PAST - PAlaeontological STatistics

Øyvind Hammer, D.A.T. Harper, P.D. Ryan

May 21, 2001

1 Introduction

Welcome to the PAST! This program is designed as a follow-up to PALSTAT, an extensive package written by P.D. Ryan, D.A.T. Harper and J.S. Whalley (Ryan *et al.* 1995). It includes a number of functions which are commonly used in palaeontology and palaeoecology.

These days, a number of large and very good statistics systems exist, including SPSS and Excel. Why yet another statistics program?

- PAST is free.
- PAST is tailor-made for palaeontology. This means both that it includes functions which are not found in off-the-shelf programs (for example cladistics, ordination and geometrical analysis), and that it does not include functions which are of little use to palaeontologists and that only make the user interface more confusing.
- PAST is easy to use, and therefore well suited for introductory courses in quantitative palaeontology.
- PAST comes with a number of example data sets, case studies and exercises, making it a complete educational package.

Further explanations of many of the techniques implemented together with case histories are located in Harper (1999).

If you have questions, bug reports, suggestions for improvements or other comments, we would be happy to hear from you. Contact us at `ohammer@toyen.uio.no.` The PAST home page is

`http://www.toyen.uio.no/~ohammer/past`

2 Installation

The basic 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. The data files for the case studies can be downloaded separately, or together in the packed file ’casefiles.zip’. This file must be unpacked with a program such as WinZip.

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

Please note: Problems have been reported for some combinations of screen resolution and default font size in Windows - the layout gets ugly and 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. We are working on solving this problem.

PAST also seems to have problems with some printers. Postscript printers from HP and Tektronix work fine.

3 Entering and manipulating data

PAST has a spreadsheet-like user interface. Data are entered in 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 can only be done 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 almost all functions will expect numbers. Please note the decimal point convention which has been chosen by Windows depending upon your nationality: A comma (,) or a dot (full stop). 'Dotted' data can be 'commatized' from the Edit menu. Absence/presence data are coded as 0 or 1, respectively. Any other positive number will be interpreted as presence.

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.

Commatize

Converts all full stops ('.') in the data matrix to commas (','). This may be necessary for the program to read decimal points correctly, depending on your nationality.

Selecting areas

Most operations in PAST are carried only 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.

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 have selected the whole array, or a smaller area as appropriate.

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 'More rows' or 'More columns' in the Edit menu. When loading large data files, rows and/or columns are added automatically as needed.

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.

Cut, copy, paste

The cut, copy and paste functions are found in the Edit menu. Note that 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.

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 has a 'Copy graphic' button. This will place the graphical image in the paste buffer for pasting into e.g. Word or Corel Draw.

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 seven attractive colors using the ’Tag rows’ option in the Edit menu. Each group is also associated with a symbol (dot, cross, square, diamond, plus, circle, triangle). This is useful for showing different groups of data in e.g. ternary and scatter plots and dendrograms.

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.

Loading and saving data

The ’Load’ function is found in the File menu. PAST uses an ASCII file format, for easy importing from other programs (e.g. Word) and easy editing in a text editor. The format is 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, which means that you must never use spaces in row or column labels. ’Oxford Clay’ is thus an illegal column label which would confuse the program.

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 6 identifying the color (symbol), and another underscore.

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). Other data sets can thus be inserted both to the right of and below your existing data.

Reading and writing Nexus files

The Nexus file format is used by many cladistics programs. PAST can read and write the Data (character matrix) block of the Nexus format. Interleaved data are not 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).

4 Massaging your data

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 Massage menu log-transforms your data using the natural logarithm (base e):

$$y = \ln(x + 1)$$

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.

Subtract mean

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

Remove trend

This function removes any linear trend from a data set (two columns with X-Y pairs). This is done by subtraction of a linear regression line from the Y values. Removing the trend can sometimes be a useful operation prior to spectral analysis.

5 Plotting functions

Graph

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

XY graph

Plots two columns containing x/y coordinate pairs.

Histogram

Plots histograms (frequency distributions) for one or more columns. The number of bins is 10 by default, but can be changed by the user.

Ternary

Ternary plot for three columns of data, normally containing proportions of compositions.

Survivorship

Survivorship curves for one or more columns of data. The data will normally consist of age or size values. A survivorship plot shows the number of individuals which survived to different ages. Assuming exponential growth (highly questionable!), size should be log-transformed to age. This can be done either in the Massage menu, or directly in the Survivorship dialogue.

