PAleontological STatistics

Version 2.16

Manual de Referência[1]

Øyvind Hammer
Natural History Museum
University of Oslo
http://folk.uio.no/ohammer/past/

1999-2012

[1] Tradução feita por Pavel Dodonov – pdodonov@gmail.com; revisado por Matheus Gonçalves dos Reis. Ambos doutorandos do Programa de Pós-Graduação em Ecologia e Recursos Naturais, Universidade Federal de São Carlos (UFSCar).

Sumário

Bem-Vinda(o) ao PAST!

Este programa foi inicialmente desenvolvido como uma sequência do PALSTAT, um pacote para análise de dados paleontológicos que foi escrito por P. D. Ryan, D. A. T. Harper e J. S. Whalley (Ryan et al. 1995).

Através de um desenvolvimento contínugo ao longo de mais de dez anos, o PAST cresceu e se tornou um pacote estatístico abrangente, usado não só por paleontólogos, mas também em muitas áreas das ciências da vida, ciências da terra e até mesmo engenharia e economia.

Explicações mais detalhadas de muitas das técnicas implementadas juntamente com estudos de caso podem ser encontradas em Harper (1999). Além disso, o livro "Palaeontological Data Analysis" (Hammer & Harper 2005) pode ser visto como um livro-companheiro do PAST.

Se você tiver perguntas, relatos de defeitos no programa (*bugs*), sugestões para melhorias ou outros comentários, nós ficaríamos felizes em ouvir você. Contacte-nos em ohammer@nhm.uio.no. Para relatos de defeitos no programa, lembre-se de mandar os dados usados, como salvos pelo PAST, juntamente com uma descrição completa das ações que levaram ao problema.

A última versão do PAST, juntamente com a documentação e um link para a lista de emails do PAST, podem ser encontrados em

http://folk.uio.no/ohammer/past

Nós seremos gratos se você citar o PAST em publicações científicas. A referência oficial é Hammer et al. (2001).

Referências

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.

Instalação

Instalar o PAST é fácil: apenas baixe o arquivo “Past.exe” e coloque-o em qualquer lugar do seu disco rígido. Clicando duas vezes no arquivo vai abrir o programa. O Windows irá considerar isso uma quebra de segurança e perguntar se você confia no provedor do programa. Se você quiser usar o programa, terá que responder que sim.

Nós sugerimos que você crie uma pasta chamada PAST em qualquer lugar do seu disco rígido e coloque nela todos os arquivos.

Note: Foram relatados alguns problemas referentes a tamanhos de fonte diferentes do padrão usados por definição (*non-standard default font size*) no Windows – o usuário pode ter que aumentar o tamanho das janelas para ver todo o texto e botões. Caso isto aconteça, por favor ajuste o tamanho da fonte para “Small fonts” (“fontes pequenas”) no painel Controle de tela (Screen control) do Windows.

Ao sair do PAST, um arquivo chamado “pastsetup” será automaticamente colocado na sua pasta de usuário (por exemplo “Meus documentos” no Windows 95/98). Este arquivo contém os últimos diretórios de arquivos que foram usados.

A ausência de uma instalação “formal” para o Windows é intencional, permitindo instalar o PAST sem ter privilégios de administrador.

A planilha e o menu *Edit* (Editar)

O PAST tem uma interface de usuário em formato de planilha. Dados são inseridos como uma matriz de células, organizada em linhas (horizontalmente) e colunas (verticalmente).

Inserindo dados

Para inserir dados em uma célula, clique na célula com o *mouse* e digite os dados. Isto só é possível quando o programa está no *"Edit mode"* ("Modo de edição"). Para selecionar o modo de edição, marque a caixa *Edit mode* acima da matriz. Quando o modo de edição estiver desligado, a matriz é bloqueada e os dados não podem ser alterados. Também é possível navegar pelas células com as teclas de seta.

Qualquer texto pode ser inserido nas células, mas a maior parte das funções espera números. Tanto a vírgula (,) quanto o ponto (.) são interpretados como separadores decimais.

Dados de presença/ausência são codificados como 0 e 1, respectivamente. Qualquer outro número positivo será interpretado como presença. Matrizes de presença/ausência podem ser visualizadas como quadrados pretos para presença ao escolher a opção "*Square mode*" ("Modo de quadrados") acima da matriz.

Dados de sequências genéticas são codificados por C, A, G, T e U (letras minúsculas também são aceitas).

Dados ausentes (*missing data*) são codificados por pontos de interrogação ("?") ou pelo valor -1. A não ser que a documentação para uma função fale explicitamente que há suporte para dados faltantes, a função ***não vai lidar corretamente com os dados ausentes***, então tome cuidado.

A convenção no PAST é que itens ocupam linhas e variáveis ocupam colunas. Três indivíduos de Brachiopoda podem então ocupar as linhas 1, 2 e 3, com seus comprimentos e larguras nas colunas A e B. Análise de agrupamento (*cluster*) sempre vai agrupar itens, ou seja, linhas. Para análise de associação de modo Q (*Q-mode analysis of association*), amostras (sítios) devem ser inseridas nas linhas e os táxons[2] (espécies) nas colunas. Para alternar entre modos Q e R, linhas e colunas podem ser facilmente intercambiadas usando a operação *Transpose* (transpor).

Selecionando áreas

A maior parte das operações no PAST só é feita em uma área da matriz que você tenha ***escolhido*** (marcado). Se você precisar rodar uma função que requer dados e nenhuma área estiver selecionada, você receberá uma mensagem de erro.

[2] Embora o plural de táxon seja taxa, traduzi como táxons para evitar confusões com taxas no sentido de frequência. (NT)

- Uma linha é selecionada clicando-se no rótulo de linha (*row label*, a coluna da extrema esquerda).
- Uma coluna é selecionada clicando-se no rótulo de coluna (*column label*, a linha superior).
- Linhas múltiplas são selecionadas clicando no rótulo da primeira linhas, segurando a tecla Shift e clicando nos rótulos das linhas adicionais. Note que você não pode selecionar as linhas "clicando e arrastando" – isso vai mover a primeira coluna (veja abaixo).
- Colunas múltiplas são selecionadas de modo similar, segurando Shift e clicando nos rótulos das colunas adicionais.
- A matriz inteira pode ser selecionada clicando no canto superior esquerdo da matriz (a célula cinza vazia) ou escolhendo a opção "*Select all*" ("Selecionar tudo") no menu *Edit* (Editar).
- Áreas menores dentro da matriz podem ser selecionadas "clicando e arrastando", mas isso só funciona quando o modo de edição (*Edit mode*) está desligado.

Importante: Infelizmente, não é possível escolher várias colunas que não sejam vizinhas. Isso quer dizer que caso você queria, por exemplo, rodar uma análise apenas na primeira e terceira coluna, você primeiro terá que mover as colunas para que elas fiquem juntas.

Movendo uma linha ou uma coluna

Uma linha ou uma coluna (incluindo o seu rótulo) pode ser movida simplesmente clicando no rótulo e arrastando para uma nova posição.

Renomeando linhas e colunas

Quando o PAST inicia, as linhas são numeradas de 1 a 99 e as colunas de A a Z. Para a sua própria referência e para uma rotulagem apropriada dos gráficos, você deveria dar às linhas e colunas nomes mais descritivos mas ainda assim curtos. Selecione a opção *"Rename columns"* ("Renomear colunas) ou *"Rename rows"* ("Renomear linhas") no menu *Edit* (Editar). Você deve selecionar a matriz inteira ou uma área menor, como for apropriado.

Uma outra forma é escolher a opção *"Edit labels"* ("Editar rótulos") acima da planilha. A primeira linha e a primeira coluna agora podem ser editadas como o resto das células.

Aumentando o tamanho da matriz

Por definição, o PAST tem 99 linhas e 26 colunas. Caso você precise de mais, você pode adicionar linhas ou colunas escolhendo as opções *"Insert more rows"* ("Inserir mais linhas") ou *"Insert more columns"* ("Inserir mais colunas") no menu *Edit* (Editar). Linhas/colunas serão inseridas depois da área marcada ou abaixo/à direita da matriz se nenhuma área estiver selecionada. Linhas e/ou colunas são adicionadas automaticamente quando um conjunto de dados grande é carregado.

Recortar, copiar, colar

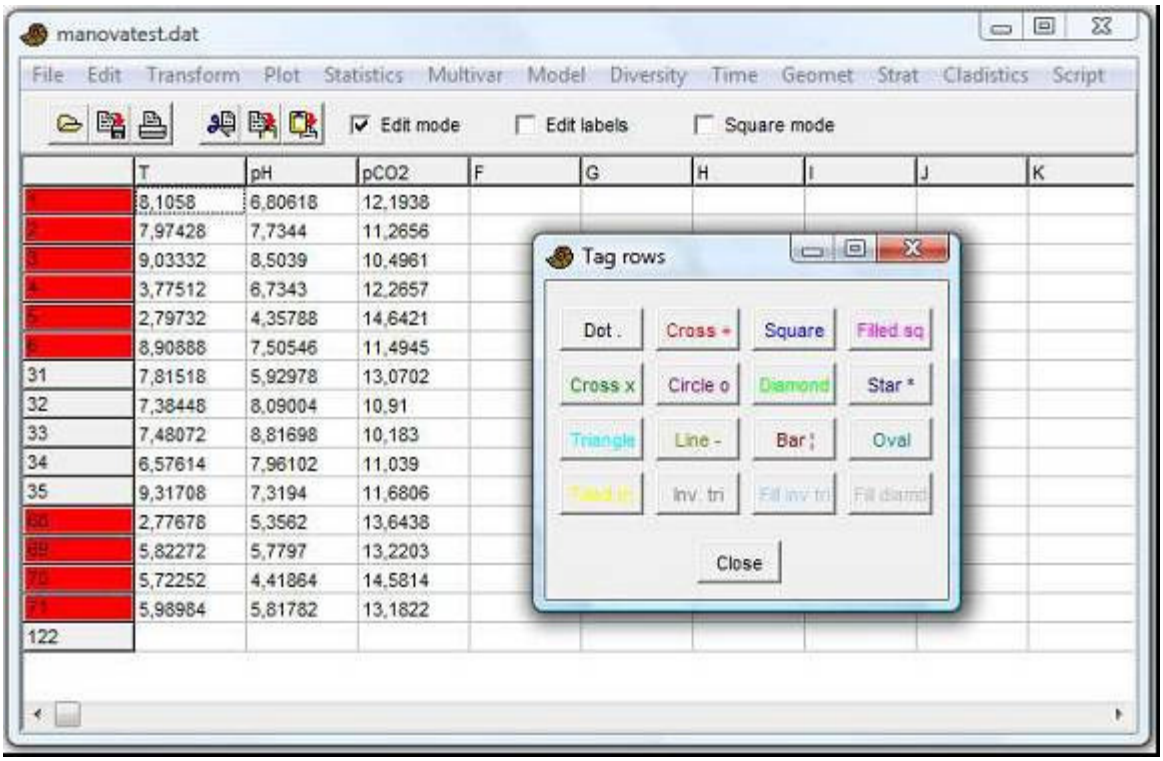
As opções para recortar, copiar e colar são encontradas no menu *Edit* (Editar). Você pode recortar/copiar dados da planilha do PAST e os colar em outros programas, por exemplo Word e Excel. Similarmente, dados de outros programas podem ser colados na planilha do PAST, contanto que estejam em formato de texto separado por tabulações.
Lembre-se que blocos locais de dados (sem serem todas as linhas ou colunas) só podem ser marcadas quando o modo de edição (*"Edit mode"*) está desligado.
Todos os módulos com *output* gráfico possuem um botão "Copiar gráfico" ("*Copy graphic*"). Este irá colocar a imagem do gráfico na área de colagem de modo que ele possa ser colado em outros programas, por exemplo um programa de desenho para edição da imagem. Gráficos são copiados no formato "*Enhanced Metafile Format*" (EMF) no Windows. Isso permite a edição de elementos individuais da imagem em outros programas. Ao colar o gráfico no Coreldraw, você precisa escolher "Colar especial" (*"Paste special"*) no menu Editar e escolher "*Enhanced metafile*". Alguns programas podem ter formas idiossincráticas de interpretar imagens EMF – cuidado com coisas engraçadas acontecendo.

Remover

A função *remover* (remover) no menu *Edit* (Editar) permite que você remova da planilha a(s) linha(s) ou coluna(s) escolhida(s). A área removida não é copiada para a área de colagem.

Agrupando (colorindo) colunas

Linhas (pontos com dados) podem ser coloridas com uma dentre 16 cores atraentes, usando a opção "*Row color/symbol*" ("Cor/símbolo de linha") no menu *Edit*. Cada grupo é também associado a um símbolo (ponto, X, quadrado, diamante, +, círculo, triângulo, linha, barra, círculo cheio, estrela, oval, triângulo cheio, triângulo invertido, triângulo invertido cheio, diamante cheio). Isso é útil para mostrar diferentes conjuntos de dados nos gráficos e é necessário para uma série de métodos de análise.
***Importante*:** Para métodos que requerem agrupamento de linhas por meio de cores, as linhas que pertencem ao mesmo grupo precisam ser consecutivas. Se são necessários mais de 16 grupos, as cores podem ser reutilizadas. No exemplo abaixo, três grupos foram marcados corretamente.



A opção *"Numbers to colors"* ("Números para cores") no menu *Edit* permite que números 1-16 em uma coluna selecionada atribuam a cor (símbolo) correspondente às colunas.

Escolhendo tipos de dados para as colunas

As colunas selecionadas podem ser marcadas com um tipo de dado (contínuo/não-especificado, ordinal, nominal ou binário – *continuous/unspecified, ordinal, nominal or binary*) usando a opção *"Column data types"* ("Tipos de dados da coluna") no menu *Edit*. Isto só é necessário se você deseja utilizar medidas mistas de similaridade/distância.

Remover linhas/colunas não-informativas – Remove uninformative rows/columns

Linhas ou colunas podem ser não-informativas especialmente no que diz respeito às análises multivariadas. Alguns tipos podem ser buscados e removidos: linhas ou colunas apenas com zeros, linhas ou colunas apenas com dados ausentes ("?") e linhas ou colunas com apenas uma célula diferente de zero (*singletons*).

Transpor - Transpose

A função *Transpose* (Transpor), no menu *Edit*, irá intercambinar linhas e colunas. Isto é usado para alternar entre modos R e Q nas análises de agrupamento (*cluster*), componentes principais (*principal components analysis*) e seriação (*seriation*).

Colunas agrupadas para dados multivariados – Grouped columns to multivar

Converte de um formato com dados multivariados apresentados em grupos consecutivos de *N* colunas para o formato do PAST, com um item por linha e todas as variáveis (*variates*) ao longo das colunas. Para *N*=2, dois espécies e quatro variáveis a-d, a conversão é de

a_1 b_1 a_2 b_2
c_1 d_1 c_2 d_2

para

a_1 b_1 c_1 d_1
a_2 b_2 c_2 d_2

Linhas agrupadas para multivariado – Grouped rows to multivar

Converte de um formato em que itens são apresentados em grupos consecutivos de *N* linhas para o formato do PAST, com um item por linha e todas as variadas (*variates*) ao longo das colunas. Para *N*=2, dois espécies e quatro variáveis a-d, a conversão é de

a_1 b_1
c_1 d_1
a_2 b_2
c_2 d_2

para

a_1 b_1c1 d_1
a_2 b_2 c_2 d_2

Empilhar linhas agrupadas em colunas – Stack colored rows into columns

Empilha horizontalmente grupos coloridos ao longo das colunas. Isso pode ser útil, por exemplo, para efetuar estatística univariada em pares de colunas entre grupos.

Eventos para amostras – Events to samples (RASC to UA)

Espera uma matriz de dados com seções/poços em linhas e táxons em colunas, com valores de FAD e LAD em colunas alternando (ou seja, duas colunas por táxon). Converte para o formato de presença/ausência de Associações Unitárias (*Unitary Associations*) com seções em grupos de linhas, amostras em linhas e táxons em colunas.

Carregando e salvando dados

A função *"Open"* ("Abrir") se encontra no menu *File* (Arquivo). Você também pode arrastar um arquivo da área de trabalho (*desktop*) para dentro da janela do PAST. O PAST utiliza um formato de texto fácil de importar de outros programas, como segue:

.	rótulo_de_coluna	rótulo_de_coluna	rótulo_de_coluna
rótulo_de_linha	dados	dados	dados
rótulo_de_linha	dados	dados	dados
rótulo_de_linha	dados	dados	dados

Células vazias (como a célula do topo à esquerda) são codificadas com um ponto (.). Células são separadas por espaço em branco. Se uma célula contém caracteres de espaço, ela precisa ser envolta em colchetes duplos, por exemplo "Argila de Oxford".
Caso a alguma célula tenha sido atribuída uma cor diferente do preto, o rótulo da linha no arquivo vai começar *underline*, um número de 0 a 15 indicando a cor (símbolo), e outro *underline*.
Caso a alguma coluna de dados tenha sido atribuído um formato que não seja o contínuo/não-especificado (*continuous/unspecified*), os rótulos das colunas no arquivo irão similarmente começar com um *underline*, um número de 0 a 3 identificando o tipo de dados (0=contúnuo/não-especificado, 1=ordinal, 2=nominal, 3=binário), e um segundo *underline*.
Adicionalmente a este formato, o PAST também consegue detectar e abrir arquivos nos seguintes formatos:

- Excel (apenas a primeira planilha)
- Nexus (veja abaixo), popular em Sistemática
- formato TPS desenvolvido por Rohlf. Os campos *landmark*, *outlines*, *curves*, *id*, *scale* e comentário têm suporte, os outros campos são ignorados
- NTSYS. Tabelas múltiplas e árvores não têm suporte. O arquivo precisa ter a extensão ".nts".
- formato de sequência molecular FASTA, especificação simplificada de acordo com NCBI.
- formato de sequência molecular PHYLIP. O arquivo precisa ter a extensão ".phy".
- formato de sequência molecular Arlequin. Para dados de genótipo os dois haplótipos são concatenados para uma única linha. Nem todas as opções têm suporte.
- formato BioGraph para bioestratigrafia (formatos SAMPLES e DATUM). Se um segundo arquivo com o mesmo nome e a extensão ".dct" for encontrado, ele será incluído como um dicionário do BioGraph.
- formato RASC para bioestratigrafia. Você precisa abrir o arquivo .DAT. O programa espera arquivos .DIC e .DEP correspondentes no mesmo diretório.
- formato CONOP para bioestratigrafia. Você precisa abrir o aquivo .DAT (*log file*). O programa espera arquivos .EVT (*event*) e .SCT (*section*) correspondentes no mesmo diretório.

A função *"Insert from file"* ("Inserir do arquivo") é útilo para concatenar conjuntos de dados. O arquivo carregado será inserido na sua planilha existente na posição escolhida (esquerda superior).

Importando dados do Excel

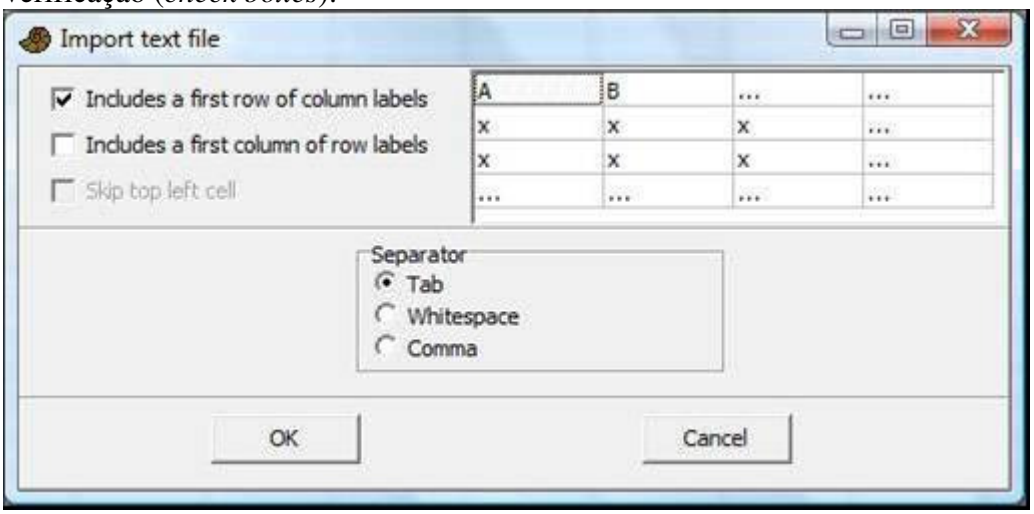
- Copie do Excel e cole dentro do PAST. Note que se você quiser que a primeira linha e coluna sejam copiadas nas células de rótulos do PAST, você precisa deixar a opção *"Edit labels"* ("Editar rótulos") ligada. Ou,
- abra o arquivo do Excel pelo PAST. A opção *"Edit labels"* opera do mesmo modo. Ou,
- deixe a primeira célula no Excel com um único ponto (.) e salve como texto separado por tabulações (*tab-separated text*) no Excel. O arquivo resultante pode ser aberto diretamente pelo PAST.

Lendo e escrevendo arquivos Nexus

O formato de arquivo Nexus é usado por muitos programas de sistemática. PAST consegue ler e escrever os blocos de Dados (matriz de caracteres) do formato Nexus. Também há suporte para *interleaved data*. Além disso, caso você tenha realizado uma análise de parsimônia e a janela *"Parsimony analysis"* se encontra aberta, todas as árvores mais curtas serão incluídas no arquivo Nexus para processamento adicional em outros programas (e.g. MacClade ou Paup). Note que no momento não há suporte para todas as opções do Nexus.

Importando arquivos de texto

Arquivos de texto separados por espaços em branco, tabulações ou vírgulas pode ser lidos usando a opção *"Import text file"* ("Importar arquivo de texto") no menu *File*. A planilha na janela ilustra o formato do arquivo a ser aberto como especificado pela caixas de verificação (*check boxes*).



Contador – Counter

Uma função de contagem (*counter function*) está disponível no menu *Edit* para usar, por exemplo, no microscópio durante a contagem de fósseis de diferentes táxons. Uma única linha (amostra) deve ser selecionada. Uma janela de contagem irá abrir com um número de contadores, um para cada coluna (táxon) selecionada. Os contadores serão inicializados com os rótulos das colunas e qualquer contagem que já esteja presente na planilha. Ao fechar a janela do contador, os valores na planilha serão atualizados.
Conte para cima (+) ou para baixo (-) com o mouse, ou para cima com as teclas 0-9 e a-z (apenas os primeiros 36 contadores). As barras representam a abundância relativa. Um registro (*log*) de eventos é fornecido à direita – role para cima ou para baixo com o mouse ou as setas do teclado. Um *feedback* auditivo opcional tem um tom específico para cada contador.

Transform menu (Transformações de dados)

Estas rotinas realizam operações matemáticas nos seus dados. Isso pode ser necessário para exibir algumas características dos dados ou pode ser um passo pré-processamento necessário para algumas análises.

Logaritmo

A função *Log* no menu *Transform* transforma os seus dados em logaritmos na base 10. Caso os dados apresentam zeros ou valores negativos, pode ser necessário adicionar uma constante (e.g. 1) antes da transformação em log (use *Evaluate Expression x+1*).
Isso é útil, por exemplo, para comparar a sua amostra com uma distribuição log-normal ou para encaixar um modelo exponencial. Além disso, dados de abundância com alguns táxons muito dominantes podem ser transformados em logaritmo para reduzir a importância desses táxons.
Há suporte para dados ausentes (*missing data*).

Remover tendência – remove trend

Esta função remove qualquer tendência linear de um conjunto de dados (duas colunas com pares X-Y ou uma coluna com valores do Y). Isso é feito subtraindo-se uma regressão linear dos valores de Y. Remover tendências pode ser uma operação prévia útil para análises de series temporais, por exemplo análise espectral (*spectral analysis*), auto-correlação e correlação cruzada, e ARMA.
Há suporte para dados ausentes.

Subtrair média – Subtract mean

Esta função subtrai, de cada coluna selecionada, o valor da média da coluna As médias não podem ser calculadas por linha.
Há suporte para dados ausentes.

Box-Cox

A transformação de Box-Cox é uma família de transformações de potência cujo objetivo é tornar os dados *x* mais similares a uma distribuição normal. A transformação tem um parâmetro λ:

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

O valor-padrão do parâmetro é calculado maximizando a função de log-verossimilhança (*lok likelihood function*)

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

onde σ^2_λ é a variância dos dados transformados. O valor ótimo pode ser mudado pelo usuário, estando limitado a $-4 \leq \lambda \leq 4$.
Há suporte para dados ausentes.

Porcentagem de linhas – Row percentage

Todos os valores são convertidos em porcentagem da somatória da linha.
Há suporte para dados ausentes.

Normalizar comprimento por linha – Row normalize length

Todos os valores são divididos pelo comprimento Euclideano do vetor da linha.
Há suporte para dados ausentes.

Abundância para presença/ausência – Abundance to presence/absence

Todos os valores positivos (diferentes de zero) são substituídos por 1's.
Há suporte para dados ausentes.

Ecaixe de Procrustes – Procrustes fitting

Transforma suas medidas de coordenadas de pontos em coordenadas Procrustes. Há uma opção no menu para coordenadas de Bookstein. Espécies vão em linhas diferentes e pontos de referência (*landmarks*) ao longo de cada linha. Se você tem três espécies com quatro pontos de referência em 2D, o seus dados devem ter a seguinte aparência:
x1 y1 x2 y2 x3 y3 x4 y4
x1 y1 x2 y2 x3 y3 x4 y4
x1 y1 x2 y2 x3 y3 x4 y4
Dados 3D são inseridos de forma similar mas com colunas adicionais para os valores de *z*.
Dados de pontos de referência (*landmarks*) neste formato podem ser analisados diretamente com os métodos multivariados do PAST, mas é recomendado padronizar para coordenadas Procrustes ao remover posição, tamanho e rotação. Uma transformação adicional dos resíduos Procrustes (coordenadas aproximadas no espaço tangente– *approximate tangent space coordinates*) pode ser feita escolhendo a opção *"Subtract mean"* ("Subtrair média") no menu *Transform*. Você precisa primeiro converter para coordenadas Procrustes para depois converter para resíduos Procrustes.

A opção *"Rotate to major axis"* ("Rotacionar para o eixo principal") coloca o resultado em uma orientação convencional, por conveniência.
A opção *"Keep size"* ("Manter tamanho") adiciona um passo final no qual a escala das formas é transformada de modo que elas voltem aos tamanhos originais dos seus centróides.
Uma descrição detalhada do coordenadas Procrustes e de espaço tangete é dada em Dryden & Mardia (1998). Os algoritmos para o encaixe Procrustes são de Rohlf & Slice (1990) (2D) e de Dryden & Mardia (1998) (3D). Deve ser notado que para 2D, o algoritmo iterativo de Rohlf & Slice (1990) frequentemente dá resultados ligeiramente diferentes do algoritmo direto de Dryden & Mardia (1998). O PAST usa o primeiro para seguir o "padrão industrial".
Dados ausentes têm suporte apenas por substituição pela média da coluna, o que pode não ser muito significativo.

Referências

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.

Encaixe de Bookstein (Bookstein fitting)

O encaixe de Bookstein tem uma função similar ao encaixe Procrustes, mas ele simplesmente padroniza tamanho, rotação e escala forçando os dois primeiros pontos de referência para as coordenadas (0,0) e (0,1). Seu uso não é comum hoje em dia. Encaixe de Bookstein só é implementado para 2D.

Projetar para espaço tangente

Depois de encaixe Procrustes ou Bookstein, alguns procedimentos estatísticos são realizados de preferência em coordenadas no espaço tangente (normalmente isso não faz nenhuma diferença, mas não nos cite para falar isso!). Sendo *d* o número de dimensões e *p* o número de pontos de referência (*landmarks*), a projeção é

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

Aqui, **X** é a matriz *n* x *dp* de *n* espécimes, **X'** é a matriz transformada, **I** é a matriz-identidade *dp* x *dp* e $\mathbf{X_c}$ é a configuração média (consenso) como um vetor de linha *dp*-elemento (*dp-element row vector*).

Remover tamanho de pontos de referência (Remove size from landmarks)

A opção "Remover tamanho de pontos de referência" ("*Remove size from landmarks*") do menu *Transform* lhe permite remover o tamanho ao dividir o valor de todas as coordenadas pelo tamanho do centróide (*centroid size*) de cada espécime (coordenadas Procrustes também são normalizadas em relação ao tamanho).
Veja Dryden & Mardia (1998), p. 23-26.

Referência

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

Transformar pontos de referência (Transform landmarks)

Permite rotação da nuvem de pontos em passos de 90 graus e espelhamento de cima para baixo e de esquerda para direta, principalmente para facilitar a plotagem. A operação de espelhamento pode ser útil para reduzir dados de um ponto de referência com simetria bilateral por meio de um encaixe de Procrustes da região esquerda à versão espelhada da região direita (e opcionalmente calculando a média dos dois).
Apenas para coordenadas 2D.

Remover tamanho de distâncias (*Remove size from distance*)

Tenta remover o compomente de tamanho de um conjunto de dados multivariados de distâncias medidas (espécimes em linhas, variáveis em colunas). Três métodos são disponíveis.

- *Método isométrico de Burnaby (Isometric Burnaby's method)* projeta o conjunto de distâncias medidas em um espaço ortogonal ao primeiro componente principal. O método de Burnaby pode (mas não necessariamente!) remover tamanho isométrico dos dados, permitindo análises futuras de dados livres de tamanho (*"size-free"*). Repare que a implementação no PAST não centra os dados dentro dos grupos – ela assume que todos os espécies (colunas) pertencem a um grupo.
- *Método alométrico de Burnaby (Allometric Burnaby's method)* transformará os dados em logaritmo antes da projeção, assim (teoricamente) removendo dos dados também a variação alométrica dependente de tamanho.
- *Alométrico vs. padrão (Allometric vs. standard)* estima coeficientes alométricos no que diz respeito a uma medida padrão (de referência) *L* tal como o comprimento total (Elliott et al. 1995). Esta variável padrão deve ser colocada na primeira coluna. Cada uma das colunas adcionais é regredida para a primeira coluna depois de transformação em logaritmo, fornecendo a inclinação (coeficiente alométrico) *b* para aquela variável. Uma medida ajustada é então calculada do valor original *M* como

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

  Referência

  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.

Ordenar crescente e decrescente (Sort ascending and descending)

Ordena as linhas na área marcada com base nos valores na coluna selecionada.
A função "Ordenar decrescente" (*"Sort descending"*) é útil, por exemplo, para plotar abundância de táxons contra seus *ranks* (isso também pode ser feito no módulo Modelo de Abundância (*Abundance Model*)).

