histogram of the values contained in the vector x, with the number of bins given.

EXAMPLES

1. Take the first two columns of the spreadsheet, and display a log-log scatter diagram:

```
begin
  x:=spreadsheetcolumn(1);
  y:=spreadsheetcolumn(2);
  openwindow;
  drawxypoints(log(x), log(y), colblack);
end;
```

2. Carry out a Principal Components Analysis, with graphical and numerical output:

```
begin
  data:=spreadsheetarray(1);
  eig:=eigen(cov(data));
  eigvectors:=array(nrows(eig), ncols(eig)-1);
  for i:=1 to nrows(eig) for j:=1 to ncols(eig)-1
    eigvectors[i,j]:=eig[i,j];
  scores:=data*eigvectors;

  openwindow;
  drawxysymbols(column(scores,1), column(scores,2), spreadsheetsymbols(0));
  for i:=1 to nrows(scores)
    drawstring(scores[i,1], scores[i,2], spreadsheetrowlabel(i), colblack);

  opennumwindow(nrows(scores)+1, ncols(scores)+1);
  for i:=1 to nrows(scores) for j:=1 to ncols(scores)
    setnumwindow(i+1, j+1, scores[i,j]);

  for i:=1 to nrows(scores) setnumwindow(i+1, 1, spreadsheetrowlabel(i));
  for i:=1 to ncols(scores) setnumwindow(1, i+1, "Axis "&i);
end;
```