6 Simple statistics

Univariate statistics

Typical application	Assumptions	Data needed
Quick statistical description of a univariate sample	None, but variance and standard deviation are most meaningful for normally distributed data	Single column of measured or counted data

Displays the following statistics: Number of entries (N), smallest value (Min), largest value (Max), mean value (Mean), population variance (that is, the variance of the population estimated from the sample), sample variance (actual variance of just the sample), population and sample standard deviations (square roots of variance), median, skewness (positive for a tail to the right) and kurtosis (positive for a peaked distribution).

Diversity statistics

Typical application	Assumptions	Data needed
Quantifying taxonomical diversity in samples	Representative samples	One or more columns, each containing counts of individuals of different taxa down the rows

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
- Total number of individuals
- Dominance=1-Simpson index. Ranges from 0 (all taxa are equally present) to 1 (one taxon dominates the community completely).
- Simpson index=1-dominance. 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.
- Menhinick’s richness index - the ratio of the number of taxa to the square root of sample size.

- Margalef's richness index: $(S - 1)/\ln(n)$, where S is the number of taxa, and n is the number of individuals.

- 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 $S = \alpha \ln(1 + n/\alpha)$ where S is number of taxa, n is number of individuals and α is the Fisher's alpha.

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

Rarefaction

Typical application	**Assumptions**	**Data needed**
Comparing taxonomical diversity in samples of different sizes	Samples are taken from the same population (!)	Single column of counts of individuals of different taxa

Given a column 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. The algorithm is from Krebs (1989). An example application in paleontology can be found in Adrain *et al*. (2000).

7 Comparing data sets

There are many different standard tests available for comparing two distributions. Here is the standard disclaimer: You can never prove that two distributions are the same. A high probability value is only consistent with a similar distribution, but does of course give an indication of the similarity between the two sample distributions. On the other hand, a very low probability value does show, to the given level of significance, that the distributions are different.

Chi-square (one sample v. normal)

Typical application	Assumptions	Data needed
Testing for normal distribution of a sample	Large sample (N>30)	Single column of measured or counted data

Tests whether a single distribution (one selected column) is normal, by binning the numbers in four compartments. This test should only be used for relatively large populations (N>30). See Brown & Rothery (1993) or Davis (1986) for details.

Shapiro-Wilk (one sample v. normal)

Typical application	Assumptions	Data needed
Testing for normal distribution of a sample	Small sample (N<50)	Single column of measured or counted data

Tests whether a single distribution (one selected column) is normal. This test is designed for relatively small populations (N<50).

F and T tests (two samples)

Typical application	Assumptions	Data needed
Testing for equality of the variances and means of two samples	Normal or almost normal distribution	Two columns of measured or counted data

Two columns must be selected. The F test compares the variances of two distributions, while the t test compares their means. The F and t statistics, and the probabilities that the variances and means of the parent populations are the same, are given. The F and t tests should only be used if you have reason to believe that the parent populations are close to normally distributed. The Chi-square test for one distribution against a normal distribution can give you an idea about this.

Also, the T test is really only applicable when the variances are the same. So if the F test says otherwise, you should be cautious about the T test. An unequal variance T statistic is also given, which should be used in this case.

Sometimes publications give not the data, but values for sample size, mean and variance for two populations. These can be entered manually using the ’F and T from parameters’ option in the menu.

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

How do I test lognormal distributions?

All of the above tests apply to lognormal distributions as well. All you need to do is to transform your data first, by taking the log transform in the Massage menu. You might want to ’backup’ your data column first, using Copy, and then get your original column back using Paste.

Chi-square (two samples)

Typical application	**Assumptions**	**Data needed**
Testing for equal distribution of compartmentalized, counted data	Each compartment containing at least five individuals	Two columns of counted data in different compartments (rows)

The Chi-square test is the one to use if your data consist of the 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 a little cautious about such comparisons if any of the bins contain less than five individuals.

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

Mann-Whitney (two samples)

Typical application	**Assumptions**	**Data needed**
Comparing the medians of two samples	None	Two columns of measured or counted data

Two columns must be selected. The (Wilcoxon) Mann-Whitney U test can be used to test whether the medians of two independent distributions are different. This test is non-parametric, which means that the distributions can be of any shape.