Ordenar por cor (Sort on color)

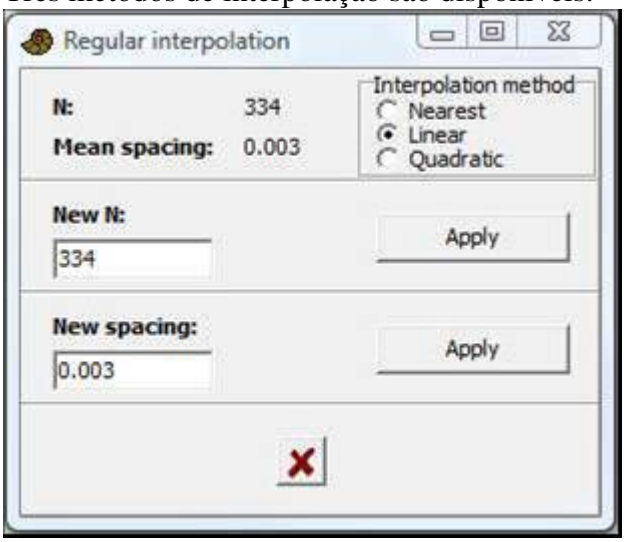
Ordena as linhas na área marcada pela cor.

Diferença entre colunas (Column difference)

Simplesmente subtrai as duas colunas selecionadas e coloca os resultados na coluna seguinte.

Interpolação regular (Regular interpolation)

Interpola uma série temporal ou transecto amostrado irregularmente (*unevenly sampled*), possivelmente multivariado, em um espaçamento regular, como é pedido por diversos métodos de análise de séries temporais. Os valores de *x* devem estar na primeira coluna selecionada. Estes serão substituídos por uma série que aumenta regularmente. Todas as colunas adicionais selecionadas serão interpoladas de maneira correspondente. Os perigos da interpolação devem ser mantidos em mente.
Você pode especificar o número total de pontos interpolados ou o novo espaçamento. Três métodos de interpolação são disponíveis.



Avaliar expressão (Evaluate expression)

Esta ferramenta poderosa permite operações matemáticas flexíveis na matriz de dados selecionada. Cada célula selecionada é avaliada e o resultado substitui o conteúdo anterior. Uma expressão matemática deve ser inserida, que pode incluir quaisquer dos operadores +, -, *, /, ^ (potência), e mod (módulo – calcula o resto da divisão de um número por outro; não confundir com abs, a seguir!). Também há suporte para parênteses ( ) e para as funções abs (valor absoluto), atan, cos, sin, exp, ln, sqrt (square root – raiz quadrada), sqr (square – quadrado), round (aproximar) e trunc.
Também são definidos os seguintes valores:

- x (o conteúdo da célula atual)
- l (a célula à esquerda se ela existe, 0 caso contrário - *left*)
- r (a célula à direita - *right*)

- u (a célula acima – *up*)
- d (a célula abaixo – *down*)
- mean (o valor médio da coluna atual)
- min (o valor mínimo)
- max (o valor máximo)
- n (número de células na coluna
- i (índice de linha)
- j (índice de coluna)
- random (número aleatório uniforme entre 0 e 1)
- normal (número aleatório Gaussiano com média 0 e variância 1)
- integral (somatória corrente – *running sum* - da coluna atual)
- stdev (desvio padrão da coluna atual)
- sum (somatória total da coluna atual)

Adicionalmente, é possível se referir a outras colunas usando o nome da coluna precedido por "c_", por exemplo c_A.

Exemplos

sqrt(x) Substitui todos os valores por suas raízes quadradas

(x-mean)/stdev Padronização por média e desvio padrão em cada coluna

x-0.5*(max+min) Centra os valores em torno de zero

(u+x+d)/3 suavização média móvel de três pontos (*three-point moving average smoothing)*

i Preenche a coluna com os números das linhas (requer células não-vazias, por exemplo todos zeros)

sin(2*3.14159*i/n) gera um período de uma função seno coluna abaixo (requer células não-vazias)

5*normal+10 Número aleatório de uma distribuição normal, com média 10 e desvio padrão 5.

Há suporte para dados ausentes.

Plot Menu (Gráficos)

Gráfico (Graph)[3]

Plota uma ou mais colunas como gráficos separados. As coordinadas *x* são estabelecidas automaticamente em 1,2,3,... Há quantro estilos de gráfico disponíveis: Gráfico (linha – *line*), pontos (*points*), linha com pontos (*line+points*) e barras (*barchart*). As opções "Legenda X" (*"X labels"*) estabelece os *labels* do eixo x com os nomes das linhas correspondentes.

A opção "Log Y" transforma em log os valores do eixo Y. O logaritmo é calculado na base 10, mas log 0 é definido como 0.

Valores faltantes são desonsiderados.

[3] Nesta seção, não traduzi os termos *Plot* (fazer um gráfico) e *Label* (legenda de um eixo ou de um ponto).

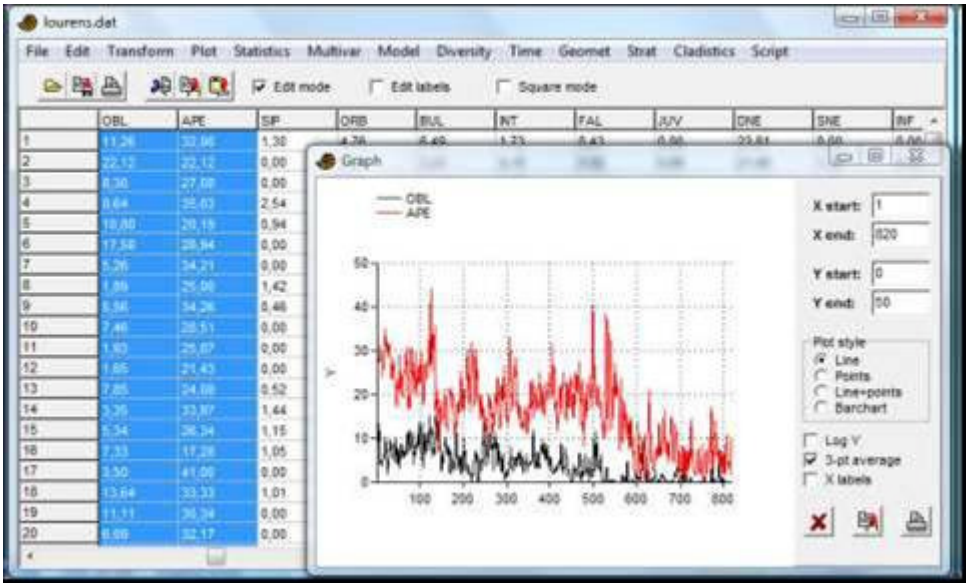


Gráfico XY (XY graph)

Plota um ou mais pares de colunas contendo pares de coordenadas x/y. A opção "log Y" transforma em logaritmo os valores de Y (se necessário, uma constante é adicionada para tornar o valor mínimo de log igual a 0). A curva também pode ser suavizada (*smoothed*) usando média móvel de 3 pontos (*3-point moving average*).

Elipses de concentração 95% podem ser plotadas na maior parte dos gráficos de dispersão no PAST, tais como os escores das análises de PCA, CA, DCA, PCO e NMDS. O cálculo destas elipses assume distribuição normal bivariada.

Envelopes convexos (*convex hulls*) também podem ser desenhados nos gráficos de dispersão para mostrar as áreas ocupadas por pontos de "cores" diferentes. O envelope convexo é o menor polígono convexo que contém todos os pontos.

A árvore de expansão mínima (*minimal spanning tree*) é o conjunto de linhas de comprimento total mínimo conectando todos os pontos. No módulo XY graph, distâncias Euclideanas 2D são usadas.

Segure o cursor do mouse sobre um ponto para ver o *label* da sua linha.

Pontos com valores ausentes em X e/ou em Y são descartados.

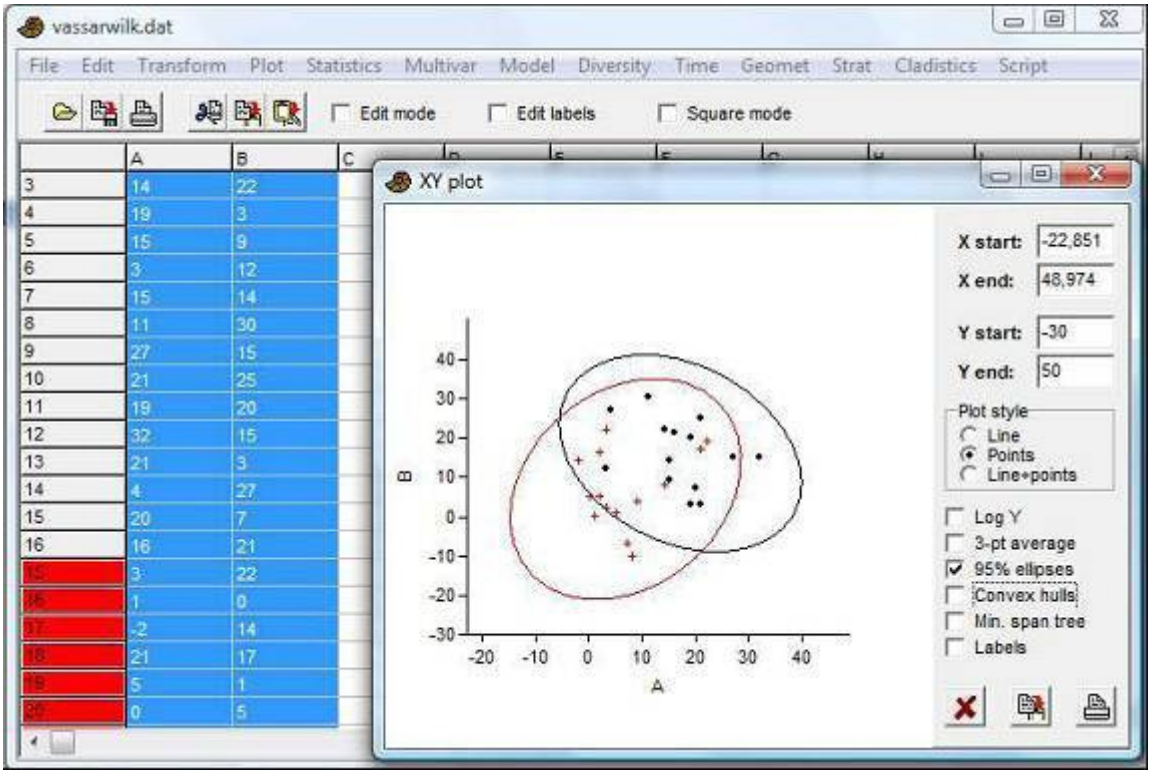
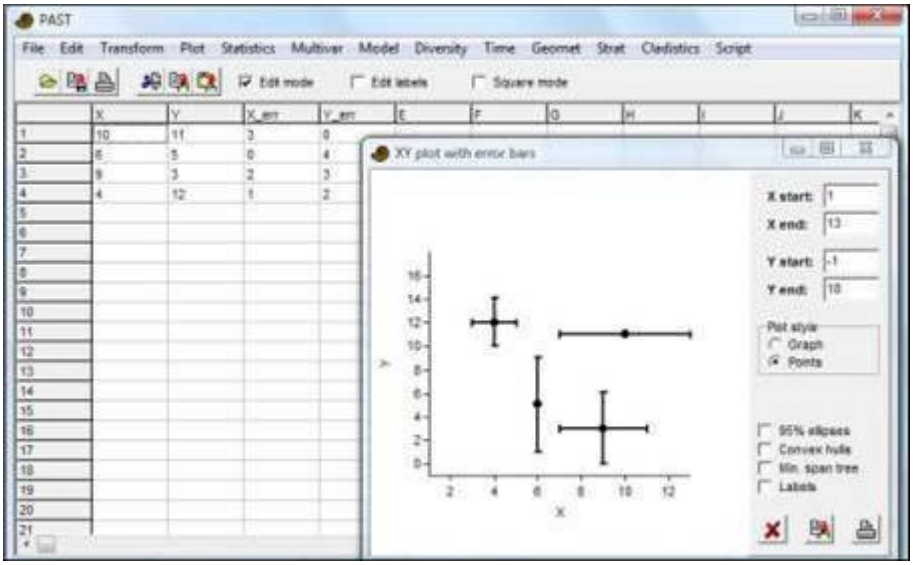


***Gráfico XY com barras de erro* (XY *graph with error bars*)**

Igual a um gráfico XY, mas espera quatro colunas (ou um múltiplo), com valores de x, y, erro de x e erro de y. Barras de erro simétricas são desenhadas ao redor de cada ponto com o semi-comprimento como espeficado. Se um valor de erro é estabelecido em zero ou não é fornecido, a barra de erro correspondente não é desenhada.
Pontos com valores ausentes de X e/ou Y são desconsiderados.



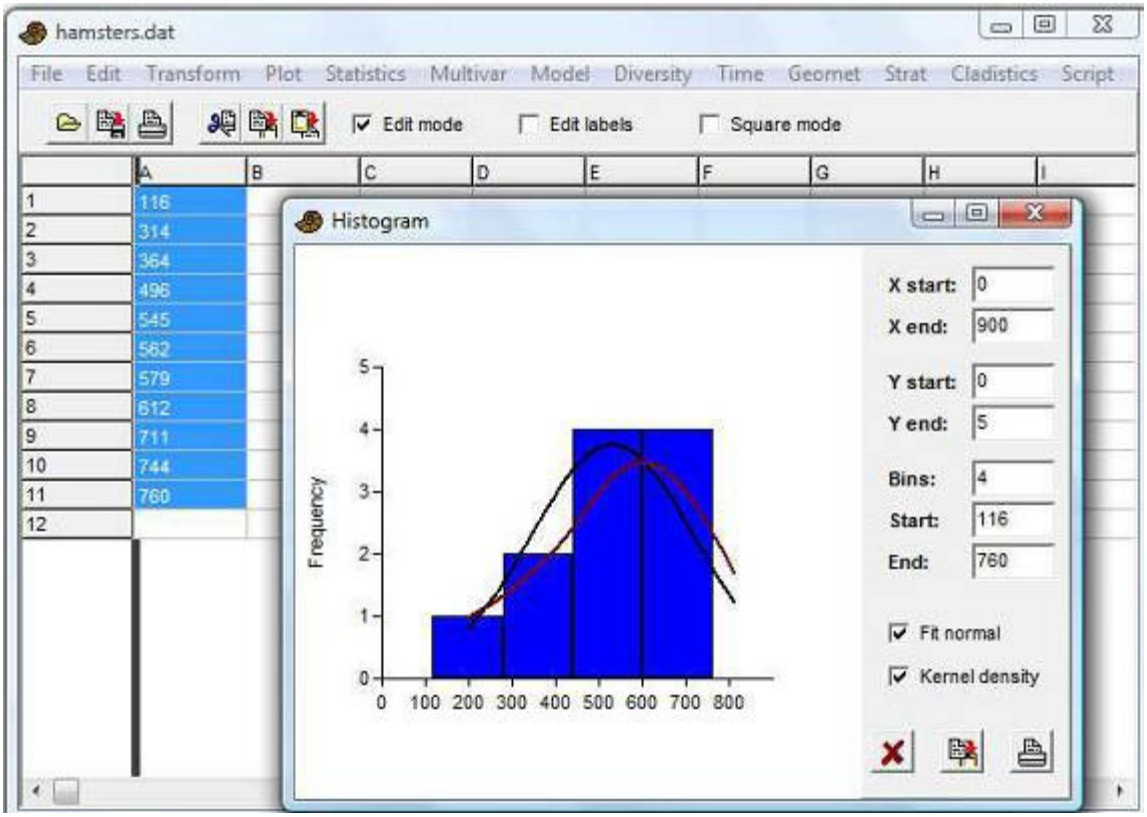
Histograma (Histogram)

Plota histogramas (distribuições de frequências) para uma ou mais colunas. O número de classes (*bins*) é definido por padrão em um número "ótimo" (a regra de fase-zero (*zero-stage rule*) de Wand 1997):

$h = 3.49\min(s, IQ/1.349)n^{-1/3}$
onde s é o desvio-padrão da amostra e IQ é a amplitude entre-quartis (*interquartile range*).
OP número de classes pode ser mudado pelo usuário. A opção *"Fit normal"* ("Ajustar normal") desenha um gráfico com uma distribuição normal ajustada (estimação Paramétrica, não por Mínimos Quadrados).

Estimação de Densidade Kernel (*Kernel Density Estimation*) é um estimador suave do histograma. PAST utiliza o Kernel Gaussiano com amplitude definida pela regra dada por Silverman (1986):

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



Valores ausentes são deletados.

Referências

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.

Gráfico de barras / boxplot (Bar chart/box plot)

Gráfico de barras ou caixas (boxplot) para uma ou mais columas (amostras) de dados univariados. Valores ausentes são deletados.

Gráfico de barras (Bar chart)

Para cada amostra, o valor médio é mostrado por uma barra. Além disso, linhas de erro podem ser mostradas. O intervalo das barras de erro representa um um-sigma ou um intervalo de confiança 95% (1.96 sigma) para a estimativa da média (baseado no erro-padrão) ou um-sigma ou intervalo de concentração de 95% (baseado no desvio padrão).

Gráfico de caixas (Box plot)

Para cada amostra, os quartis de 25-75% são desenhados usando uma caixa. A mediana é mostrada com uma linha horizontal dentro da caixa. Os valores máximo e mínimo são mostrados com linhas horizontas curtas (*"whiskers"*).

Se a caixa *"Outliers"* ("Pontos extremos) for selecionada, uma outra convenção de box plot é usada. Os *whiskers* são desenhados do topo da caixa até o maior ponto que esteja a menos do que 1.5 vezes a altura da caixa acima da caixa(*upper outer fence*) e similarmente abaixo da caixa. Valores fora dos limites internos são mostrados como círculos, valores mais longe do que três alturas da caxa da caixa ("limites externos" – *"outer fences"*) são mostrados como estrelas.

Os métodos dos quartis (arredondamento ou interpolação) são descritos em "Percentis" (*"Percentiles"*) abaixo.

Jitter plot

Cada valor é plotado como um ponto. Para mostrar pontos sobrepostos mais claramente, eles podem ser deslocados usando um valor de *"jitter"* aleatório controlado por uma barra deslizante.

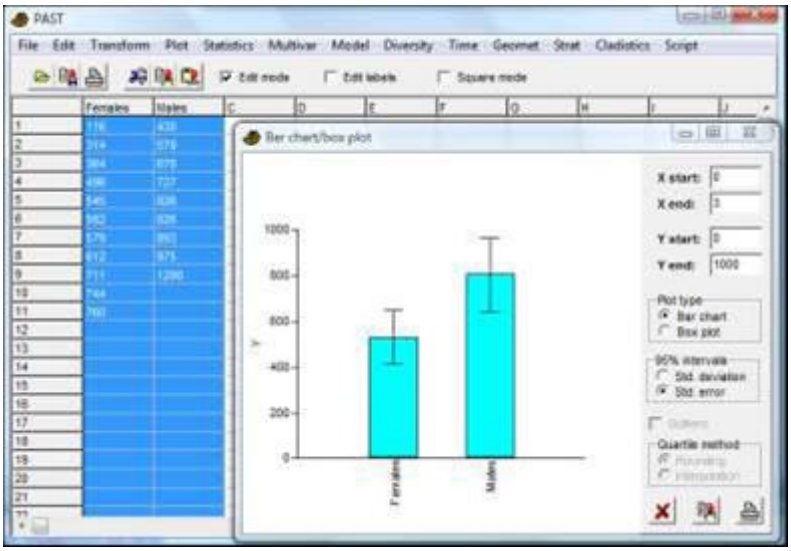
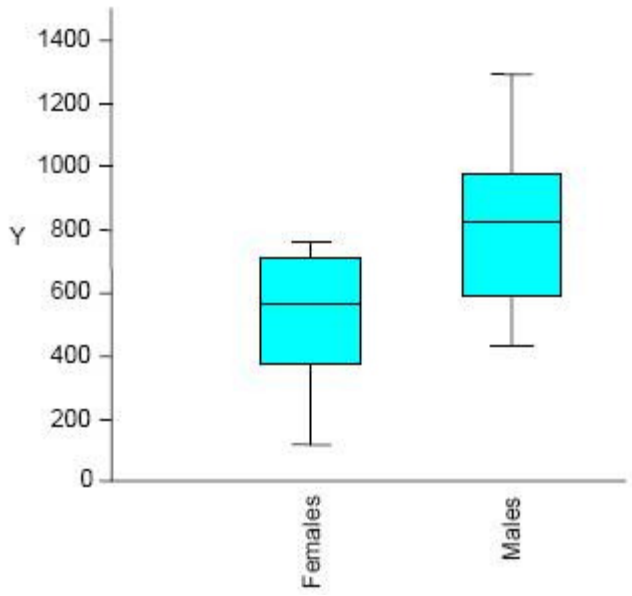


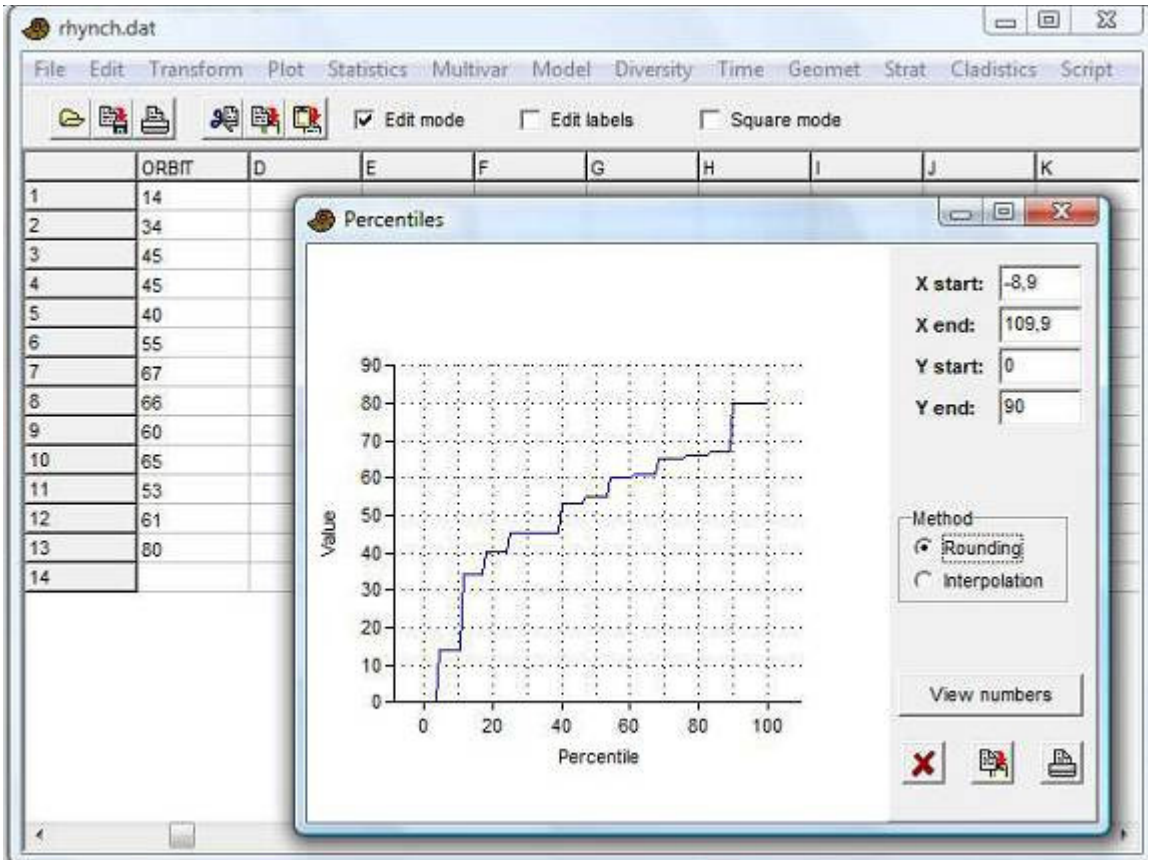
Gráfico de barras (*Bar chart*)



Box plot

Percentis (Percentiles)

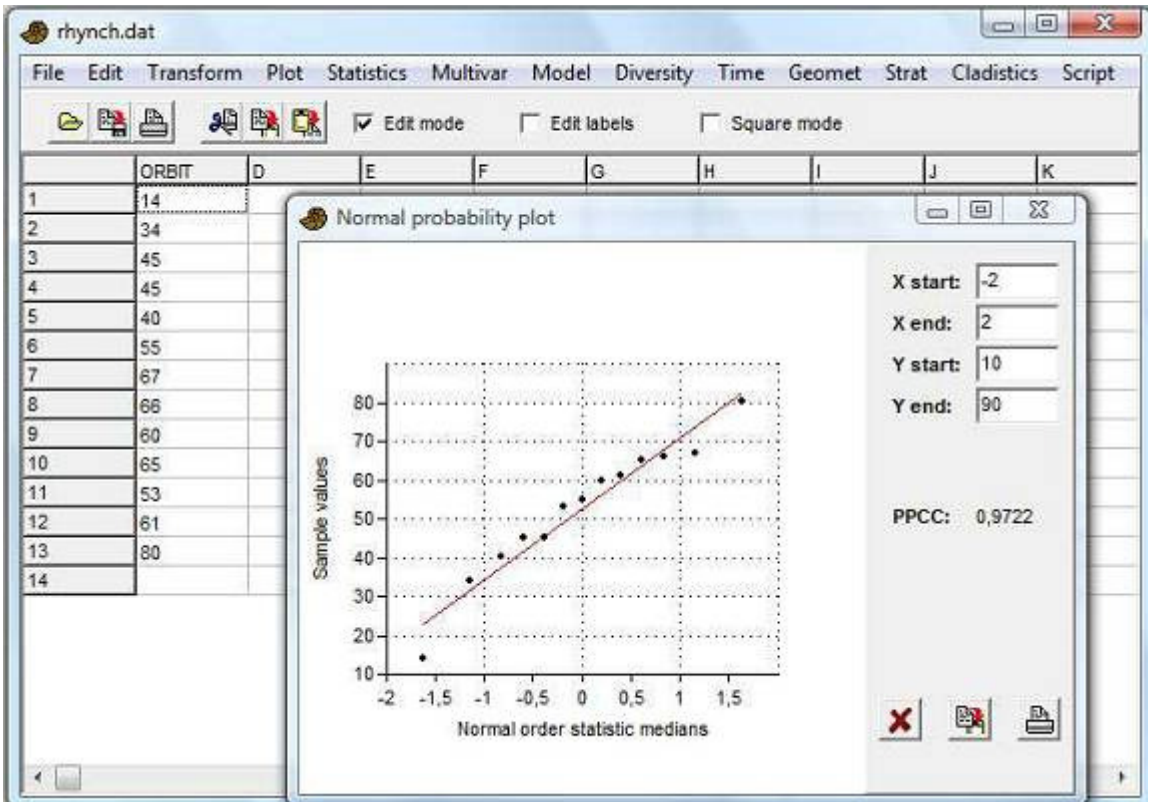
Para cada percentil *p*, plota o valor de *y* tal que *p* porcento dos pontos são menores do que *y*. Dois métodos populares são inclusos. Para um percentil *p*, o *rank* é calculado de acordo com $k=p(n+1)/100$, e o valor correspondente àquele *rank* é tomado. No método de arredondamento, *k* é arredondado até o número inteiro mais próximo; já no método de interpolação, *rank*s não-inteiros são tratados por interpolação entre os dois *rank*s mais próximos.



Valores ausentes são deletados.

Gráfico de probabilidade normal (Normal probability plot)

Plota um gráfico de probabilidade normal (QQ normal) para uma coluna de dados. Uma distribuição normal irá formar uma linha reta. Para comparação, é fornecida uma linha de regressão RMA juntamente com o Coeficiente de Correlação do Gráfico de Probabilidade (*Probability Plot Correlation Coefficient*).



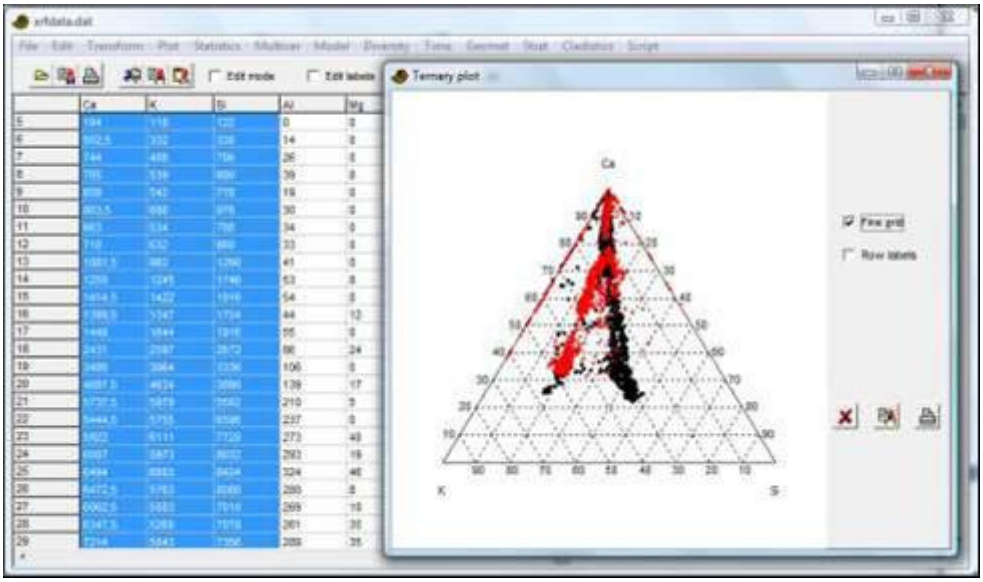
Dados ausentes são deletados.