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

Kolmogorov-Smirnov (two samples)

Typical application	**Assumptions**	**Data needed**
Comparing the distributions of two samples	None	Two columns of measured data

Two columns must be selected. The K-S test can be used to test whether two independent distributions of continuous, unbinned numerical data are different. The K-S test is non-parametric, which means that the distributions can be of any shape. If you want to test just the locations of the distribution (medians), you should rather use the Mann-Whitney U test.

See Davis (1986) for details.

Spearman's rho and Kendall's tau (two samples)

Typical application	Assumptions	Data needed
Testing whether two variables are correlated	None	Two columns of measured or counted paired data (such as *x/y* pairs)

These non-parametric rank-order tests are used to test for correlation between two variables.

Dice and Jaccard similarity indices

Typical application	Assumptions	Data needed
Comparing two or more presence/absence samples	Equal sampling conditions	Two or more columns of presence/absence (1/0) data with taxa down the rows

The Dice and Jaccard similarity indices are used to compare associations, limited to absence/presence data (any positive number is interpreted as presence). When comparing two columns (associations), a match is counted for all taxa with presences in both columns. Using 'M' for the number of matches and 'N' for the the total number of taxa with presences in just one column, we have

Dice similarity = 2M / (2M+N)

Jaccard similarity = M / (M+N)

Both these indices range from 0 (no similarity) to 1 (identity). A matrix is presented with the comparisons between all pairs of associations. Dice indices are given in the upper triangle of the matrix (above and to the right of the diagonal), and Jaccard indices are given in the lower.

See Harper (1999) for details.

Raup-Crick similarity index

Typical application	Assumptions	Data needed
Comparing two or more presence/absence samples	Equal sampling conditions	Two or more columns of presence/absence (1/0) data with taxa down the rows

The Raup-Crick similarity index is used to compare associations, limited to absence/presence data (any positive number is interpreted as presence). This index ranges from 0 (no similarity) to 1 (identity). A matrix is presented with the comparisons between all pairs of associations.

The Raup-Crick 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 200 random replicates.

Correlation matrix

Typical application	**Assumptions**	**Data needed**
Quantifying correlation between two or more variables	Normal distribution	Two or more columns of measured or counted variables

A matrix is presented with the correlations between all pairs of columns. Correlation values (Pearson's r) are given in the upper triangle of the matrix, and the probabilities that the columns are uncorrelated are given in the lower.

Contingency table analysis

Typical application	**Assumptions**	**Data needed**
Testing for dependence between two variables	None	Matrix of counted data in compartments

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. A measure and probability of association of the two variables (based on Chi-square) is then given.

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. For details, see Press *et al.* (1992).

One-way ANOVA

Typical application	**Assumptions**	**Data needed**
Testing for equality of the means of several univariate samples	Normal distribution and similar variances and sample sizes	Two or more columns of measured or counted data

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

The data sets are required to be close to normally distributed.
See Brown & Rothery (1993) or Davis (1986) for details.

8 Multivariate statistics

Principal components analysis

Typical application	Assumptions	Data needed
Reduction and interpretation of large multivariate data sets with some underlying linear structure	Debated	Two or more rows of measured data with three or more variables

Principal components analysis (PCA) is a procedure for finding hypothetical variables (components) which account for as much of the variance in your multidimensional data as possible (Davis 1986, Harper 1999). These new variables are linear combinations of the original variables. PCA has several applications, two of them are:

- Simple reduction of the data set to only two variables (the two most important components), for plotting and clustering purposes.
- More interestingly, you might try to hypothesize that the most important components are correlated with some other underlying variables. For morphometric data, this might be simply age, while for associations it might be a faunal gradient (e.g. latitude or position across the shelf).

The PCA routine finds the eigenvalues and eigenvectors of the variance-covariance matrix or the correlation matrix. Choose var-covar if all your variables are measured in the same unit (e.g. centimetres). Choose correlation (normalized var-covar) if your variables are measured in different units; however, all variables will be normalized. The eigenvalues, giving a measure of the variance accounted for by the corresponding eigenvectors (components) are given for the first four most important components (or fewer if there are fewer than four variables). 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.

The ’View scatter’ option allows you to see all your data points (rows) plotted in the coordinate system given by the two most important components. If you have tagged (grouped) rows, the different groups will be shown using different symbols and colours.