As medianas das estatísticas de ordem da normal (*normal order statistic medians*) são calculadas como $N(i) = G(U(i))$, onde G é o inverso da função de distribuição cumulativa da normal e U são as medianas das estatísticas de ordem da uniforme (*uniform order statistic medians*):

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

Ternário (Ternary)

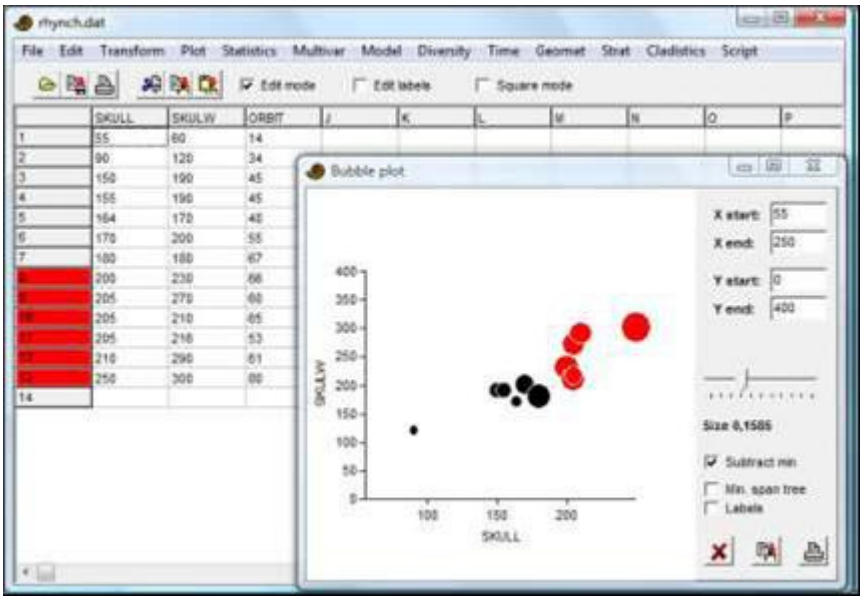
Gráfico ternário para três colunas de dados, normalmente contendo proporções de composições. Se uma quarta coluna for incluída, ela será apresentada por meio de uma representação de bolhas ou um mapa colorido/escala de cinza.



Linhas com valor(es) ausente(s) em qualquer coluna são deletadas. Quando utilizar a opção de mapa colorido, as linhas com apenas a quarta coluna ausente são incluídas no gráfico, mas não contribuem com o mapa.

Gráfico de bolhas (Bubble plot)

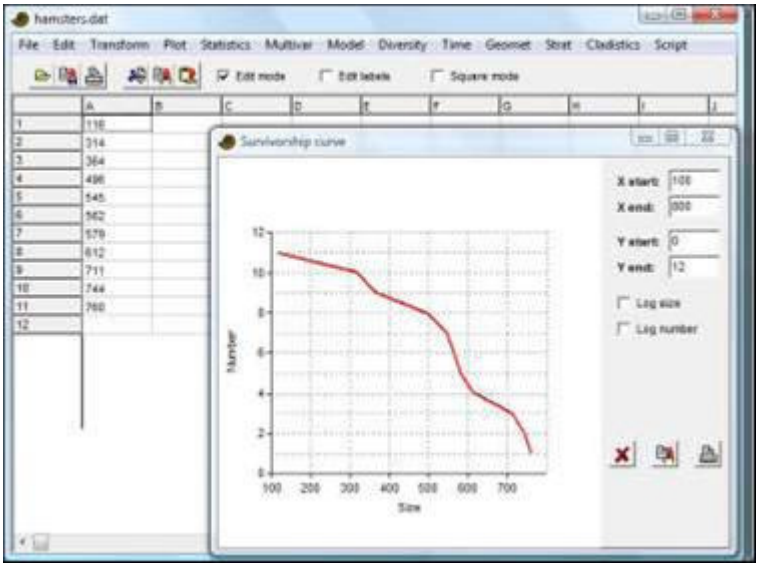
Plota dados 3D (três colunas) mostrando o terceiro eixo como tamanho dos discos. Selecione "Subtract min" para subtrair o o menor valor do terceiro eixo de todos os valores – isso vai forçar os dados a ficarem positivos. A barra deslizante *"Size"* ("tamanho") muda a escala das bolhas em relação à unidade escala das unidades do eixo x.



Linhas com valor(es) ausente(s) em qualquer das colunas são deletadas.

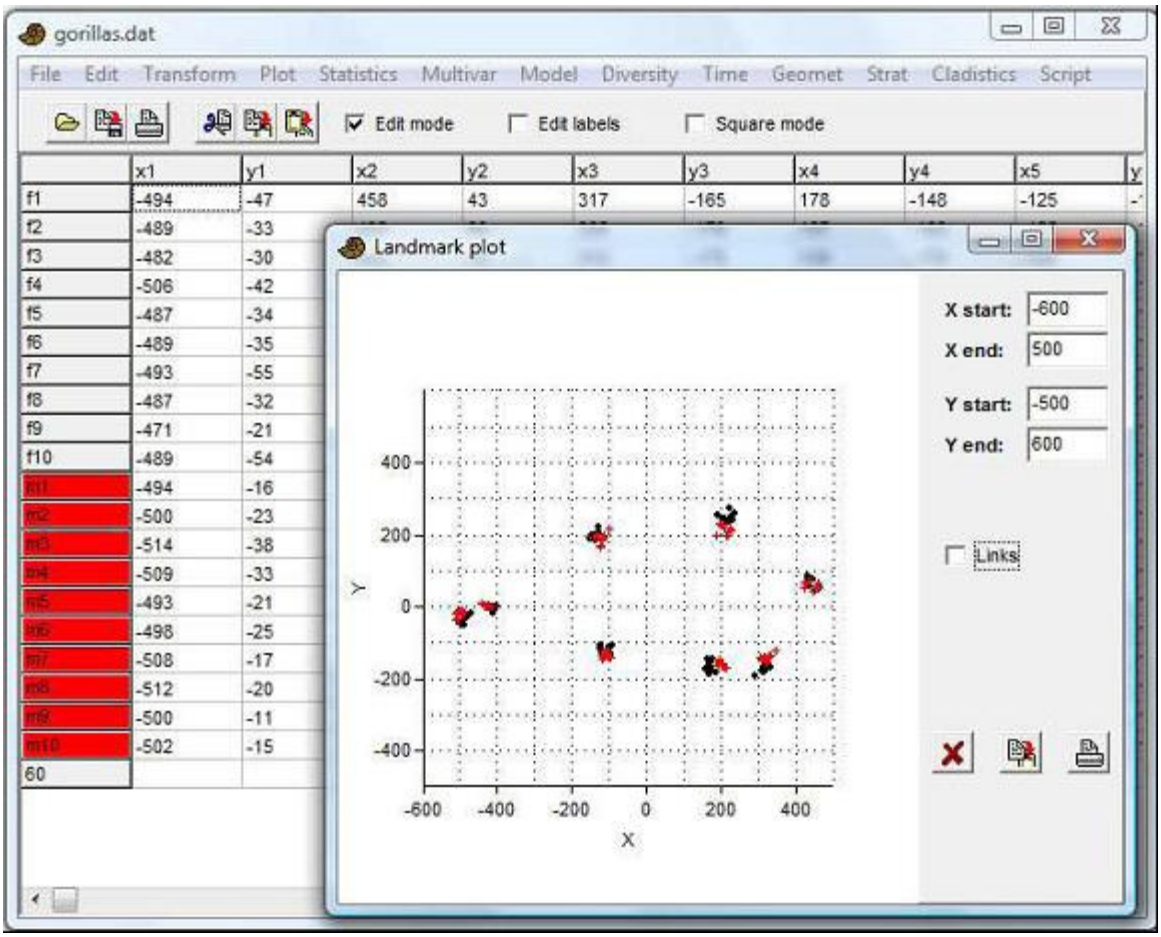
Sobrevivência (Survivorship)

Curvas de sobrevivência para uma ou mais colunas de dados. Os dados podem consistir de valores de idade ou tamanho. O gráfico mostra o número de indivíduos que sobreviveram até diferentes idades. Assumindo crescimento exponencial (altamente questionável!), tamanho pode ser transformado, por logaritmo, em idade. Isso pode ser feito no menu Transform ou diretamente no diálogo do *Survivorship*. Veja também Análise de sobrevivência (*Survival analysis*) no menu Statistics.
Valores ausentes são deletados.



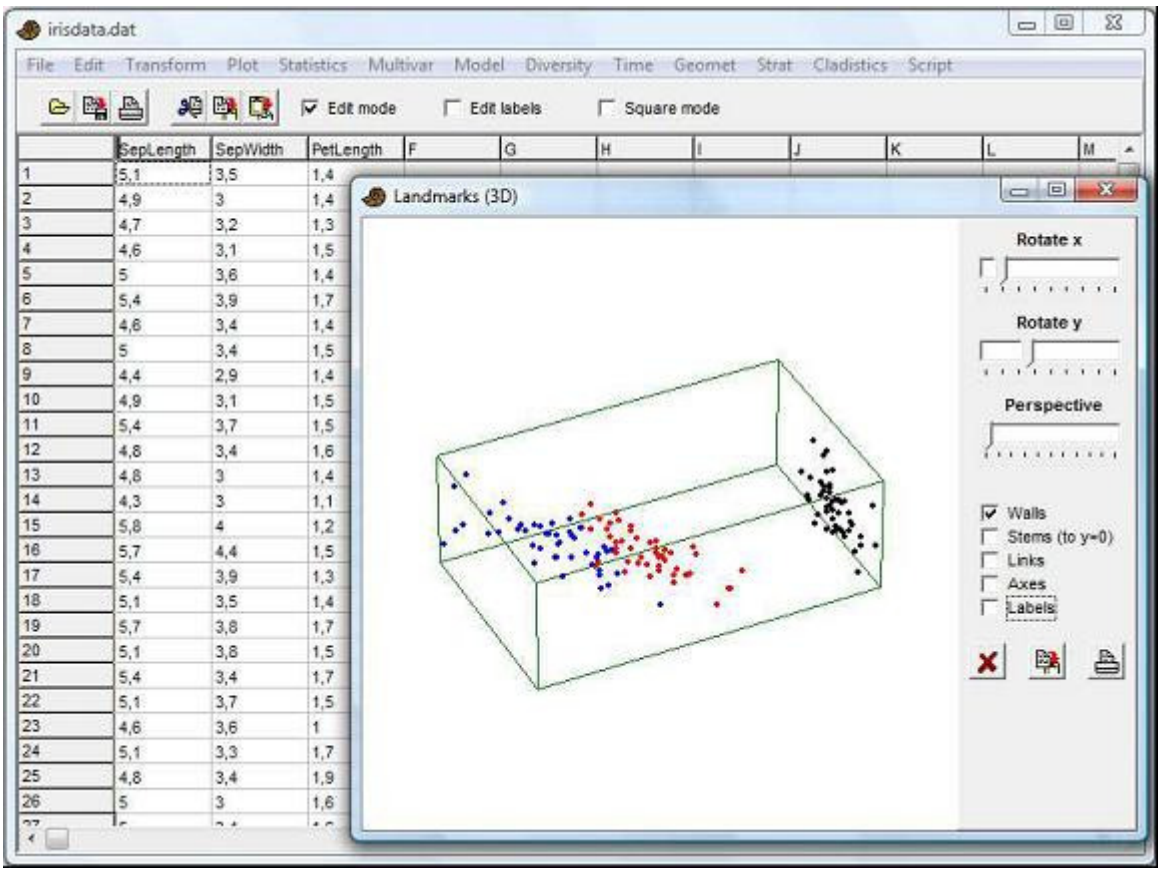
Pontos de referência (Landmarks)

Essa função é muito similar ao "gráfico XY", a única diferençe é que todos os pares XY de cada linha são plotados com a cor e símbolo apropriados. Ele também força relação de aspecto igual a um (*unit aspect ration*), e é bastante apropriada para plotar dados de pontos de referência. A opção *"links"* plota linhas entre os pontos de referência, como especificado pela opção *"Landmark linking"* no menu Geomet.
Pontos com valores ausentes em X e/ou em Y são desconsiderados.



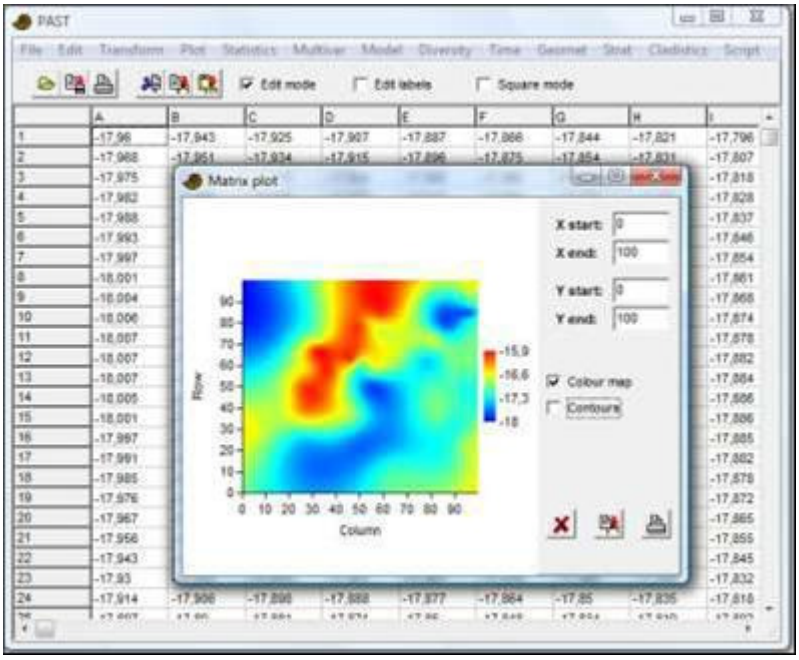
Pontos de referência 3D (Landmarks 3D)

Plotagem de pontos em 3D (XYZ). Especialmente adequado para dados em 3D de pontos de referência (*landmarks*), mas também pode ser usado, e.g., para gráficos de dispersão de PCA com três componentes principais. A nuvem de pontos pode ser rotacionada ao redor dos eixos *x* e *y* (observe: sistema de coordenadas mão-esquerda (*left-handed*)). O deslizador "*Perspective*" ("Perspectiva") normalmente não é usado. A opção *"Stems"* (*"Caules"*) desenha desenha uma linha de cada ponto até o plano de baixo, o que às vezes pode melhorar a informação 3D. "Lines" ("Linhas"), desenha linhas entre pontos de referência consecutivos dentro de cada espécime (linha) separado. "Axes" ("Eixos"), mostra os três eixos de coordenadas com o centróide dos pontos como origem.
Pontos com valores ausentes em X, Y ou Z são desconsiderados.



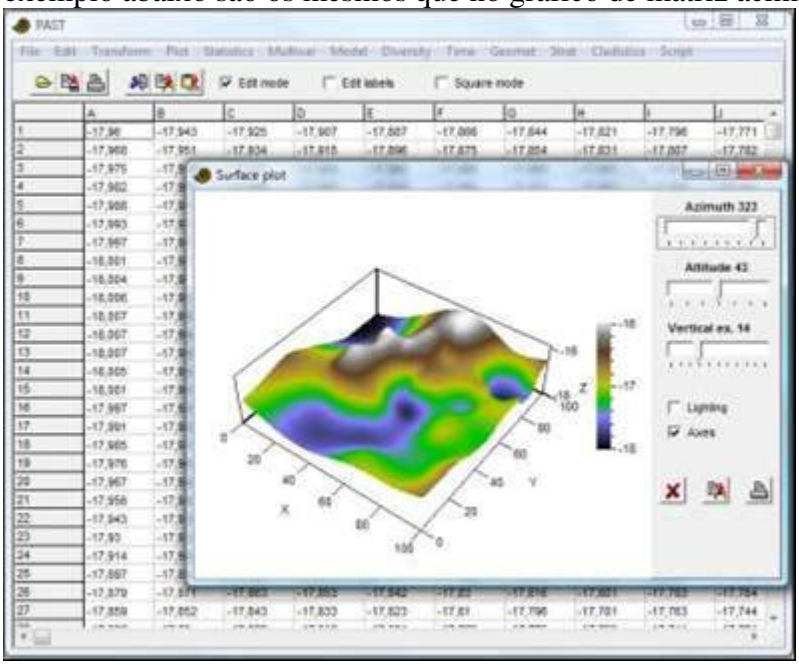
Matriz (Matrix)

Gráfico bidimensional da matriz de dados, usando uma escala de cinza com branco para o valor mais baixo e preto para o valor ou mais alto, ou uma escala de cores. Use para ter uma visão geral de uma matriz de dados grande. Valores ausentes são plotados como vazios (permitindo buracos e limites não-quadrados).



Superfície (Surface)

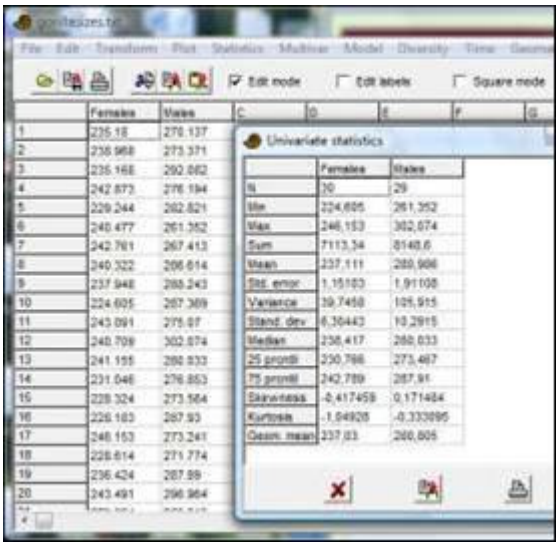
Gráfico de paisagem tridimensional de uma matriz de dados com valores de elevação. Cores são atribuídas de acordo com a elevação, ou a superfície pode ser preenchida com tons de cinza usando um modelo de luz com uma fonte de iluminação fixa. Os dados no exemplo abaixo são os mesmos que no gráfico de matriz acima.



Statistics Menu (Estatística univariada)

Univariada (Univariate)

Essa função calculca uma série de estatística descritivas básicas para uma ou mais amostras de dados univariados. Cada amostra deve ter ao menos 3 valores, e ocupar uma coluna na planilha. As colunas não precisam conter o mesmo número de valores. O exemplo abaixo usa duas amostras: os tamanhos, em mm, dos crânios de 30 gorilas fêmeas e 29 gorilas machos. Para rodar a análise, as duas colunas (ou a planilha inteira) devem ser selecionadas.



Os seguintes valores são mostrados para cada amostra:

N:	O número de valores *n* na amostra
Min:	O valor mínimo
Max:	O valor máximo
Sum:	A soma
Mean:	A estimativa da média, calculada por $\bar{x} = \frac{\sum x_i}{n}$
Std. error:	O erro padrão da estimativa da média, calculado por $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$ onde *s* é a estimativa do desvio padrão (ver abaixo).
Variance:	A variância da amostra, calculada por $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$.
Stand. dev.:	O desvio padrão da amostra, calculado por $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$.
Median:	A mediana da amostra. Para *n* ímpar, o valor fornecido é tal que há tantos valores acima quanto abaixo dele. Para *n* par, é a média dos dois valores centrais.
25 prcntil:	O 25º percentil, ou seja o valor tal que 25% da amostra está abaixo dele e 75% está acima. O método de "interpolação" é usado (ver Gráfico de Percentis – *Percentile Plot* acima).
75 prcntil:	O 75º percentil, ou seja o valor tal que 75% da amostra está abaixo dele e 35% está acima. O método de "interpolação" é usado (ver

Gráfico de Percentis – *Percentile Plot* acima).

Skewness: A assimetria da amostra, zero para uma distribuição normal, positiva para uma distribuição com cauda para a direita.

Calculada por $G1 = \frac{n}{(n-1)(n-2)} \frac{\sum (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}\right)^3}$. Observe que há diversas versões desta fórmula – o Past usa a mesma equação que SPSS e Excel. Resultados ligeiramente diferentes podem ocorrer em outros programas, especialmente para tamanhos amostrais pequenos.

Kurtosis:

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

Novamente o Past usa a mesma equação que SPSS e Excel.

Geom. mean: A média geométrica, calculada como $(x_1 x_2 ... x_n)^{1/n}$.

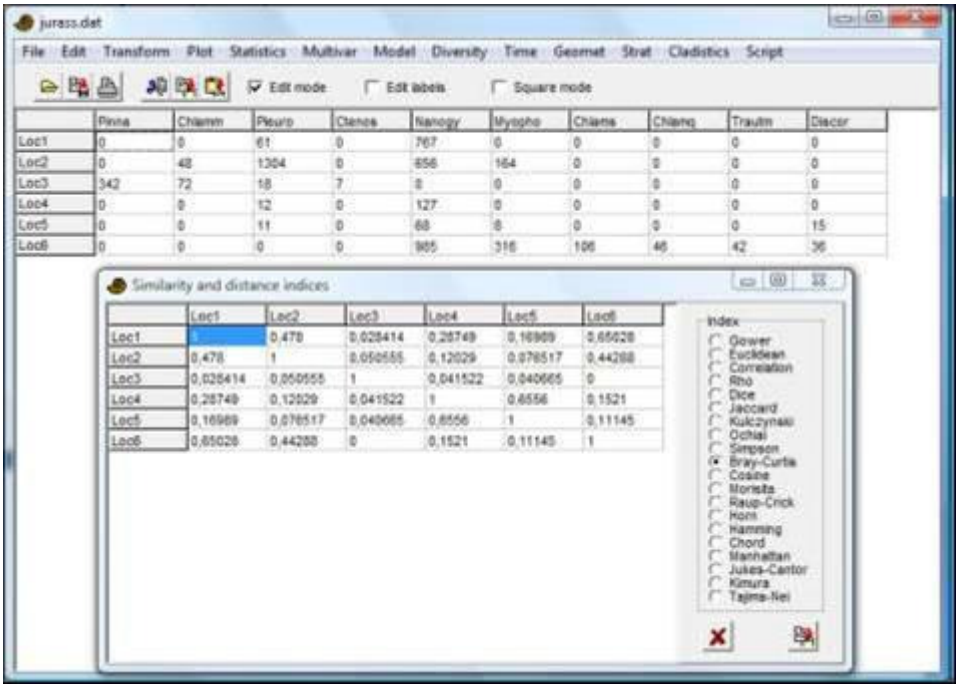
Bootstrapping

Selecionando a opção *bootstrapping* irá calcular os limites superior e inferior dos intervalos de confiança de 95% por meio de 9999 réplicas *bootstrap*. Intervalos de confiança para os valores mínimo e máximo não são fornecidos, porque sabe-se que o *bootstrap* não funciona bem para essas estatísticas.

Dados ausentes: suportados por deleção.

Índices de similaridade e distância (Similarity and distance indices)

Calcula uma série de medidas de similaridade ou distância entre todos os pares de linhas. Os dados podem ser univariados ou (mais comumente) multivariados, com as variáveis em colunas. Os resultados são fornecidos como uma maitrz simétrica de similaridade/distância. Este módulo é raramente usado porque matrizes de similaridade/distância normalmente são computados automaticamente em módulos como PCO, NMDS, análise de agrupamento (*cluster analysis*) e ANOSIM no Past.



Gower

Uma medida de distância que calcula a média da diferença entre todas as variáveis, sendo cada termo normalizado para a amplitude daquela variável:

$$d_{jk} = \frac{1}{n}\sum \frac{\left|x_{ji} - x_{ki}\right|}{\max\limits_{s} x_{si} - \min\limits_{x} x_{si}} .$$

A medida de Gower é similar à distância de Manhattan (ver abaixo) mas com normalização de amplitude. Quando usando tipos mistos de dados (ver abaixo), esta é a medida-padrão para dados contínuos e ordinais.

Euclidean

Distância Euclideana básica. Nas primeiras versões do Past, era normalizada para o número de variáveis (o valor ainda é ajustado para dados ausentes).

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

Mahalanobis

Uma medida de distância que leva em conta a estrutura de covariância dos dados, sendo **S** a matriz de variância-covariância:

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

Geographical

Distância em metros a longo de um grande círculo entre dois pontos na superfície da Terra. Requer exatamente duas variáveis (colunas), com latitudes e longitudes em graus decimas (e.g. 58 graus 30 minutos Norte é 58.5). Espera-se que as coordenadas estejam no datum WGS84, e a distância é calculada de acordo com o elipsóide WGS84. O uso de outros datums irá resultar em erros muito pequenos.

A acurácia do algoritmo usado (Vicenty 1975) é da ordem de 1 mm com relação a WGS84.

Correlation

O complemento 1-*r* do coeficiente *r* de correlação de Pearson entre as variáveis:

$$d_{jk} = 1 - \frac{\sum_i (x_{ij} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{\sqrt{\sum_i (x_{ji} - \bar{x}_j)^2 \sum_i (x_{ki} - \bar{x}_k)^2}}.$$

Usar o complemento faz disso uma medida de distância. Veja também o módulo Correlação (*Correlation*), onde o *r* de Pearson é fornecido diretamente e com testes de significância.

Rho

O complemento 1-r_s do rho de Spearman, que é um coeficiente de correlação de *ranks*. Veja também o módulo Correlação (*Correlation*), onde o rho é dado diretamente e com testes de significância.

Dice

Também conhecido como coeficiente de Sorensen. Para dados binários (presença-ausência), codificados como 0 ou 1 (qualquer número positivo é tratado como 1). A similaridade de Dice põe mais peso em ocorrências conjuntas do que em ocorrências disjuntas (*mismatches*).

Quando comparado duas linhas, uma ocorrência conjunta (*match*) é contada para todas as colunas com presença em ambas as linhas. Usando *M* para o número de ocorrências conjuntas e *N* para o número total de colunas com presença em apenas uma linha, temos $d_{jk} = 2M/(2M+N)$.

Jaccard

Um índice de similaridade para dados binários. Com a mesma notação usada para o índice de Dice acima, temos

$d_{jk} = M/(M+N)$.

Kulczynski

Um índice de similaridade para dados binários. Com a mesma notação dada para a similaridade de Dice acima (com N_1 e N_2 se referindo às duas colunas), temos

$$d_{jk} = \frac{\frac{M}{M+N_1} + \frac{M}{M+N_2}}{2}.$$

Ochiai

Um índice de similaridade para dados binários, comparável à similaridade de coseno (*cosine*) para outros tipos de dados:

$$d_{jk} = \sqrt{\frac{M}{M+N_1}\frac{M}{M+N_2}}.$$

Simpson

O índice de Simpson é definido simplesmente como M/N_{min}, onde N_{min} é o menor dos números de presenças nas duas linhas. Esse índice trata as linhas como idênticas caso uma seja um subconjunto da outra, o que o torna útil para dados fragmentários (*fragmentary data*).

Bray-Curtis

Bray-Curtis é um índice de similaridade popular para dados de abundância. O Past calcula a similaridade de Bray-Curtis da seguinte maneira:

$$d_{jk} = 1 - \frac{\sum_i \left| x_{ji} - x_{ki} \right|}{\sum_i \left(x_{ji} + x_{ki} \right)}.$$

Isso é algebricamente equivalente à fórmula dada originalmente por Bray e Curtis (1957):

$$d_{jk} = 2 \frac{\sum_i \min(x_{ji}, x_{ki})}{\sum_i (x_{ji} + x_{ki})}.$$

Muitos autores usam uma distância de Bray-Curtis, que é simplesmente 1-*d*.

Cosine

O produto interno das abundâncias, cada uma normalizada à norma unitária (*normalised to unit norm*), i.e. o coseno do ângulo entre os vetores.

$$d_{jk} = \frac{\sum_i x_{ji} x_{ki}}{\sqrt{\sum_i x_{ji}^2} \sqrt{\sum_i x_{ki}^2}}$$

Morisita

Para dados de abundância.

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

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

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

Raup-Crick

Índice de Raup-Crick para dados de presença-ausência. Este índice (Raup & Crick 1979) usa um procedimento de aleatorização (Monte Carlo) comparando o número observado de espécies que ocorrem em ambas as associações com a distribuição de co-ocorrências a partir de 1000 réplicas aleatórias do conjunto (*pool*) de amostras.

Horn

Índice de sobreposição de Horn para dados de abundância (Horn 1966).

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

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

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

Hamming

Distância de Hamming para dados categóricos codificados como números inteiros (ou dados de sequência genética codificados como CAGT). A distância de Hamming é o número de diferenças (*mismatches* ou ocorrências disjuntas), de modo que a distância entre (3,5,1,2) e (3,7,0,2) é igual a 2. No Past, ela é normalizada para a amplitude [0,1], a qual é conhecida por geneticistas como "*p-distance*".

Chord

Distância Euclideana entre vetores normalizados. Comumente usada para dados de abundância. Pode ser escrita como

$$d_{jk} = \sqrt{2 - 2\frac{\sum_i x_{ji} x_{ki}}{\sqrt{\sum_i x_{ji}^2 \sum_i x_{ki}^2}}}.$$

Manhattan

A somatória das diferenças em cada variável:

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

Jukes-Cantor

Medida de distância para dados de sequência genética (CAGT). Similar à distância *p* (ou Hamming), mas leva em conta a probabilidade de reversões (*reversals*):

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right).$$

Kimura

A medida de distância de 2 parâmetros de Kimura para dados de sequência genética (CAGT). Similar à distância de Jukes-Cantor, mas leva em conta diferentes probabilidades de transições vs. transversões de nucleotídeos (Kimura 1980). Sendo P a proporção observada de transições e Q o número observado de transversões, temos

$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q).$$

Tajima-Nei

Medida de distância para dados de sequência genética (CAGT). Similar à distância de Jukes-Cantor, mas não assume frequências iguais de nucleotídeos.

Similaridade definida por usuário (*User-defined similarity*)

Espera uma matriz simétrica de similaridade ao invés de dados originais. Sem verificação de erros!

Distância definida por usuário (*User-defined distance*)

Espera uma matriz simétrica de distância ao invés de dados originais. Sem verificação de erros!

Mixed (mista)

Esta opção requer que tipos de dados sejam atribuídos às colunas (veja *Inserindo e manipulando dados*). Uma janela *pop-up* irá perguntar a medida de similaridade/distância a ser usada para cada tipo de dados. Estas serão combinadas usando uma média ponderada pelo número de variáveis de cada tipo. As opções-padrão correspondem às sugeridas por Gower, mas outras combinações podem funcionar melhor. A opção "Gower" é uma distância de Manhattan normalizada pela amplitude (*range-normalised*).

Colunas só com zeros: Algumas medidas de similaridade (Dice, Jaccard, Simpson etc.) são indefinidas quando linhas contendo apenas zeros são comparadas. Para evitar erros, especialmente quando fazendo *bootstrap* em conjuntos de dados com poucos valores, a similaridades nestes casos é definida como zero.

Dados ausentes: A maio parte dessas medidas trata os dados ausentes (codificados por "?") por deleção par-a-par, significando que se um valor está ausente em uma das variáveis de um par de linhas, esta variável é omitida do cálculo de distâncias entre essas duas linhas. As excessões são: distância rho, a qual usa substituição pela média da coluna (*column average substitution*), e Raup-Crick, que não aceita dados ausentes.

Referências

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.

Tabela de correlação (Correlation table)

Apresenta uma matriz com os coeficientes de correlação entre todos os pares de colunas. Valores de correlação são fornecidos no triângulo inferior da matriz, e as probabilidades bicaudais de que as colunas não estejam correlacionadas (*columns are uncorrelated*) são apresentadas no triângulo superior. Coeficientes e testes tanto paramétricos (Pearson) quanto não-paramétricos (Spearman) são disponíveis. Algoritmos seguem Press et al. (1992), com a excessão de que a significância do coeficiente de Spearman é calculada por um teste exato para n<=9 (veja a seção sobre correlação de rank/ordinal, abaixo).
O *r* de Perason é dado por

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

A significância é calculada por meio de um teste *t* bicaudal com 2 graus de liberdade e

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

Dados ausentes: suportados por deleção.

Correlação linear parcial

Usando essa opção, é calculada, para cada par de colunas, a correlação linear controlando todas as colunas remanescentes. Por exemplo, com três colunas A, B, C a correlação AB é controlada para C; AC é controlada para B; BC é controlada para A. A correlação parcial linear pode ser definida como a correlação dos resíduos depois de calcular a regressão com a(s) variável(is) controlada(s). A significância é estimada com um teste *t* com *n*-2-*k* graus de liberdade, onde *k* é o número de variáveis controladas:

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

Dados ausentes: suporte por deleção.

Referência

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

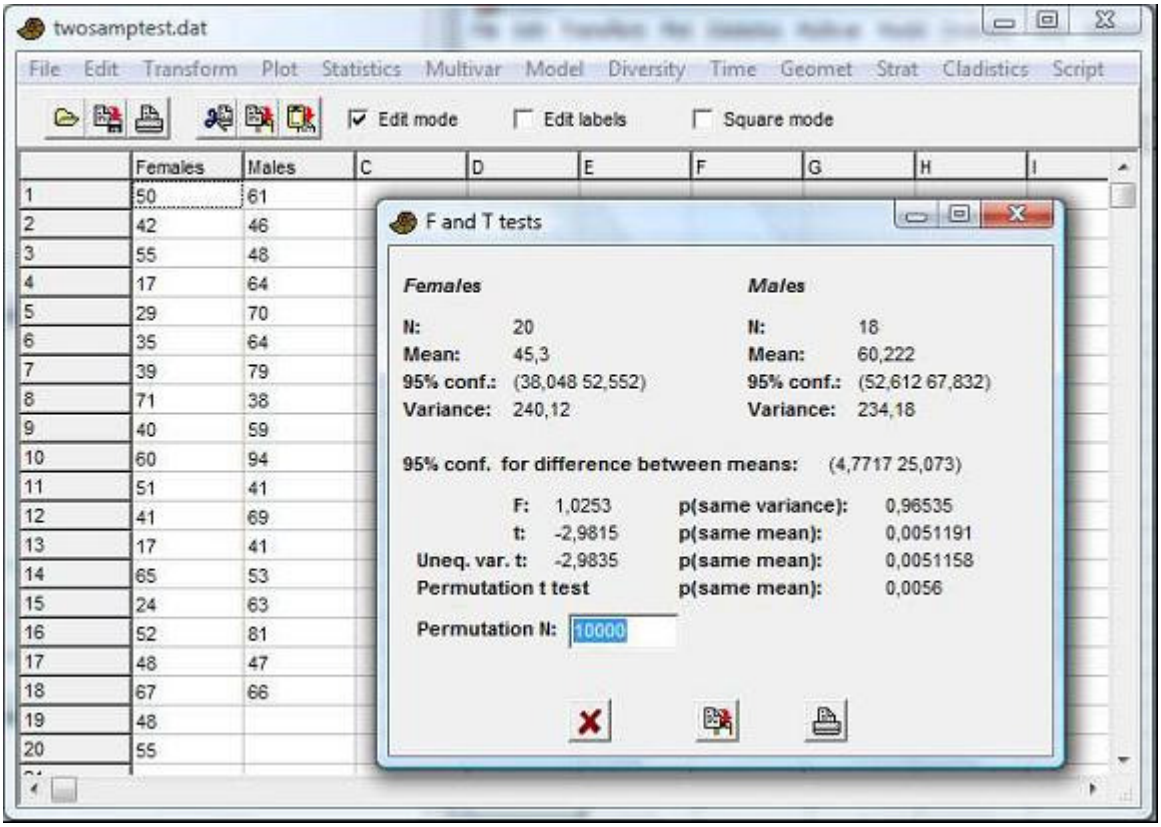
Var-covar

Apresenta uma matriz simétrica com as variâncias e covariâncias entre todos os pares de colunas.

Dados ausentes: suporte por deleção.

Testes F e t (duas amostras) (F and t tests (two samples))

Uma série de testes paramétricos clássicos e testes para comparão as médias e variâncias de duas amostras univariadas (em duas colunas). Assume-se distribuição normal.



Estatísticas da amostra

Média e variância são estimadas como descrito acima, sob *Estatística univariada*. O intervalo de confiança de 95% para a média é baseado no erro padrão para a estimatitva de média e na distribuição *t*. Sendo *s* a estimativa do desvio padrão, o intervalo de confiança é

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

Aqui, *t* tem *n*-1 graus de liberdade, e 1-α=0.95 para um intervalo de confiança 95%.
O intervalo de confiança 95% para a diferença entre as médias aceita tamanhos amostrais desiguais:

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

onde

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

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

$$MSE = SSE / df$$

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

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

O intervalo de confiança é calculado para a média maior menos a menor, i.e. o centro do IC sempre deve ser positivo. O intervalo de confiança para a diferença das médias é também estimado por *bootstrap*, com 9999 replicações.

Teste *F* (*F* test)
O teste *F* tem como hipótese nula
H_0: As duas amostras são tomadas de populações com variância igual.
A estatística *F* é a razão da maior variância pela menor variância. A significância é bicaudal, com n_1 e n_2 graus de liberdade.

Teste *t* (*t* test)
O teste *t* tem a hipótese nula
H_0: As duas amostras são tomadas de populações com médias iguais.
A partir do erro padrão s_D da diferença das médias dadas acima, a estatística de teste é

$$t = \frac{\bar{x} - \bar{y}}{\bar{s}_D}.$$

Teste *t* para variâncias desiguais (Unequal variance *t* test)
O teste *t* para variâncias desiguais também é conhecido como o teste de Welch. Pode ser usado como alternativa para o teste *t* básico quando as variâncias são muito diferentes, embora pode ser argumentado que o teste para a diferenças nas médias neste caso é questionável. A estatística de teste é

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

O número de graus de liberdade é

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

Teste por permutação (*Permutation test*)
O teste por permutação para igualdade das médias usa a diferença absoluta nas médias como estatística do teste. O teste por permutação é não-paramétrico com poucas premissas. O número de permutações pode ser definido pelo usuário. O poder do teste é limitado pelo tamanho amostral – significância no nível de $p<0.05$ só pode ser conseguida para $n>3$ em cada amostra.

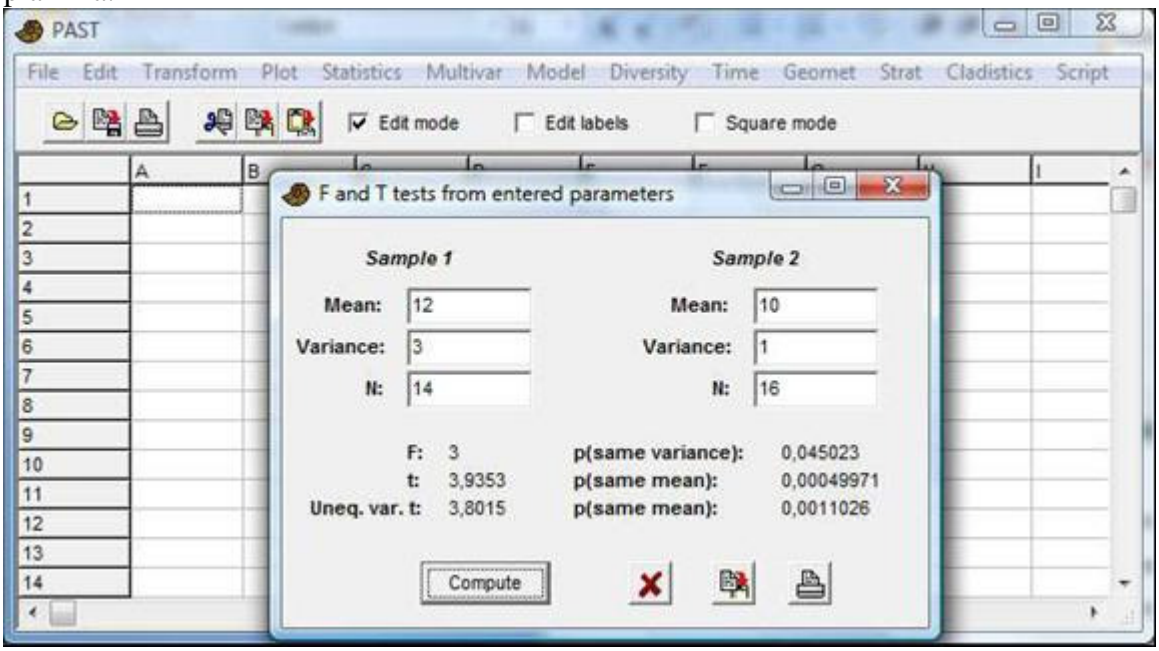
Dados ausentes: suporte por deleção.

Teste t (uma amostra) (t test (one sample))

O teste *t* de uma amostra (*one-sample t test*) é usado para investigar se é provável que uma amostra tenha sido retirada de uma população com uma dada média (teórica).
O intervalo de confiança de 95% para a média é calculado por meio da distribuição *t*.
Dados ausentes: suporte por deleção.

Testes F e t a partir de parâmetros (F and t tests from parameters)

Às vezes, as publicações não fornecem os dados, mas fornecem valores para tamanhos amostrais, média e variância de duas amostras. Estes podem ser inseridos manualmente usando a opção "*F and t from parameters*" no menu. Esse módulo não usa dados da planilha.

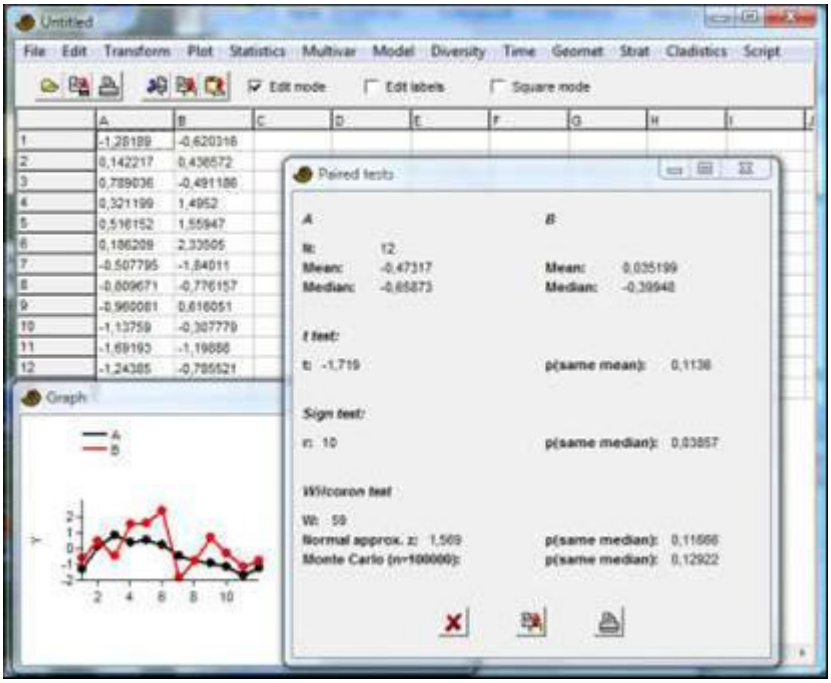


Testes pareados (t, sinal, Wilcoxon) (Paired tests (t, sign, Wilcoxon)

Três testes estatísticos (um paramétrico, dois não-paramétricos) para duas amostras (colunas) de dados univariados. Os pontos de dados são pareados, significando que os dois valores de cada linha são associados. Por exemplo, o teste pode ser usado para comparar o comprimento o braço esquerdo vs. braço direito de um grupo de pessoas, ou a diversidade no verão vs. no inverno de uma série de sítios. Controlado por um "fator de ruído" (*"nuisance factor"*) (pessoa, sítio), aumenta-se assim o poder do teste. A hipótese nula é:

H_0: A média (teste *t*) ou mediana (teste de sinal, teste de Wilcoxon) da diferença é zero.

Todos os valores de *p* relatados são bicaudais.



Teste *t* (*t test*)
Testa se a diferença média é igual a zero por meio de um teste *t* comum de uma amostra. Sendo $d_i = x_i - y_i$, temos

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

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

Há n-1 graus de liberdade. O teste assume distribuição normal das diferenças.

Teste de sinal (Sign test)
O teste de sinal (binomial) simplesmente conta o número de casos n_1 em que $x_i > y_i$ e n_2 em que $y_i > x_i$. O valor de p é exato, calculado a partir da distribuição binomial. O teste de sinal tipicamente terá menor poder explanatório do que os outros testes pareados, mas apresenta menos premissas.

Teste de ranks com sinal de Wilcoxon (*Wilcoxon signed rank test*)
Um teste não-paramétrico de ranks que não assume distribuição normal. A hipótese nula é de que não há deslocamente da mediana (sem diferenças).
Inicialmente, todas as linhas com diferença de zero são removidas pelo programa. Então os valores absolutos das diferenças $|d_i|$ são ranqueados (R_i), com ranks médios atribuídos a valores repetidos (*ties*). A somatória dos ranks para os pares em que d_i é positivo é W^+. A somatória dos ranks para pares em que d_i é negativo é W^-. A estatística relatada é
$W = \max(W^+, W^-)$
(repare que existem outras versões deste teste que são equivalentes a esta, mas que relatam outras estatísticas).

Para n grande (digamos n>10), a aproximação do p para grandes amostras (*large-sample approximation to p*) pode ser usada. Isso depende da distribuição normal da estatística de teste W:

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

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

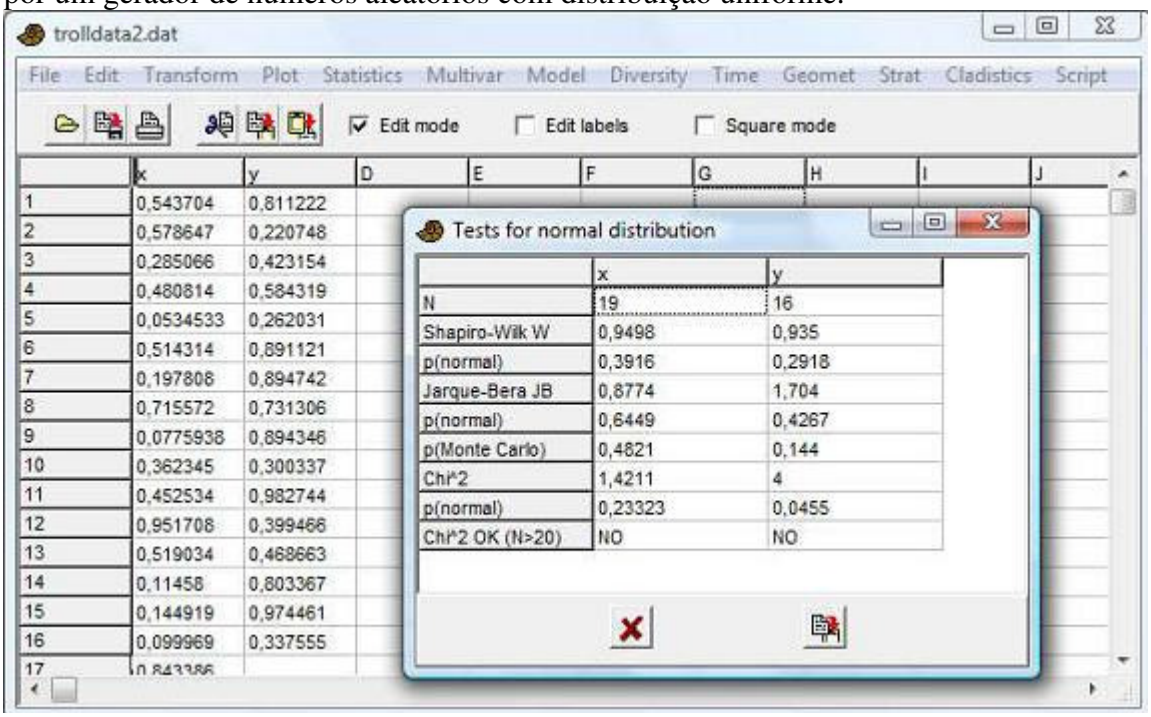
O último termo é uma correção para valores repetidos, onde f_g é o número de elementos no conjunto de valores repetidos g. O z resultante é relatado juntamente com o valor de p. O valor de significância de Monte Carlo é baseado em 99 999 remanejamentos aleatórios de valores entre as colunas dentro de cada par. Este valor será praticamente idêntico ao valor exato do p.

Para n<26, um valor exato de p é calculado por enumeração completa de todos os remanejamentos possíveis (há 2^n remanejamentos possível, i.e. mais de 33 milhões n=25). Este é o valor preferível quando disponível.

Dados ausentes: suporte por deleção da linha.

Testes de normalidade (Normality tests)

Quatro testes estatísticos para distribuição normal de apenas uma ou de uma série de amostras univariadas de dados, fornecidos em colunas. Os dados abaixo foram gerados por um gerador de números aleatórios com distribuição uniforme.



	x	y
1	0,543704	0,811222
2	0,578647	0,220748
3	0,285066	0,423154
4	0,480814	0,584319
5	0,0534533	0,262031
6	0,514314	0,891121
7	0,197808	0,894742
8	0,715572	0,731306
9	0,0775938	0,894346
10	0,362345	0,300337
11	0,452534	0,982744
12	0,951708	0,399466
13	0,519034	0,468663
14	0,11458	0,803367
15	0,144919	0,974461
16	0,099969	0,337555
17	0,843386	

	x	y
N	19	16
Shapiro-Wilk W	0,9498	0,935
p(normal)	0,3916	0,2918
Jarque-Bera JB	0,8774	1,704
p(normal)	0,6449	0,4267
p(Monte Carlo)	0,4821	0,144
Chi^2	1,4211	4
p(normal)	0,23323	0,0455
Chi^2 OK (N>20)	NO	NO

Para os quatro testes, a hipótese nula é

H_0: A amostra foi retirada de uma população com distribuição normal.

Se o p(normal) fornecido for menor do que 0.05, distribuição normal pode ser rejeitada. Dos quatro testes fornecidos, os de Shapiro-Wilk e de Anderson-Darlink são considerados os mais exatos, e os outros dois testes (Jarque-Bera e um teste por qui-quadrado (*chi-square*)) são fornecidos como referência. Existe um tamanho amostral máximo de n=5000, enquanto o tamanho amostral mínimo é 3 (é claro que os testes terão poder muito pequeno para um n tão baixo).
Lembre-se da questão dos testes múltiplos caso você analise mais de uma amostra por esses testes – uma correção de Bonferroni ou uma outra pode ser apropriada.

Teste de Shapiro-Wilk (*Shapiro-Wilk test*)
O teste de Shapiro-Wilk (Shapiro & Wilk 1965) retorna uma estatística de teste W, que é pequena para amostras não-normais, e um valor de p. A implementação é baseada no código padrão "AS R94" (Royston 1995), corrigindo uma inacurácia para tamanhos amostrais grandes no algoritmo interior "AS 181".

Teste de Jarque-Bera (*Jarque-Bera test*)
O teste de Jarque-Bera (Jarque & Bera 1987) é baseado na assimetria S e na curtose K. A estatística de teste é

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

Neste contexto, a assimetria e a curtose usadas são

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

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

Repare que estas equações contêm estimadores mais simples de G_1 e G_2 do que os fornecidos acima, e que a curtose aqui será igual a 3, não a zero, para uma distribuição normal.
Assimptoticamente (para tamanhos amostrais grandes), a estatística de teste tem uma distribuição de qui-quadrado com dois graus de liberdade, e isso forma a base do valor de p fornecido pelo Past. Sabe-se que essa abordagem funciona bem apenas para tamnhos amostrais grandes, e o Past também inclui um teste de significância baseado numa simulação de Monte Carlo, com 10 000 valores aleatórios tomados de uma distribuição normal.

Teste de qui-quadrado (*Chi-square test*)
O teste de qui-quadrado usa uma distribuição normal esperada com quatro classes (*bins*) com base na média e no desvio padrão estimados da amostra, e construída de modo a ter frequências esperadas iguais em todas as classes. O limite superior de todas as classes e as frequências observadas e esperadas são mostradas. Uma mensagem de aviso é dada se n<20, i.e. frequência esperada em cada classe menor do que 5. Há 1 grau de liberdade.

Esse teste é questionável teoricamente e tem baixo poder, e não é recomendado. É incluído para referência.

Teste de Anderson-Darling (*Anderson-Darling test)*
Os dados X_i são ordenados em ordem crescente e normalizados para média e desvio padrão:

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

Sendo F a função de distribuição cumulativa (*cumulative distribution function* - CDF) da normal, a estatística do teste é

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

A significância é estimada de acordo com Stephens (1986). Inicialmente, uma correção para tamanho amostral pequeno é aplicada:

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

O valor de p é estimado por

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

Dados ausentes: suporte por deleção.

Referências
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.

Qui^2 (Chi^2)

O Qui-quadrado (*Chi-square*) espera duas colunas com números de elementos em diferentes classes (compartimentos). Por exemplo, esse teste pode ser usado para comparar duas associações (colunas) com o número de indivíduos de cada táxon organizado nas linhas. Você deve ter cautela com esse teste caso alguma(s) das células tenha(m) menos de cinco indivíduos (ver teste exato de Fisher abaixo).
Há duas opções que devem ser selecionadas ou não para obter resultados corretos. "*Sample vs. expected*" ("Amostra vs. esperado") deve ser selecionado se a sua segunda coluna consiste de valores retirados de uma distribuição teórica (valores esperados) com

barras de erro iguais a zero. Se seus dados são de duas amostras de contagem, cada uma com barras de erro, deixe esta caixa desmarcada. Isso *não é* uma correção para amostra pequena.
"*One constraint*" ("Uma restrição") deve ser marcada se os valores esperados foram normalizados para se ajustar ao número total de eventos observados, ou se as duas amostras contadas têm necessariamente os mesmos valores totais (por exemplo, por serem porcentagens). Isso irá reduzir em um o número de graus de liberdade.
Quando a opção "*one constraint*" está selecionada, um teste de permutação é disponibilizado, com 10000 réplicas aleatórias. Para "*Sample vs. expected*" essas réplicas são geradas mantendo os valores esperados fixos, enquanto os valores da primeira coluna são aleatórios com probabilidades relativas como especificado pelos valores esperados e com somatória constante. Para duas amostras, todas as células são aleatórias mas com somatórias constantes de linhas e colunas.
Veja e.g. Brown & Rothery (1993) ou Davis (1986) para detalhes.
Com uma restrição, o teste exato de Fisher (bicaudal) também é fornecido. Quando disponível, o teste exato de Fisher pode ser muito melhor do que o qui-quadrado. Para grandes tabelas ou grandes contagens, o tempo de cálculo pode ser proibitivo e se esgotará depois de um minuto. Nesses casos o teste paramétrico é provavelmente aceitável de qualquer modo. O procedimento é complexo e baseado no algoritmo de rede de Mehta & Patel (1986).

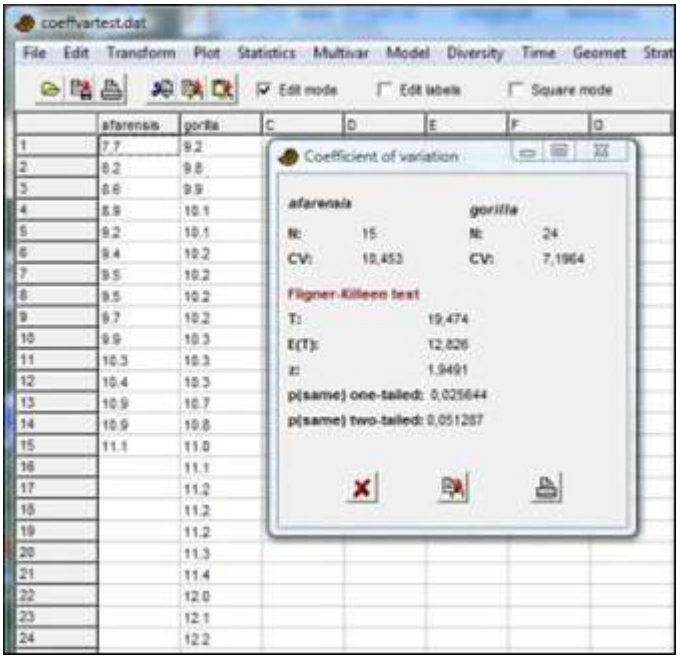
Dados ausentes: Suporte por deleção de linha.

Referências

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.

Coeficiente de variação (Coefficient of variation)

Este módulo testa se duas amostras, fornecidas em duas colunas, têm coeficiente de variação igual.



O coeficiente de variação (ou variação relativa) é definido como a razão do desvio padrão e da média em porcentagem, e é calculado por:

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

Os intervalos de confiança de 95% são estimados por *bootstrap*, com 9999 réplicas.
A hipótese nula do teste estatístico é:
H_0: As amostras foram retiradas de populações com o mesmo coeficiente de variação.
Se o valor de *p*(same) fornecido for menor do que 0.05, coeficientes de variação iguais podem ser rejeitados. Donnelly & Kraem (1999) descrevem o coeficiente de variação e revisam uma série de testes estatísticos para a comparação de duas amostras. Eles recomendam o teste de Fligner-Killeen (Fligner & Killeen 1976), como implementado no Past. Este teste é poderoso e relativamente insensível à distribuição dos dados. As seguintes estatísticas são relatadas:

T:	A estatística de teste de Fligner-Killeen, correspondente à somatória das posições ranqueadas e transformadas da amostra menor dentro da amostra agrupada (veja Donnelly & Kramer 1999 para detalhes).
E(*T*):	O valor esperado de *T*.
z:	A estatística *z*, baseada em *T*, Var(*T*) e E(*T*). Observe que isso é uma aproximação de amostra grande.
p:	O valor de $p(H_0)$. São fornecidos os valores unicaudal e bicaudal. Para a hipótese alternatica de diferença em qualquer direça, o valor bicaudal deve ser usado. No entanto, o teste de Fligner-Killeen já foi usado para comparar a variação dentro de uma amostra de fósseis com a variação dentro de uma espécie moderna com parentesco próximo, para testar se havia múltiplas espécies fósseis (Donnekky & Kramer 1999). Neste caso, a hipótese alternativa poderia ser a de que o CV é maior na população fóssil; neste caso um teste unicaudal pode ser usado para ter aumentar o poder do teste.

A imagem de tela acima reproduz o excemplo de Donnelly & Kramer (1999), mostrando que a variação relativa dentro de *Australopithecus afarensis* é significativamente maior do que em *Gorilla gorilla*. Isso poderia indicar que *A. afarensis* representa mais de uma espécie.

Dados ausentes: Suporte por deleção.

Referências

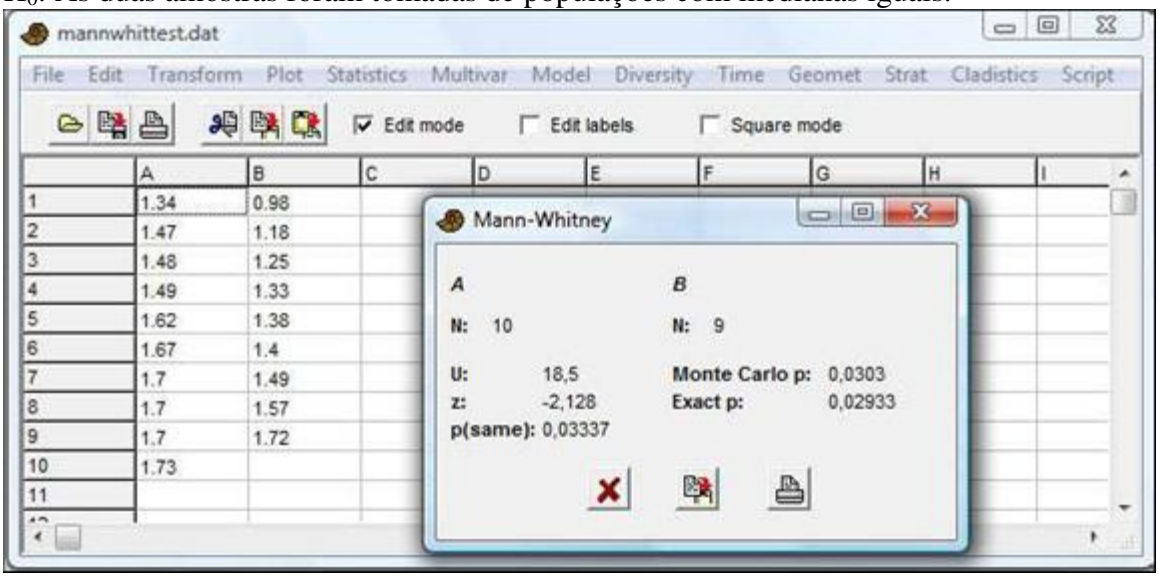
Donnelly, S.M. & Kramer, A. 1999. Testing for multiple species in fossil samples: An evaluation and comparison of tests for equal relative variation. *American Journal of Physical Anthropology* 108:507-529.

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

Teste de Mann-Whitney (Mann-Whitney test)

O teste bicaudal U de Mann-Whitney (Wilcoxon) pode ser usado para testar se as medianas de duas amostras independentes são diferentes. Ele não assume distribuição normal, mas assume que as distribuições dos dois grupos têm a mesma forma. A hipótese nula é

H_0: As duas amostras foram tomadas de populações com medianas iguais.



O teste é não-paramétrico, o que significa que as distribuições podem ter qualquer forma. Para cada valor na amostra 1, conte o número de valores na amostra 2 que são menores do que ele (valores repetidos contam como 0.5). O total dessas contagens é a estatística de teste *U* (às vezes chamada de *T*). Se o valor de *U* for menor quando a ordem das amostras é revertida, este valor é escolhido no seu lugar (pode ser mostrado que $U_1+U_2=n_1n_2$).