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.

Bruton & Owen (1988) describe a typical morphometrical application of PCA.

Principal coordinates

Typical application	Assumptions	Data needed
Reduction and interpretation of large multivariate data sets with some underlying linear structure	Unknown	Two or more rows of measured data with three or more variables

Principal coordinates analysis (PCO) is another ordination method, somewhat similar to PCA. The algorithm is taken from Davis (1986).

The PCO routine finds the eigenvalues and eigenvectors of a matrix containing the distances between all data points. You can choose between the Gower distance measure or the Euclidean distance. The Gower measure will normally be used - Euclidean distance gives results similar to PCA. 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 'View scatter' option allows you to see all your data points (rows) plotted in the coordinate system given by the PCO. If you have tagged (grouped) rows, the different groups will be shown using different symbols and colours.

Correspondence analysis

Typical application	Assumptions	Data needed
Reduction and interpretation of large multivariate ecological data sets with environmental or other gradients	Unknown	Two or more rows of counted data in three or more compartments

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. The algorithm is taken from Davis (1986).

The CA routine finds the eigenvalues and eigenvectors of a matrix containing the Chi-squared distances between all data points. The eigenvalues, giving a measure of the similarity accounted for by the corresponding eigenvectors, are given for the first four most important eigenvectors (or fewer if there are fewer than four variables). 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 tagged (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.

Detrended correspondence analysis

Typical application	Assumptions	Data needed
Reduction and interpretation of large multivariate ecological data sets with environmental or other gradients	Unknown	Two or more rows of counted data in three or more compartments

The Detrended Correspondence (DCA) module uses the same algorithm as Decorana (Hill & Gauch 1980). It is specialized for use on ’ecological’ data sets with abundance data (taxa in rows, localities in columns). When the ’Detrending’ option is switched off, a basic Reciprocal Averaging will be carried out. The result should be similar to Correspondence Analysis (see above) plotted on the second and third axes.

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.

Cluster analysis

Typical application	Assumptions	Data needed
Finding hierarchical groupings in multivariate data sets	None	Two or more rows of counted, measured or presence/absence data in one or more variables or categories

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), single linkage (nearest neighbour) and Ward’s method. One 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 eight different measures:

- The Euclidean distance (between rows) is a robust and widely applicable measure.
- Correlation (of the variables along rows) using Pearson's r. A little meaningless if you have only two variables.
- Correlation using Spearman's rho (basically the r value of the ranks). Will often give the same result as correlation using r.
- Dice coefficient for absence-presence (coded as 0 or positive numbers). Puts more weight on joint occurences than on mismatches.
- Jaccard coefficient for absence-presence data.
- Bray-Curtis measure for abundance data.
- Chord distance for abundance data. Recommended!
- Morisita's index for abundance data. Recommended!
- Raup-Crick index for absence-presence data. Recommended!

See Harper (1999) or Davis (1986) for details.

Seriation

Typical application	**Assumptions**	**Data needed**
Stratigraphical or environmental ordering of taxa and localities	None	Presence/absence (1/0) matrix with taxa in rows

Seriation of an absence-presence matrix using the algorithm described by Brower and Kyle (1988). This method is typically applied to an association matrix with taxa (species) in the rows and populations 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' biozonation, 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 informnative 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.

Discriminant analysis

Typical application	Assumptions	Data needed
Testing for separation of multivariate data sets	Multivariate normality	Two multivariate data sets of measured data, marked with different colors

Given two sets of multivariate data, an axis is constructed which maximizes the difference between the sets. The two sets are then plotted along this axis using a histogram.

This module expects the rows in the two data sets to be tagged with dots (black) and crosses (red), respectively. The histogram may not show the entire discriminant axis, so the start and end values for the histogram may have to be set manually.

Equality 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.

Discriminant analysis is a standard method for visually confirming or rejecting the hypothesis that two species are morphologically distinct.

See Davis (1986) for details.

9 Fitting data to functions

Linear

Typical application	Assumptions	Data needed
Fitting data to a straight line, or exponential or power function	None	Two columns of counted or measured data

Two columns must be selected (x and y values). A straight line $y = ax + b$ is fitted to the data. There are two different algorithms available: Standard regression and Reduced Major Axis (the latter is selected by ticking the box). Standard regression keeps the x values fixed, and finds the line which minimizes the squared errors in the y values. Use this if your x values have very small errors associated with them. Reduced Major Axis tries to minimize both the x and the y errors.

Also, both x and y values can be log-transformed, 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, Pearson's r correlation, and the probability that the columns are not correlated are given.

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 Massage menu) and then performing a straight-line fit.

Sinusoidal

Typical application	Assumptions	Data needed
Fitting data to a set of periodic, sinusoidal functions	None	Two columns of counted or measured data

Two columns must be selected (x and y values). A sum of up to six 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 and 1/6 of the fundamental period). These values can be changed, and need not be in harmonic proportion.

With a little effort, frequencies can also be estimated by trial and error, by adjusting the frequency so that amplitude is maximized (this procedure is difficult with more than a single sinusoidal).

It is not meaningful to specify periodicities that are smaller than two times the typical spacing of data points.

Logistic

Typical application	Assumptions	Data needed
Fitting data to a logistic or von Bertalanffy growth model	None	Two columns of counted or measured data

Attempts to fit the data to the logistic equation $y = a/(1 + b * e^{-cx})$. For numerical reasons, the x axis is normalized. The algorithm is a little complicated. The value of a is first estimated to be the maximal value of y. The values of b and c are then estimated using a straight-line fit to a linearized model.

Though acceptable, this estimate can optionally be improved by using the estimated values as an initial guess for a Levenberg-Marquardt nonlinear optimization (tick the box). This procedure can sometimes improve the fit, but due to the numerical instability of the logistic model it often fails with an error message.

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.

Von Bertalanffy

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

B-splines

Typical application	Assumptions	Data needed
Smoothing noisy data	None	Two columns of counted or measured data

Two columns must be selected (x and y values). The data are fitted with a least-squares criterion to a B-spline, which is a sequence of third-order polynomials, continuous up to the second derivative. A typical application of this is the construction of a smooth curve going through a noisy data set.

A decimation factor is set by the user, and controls how many data points contribute to each polynomial section. Larger decimation gives a smoother curve.

Note that sharp jumps in your data can give rise to oscillations in the curve, and that you can also get large excursions in regions with few data points.

10 Time series analysis

Spectral analysis

Typical application	**Assumptions**	**Data needed**
Finding periodicities in counted or measured data	Time series long enough to contain at least four cycles	One or two columns of counted or measured data

Two columns must be selected (x and y values). Since palaeontological data are often unevenly sampled, the FFT algorithm can be difficult to use. PAST therefore includes the Lomb periodogram algorithm for unevenly sampled data, with time values given in the first column.

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.

You may want to remove any linear trend in the data (Edit menu) before applying the Lomb periodogram. Failing to do so can produce annoying peaks at low frequencies.

Autocorrelation

Typical application	**Assumptions**	**Data needed**
Finding periodicities in counted or measured data	Time series long enough to contain at least two cycles. Even spacing of data points.	One column of counted or measured data

Autocorrelation (Davis 1986) is carried out on separate column(s) of evenly sampled temporal/stratigraphic data. Lag times 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.

11 Geometrical analysis

Directional analysis

Typical application	Assumptions	Data needed
Displaying and testing for random distribution of directional data	See below	One column of directional data in degrees (0-360)

Plots a rose diagram (polar histogram) of directions given in a column of degree values (0 to 360). Used for plotting current-oriented specimens, orientations of trackways, orientations of morphological features (e.g. terrace lines), etc.

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 mean angle, together with the R value (Rayleigh's spread), are given. 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 distributed data - the test is not appropriate for bidirectional data. Also, the test is not accurate for N>50; it will then report a too high p value.

A four-bin chi-square test is also available, giving the probability that the directions are randomly and evenly distributed.

Point distribution

Typical application	Assumptions	Data needed
Testing for clustering or overdispersion of two-dimensional position values	Elements small compared to their distances, mainly convex domain, N>50.	Two columns of *x/y* positions

Point distribution statistics using nearest neighbour analysis (modified from Davis 1986). The area is estimated using the convex hull, which is the smallest convex polygon enclosing the points. This is inappropriate for points in very concave domains. Also, there is no correction for boundary effects, meaning that the statistics are reasonably valid only for large N (N>50).

The probability that the distribution is random (Poisson process, giving an exponential nearest neighbour distribution) is presented, together with the R value. Clustered points give R<1, Poisson patterns give R 1, while overdispersed points give R>1.