Na coluna da esquerda é dada uma aproximação assimptótica ao *p* com base na distribuição normal (bicaudal), que só é válida para *n* grandes. Ela inclui uma correção para continuidade e uma correção para valores repetidos:

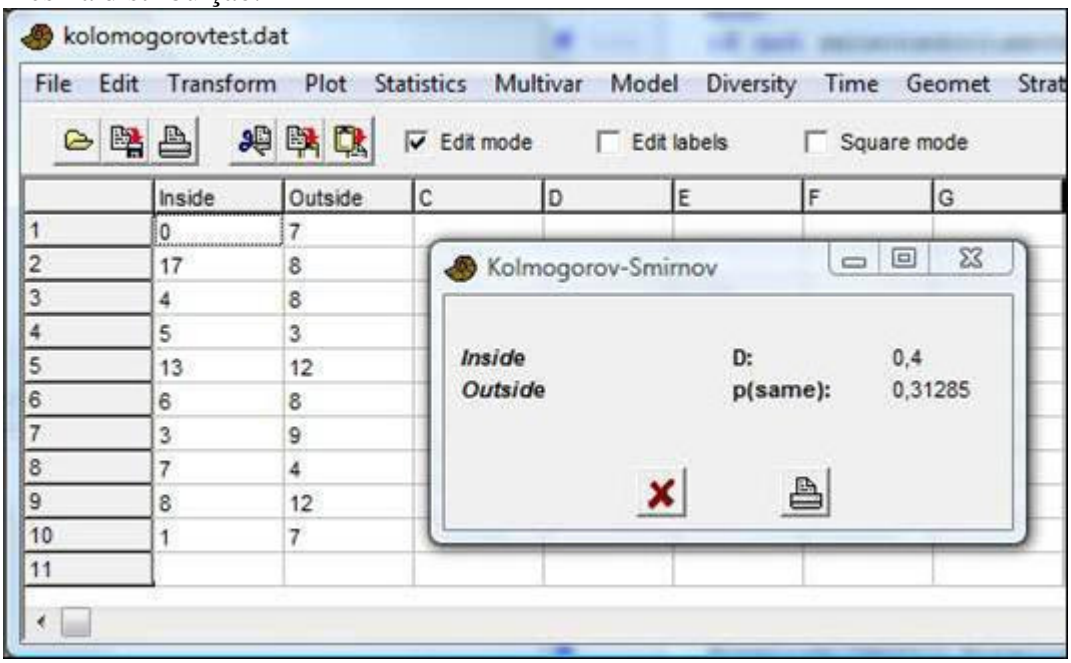
$$z = \frac{U - n_1 n_2 / 2 + 0.5}{\sqrt{\frac{n_1 n_2 \left(n^3 - n - \sum_g f_g^3 - f_g \right)}{12n(n-1)}}}$$

onde $n=n_1n_2$ e f_g é o número de elementos no conjunto de elementos repetidos (*tie*) g. Para $n_1+n_2<=30$ (e.g. 15 valores em cada grupo), um valor exato de p é fornecido, baseado em todas as combinações possíveis de realocações de elementos entre os grupos. Sempre use este valor exato se ele está disponível. Para amostras grandes, a aproximação asimptótica é bastante precisa. Um valor por Monte Carlo baseado em 10 000 realocações aleatórias também é fornecido – o principal objetivo disso é servir de controle ao valor assimptótico.

Dados ausentes: suporte por deleção.

Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov é um teste não paramétrico que testa se duas distribuições univariadas apresentam a mesma distribuição geral. Em outras palavras, este teste não testa especificamente a igualdade de média, variância ou qualquer outro parâmetro. A hipótese nula é H_0: As duas amostras foram tomadas de populações com a mesma distribuição.



Na versão do teste que é fornecida no Past, ambas as colunas devem representar amostras. Você não pode testar uma amostra contra uma distribuição teórica (teste de uma amostra – *one-sample test*).

A estatística de teste é a diferença absoluta máxima entre duas funções de distribuição cumulativas empíricas:

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

O algoritmo é baseado em Press et al. (1992), com a significância estimada de acordo com Stephens (1970).
Defina a função

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

Sendo $N_e = N_1 N_2/(N_1+N_2)$, a significância é calculada por

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

O teste por permutação usa 10 000 permutações. Use o valor de p por permutação para $N<30$ (ou no geral).

Dados ausentes: suporte por deleção.

Referências

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.

Correlação ordinal/de rank (Rank/ordinal correlation)

Essas correlações e testes de ordem de rank (*rank-order correlations*) são usadas para investigar a correlação entre duas variáveis, fornecidas em duas colunas.
O coeficiente de correlação (não-paramétrica) de ordem de rank de Spearman é o coeficiente de correlação linear (r de Pearson) dos ranks. De acordo com Press et al. (1992), é calculado por

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

Aqui, D é a soma do quadrado da diferença dos ranks (ranks intermediários para valores repetidos):

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

Os f_k é o número de valores repetidos no $k$ésimo grupo de valores repetidos entre os R_i's, e os g_m são os números de valores repetidos no $m$ésimo grupo de valores repetidos entre os S_i's.
Para $n>9$, a probabilidade de r_s diferente de zero (bicaudal) é calculada por meio de um teste t com n-2 graus de liberdade:

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

Para *n* pequeno, essa aproximação é imprecisa, e para $n<=9$ o programa portanto alterna automaticamente para um teste exato. Esse teste compara o r_s observado com os valores obtidos com todas as permutações possíveis da primeira coluna.
O teste por Monte Carlo é baseado em 9999 amostras aleatórias.
Essas estatísticas também estão disponíveis no módulo "*Correlation*", mas sem a opção de permutação.

Dados ausentes: suporte por deleção.

Correlação poliserial (*Polyserial correlation*)

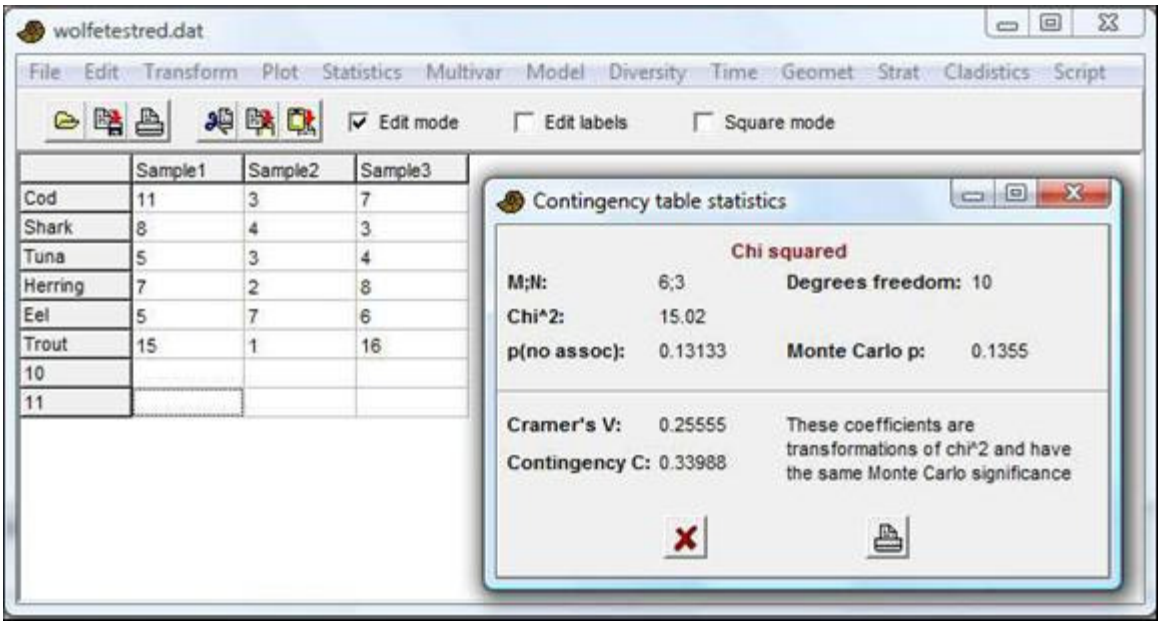
Essa correlação só é calculada se a segunda coluna consiste de valores inteiros com uma amplitude menor do que 1000. É delineada para correlacionar uma variável contínua/de intervalo com distribuição normal (primeira coluna), com uma variável ordinal (segunda coluna) cujas classes representam uma variável de distribuição normal. Por exemplo, a segunda coluna poderia conter os números 1-3 codificando "pequeno", "médio" e "grande". Tipicamente, haveria mais valores "médio" do que "pequeno" ou "grande" devido à distribuição normal de tamanhos que está por trás da amostra.
O Past uso o algoritmo de dois passos de Olsson et al. (1982). Esse algoritmo é mais preciso do que o estimador "ad hoc" desses autores, e quase tão preciso quanto o algoritmo multivariado de máxima verossimilhança completo (*full multivariate ML algoritm*). O algoritmo de dois passos foi escolhido por causa da velocidade, permitindo um teste por permutação (mas apenas para $N<100$). Para *N* maiores, o teste assimptótio (teste de *log-ratio*) tem precisão.

Referências

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.

Tabela de contingência (Contingency table)

Uma tabela de contingência é o *input* dessa rotina. Linhas representam os diferentes estados de uma variável nominal, colunas representam os estados de outra variável nominal, e células contêm as contagens de ocorrências daquele estado específico (linha, coluna) das duas variáveis. A significância da associação entre as duas variáveis (com base em qui-quadrado) é então fornecida, com valores de *p* a partir de uma distribuição de qui-quadrado e de um teste por permutação com 9999 replicações.
Por exemplo, linhas podem representar táxons e colunas amostras, como usual (com contagens de espécimes nas células). A análise da tabela de contingência então fornece informações se as duas variáveis de táxon e local estão associadas. Se não estiverem, a matriz de dados não é muito informativa.



Duas medidas adicionais de associação são fornecidas. Ambas são transformação do qui-quadrado (Press et al. 1992). Sendo n a somatória total das contagens, M o número de linhas e N o número de colunas:

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

Coeficiente de contingência C: $$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Note que para tabelas nx2, o teste exato de Fisher (*Fisher's exact test*) é disponibilizado no módulo Chi^2.

Não há suporte para dados ausentes.

Referência

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

ANOVA Uni-fatorial (One-way ANOVA)

ANOVA (análise de variância) unifatorial é um procedimento estatístico para testar a hipótese nula de que uma série de amostras univariadas (em colunas) são tomadas de populações com a mesma média. Assume-se que as amostras têm distribuição próxima da normal e variâncias similares. Se os tamanhos amostrais são iguais, essas premissas não são críticas. Caso as premissas sejam seriamente violadas, o teste não-paramétrico de Kruskal-Wallis deve ser usado ao invés da ANOVA.

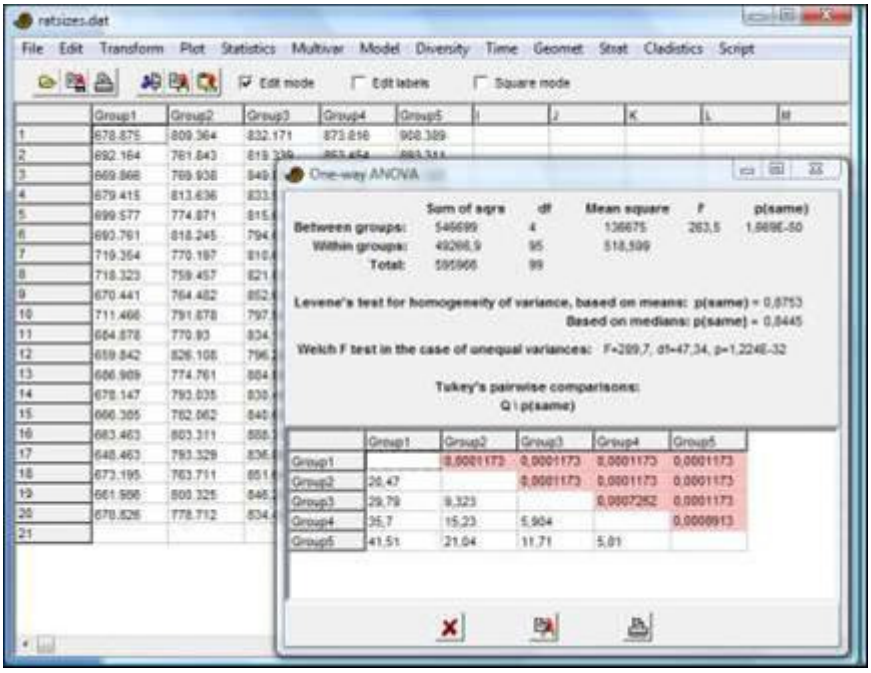


Tabela da ANOVA

A soma dos quadrados entre-grupos (*between-groups sum of squares*) é dada por:

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

onde n_g é o tamanho do grupo g e as métidas são médias total e de grupo. A soma entre-grupos tem um número associado de graus de liberdade, $\mathrm{df}_{\mathrm{bg}}$, igual ao número de grupos menos 1.

A soma dos quadrados intra-grupos (*within-groups sum of squares*) é

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

onde x_i são aqueles do grupo g. A soma dos quadrados intra-grupos tem um número associado de graus de liberdade, $\mathrm{df}_{\mathrm{wg}}$, igual ao número total de valores menos o número de grupos.

Os quadrados médios (*mean squares*) entre e intra-grupos são dados por

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

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

Finalmente, a estatística F é calculada por

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

O valor de p é baseado no F com $\mathrm{df}_{\mathrm{bg}}$ e $\mathrm{df}_{\mathrm{wg}}$ graus de liberdade.

Omega quadrado

O ômega quadrado é uma medida da intensidade do efeito (*effect size*), variando de 0 a 1 (não disponível para ANOVA de medida repetida):

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

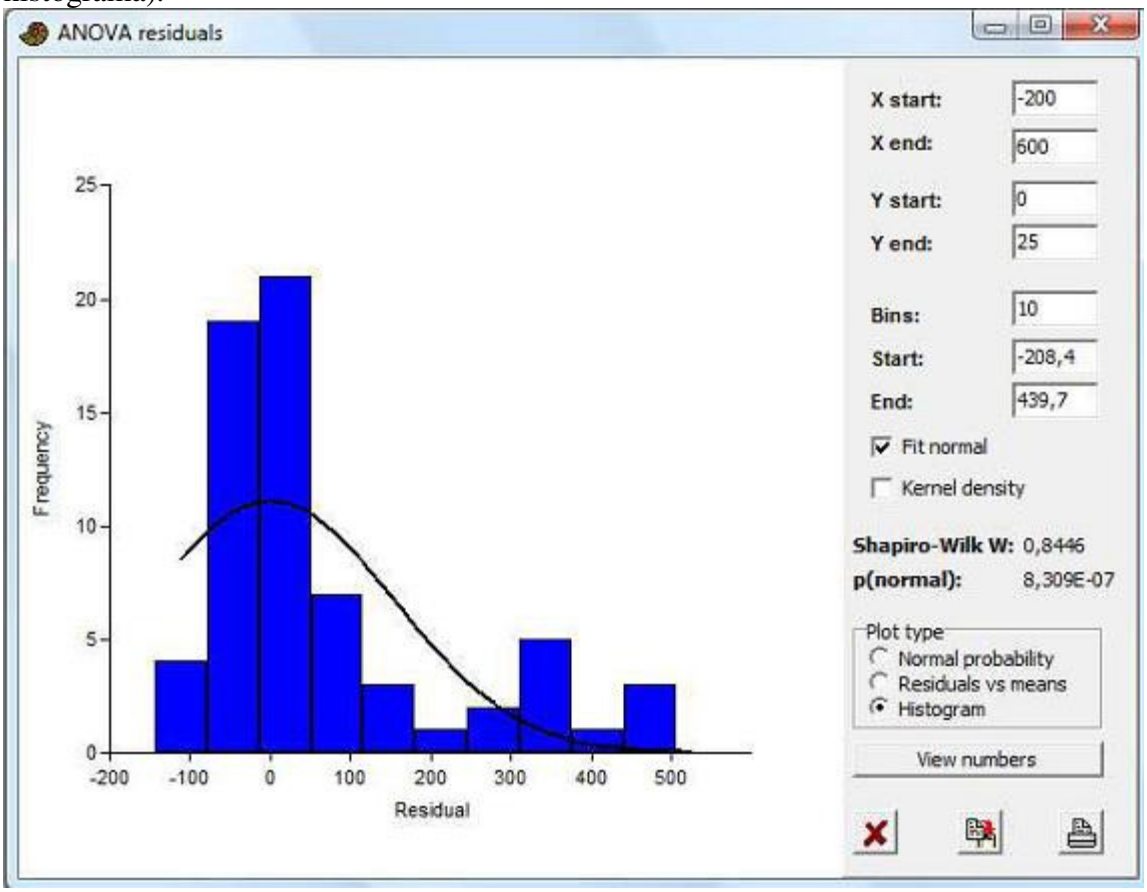
Teste de Levene (*Levene's test*)

Se o teste de Levene é significativo, significando que as amostras apresentam variâncias desiguais, pode ser usado a versão de ANOVA para variâncias desiguais (Welch), com os valores correspondentes de *F*, df e *p*.

Análise dos resíduos

O botão *"Residuals"* abre uma janela para analisar as propriedades dos resíduos para avaliar algumas premissas da ANOVA, tais como uma distribuição normal e homoscedástica dos resíduos.

É fornecido o teste de Shapiro-Wilk para distribuição normal, juntamente com alguns gráficos comuns de resíduos (probabilidade normal, resíduos vs. médias dos grupos, e histograma).



Testes par-a-par *post-hoc*

Se a ANOVA mostra desigualdade significativa das médias (*p* pequeno), você pode partir para a análise da tabela de comparações par-a-par "post-hoc", com base na DHS (Diferença Honestamente Significativa – *Honestly Significant Difference*) de Tukey. A *Studentized Range Statistic Q* é fornecida no triângulo inferior esquerdo da matriz, e as

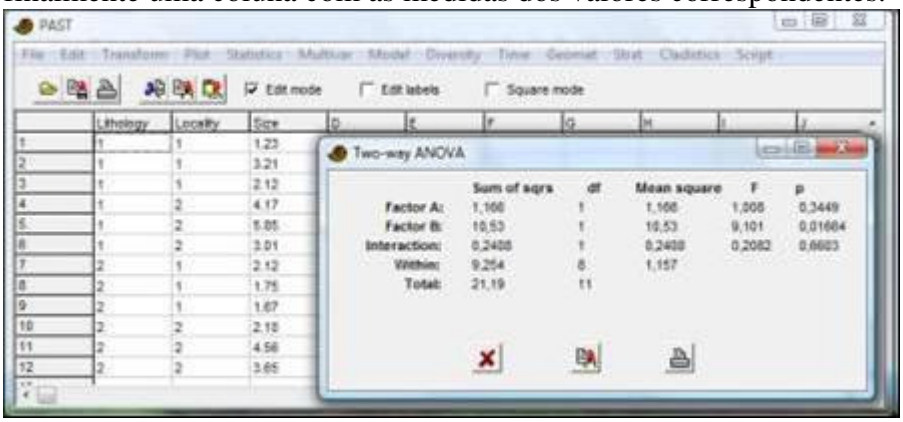
probabilidades *p(igual)* são fornecidas no triângulo superior direito. Tamanhos amostrais não precisam ser iguais para a versão do teste de Tukey utilizada.

ANOVA de medidas repetidas (intra-sujeitos) (*Repeated measures (within-subjects) ANOVA*)
Marcando a caxa "*Repeated measures*", seleciona-se um outro tipo de ANOVA, na qual os valores em cada coluna são observações do mesmo "sujeito". ANOVA de medida repetida é a extensão do teste *t* pareado para várias amostras. Cada coluna (amostra) precisa conter o mesmo número de valores.

Valores ausentes: suporte por deleção, exceto para ANOVA de medidas repetidas, na qual não há suporte para valores ausentes.

ANOVA bifatorial (Two-way ANOVA)

A ANOVA (análise de variância) bifatorial é uma medida estatística para testar a hipótese nula de que uma série de amostras univariadas têm a mesma média em relação a cada um de dois fatores, e que não há dependências (interações) entre fatores. Assume-se que as amostras têm distribuição próxima da normal e variâncias similares. Se os tamanhos amostrais forem iguais, essas premissas não são críticas. O teste assume um delineamento de fator fixo (*fixed-factor design*) (o caso mais comum).
Três colunas são necessárias. Primeiro, uma coluna com os níveis do primeiro fator (codificadas como 1, 2, 3 etc), depois uma coluna com os níveis do segundo fator, e finalmente uma coluna com as medidas dos valores correspondentes.



O algoritmo utiliza médias ponderadas para delineamentos não-balanceados.

ANOVA de medidas repetida (intra-sujeitos) (*Repeated measures (within-subjects) ANOVA*)
Selecionando a caixa "Repeated measures" seleciona um outro tipo de ANOVA bifatorial, na qual cada um de uma série de "sujeitos" tenha recebido uma série de tratamentos. O formato dos dados é como acima, mas é preciso que todas as medidas do primeiro sujeito sejam dadas nas primeiras linhas, depois todas as medidas do segundo sujeito, etc. Cada sujeito deve ter recebido todas as combinações de tratamentos, e cada combinação de tratamentos deve ter sido fornecida uma única vez. Isso significa que para, por exemplo, dois fatores, com 2 e 3 níveis, cada sujeito deve ocupar exatamente

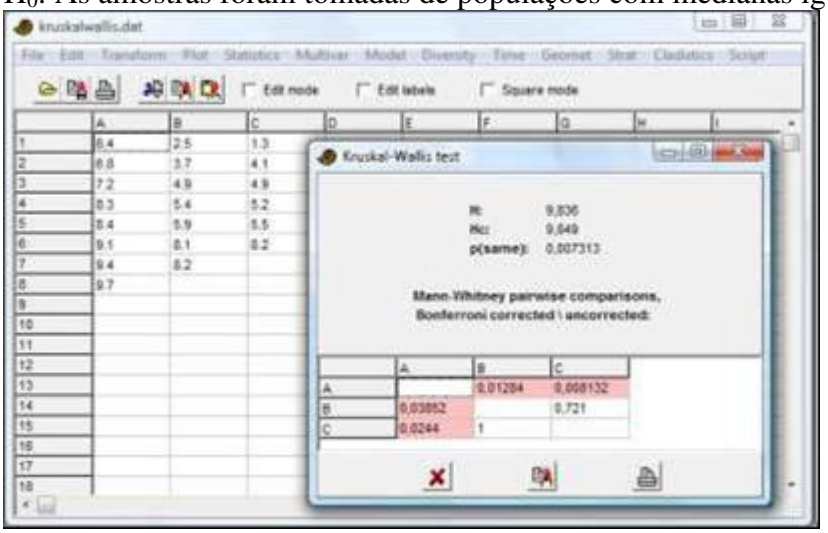
2x3=6 linhas. O programa automaticamente calcula o número de sujeitos pelo número de combinações de níveis e o número total de linhas.

Valores ausentes: linhas com valores ausentes são deletadas.

Kruskal-Wallis

O teste de Kruskal-Wallis é uma ANOVA não-paramétrica, que compara as médias de uma série de grupos univariados (fornecidos em colunas). Pode ser considerado uma extensão do teste de Mann-Whitney para vários grupos (Zar 1996). Não assume distribuição normal, mas assume que todos os grupos tenham a mesma distribuição. A hipótese nula é

H_0: As amostras foram tomadas de populações com medianas iguais.



A estatística de teste H é calculada da seguinte maneira:

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

sendo n_g o número de elementos no grupo g, n o número total de elementos, e T_g a soma de *ranks* no grupo g.

A estatística de teste H_c é ajustada para valores repetidos (*ties*):

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

onde f_i é o número de elementos no grupo i de elementos repetidos.

Sendo G o número de grupos, o valor de p é aproximado a partir de H_c por meio da distribuição de qui-quadrado com G-1 graus de liberdade. A precisão dessa aproximação é menor se algum n_g<5.

Testes par-a-par post-hoc (*Post-hoc pairwise tests*)

Valores de p de testes par-a-par de Mann-Whtney são fornecidos para todos os N_p=G(G-1)/2 pares de grupos, no triângulo superior direito da matriz. O triângulo inferior

esquerdo fornece os valores de p correspondentes, mas multiplicados por N_p como uma correção conservativa para testes múltiplos (correção de Bonferroni). Os valores usam a aproximação assimptótica descrita para Mann-Whitney. Caso as amostras sejam muito pequenas, pode ser útil usar o teste exato disponível em Mann-Whitney no lugar destas comparações.

Dados ausentes: suporte por deleção.

Referência

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

Teste de Friedman (Friedman test)

O teste de Friedman é um teste não-paramétrico para igualdade de médias em uma série de grupos univariados com medidas repetidas. Pode ser considerado uma versão não-paramétrica da ANOVA de medidas repetidas (*repeated-measures ANOVA*) ou a versão para medidas repetidas do teste de Kruskal-Wallis. Os grupos (tratamentos) são dados em colunas, e os sujeitos em linhas.

O teste de Friedman é feito de acordo com Bortz et al. (2000). A estatística de teste básica é

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

sendo n o número de linhas, k o número de colunas e T_j as somatórias das colunas da tabela de dados.

O valor de χ^2 é então corrigido para valores repetidos (caso existam):

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

sendo m o número total de grupos de valores repetidos e t_i o número de valores em cada grupo de valores repetidos.

Para k=2 é recomendado usar um dos testes pareados (e.g. testes de sinal ou de Wilcoxon) ao invés do teste de Friedman. Para conjuntos de dados pequenos com k=3 e n<10 ou k=4 e n<8, o valor de χ^2 com correção para valores repetidos é encontrado em um tabela de valores "exatos" de p. Quando disponível, é o valor de p preferível.

O valor assimptótico de p (usando a distribuição de χ^2 com k-1 graus de liberdade) é razoavelmente precisa para conjuntos grandes de dados. Ela é calculada a partir de uma versão de χ^2 com correção para continuidade:

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

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

Este valor de χ^2 é também corrigido para valores agrupados usando a equação acima.

Os testes “post hoc” são simplesmente comparações par-a-par de Wilcoxon, exatos para $n<20$ e assimptóticos para $n>=20$. Estes testes têm poder maior do que o teste de Friedman.

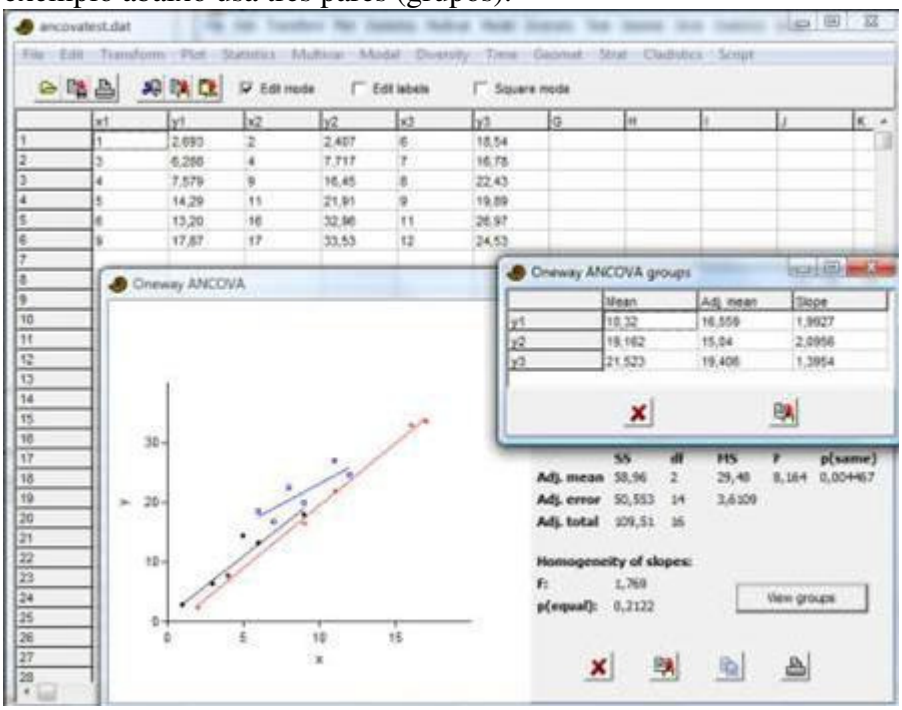
Não há suporte para valores ausentes.

Referência

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

ANCOVA unifatorial (One-way ANCOVA)

Testes de ANCOVA (análise de covariâncias) para igualdade de médias de uma série de grupos univariados com ajuste para covariância com outra variável. ANCOVA pode ser comparada a ANOVA, mas tem a característica adicional de que, para cada grupo, é removida a variância que pode ser explicada por uma covariável de “ruído” (x). Este ajuste pode aumentar substancialmente o poder do teste.
O programa espera dois ou mais pares de colunas, sendo cada para (grupo) é um conjunto de dados correlacionados x-y (médias são comparados para y, sendo x a covariável). O exemplo abaixo usa três pares (grupos).



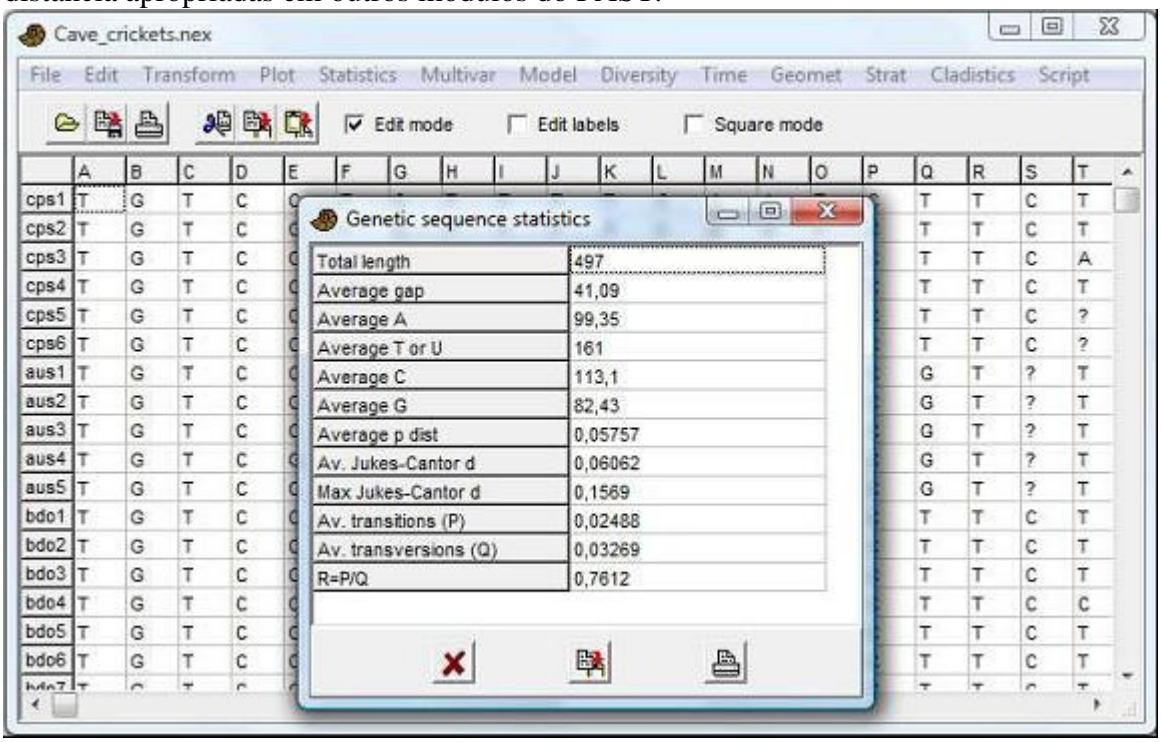
O programa apresenta um gráfico de dispersão e linhas de regressão linear para todos os grupos. A tabela de resumo, parecida com a tabela da ANOVA, contém soma-de-quadrados (*sum-of-squares*) etc., para as médias ajustadas (efeito entre-grupos) e para o erro ajustado (intra-grupo – *within-groups*), juntamente com um teste F para as médias ajustadas. Um teste F para e igualdade das inclinações da regressão (como assumido pela ANCOVA) também é fornecido. No exemplo, a igualdade das médias ajustadas nos três grupos pode ser rejeitada com $p<0.005$. Igualdade das inclinações não pode ser rejeitada ($p=0.21$).

"*View groups*" ("Ver grupos") fornece as estatísticas de resumo para cada grupo (média, média ajustada e inclinação da regressão).
Premissas incluem inclinações de regressão (*regression slopes*) similares em todos os grupos, distribuições normais, variância similar e tamanhos amostrais similares.

Dados ausentes: pares *x*-*y* com *x* ou *y* ausente são descartados.

Estatísticas de sequência genética (Genetic sequence stats)

Um número de estatísticas simples de dados de sequência genética (DNA ou RNA). O módulo espera um número de linhas, cada uma com uma sequência. Espera-se que as sequências estejam alinhadas e tenham o mesmo comprimento, incluíndo vazios (*gaps*) (codificados por "?"). Algumas destas estatísticas são úteis para selecionar medidas de distância apropriadas em outros módulos do PAST.



Comprimento total – *Total length*	O comprimento total, incluindo *gaps*, de uma sequência
Gap* médio – *Average gap	O número médio de posições com *gaps* em todas as sequências
Média de (*Average*) A, T/U, C, G	O número médio de posições contendo cada um dos nucleotídeos.
d* de Jukes-Cantor média – *Average Jukes-Cantor d	A distância *d* de Jukes-Cantor entre duas sequência, sendo feita a média entre todos os pares de sequências. $d = -3\ln(1-4p/3)/4$, sendo *p* a distância *p*.
d* de Jukes-Cantor máxima – *Maximal	A distância de Jukes-Cantor máxima entre quaisquer duas sequências

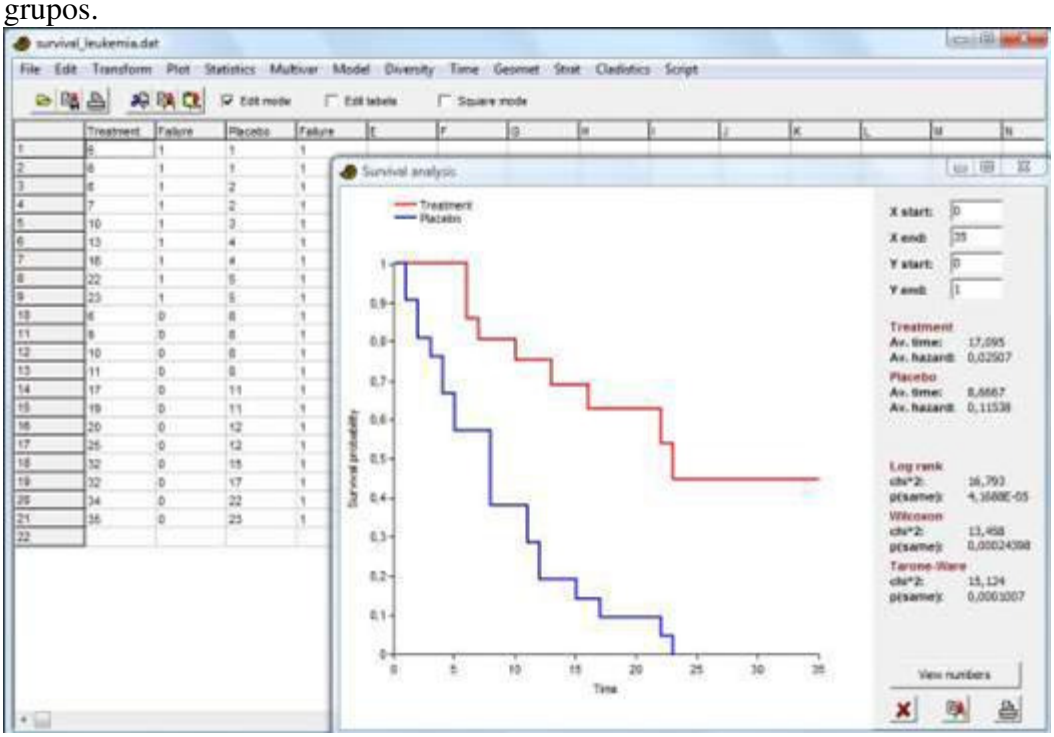
Jukes-Cantor d	
Média de transições (*P*) – *Average transitions (P)*	Número médio de transições (a↔g, c↔t, ou seja, dentro de purinas ou pirimidinas)
Transversões médias (*Q*) – *Average transversions (Q)*	Número médio de transversões (a↔t, a↔c, c↔g, t↔g, ou seja, purina para pirimidina ou pirimidina para purina)
R=P/Q	A relação transições/transversões

Dados ausentes: Tratados como *gaps*.

Análise de sobrevivência (curvas de Kaplan-Meier, teste log-rank etc) (Survival analysis (Kaplan-Meier curves, log-rank test etc.)

Análise de sobrevivência para dois grupos (tratamentos) com provisão para censura à direita (*with provision for right censoring*). O módulo desenha curvas de sobrevivências de Kaplan-Meier para os dois grupos e calcula três testes distintos de equivalência entre as curvas. O programa espera quatro colunas. A primeira coluna contém tempos até falha (morte) ou censura (tempo até o qual a falha não foi observada) para o primeiro grupo, a segunda coluna indica falha (1) ou censura (0) para os indivíduos correspondentes. As duas últimas colunas contêm dados para o segundo grupo. Tempos até falha devem ser maiores do que zero.

O programa também aceita um único tratamento (fornecido em duas colunas), ou mais de dois tratamentos em pares de colunas consecutivos, plotando uma ou várias curvas de Kaplan-Meier. Os testes estatísticos, no entanto, comparam apenas os dois primeiros grupos.



As curvas de Kaplan-Meier e os testes de log-rank, Wilcoxon e Tarone-Ware são calculados de acordo com Kleinbaum & Klein (2005).

Tempo médio até falha inclui os dados censurados. Risco (*hazard*) médio é o número de falhas dividido pela soma dos tempos até falha ou censura.
O teste log-rank é calculado por qui-quadrado no segundo grupo:

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

Aqui, n_{ij} é o número de indivíduos sob risco, e m_{ij} o número de falhas, no grupo i no tempo de falha j. O número esperado de falhas no grupo 2 no tempo de falha j é

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

O qui-quadrado tem um grau de liberdade.
Os testes de Wilcoxon e Tarone-Ware são versões ponderadas do teste log-rank, nos quais os termos nas fórmulas de soma para O_2-E_2 e var(O_2-E_2) recebem pesos de n_j e $\sqrt{n_j}$, respectivamente. Estes testes, portanto, dão mais peso a tempos curtos de falha (*early failure times*). Eles não são de uso comum se comparados ao teste log-rank.
Este módulo não é estritamente necessário para análise de sobrevivência sem censura à direita – o teste de Mann-Whitney pode ser suficiente para este caso mais simples.

Dados ausentes: Pontos de dados com valores ausentes em uma ou ambas as colunas são desconsiderados.

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

Riscos / probabilidades (Risks / odds)

Este módulo compara as contagens de um resultado binário sujeito a dois tratamentos distintos, com estatísticas que são de uso comum na medicina. Os dados são inseridos em uma tabela 2x2, com tratamento em linhas e contagens dos diferentes resultados (*outcomes*) em colunas.
O exemplo abaixo mostra os resultados de um teste de vacinação em 460 pacientes:

	Contraiu influenza	Não contraiu influenza
Vacina	20	220
Placebo	80	140

No geral, os dados apresentam o seguinte formato:

	Resultado 1	Resultado 2
Tratamento 1	d_1	h_1
Tratamento 2	d_0	h_0

Sejam $n_1=d_1+h_1$, $n_0=d_0+h_0$ e $p_1=d_1/n_1$, $p_0=d_0/n_0$. A estatísticas são então calculadas da seguinte maneira:

Diferença de risco (*Risk difference*): $RD=p_1-p_0$

Intervalo de confiança de 95% para a diferença de risco (qui-quadrado de Pearson):

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

Intervalo: $RD - 1.96\ s_e$ até RD + $1.96\ s_e$.

Teste *Z* da diferença de risco (bicaudal):

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

Razão de risco (*Risk ratio*): $RR = p_1 / p_0$

Intervalo de confiança de 95% da razão de risco ("método delta"):

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

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

Intervalo: *RR*/*EF* até *RR* x *EF*.

Teste *Z* da razão de risco (bicaudal):

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

Razão de probabilidades (*Odds ratio*): $OR = \frac{d_1 / h_1}{d_0 / h_0}$

Intervalo de confiança de 95% da razão de probabilidades ("fórmula de Woolf"):

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

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

Intervalo: *OR* / *EF* até *OR* x *EF*.

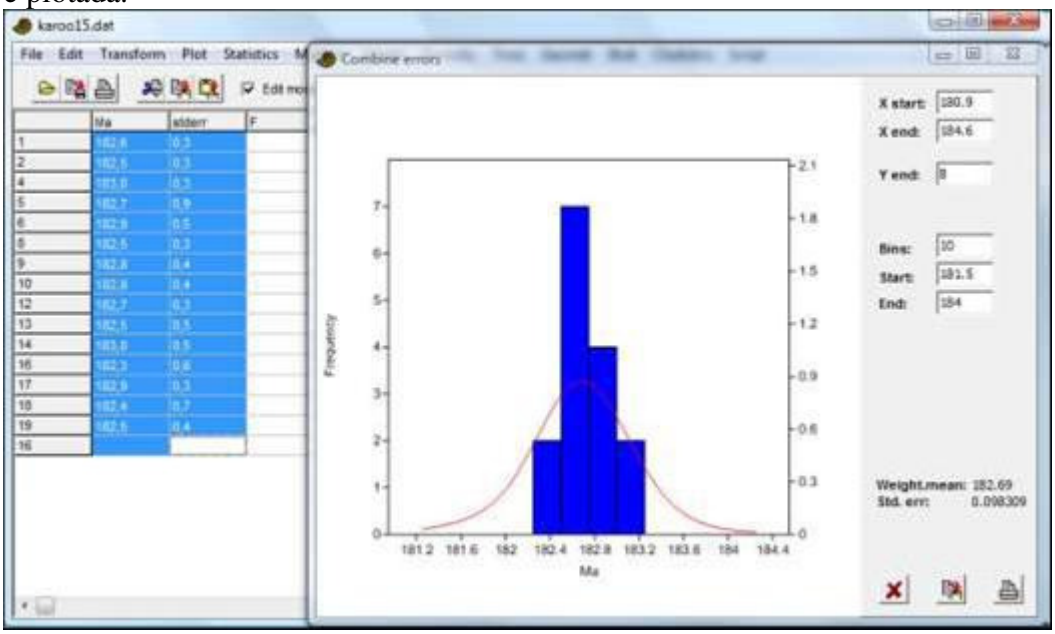
Repare que atualmente não há correção para continuidade.

Dados ausentes não são permitidos e resultam em mensagem de erro.

Combinar erros (Combine errors)

Um módulo simples para produzir uma média ponderada e seu desvio padrão a partir de uma série de medidas com erros (um sigma). Espera duas colunas: os dados *x* e seus erros um-sigma (one-sigma errors) σ. A soma das distribuições gaussianas individuais também

é plotada.



A média ponderada e seu desvio padrão são calculados por

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

Este é o estimador de máxima verossimilhança para a média, assumindo que todas as distribuições individuais são normais com a mesma média.

Dados ausentes: Linhas com dados ausentes em uma ou ambas as colunas são deletadas.

Multivar menu (Multivariada)

Componentes principais (Principal components)

Análise de componentes principais (*Principal componentes analysis – PCA*) encontra variáveis hipotéticas (componentes) que agregam o máximo possível da variância presente nos seus dados multivariados (Davis 1986, Harper 1999). Estas novas variáveis são combinações lineares das variáveis originais. A PCA pode ser usada para reduzir o conjunto de dados a apenas duas variáveis (os dois primeiros componentes) para fazer gráficos. Também pode ser hipotetizado que os componentes mais importantes estejam correlacionados com outras variáveis. Para dados morfológicos, pode ser o tamanho, enquanto para dados ecológicos pode ser um gradiente físico (e.g. temperatura ou profundidade). Bruton & Owen (1988) descrevem uma aplicação típica de PCA para dados morfométricos.

O *input* (entrada) é uma matriz de dados multivariados, com itens nas linhas e variáveis nas colunas. Não é feita centragem (*centering*) separada dos grupos antes da análise – portanto, grupos não são levados em conta.

A rotina PCA encontra os autovalores (*eigenvalues*) e os autovetores (*eigenvectors*) da matriz de variância-covariância ou da matriz de correlação. Use var-covar se todas as variáveis são medidas nas mesmas unidades (e.g. centímetros). Use correlação (var-covar normalizada) se as variáveis são medidas em unidades diferentes; isso implica normalizar todas as variáveis, dividindo-as por seus desvios padrões. Os autovalores fornecem uma medida da variância que é legava em conta por cada autovalor (componente) correspondente. As porcentagens da variância levada em conta por estes componentes também é fornecida. Se a maior parte da variância for levada em conta pelos dois primeiros componentes, a análise foi um sucesso, mas se a variância estiver distribuída de forma mais ou menos uniforme entre os componentes, a PCA foi, de um certo modo, pouco bem-sucedida.

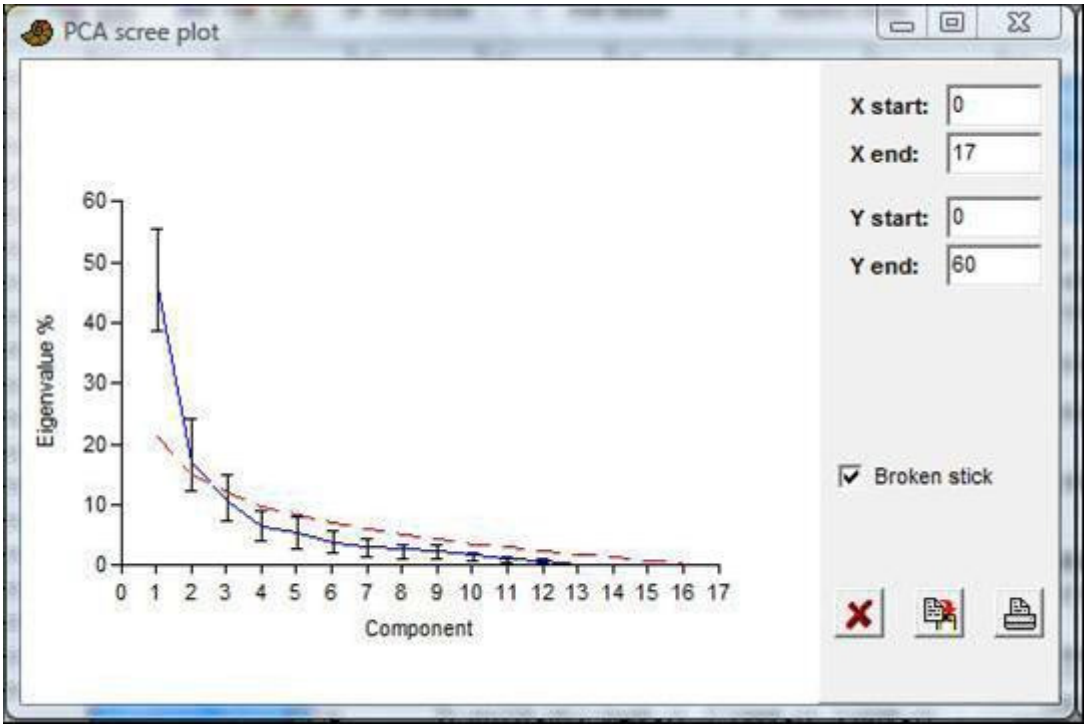
Grupos: se grupos forem especificados por cores de linhas, a PCA pode ser opcionalmente feita *dentro-de-grupos* ou *entre-grupos* (*within-group* ou *between-group*). Na PCA dentro-de-grupos, a média de cada grupo é subtraída antes da auto-análise (*eigenanalysis*), essencialmente removendo as diferenças entre os grupos. Na PCA entre-grupos, a auto-análise é feita sobre as médias dos grupos (ou seja, os itens analisados são os grupos, não as linhas). Para a análise tanto dentro-de-grupo quanto entre-grupos, os escores (*scores*) da PCA são computados usando produtos vetoriais com os dados originais.

No exemplo abaixo (pontos de referência de crânios de gorilas), o componente 1 é forte, explicando 45.9% da variância. Os intervalos de confiança por *bootstrap* não são mostrados a não ser que o valor de "*Boot N*" seja diferente de zero.

O valor de ponto-de-corte de Jolliffe (*Jolliffe cut-off value*) pode indicar o número de componentes principais significativos (Jolliffe, 1986). Componentes com autovalores menores do que este valor podem ser considerados insignificantes, mas não deve ser colocado muito peso neste critério.

Bootstrap por linhas (*row-wise bootsrapping*) é reaizado se um número positivo de réplicas por *bootstrap* (e.g. 1000) for fornecido na caixa "*Boot N*". Os componentes *bootstrap*ados são reordenados e revertidos de acordo com Peres-Neto et al. (2003) para aumentar a correspondência com os eixos originais. São fornecidos intervalos de confiança de 95% por *bootstrap* para os autovalores.

O "*Scree plot*" (gráfico simples de autovalores) também pode indicar o número de componentes significativos. Depois que esta curva começa a se endireitar, os componentes podem ser considerados como insignificantes. Intervalos de confiança de 95% são mostrados caso tenha sido feito *bootstrap*. Os autovalores esperados em um modelo aleatório (*Broken Stick*) podem ser plotados opcionalmente – autovalores debaixo desta curva podem indicar componentes não-significativos (Jackson 1993).



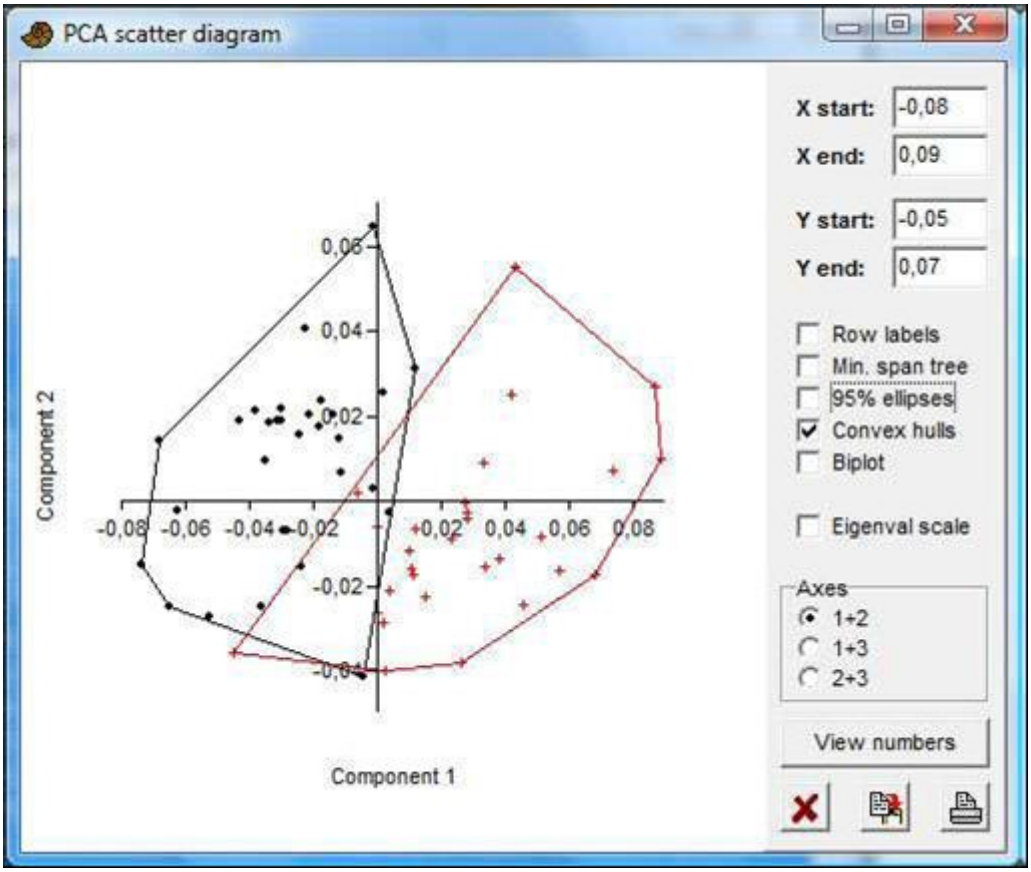
No exemplo dos gorilas acima, os autovalores dos 16 componentes (linha azul) ficam acima dos valores do model *broken stick* (linha vermelha tracejada) para os primeiros dois componentes, embora o modelo *broken stick* esteja dentro do intervalo de 95% do segundo componente.

A opção "*View scatter*" ("Ver dispersão") mostra todos os pontos de dados (linhas) plotados no sitema de coordenadas dado por dois dos componentes. Caso você tenha linhas coloridas (agrupadas), os grupos serão mostrados com símbolos e cores diferentes.

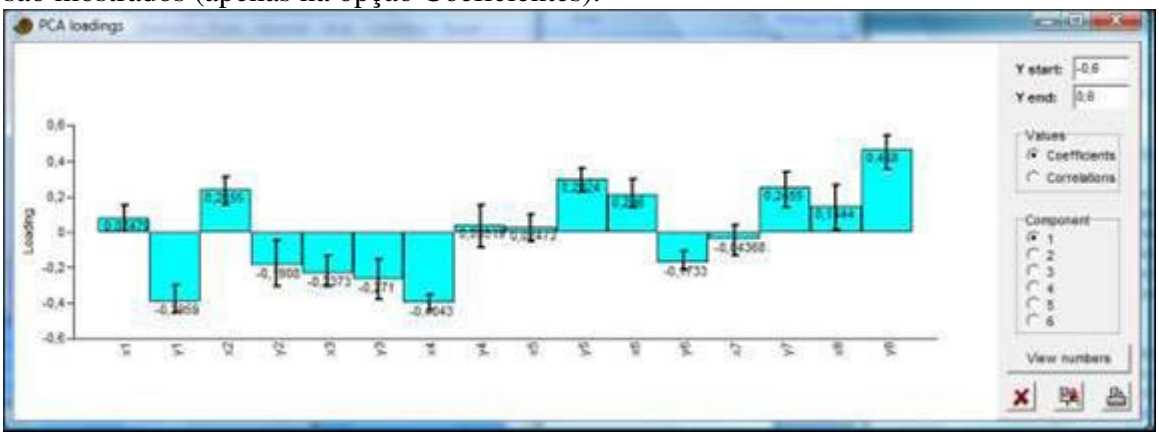
A Árvore de Menor Percurso (*Minimal Spanning Tree*) é o conjunto mais curto possível de linhas conectando todos os pontos. Ela pode ser usado como auxílio visual para agrupar pontos próximos. A MST é baseada em medida de distância Euclideana dos pontos originais, e tem mais significado quanto todos os pontos usam a mesma unidade.

A opção "Biplot" mostra uma projeção dos eixos originais (variáveis) no gráfico de dispersão. Essa é outra visualização dos pesos (*loadings*) ou coeficientes da PCA - veja abaixo.

Se a opção "*Eigenval scale*" ("Escala de autovalor") for selecionada, os pontos de dados sofrerão um reajuste de escala de $1/\sqrt{d_k}$, e os autovetores do biplot de $\sqrt{d_k}$ - este é o biplot de correlação de Legendre & Legendre (1998). Se esta opção não for selecionada, os pontos de dados não sofrem reajuste de escala, enquanto os autovetores do biplot são normalizados para terem o mesmo comprimento (não unitário, por motivos gráficos) – este é o biplot de distância.

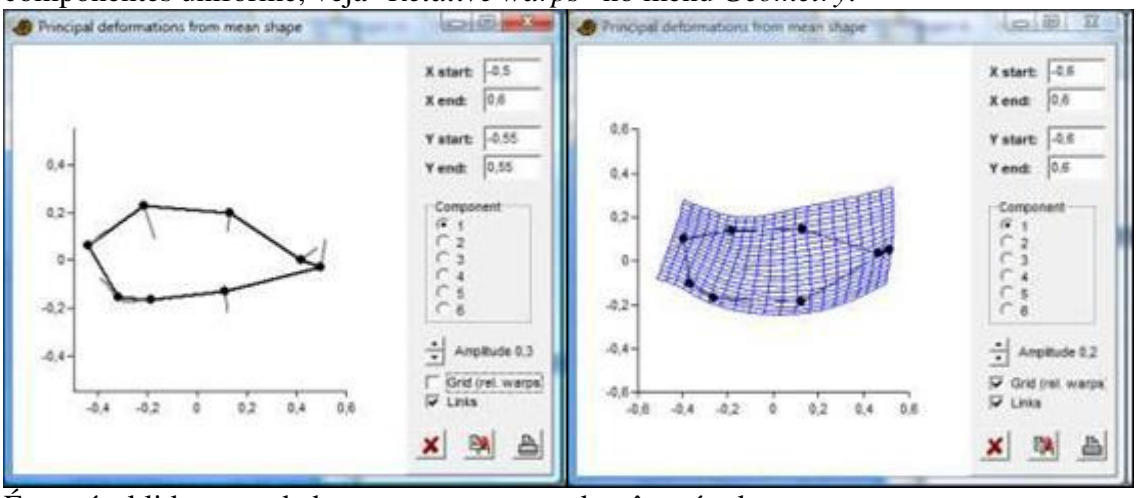


A opção "*View loadings*" ("Ver pesos") mostra até que grau as variáveis originais (mostradas na ordem original ao longo do eixo x) entram nos diferentes componentes (como escolhido no menu de botões ao lado). Os pesos dos componentes são importantes para tentar interpretar o "significado" dos componentes. A opção "Coeficientes" ("*Coefficients*") fornece os coeficientes dos componentes principais, enquanto a opção "Correlação" ("*Correlation*") fornece a correlação entre a variável e os escores dos componentes principais. Caso tenha sido feito *bootstrap*, intervalos de confiança de 95% são mostrados (apenas na opção Coeficientes).



A opção "SVD" força o algoritmo superior de Decomposição em Valores Singulares (*Singular Value Decomposition*) no lugar da autoanálise "clássica". Os dois algoritmos normalmente dão resultados praticamente idênticos, mas os eixos podem ser invertidos.

A opção "*Shape deform*" ("Deformar forma") foi delineada para dados de posição de pontos de referência em 2D. O gráfico padrão da Deformação de Forma é um "gráfico-pirulito" (*"lollipop plot"*), com a forma média mostrada como pontos e vetores (linhas) apontando nas direções dos pesos dos eixos. A opção *"Grid"* ("Grade") mostra as grades de deformação suave de placa fina (*thin-plate spline deformation grids*) correspondentes aos diferentes componentes. Este é, na prática, uma análise de "deformações relativas" (*"relative warps"*), incluindo o componente uniforme. Para deformações relativas sem o componentes uniforme, veja "*Relative warps*" no menu *Geometry*.



É possível lidar com dados ausentes por um de três métodos:

1. Imputação do valor médio (*Mean value imputation*): Valores ausentes são substituídos pelas médias das suas colunas. Não é recomendado.
2. Imputação interativa (*Iterative imputation*): Valores ausentes são inicialmente substituídos pelas médias das suas colunas. Uma primeira rodada de PCA é então usada para calcular valores de regressão para dados ausentes. O procedimento é iterado até que hava convergência. Este é normalmente o método indicado, mas pode causar uma certa superestimativa da força dos componentes (veja Ilin & Raiko 2010).
3. Deleção par-a-par (*Pairwise deletion*): Deleção par-a-par na matriz de var/covar ou de correlação. Pode funcionar quando o número de valores ausentes for pequeno. Esta opção irá forçar o método de decomposição em autovalores (i.e. não SVD).

Referências

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.

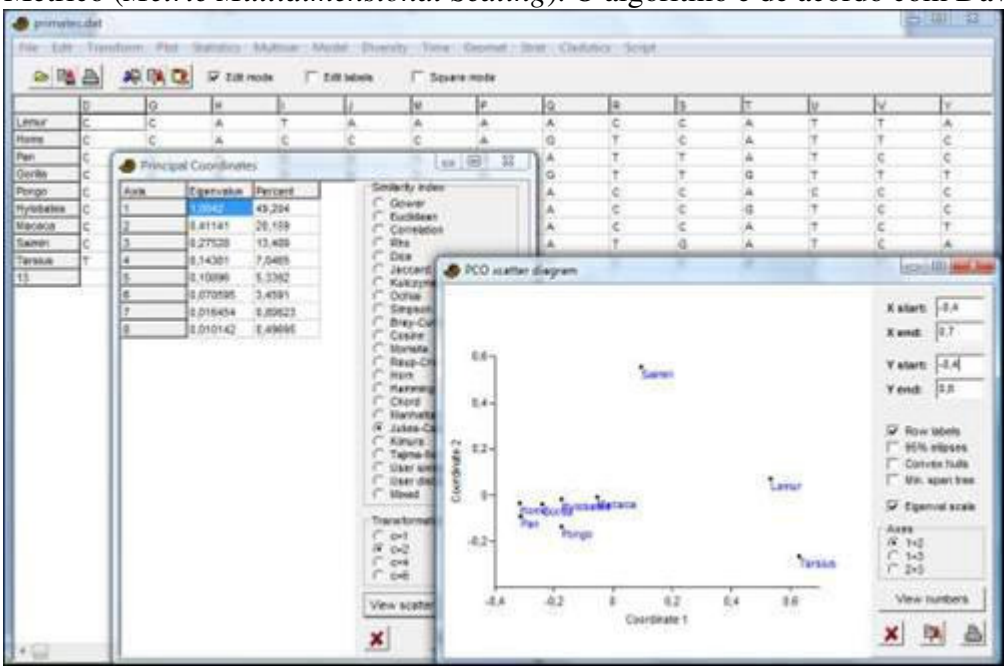
Jackson, D.A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74:2204-2214.