Applications of this module include spatial ecology (are in-situ brachiopods clustered) and morphology (are trilobite tubercles overdispersed).

Fourier shape analysis

Typical application	Assumptions	Data needed
Analysis of fossil outline shape	Shape expressible in polar coordinates, sufficient number of digitized points to capture featues.	Two columns of digitized *x/y* positions around an outline

Accepts $X-Y$ coordinates digitized around an outline. More than one shape can be simultaneously analyzed by giving more than one pair of columns. 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 origin for the polar coordinate system is found by numerical approximation to the centroid. 64 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 ten 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).

Elliptic Fourier shape analysis

Typical application	Assumptions	Data needed
Analysis of fossil outline shape	Sufficient number of digitized points to capture featues.	Two columns of digitized *x/y* positions around an outline

Elliptic Fourier shape analysis is in some 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 in Ferson et al. (1985).

Cosine and sine components of x and y increments along the outline for the first 10 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 coefficients can be copied to the main spreadsheet for further analysis (e.g. by PCA).

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

12 Cladistics

Typical application	Assumptions	Data needed
Semi-objective analysis of relationships between taxa from morphological or genetic evidence	Many! See Kitchin *et al.* (1998)	Character matrix with taxa in rows, outgroup in first row

The cladistics package in PAST is fully operational, but lacking in comprehensive functionality. For example, there is no character reconstruction (plotting of steps on the cladogram). This means that PAST could be used for educational purposes and for initial data exploration, but perhaps not for more 'serious' work. Maybe in a later version?

Algorithms are from Kitchin et al. (1998).

Parsimony analysis

Character states should be coded using integers in the range 0 to 255. 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 three 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 1000 trees are saved. You should not use the branch-and-bound algorithm 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 Kitchin 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.

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.

Character optimization criteria

Three different optimality 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.

Dollo

Characters are irreversible and 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 1000 trees. If bootstrapping has been performed, a bootstrap value is given at the root of the subtree specifying each group.

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 percent of the trees).

13 Unitary associations

Typical application	Assumptions	Data needed
Quantitative biostratigraphical correlation	None	Presence/absence (1/0) matrix with horizons in rows, taxa in columns

Unitary Associations analysis (Guex 1991) is a method for biostratigraphical correlation (See Angiolini & Bucher 1999 for an example 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) are tagged with 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 (see alveolinid.dat for an example).

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 does not include all the features found in the standard program, called BioGraph (Savary & Guex 1999), and advanced users are referred to that package. 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 in all levels between the first and last appearance in any section. Then, any samples with a set of taxa that is contained in some other 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 relationship: 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.

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.

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 (PAST does not carry out this procedure automatically in the present version). The reproducibility matrix should be inspected to identify such unitary associations.

14 Acknowledgments

PAST was inspired by and includes many functions found in PALSTAT, which was programmed by P.D. Ryan with assistance from J.S. Whalley. Harper thanks the Danish Science Research Council (SNF) for support.

15 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.

Angiolini, L. & H. Bucher 1999. Taxonomy and quantitative biochronology of Guadalupian brachiopods from the Khuff Formation, Southeastern Oman. *Geobios* 32:665-699.

Brower, J.C. & K.M. Kyle 1988. Seriation of an original data matrix as applied to palaeoecology. *Lethaia* 21:79-93.

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

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.

Ferson, S.F., F.J. Rohlf & R.K. Koehn 1985. Measuring shape variation of two-dimensional outlines. *Systematic Zoology* 34:59-68.

Guex, J. 1991. Biochronological Correlations. Springer Verlag.

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

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

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

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

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

Raup, D. & R.E. Crick. 1979. Measurement of faunal similarity in paleontology. *Journal of Paleontology* 53:1213-1227.

Ryan, P.D., Harper, D.A.T. & Whalley, J.S. 1995. PALSTAT, Statistics for palaeontologists. Chapman & Hall (now Kluwer Academic Publishers).

Savary, J. & J. Guex. 1999. Discrete Biochronological Scales and Unitary Associations: Description of the BioGraph Computer Program. *Memoires de Geologie (Lausanne)* 34.

Sepkoski, J.J. 1984. A kinetic model of Phanerozoic taxonomic diversity. *Paleobiology* 10:246-267.