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.

Coordenadas principais (Principal coordinates)

A análise de coordenadas principais (*Principal coordinates analysis – PCO*) é outro método de ordenação, também conhecido como Escalonamento Multidimensional Métrico (*Metric Multidimensional Scaling*). O algoritmo é de acordo com Davis (1986).



A rotina PCO encontra os autovalores e autovetores de uma matriz contendo as distâncias ou similaridades entre todos os pontos de dados. A medidade de Gower normalmente será usada ao invés de distância Euclideana, o que dá resultados similares à PCA. Onze medidas adicionais de distância são disponíveis – estas são explicadas em Análise de Agrupamento (*Cluster Analysis*). Os autovalores, que fornecem uma medida da variância que é levada em conta pelos autovetores (coordenadas) correspondentes, são fornecidos para as primeiro coordenadas mais importantes (ou menores casa haja menos do que quatro pontos de dados). As porcentagens da variância que cada um destas coordenadas leva em conta também são fornecidas.

Os valores de similaridade/distância são elevados à potência de *c* (o "Exponente de transformação") antes da autoanálise (*eigenanalysis*). O valor padrão é *c*=2. Valores mais elevados (4 ou 6) podem dimuir o efeito "ferradura" (Podani & Miklos 2002).

A opção *"View scatter"* ("Ver dispersão") permite ver todos os pontos de dados (linhas) plotados no sistema de coordenadas dado pela PCO. Caso haja linhas coloridas (agrupadas), os diferentes grupos serão mostrados com diferentes símbolos e cores. A opção *"Eigenvalue scaling"* ("Escalonamento de autovalores") muda a escala de cada eixo usando a raiz quadrada do autovalor (recomendado). A opção de árvore de menor percurso (*minimal spanning tree*) é baseada no índice de distância ou similaridade escolhido no espaço original.

Há suporte para dados ausentes por deleção par-a-par (não para índices de Raup-Crick, Rho ou definido pelo usuário).

Referências

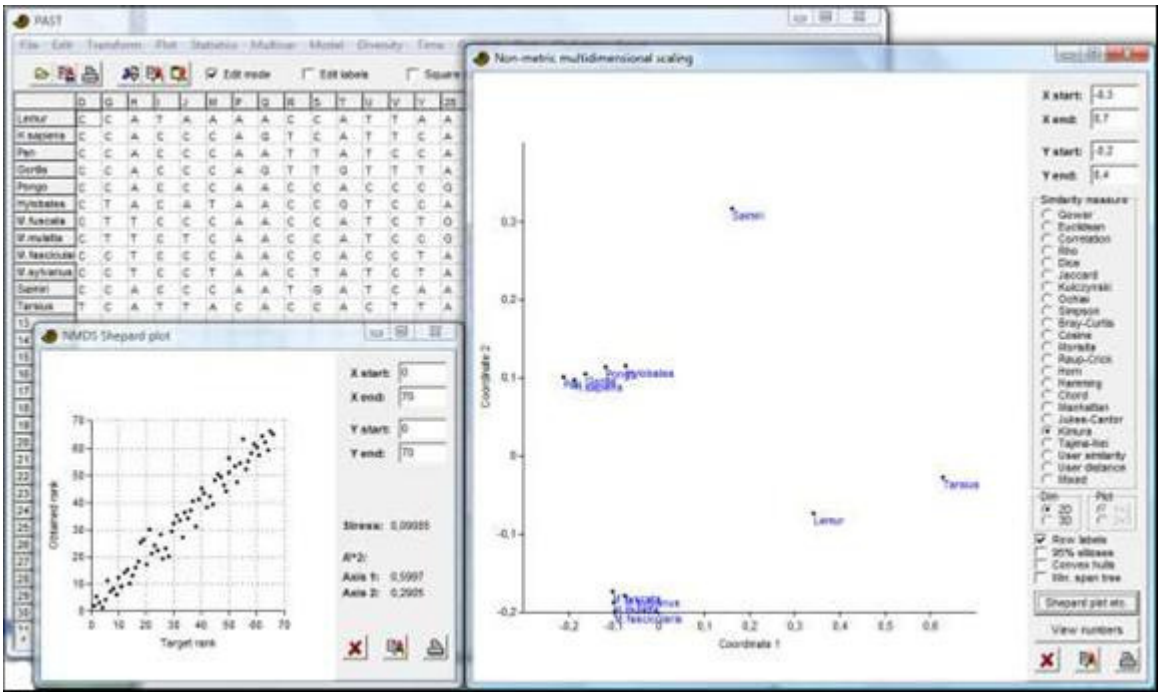
Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.

Podani, J. & I. Miklos. 2002. Resemblance coefficients and the horseshoe effect in principal

coordinates analysis. *Ecology* 83:3331-3343.

Escalonamento multidimensional não-métrico (Non-metric MDS)

Escalonamento multidimensional não-métrico (*Non-metric multidimensional scaling – NMDS*) é baseado em uma matriz de distância calculada com qualquer uma das 21 medidas de distância com suporte no Past, como explicado em Índices de Similaridade e Distância acima. O algoritmo então tenta colocar os pontos de dados num sistema bi- ou tri-dimensional de coordenadas de tal modo que as *diferenças ranqueadas* sejam preservas. Por exemplo, se a distância original entre os pontos 4 e 7 é a nona maior de todas as distância entre quaisquer dois pontos, pontos 4 e 7 serão idealmente colocados de tal modo que a distância Euclideana entre eles no plano 2D ou no espaço 3D continua sendo a nona maior. Escalonamento multidimensional não-métrico não leva em conta distâncias absolutas.



O programa pode convergir em uma solução diferente em cada rodada, dependendo das condições iniciais aleatórias. Cada rodada é na verdade uma sequência de 11 testes, dos quais é escolhido o teste com o menor *stress*. Um destes testes usa o PCO como condição inicial, mas isso raramente dá a melhor solução. A solução é automaticamente rotacionada para os eixos principais (2D e 3D).

O algoritmo implementado no Past, que parece funcionar muito bem, é baseado em uma nova abordagem desenvolvida por Taguchi & Oono (no prelo).

A árvore de menor percurso é baseada no índice de similaridade ou distância, escolhido no espaço original.

Variáveis ambientais (Environmental variables): é possível incluir uma ou mais colunas iniciais contendo variáveis "ambientais" adicionais para a análise. Estas variáveis não são incluídas na ordenação. Os coeficientes de correlação entre cada variável ambiental e os escolres do NMDS são apresentados como vetores partindo da origem. O comprimento dos vetores é ajustado a uma escala arbitrária para tornar o biplot visível, de modo que apenas suas direções e comprimentos relativos devem ser considerados.

Gráfico de Shepard (Shepard plot): Este gráfico de ranks obtidos versus observados (alvo) indica a qualidade do resultado. Idealmente, todos os pontos devem ser colocados em uma linha reta ascendente (x=y). Os valores de R^2 são os coeficientes de determinação entre as distâncias ao longo de cada eixo da ordenação e as distância originais (talvez um valor sem muito significado, mas ele é relatado por outros programas de NMDS e, portanto, é incluído também no Past).
Dados ausentes: suporte por deleção par-a-par (não para índices de Raup-Crick, Rho e definido pelo usuário). Para variáveis ambientais, valores ausentes não são incluídos no cálculo das correlações.

Análise de correspondência (Correspondence analysis)

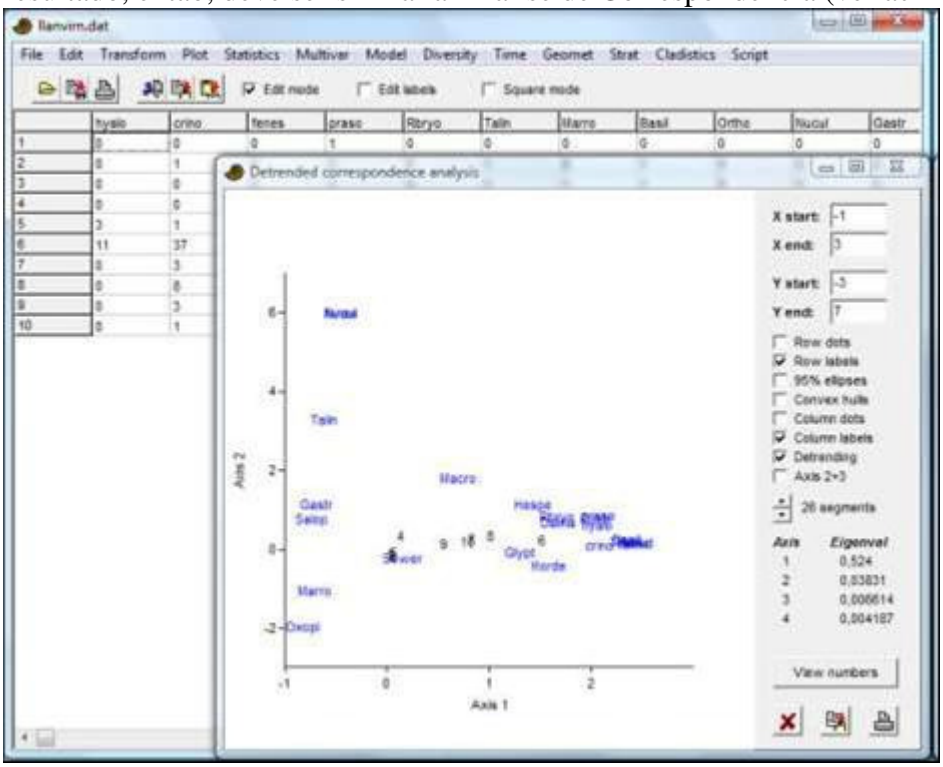
A análise de correspondência (*Correspondence analysis – CA*) é mais um método de ordenação, de certo modo similar à PCA, mas para *dados de contagem*. Para comparar associações (colunas) contendo contagens de táxons, ou contagens de táxons (linhas) entre associações, a CA é o algoritmo mais apropriado. Além disso, a CA é mais apropriada se você espera que as espécies tenham respostas unimodais aos parâmetros subjacentes, ou seja, que elas sejam favorecidas por uma certa faixa de valores do parâmetro, se tornando mais raras para valores mais baixos ou mais altos (em contraste com a PCA, que assume uma resposta linear).
A rotina da CA encontra os autovalores e autovetores de uma matriz contendo as distâncias de qui-quadrado entre todas as linhas (ou colunas, se for mais eficiente – o resultado é o mesmo). O autovalor, fornecendo uma medidade da similaridade que é levada em conta pelo autovetor correspondente, é dado para cada autovetor. As porcentagens de similaridade que são explicadas por estes componentes também são fornecidas.
A opção "Ver dispersão" (*"View scatter"*) permite ver todos os pontos de dados (linhas) plotados no sistema de coordenadas dado pela CA. Caso haja linhas coloridas (agrupadas), os diferentes grupos serão mostrados usando diferentes símbolos e cores. Adicionalmente, as variáveis (colunas, ou associações) podem ser plotadas no mesmo sitema de coordenadas (mode Q), incluindo opcionalmente os nomes das colunas. Se o seus dados forem "bem-comportados", táxons típicos de uma associação deverão aparecer próximos àquela associação no gráfico.
O PAST atualmente usa escalonamento assimétrico (*"Benzecri scaling"*).
Caso haja mais de duas colunas no conjunto de dados, é possível ver um gráfico de dispersão dos eixos dois e três.
Relay plot: É um diagrama composto com um gráfico por coluna. Os gráficos são ordenados de acordo com os escores das colunas da CA. Cada ponto de dados é plotado com os escores das linhas do primeiro eixo da CA no eixo vertical, e o ponto de dados original (abundância) na coluna correspondente no eixo horiontal. Isso pode ser mais útil quando as amostras estão em linhas e os táxons em colunas. O *relay plot* então irá mostrar os táxons ordenados de acordo com suas posições ao longo dos gradientes, e para cada táxon, o gráfico correspondente deve idealmente mostrar um pico unimodal, parcialmente sobreposto ao pico do próximo táxon ao longo do gradiente (ver Hennebert & Lees 1991 para um exemplo da sedimentologia).
Dados ausentes: suporte por substituição pela média da coluna.

Referências

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.

Análise de correspondência destendenciada (Detrended correspondence analysis)

O módulo Correspondência Destendêciada (*Detrended Correspondence – DCA*) usa o mesmo algoritmo que Decorana (Hill & Gauch 1980), com modificações de acordo com Oxanen & Minchin (1997). É especializado para ser usado em conjuntos de dados "ecológicos" com dados de abundância; amostras em linhas, táxons em colunas (vice-versa antes da v. 1.79). Quando a opção *"Detrending"* ("Destendenciamento") é desligada, uma análise por Média Recíproca (*Reciprocal Averaging*) básica será feita. O resultado, então, deve ser similar à Análise de Correspondência (ver acima).



Autovalores para os três primeiros eixos da ordenação são fornecidos como na CA, indicando sua importância relativa na explicação do espalhamento dos dados. Destendenciamento é uma espécie de procedimento de normalização em dois passos. O primeiro passo envolve uma tentativa de "endireitar" pontos que distribuídos em arco, um acontecimento comum. O segundo passo envovle "espalhar os pontos" de modo a evitar agrupamento de pontos nas bordas do gráfico. Destendenciamento pode parecer um procedimento arbitrário, mas pode ser um auxílio útil à interpretação.

Dados ausentes: suporte por substituição pela média da coluna.

Referências

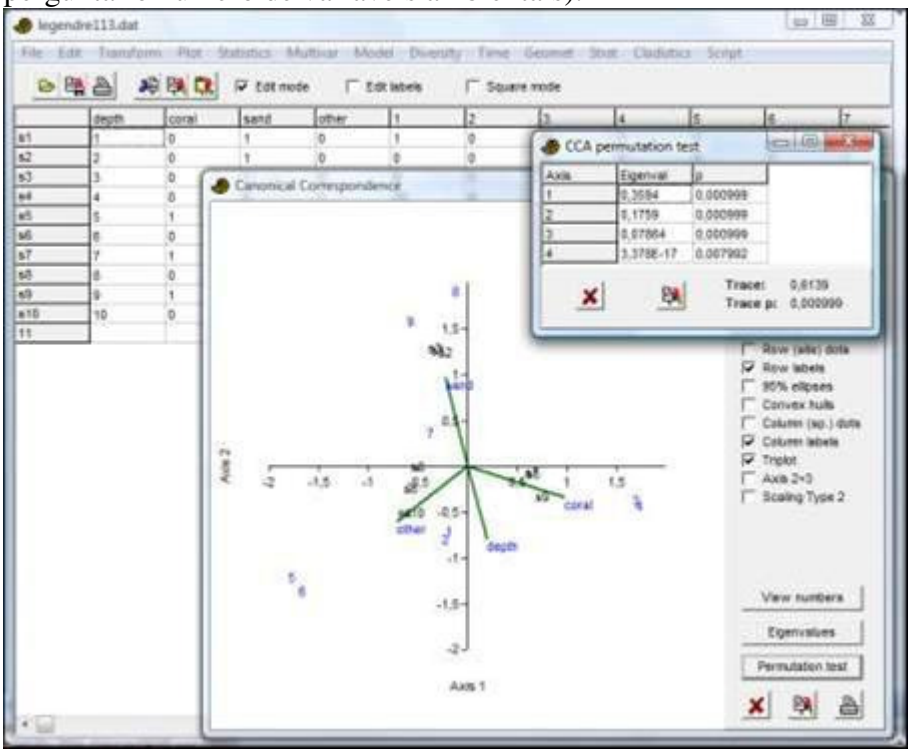
Hill, M.O. & H.G. Gauch Jr. 1980. Detrended Correspondence analysis: an improved ordination technique. *Vegetatio* 42:47-58.

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.

Correspondência canônica (Canonical correspondence)

Análise de Correspondência Canônica (*Canonical Correspondence Analysis – CCA*; Legendre & Legendre 1998) é a análise de correspondência de uma matriz sítio/espécie, onde cada sítio tem valores correspondentes de uma ou mais variáveis ambientais (temperatura, profundidade, tamanho de grãos etc). Os eixos da ordenação são combinações lineares das variáveis ambientais. CCA é, portanto, um exemplo de análise direta de gradiente, onde o gradiente é conhecido *a priori* e as abundâncias (ou presenças/ausências) das espécies são consideradas como sendo uma resposta ao gradiente.

Cada sítio deve ocupar uma linha na planilha. As variáveis ambientias devem ser inseridas nas primeiras colunas, seguidas pelos dados de abundância (o programa irá perguntar o número de variáveis ambientais).



A implementação no PAST segue o algoritmo de autoanálise (*eigenanalysis*) fornecido por Legendre & Legendre (1998). As ordenações são dadaos como escores dos sítios – escores ajustados de sítios atualmente não são disponíveis. Ambos os escalonamentos (*scalings*) (tipos 1 e 2) de Legendre & Legendre (1998) estão disponíveis. Escalonamento 2 enfatiza as relações entre espécies.

Valores ausentes: suporte por substituição pela média da coluna.

Referência

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

Análise de fator CABFAC (CABFAC factor analysis)

Este módulo implementa o método clássico de Imbrie & Kipp (1971) de análise de fatores e regressão ambiental (CABFAC e REGRESS, veja também Klovan & Imbrie 1971).

O programa pergunta se a primeira coluna contém dados ambientais. Caso não contenha, uma análise simples de fator com rotação Varimax será calculada em dados normalizados por linha.

Se dados ambientais forem incluídos, será feita uma regressão dos fatores pelas variáveis ambientais usando o método de segunda ordem (parabólico) de Imbrie & Kipp, com termos cruzados. O PAST então relata a regressão RMA dos valores ambientais originais contra valores reconstruídos da função de transferências. Métodos diferentes de validação cruzada (deixe-um-fora e *k*-vezes – *leave-one-out* e *k-fold*) são disponíveis. Você também pode salvar a função de transferência como um arquivo de texto que pode ser usado posteriormente para reconstrução do paleoambiente (ver abaixo). O arquivo contém:

- Número de táxons
- Número de fatores
- Escores de fatores para cada táxon
- Número de coeficientes de regressão
- Coeficientes de regressão (termos de segunda e primeira ordem, e intercepto)

Valores ausentes: suporte por substituição pela média da coluna.

Referências

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.

Mínimos quadrados parciais de dois blocos (Two-block PLS)

Mínimos qudrados parciais (*partial least squares – PLS*) de dois blocos podem ser vistos como um método de ordenação comparável com a PCA, mas o objetivo é maximizar a covariância entre dois cojuntos de variáveis na mesma linha (espécies, sítios). Por exemplo, dados morfométricos e ambientias coletados para os mesmos espécimes podem ser ordenados para estudar a covariação entre os dois.

O programa irá perguntar o número de colunas que pertencem ao primeiro bloco. As colunas restantes serão atribuídas ao segundo bloco. Há opção para plotar escores PLS dentro e entre blocos, assim como pesos (*loadings*) PLS.

O algoritmo segue Rohlf & Corti (2000). Testes de permutação e *biplots* ainda não estão implementados.

Particione a matriz de dados nxp $\mathbf{Y}$ em $\mathbf{Y}_1$ e $\mathbf{Y}_2$ (os dois blocos), com p_1 e p_2 colunas. A matriz de correlação ou covariânica $\mathbf{R}$ de $\mathbf{Y}$ pode então ser parcionada como

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}$$

O algoritmo procede por decomposição em valores singulares (*singular value decomposition*) da matriz $\mathbf{R}_{12}$ de correlações entre os blocos:

$$\mathbf{R_{12}} = \mathbf{F_1 D F_2^t}.$$

A matriz **D** contém os valores singulares λ_i ao longo da diagonal. $\mathbf{F}_1$ contém os pesos do bloco 1, e $\mathbf{F}_2$ contém os pesos do bloco 2 (cf. PCA).

O *"Squared covar %"* ("Quadrado da covar %") é uma medida do quadrado da covariância geral entre os dois conjuntos de variáveis, em porcentagem relativa ao máximo possível (todas as correlações iguais a 1) (Rohlf & Corti p. 741). As "% covar" dos eixos são as quantidades de variância que são explicadas para cada eixo da PLS, em porcentagem da covariância total. Eles são calculados como $100\frac{\lambda_i^2}{\sum \lambda_i^2}$.

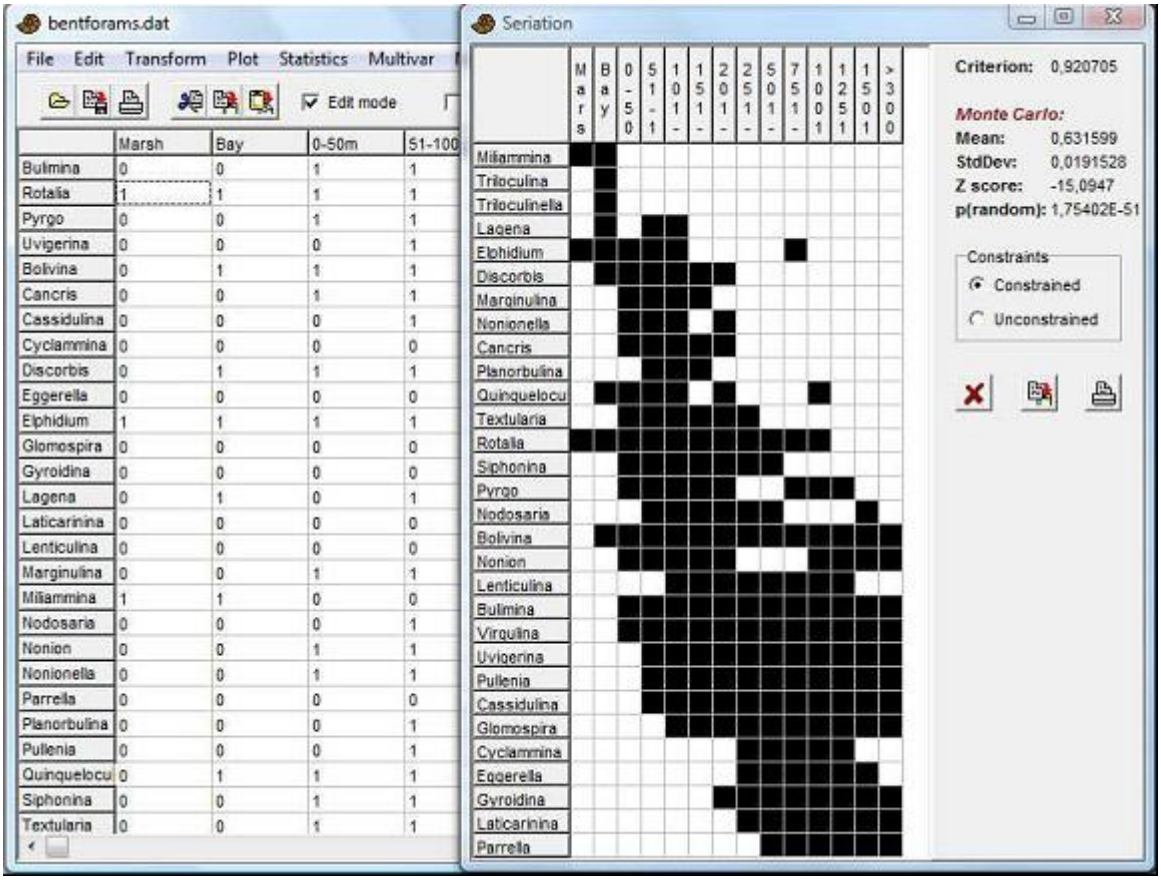
Dados ausentes: suporte por substituição pela média da coluna.

Referência

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

Seriação (Seriation)

Seriação de uma matriz de ausência-presença (0/1) usando o algoritmo descrito por Brower & Kile (1988). O método é tipicamente aplicado a uma matriz de associação com táxons (espécies) em linhas e amostras em colunas. Para seriação restrita (*constrained seriation* – ver abaixo), as colunas devem ser ordenadas de acordo com algum critério, normalmente nível estratigráfico ou posição ao longo de um gradiente faunal presumido.



A rotina de seriação tenta reorganiar a matriz de dados de tal modo que as presenças estejam concentradas ao longo da diagonal. Há dois algoritmos: otimização restrita e irrestrita (*contstrained* e *unconstrained*). Em otimização restrita, apenas as linhas (táxons) podem ser movimentados. Dado que as colunas estejam dispostas em uma certa ordem, este procedimento encontra a ordem "ótima" das linhas, ou seja, a ordem de táxons que dá o gráfico de amplitude mais bonito. Além disso, no modo restrito, o programa roda uma simulação "Monte Carlo", gerando e seriando 30 matrizes aleatórias com o mesmo número de ocorrências dentro de cada táxon, e compara estas à matriz original para ver se a matriz original é mais informativa do que uma aleatória (este procedimento gasta muito tempo para conjuntos grandes de dados).

No modo irrestrito, tanto as linhas quanto as colunas podem ser movidas.

Dados ausentes são tratados como ausências.

Referência

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

Análise de agrupamento (Cluster analysis)

A rotina de agrupamento hierárquico (*hierarchical clustering*) produz um "dendrograma" mostrando como os pontos de dados (linhas) podem ser agrupados. Para agrupamento de modo "R", colocando peso em agrupamentos de táxons, os táxons devem ser colocados em linhas. Também é possível encontrar agrupamentos de variáveis ou associações (modo Q), colocando táxons em colunas. A mudança entre os dois modos é feita pela transposição da matriz (no menu *Edit*).

Três algoritmos distintos são disponíveis:

- Média de pares de grupos não ponderados (*Unweighted pair-group average – UPGMA*). Grupos são juntados com base na distância média entre todos os membros dos dois grupos.
- Ligação simples ou vizinho mais próximo (*Single linkage or nearest neighbour*). Grupos são juntados com base na menor distância entre os dois grupos.
- Método de Ward (*Ward's method*). Grupos são juntados de tal modo que o aumento da variância dentro-de-grupo (*within-group variance*) é minimizado.

Um método não é necessariamente melhor do que outro, embora a ligação simples não seja recomendada por alguns. Pode ser útil comparar os dendrogramas produzidos por diferentes algoritmos para verificar informalmente a robustez dos agrupamentos. Caso um agrupamento seja modificado quando se tenta um outro algoritmo, talvez este agrupamento não seja confiável.
Para o método de Ward, uma medida de distância Euclideana é inerente ao algoritmo. Para UPGMA e ligação simples, a matriz de distância pode ser calculada usando 20 índices diferentes, como descrito no menu *Statistics* (Índices de similaridade e distância).

Dados ausentes: O algoritmo de análise de agrupamento pode lidar com dados ausentes, codificados por ponto de interrogação (?). Isso é feito usando deleção par-a-par, mostrando que quando a distância é calculada entre dois pontos, qualquer variável que esteja ausente é ignorada no cálculo. Para Raup-Crick, valores ausentes são tratados como ausência. Dados ausentes não têm suporte no método de Ward e nem na medida de similaridade Rho.

Agrupamento bifatorial (Two-way clustering): A opção *two-way* permite agrupamento simultâneo nos modos R e Q.

Agrupamento restrito estratigraficamente (Stratigraphically constrained clustering): Essa opção permite que apenas linhas ou grupos de linhas adjacentes sejam juntadas durante o procedimento de agrupamento. Isso pode produzir dendrogramas de aparência estranha (mas corretos).

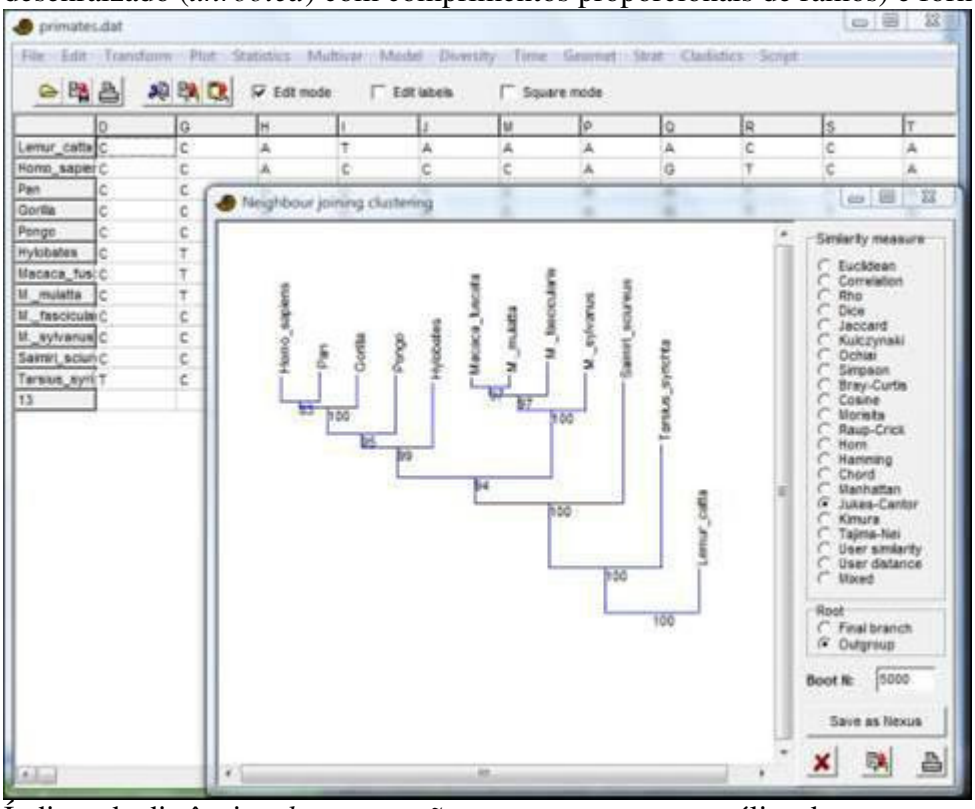
Bootstrap: Se um número de réplicas para o bootstrap for fornecido (e.g. 100), as colunas são sujeitas a reamostragem. Pressione Enter depois de atualizar o valor na caixa "Boot N". A porcentagem de réplicas onde cada nó ainda tem suporte é mostrada no dendrograma.

Nota sobre o método de Ward: o Past produz dendrogramas de Ward idênticos àqueles feitos pelo Stata, mas um tanto diferentes dos produzidos pelo Statistica. A razão desta discrepância não é conhecida.

Agrupamento de vizinho (Neighbour joining)

Agrupamento *Neighbour joining* (Saitou & Nei 1987) é um método alternativo para análise de agrupamento hierárquico. Este método foi originalmente desenvolvido para análise filogenética, mas pode ser superior à UPGMA também para dados ecológicos.

Diferentemente da UPGMA, dois ramos com o mesmo nó interno não precisam necessariamente ter os mesmos comprimentos de ramo. Um filograma (dendrograma desenraizado (*unrooted*) com comprimentos proporcionais de ramos) é fornecido.



Índices de distância e *bootstrap* são como para a outra análise de agrupamento (ver acima). Para fazer a análise de *bootstrap*, digite um número de réplicas *bootstrap* requeridas (e.g. 1000, 10000) na casa "Boot N" e aperte Enter para atualizar o valor. Comprimentos de braço negativos são forçados a zero e transferidos ao braço adjacente, de acordo com Kuhner & Falsenstein (1994).
A árvore é, por definição, enraizada no último braço adicionado durante a construção da árvore (isso não é enraizamento por ponto médio (*midpoint rooting*)). Opcionalmente, a árvore pode ser enraizada na primeira linha da matriz de dados (grupo externo – *outgroup*).

Dados ausentes recebem suporte por deleção par-a-par.

Referências

Saitou, N. & M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Agrupamento por K-medias (K-means clustering)

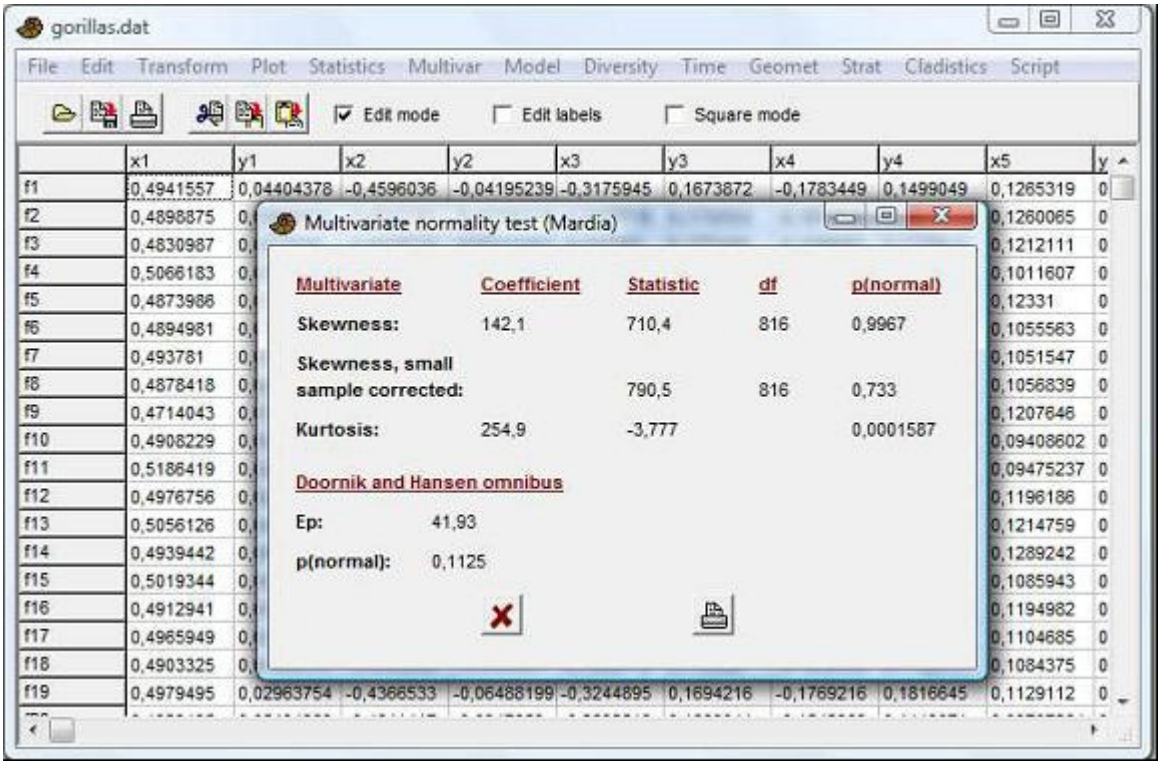
Agrupamento por K-médias (*K-means clustering*) (e.g. Bow 1984) é um método de agrupamento não-hierárquico. O número de grupos a serem usados é especificado pelo usuário, normalmente de acordo com alguma hipótese tal como a existência de dois sexos, quatro regiões geográficas ou três espécies no conjunto de dados.

As atribuições aos grupos incialmente são aleatórias. Em um procedimento iterativo, ítens são então movidos ao grupo que tem a média de grupo mais próxima, e as médias dos grupos são atualizadas de acordo. Isso continua até que elementos não mais estejam se movendo entre grupos. O resultado do agrupamento é até um certo nível dependente da ordem aleatória inicial, e elementos podem pertencer a diferentes grupos em diferentes rodadas da análise. Isso não é um erro, e sim comportamento normal do agrupamento por k-médias.
As atribuições de elementos a grupos podem ser copiados e colados dentro da planilha principal, e cores (símbolos) correspondentes podem ser atribuídos a eles usando a opção "*Numbers to colors*" no menu *Edit*.
Dados ausentes: suporte por substituição pela média da coluna.

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

Normalidade multivariada (Multivariate normality)

Normalidade multivariada é assumida por uma série de testes multivariados. o PAST calcula a assimetria (*skewness*) e curtose (*kurtosis*) multivariada de Mardia, com testes baseados em distribuições de qui-quadrado (assimetria) e normal (curtose). Um poderoso teste omnibus (geral), de Doornik & Hansen (1994), também é fornecido. Se ao menos um destes testes mostrar desvios da normalidade (pequeno valor de *p*), a distribuição é significativamente não-normal. Tamanho amostral deve ser razoavelmente grande (>50), embora uma tentativa de correção para tamanho amostral pequeno seja feito no teste de assimetria.



Dados ausentes: suporte por substituição pela média da coluna.

Referências

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.

Discriminantes (Discriminant)/Hotelling

Dados dois conjuntos de dados multivariados, é construído um eixo que maximiza a diferença entre os conjuntos (e.g. Davis 1986). Os dois conjuntos de dados são então plotados ao longo deste eixo por meio de um histograma. O módulo espera que as linhas dos dois conjuntos de dados sejam agrupadas em dois grupos, colorindo as linhas, e.g. com preto (pontos) e vermelho (cruzes).
A igualdade das médias dos dois grupos é testada por um análogo multivariado do teste *t*, conhecido como T-quadrado de Hotelling (*hotelling's T-squared*), e é fornecido o valor de *p* para esse teste. As variáveis precisam ter distribuição normal, e pelo menos duas vezes mais observações do que variáveis.
Número de restrições (Number of constraints): Para calcular corretamente o valor de *p*, o número de variáveis dependentes (*constraints* ou restrições) precisa ser especificado. Normalmente deve ser deixado em 0, mas use 4 (para 2D) ou 6 (para 3D) no caso de dados de pontos de referências ajustados por Procrustes.
A análise de discriminantes pode ser usada para confirmar ou rejeitar visualmente a hipótese de que duas espécies são morfologicamente distintas. Usando um ponto de corte (*cutoff*) de zero (o ponto médio entre as médias dos escores de discriminantes para os dois grupos), uma classificação nos dois grupos é mostrada na opção *"View numbers"* ("Ver números"). A porcentagem de itens classificados corretamente também é mostrada.
Função discriminante (Discriminant function): Novos espécimes podem ser classificados de acordo com a função discriminante. Pegue o produto interno entre as medidas do novo espécime e os fatores da função discriminantes fornecida e subtraia o valor de *offset* fornecido.
Deixar um fora (avaliação cruzada) (Leave on out – cross-evaluation): Existe a opção de deixar fora da análise uma linha (espécime) por vez, re-calcular a análise de discriminantes com os espécimes restantes, e classificar de acordo com ela a linha que foi deixada fora (como ditado pelo valor de escore (*Score*)).

Dados ausentes: suporte por substituição pela média da coluna.

Deformação de pontos de referência (*Landmarks warps*)

Esta função deve ser usada apenas se a análise foi feita sobre dados de pontos de referência em 2D. Permite uma plotagem interativa das deformações de forma como função da posição ao longo do eixo discriminantes, como gráficos-pirulito (*lollipop-plots*) (vetores para fora das posições médias dos pontos de referências) ou como deformações suavizadas de placa fina (*thin-plate spline deformations*). REMOVIDO TEMPORARIAMENTE (?) POR FALTA DE ESTABILIDADE.

Deformações (*warps*) EFA

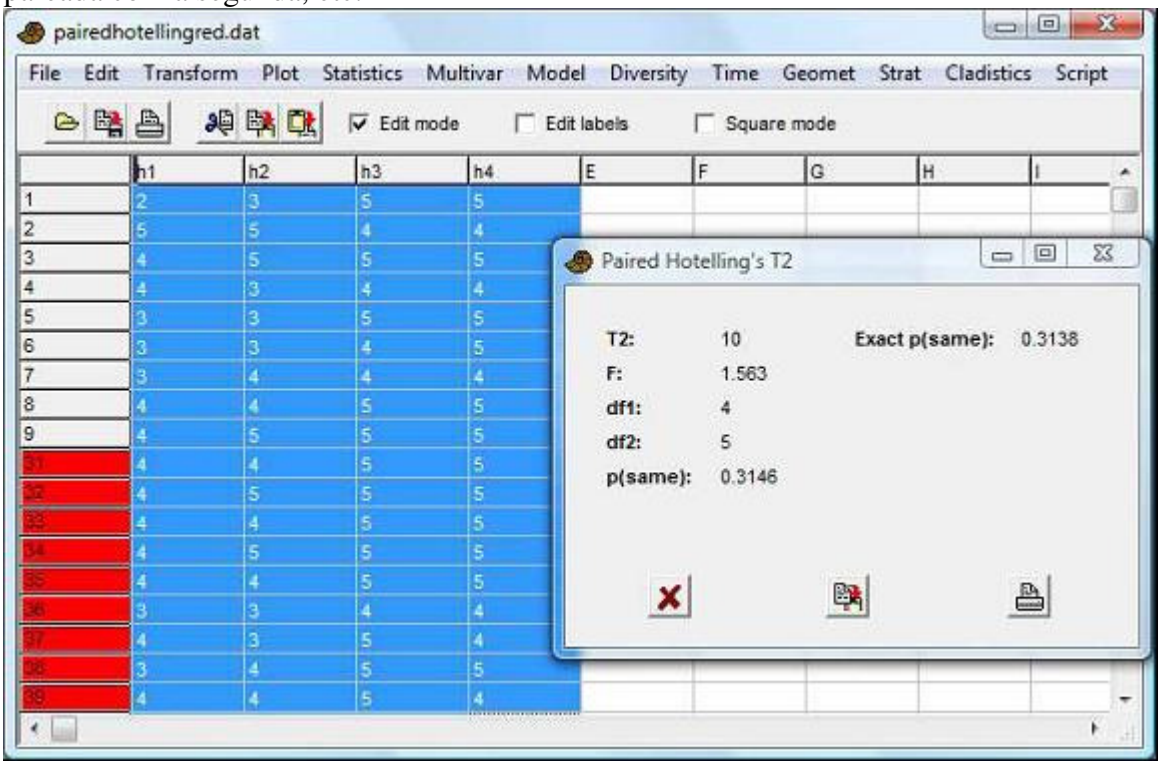
Esta função deve ser usada apenas se a análise de discriminantes foi rodada em coeficientes calculadas pelo módulo de Análise Elíptica de Fourier (*Elliptic Fourier Analysis*). Permite uma plotagem interativa dos contornos como uma função da posição ao longo do eixo discriminantes. REMOVIDO TEMPORARIAMENTE (?) POR FALTA DE ESTABILIDADE.

Referência

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

Hotelling pareado (Paired hotelling)

O teste pareado de Hotelling espera dois grupos de dados multivariados, marcados com cores diferentes. As linhas de cada grupo devem ser consecutivas. A primeira linha do primeiro grupo é pareada com a primeira linha do segundo grupo, a segunda linha é pareada com a segunda, etc.



Sendo n o número de pares e p o número de variáveis:

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

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

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

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

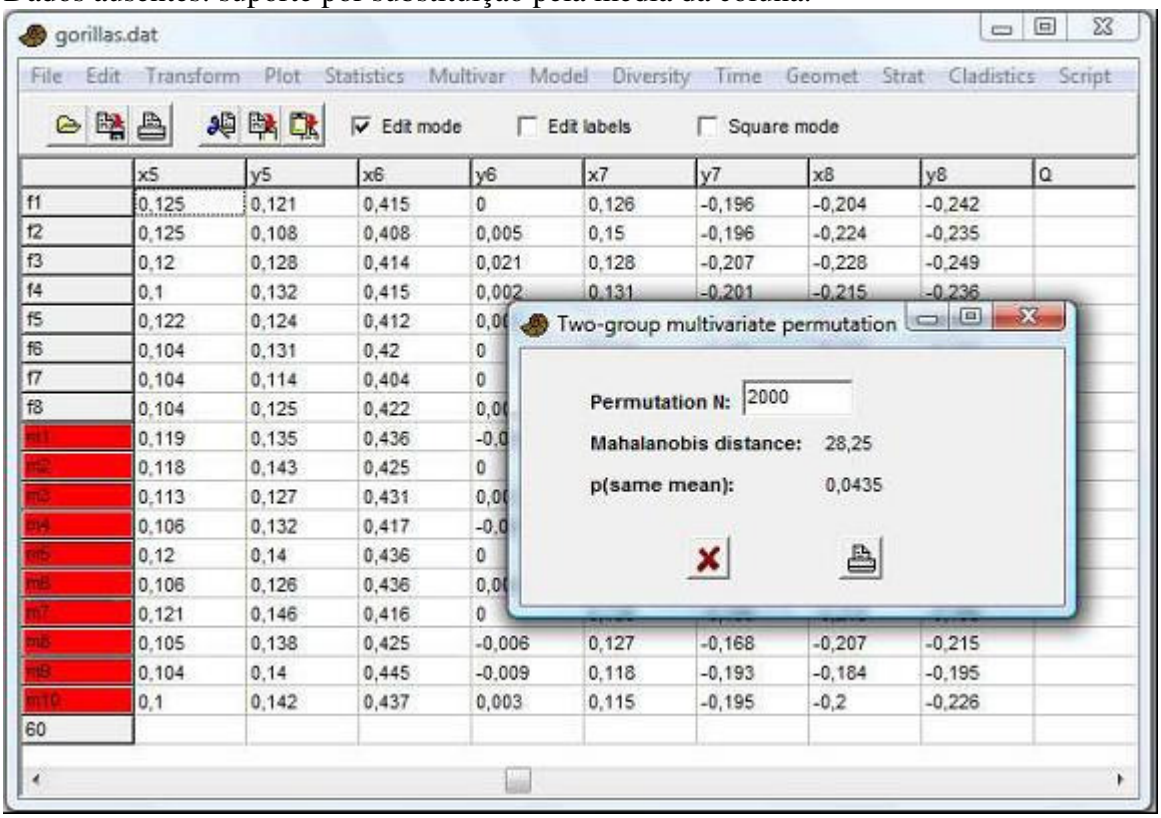
$$F = \frac{n-p}{p(n-1)}T^2$$

O F tem p e n-p graus de liberdade.

Para $n \leq 16$, o programa também calcula um valor exato de p baseado na estatística T^2 avaliada para todas as permutações possíveis.
Dados ausentes: suporte por substituição pela média da coluna.

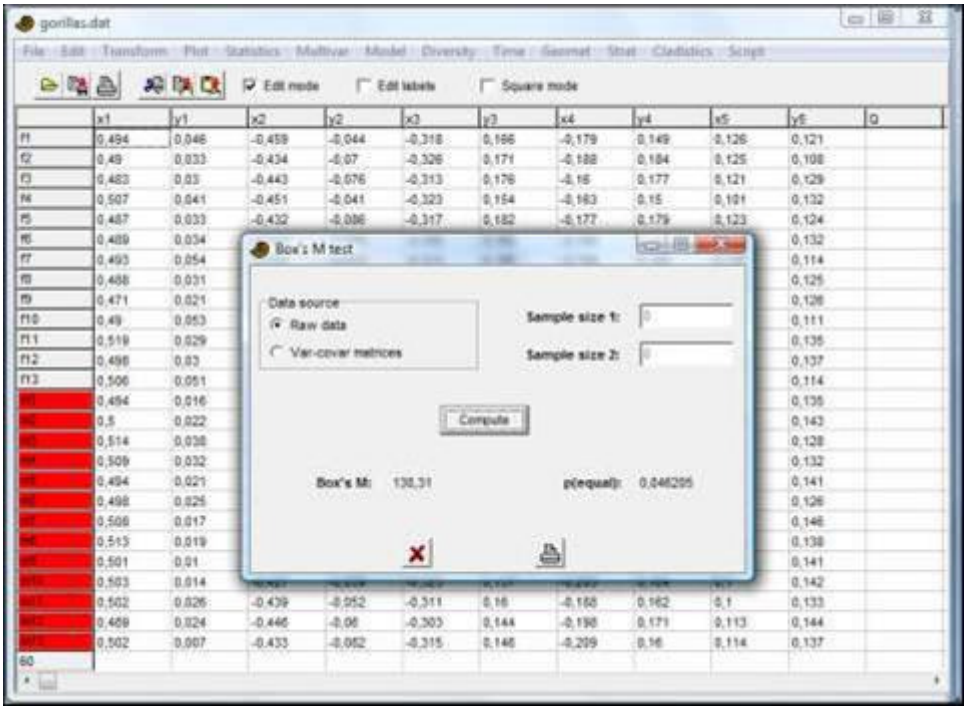
Permutação de dois grupos (Two-group permutation)

Este módulo espera que as linhas dos dois conjuntos de dados sejam agrupadas em dois conjuntos colorindo as linhas, e.g. com preto (pontos) e vermelho (cruzes).
Igualdade das médias dos dois grupos é testada por permutação com 2000 réplicas (pode ser alterado pelo usuário) e a distância de Mahalanobis elevada ao quadrado é medida. O teste de permutação é uma alternativa ao teste de Hotelling quando as premissas de distribuição com normalidade multivariada e com matrizes de covariância iguais são violadas.
Dados ausentes: suporte por substituição pela média da coluna.



M de Box (Box's M)

Teste para equivalência das matrizes de covariância de duas amostras multivariadas marcadas com cores diferentes. É um teste de homoscedasticidade, como assumida pela MANOVA. Você pode usar duas amostras multivariadas originais, cujas matrizes de covariância são calculadas automaticamente, ou duas matrizes de variância-covariância. No último caso você também deve especificar os tamanhos (número de indivíduos) das duas amostras.



A estatística *M* de Box é fornecida juntamente com o valor de significância basedo em uma aproximação por qui-quadrado. Repare que esse teste é supostamente muito sensível. Isso significa que um valor alto de *p* será um bom, embora informal, indicador de igualdade, embora um resultado altamente significativo (baixo valor de *p*) pode ser, em termos práticos, um indicador um tanto sensível demais de desigualdade.
A estatística é calculada da seguinte maneira – repare que isso é igual ao "-2 ln *M*" de alguns textos (Rencher 2002):

$$M = (n-2)\ln|\mathbf{S}| - (n_1 - 1)\ln|\mathbf{S_1}| - (n_2 - 1)\ln|\mathbf{S_2}|,$$

onde $\mathbf{S}_1$ e $\mathbf{S}_2$ são as matries de covariância, $\mathbf{S}$ é a matriz de covariância agrupada, $n=n_1+n_2$ e $|\bullet|$ representa o determinante.

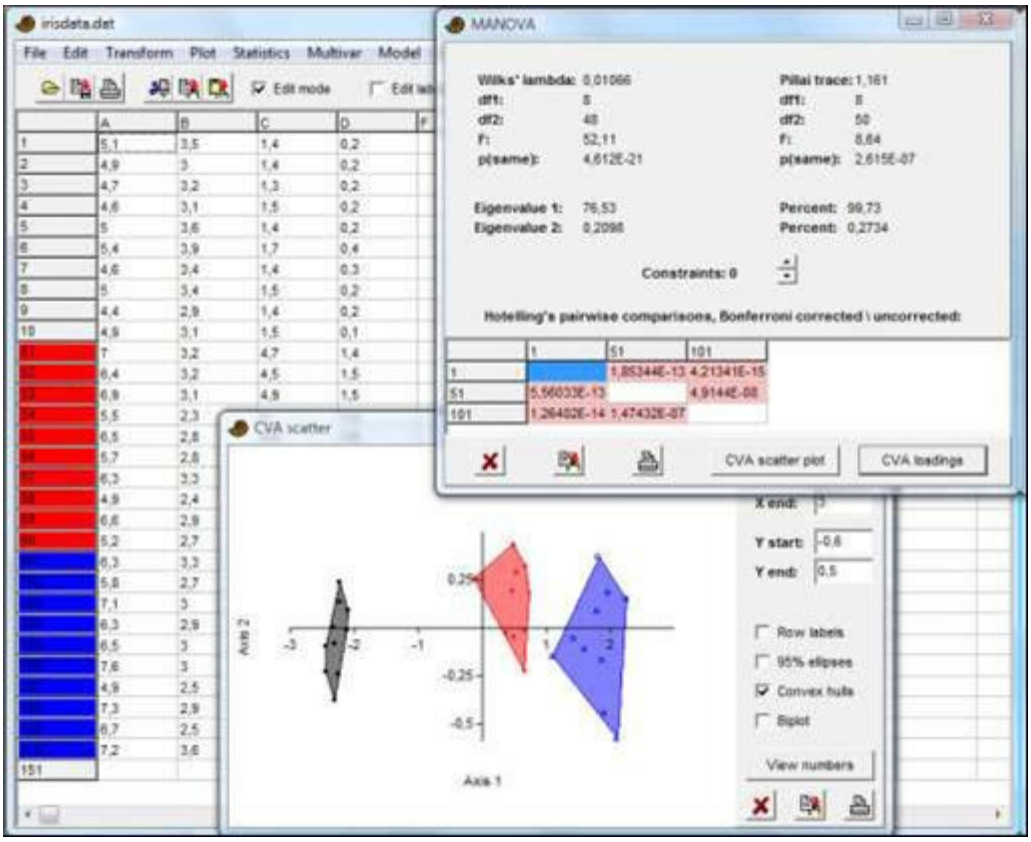
O teste de Monte carlo é baseado em 999 permutações aleatórias.
Dados ausentes: suporte por substituição pela média da coluna.

Referência

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

MANOVA/CVA

MANOVA (ANálise De VAriância Multivariada – *Multivariate ANalysis Of VAriance*) unifatorial é a versão multivariada da ANOVA univariada, que testa se uma série de amostras têm a mesma média. Caso você só tenha duas amostras, o teste de T^2 de Hotelling de duas amostras pode ser usado no lugar.



Duas estatísticas são fornecidas: lambda de Wilk com seu valor associado *F* de Rao, e o traço de Pillai (*Pillai trace*) com seu *F* aproximado. O lambda de Wilk provavelmente é mais comumente usado, mas o traço de Pillai pode ser mais robusto.

Número de restrições (Number of constraints): Para cálculo correto dos valores de *p*, o número de variáveis dependentes (restrições ou *constraints*) deve ser especificado. Normalmente, deve ser deixado em 0, mas para dados de pontos de referência (*landmarks*) com ajuste de Procrustes use 4 (para 2D) ou 6 (para 3D).

Comparaçõe par-a-par (post-hoc): Caso a MANOVA mostre diferença geral significativa entre os grupos, a análise pode proceder por comparações par-a-par. No PAST, a análise post-hoc é bem simples, feita por testes de Hotelling par-a-par. Na tabela post-hoc, os grupos são nomeados de acordo com o nome da linha do primeiro item do grupo. Os seguintes valores podem ser mostrados na tabela:

- Valores de *p* de Hotelling, sem correção para testes múltiplos. Marcados em rosa se significativos ($p<0.05$).
- Os mesmos valores de *p*, mas a significância é verificada usando o esquema sequência de Bonferroni.
- Valores de *p* corrigidos por Bonferroni (multiplicados pelo número de comparação par-a-par). A correção de Bonferroni resulta em um poder de teste muito baixo.
- Distâncias de Mahalanobis elevadas ao quadrado.

*Observação:*Estas comparações par-a-par usam a matriz de covariância intra-grupo agrupada que abrange todos os grupos envolvidos na MANOVA. Assim, elas podem diferir dos valores fornecidos pelos módulos “Permutação de dois grupos” e

"Discriminante", os quais agrupam apenas matrizes de covariância dos dois grupos que estão sendo comparados.

Dados ausentes: suporte por substituição pela média da coluna.

Análise de Variáveis Canônicas (*Canonical Variates Analysis*)
Uma opção em MANOVA, CVA, produz um gráfico de dispersão dos espécimes ao longo dos dois primeiros eixos canônicos, produzindo a separação máxima e segunda máxima entre todos os grupos (análise de discriminantes para grupos múltiplos). Os eixos são combinações lineares das variáveis originais como na PCA, e os autovalores (*eigenvalues*) indicam a quantidade de variação que é explicada pelos eixos.

Classificador (Classifier)
Classifica os dados, atribuindo cada ponto ao grupo que resulta na menor distância de Mahalanobis até a média do grupo. A distância de Mahalanobis é calculada a partir da matriz de covariância intra-grupo agrupada, fornecendo um classificador discriminante linear. Os grupos a quais os dados pertencem e aos quais foram atribuídos pela análise (i.e. grupos dados – *given* e estimados – *estimated*) são listados para cada ponto. Além disso, cada grupo é validado por um procedimento de validação cruzada deixe-um-de-fora (*jackknife*).

Matriz de confusão (Confusion matrix)
Uma tabela com o número de pontos de cada grupo (linhas) que são atribuídos aos diferentes grupos (colunas) pelo classificador. Idealmente, cada ponto deve ser atribuído ao seu respectivo grupo, resultando em uma matriz de confusão diagonal. Contagens fora da diagonal indicam o grau de falha da classificacão.

Deformações de pontos de referência (Landmark warps)
Esta função só deve ser usada se análise CVA foi feita sobre dados de pontos de referência 2D. Ela permite a plotagem interativa de deformações de forma (*shape deformations*) como uma função da posição ao longo do eixo discriminante, como gráficos-pirulilo (*lollipop plots*) (vetores para fora da posição média do ponto de referência) ou como deformações suavizadas de placa fina (*think-plate spline deformations*).

Deformações EFA
Esta função só deve ser usada se a CVA foi rodada em coeficientes calculados pelo módulo de Análise Elíptica de Fourier. Ele permite a plotagem interativa de contornos como função da posição ao longo do eixo discriminante.

Detalhes computacionais da CVA

Softwares diferentes usam versões diferentes da CVA. O cálculo usado pelo Past é fornecido abaixo.

Seja **B** os dados fornecidos, com *n* itens em linhas e *k* variáveis em colunas, centradas nas médias gerais das colunas (subtraindo as médias das colunas). Seja *g* o número de grupos, n_i o número de itens no grupo *i*. Calcule a matrix *g*x*k* das médias ponderadas dos resíduos intra-grupo, para grupo *i* e variável *j*

$$\mathbf{X}_{ij} = \sqrt{n_i}\,\overline{\mathbf{B}}_{ij},$$

onde $\overline{\mathbf{B}}_{ij}$ é a média da coluna dentro do grupo *i*. Calcule $\mathbf{B}_2$ a partir de **B** centrando dentro de grupos. Agora calcule **W** e a matriz de covariância intra-grupo normalizada e agrupada $\mathbf{W_{cov}}$:

$$\mathbf{B} = \mathbf{B}_2'\mathbf{B_2}$$

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

e e **U** são os autovalores e autovetores de **W**; $\mathbf{e_c}$ e $\mathbf{U_c}$ são os autovalores e autovetores de $\mathbf{W_{cov}}$. Então,

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

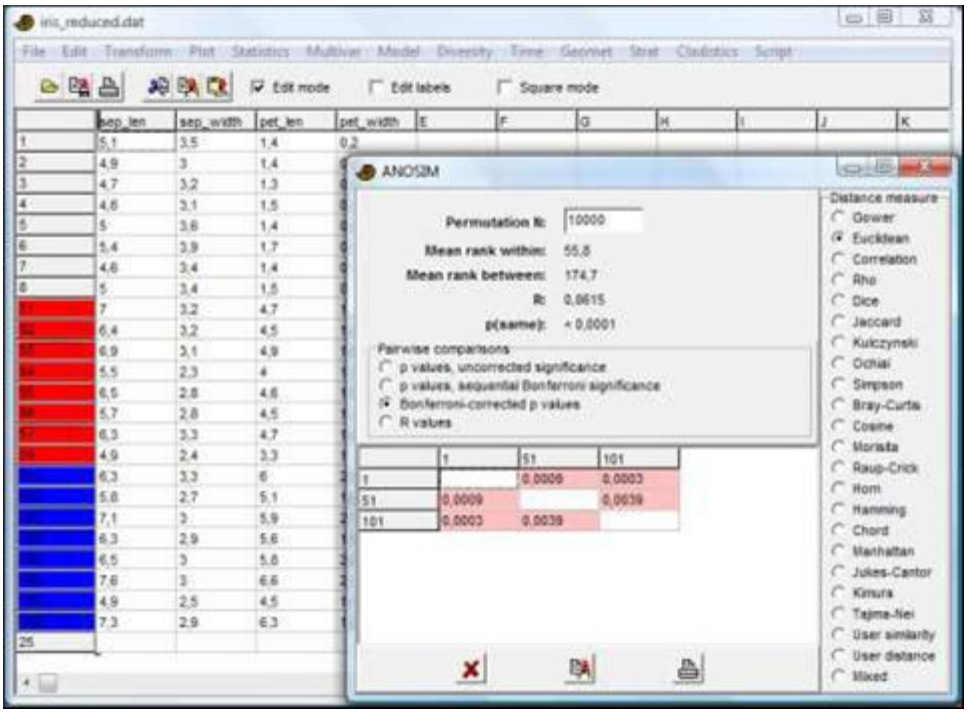
a e **A**são os autovalores e autovetores de **Z'Z**. Nós pegamos apenas os *g*-1 primeiros autovetores (colunas de **A**), e o resto será zero. As variáveis canônicas agora são

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

Os escores da CVA são, então, **BC**. A visualização das deformações de forma é mostrada ao longo de vetores $\mathbf{B_{cov}C}$.

ANOSIM unifatorial (One-way ANOSIM)

ANOSIM (ANálise De SIMilaridades – *ANalysis Of SIMilarities*) é um teste não-paramétrico de diferença significativa entre dois ou mais grupos com base em uma medida qualquer de distância (Clarke 1993). As distâncias são convertidas em ranks. ANOSIM é normalmente usada para dados de táxons-em-amostras, onde grupos são amostras que precisam ser comparadas. Ítens vão em linhas, variáveis vão em colunas, e os grupos devem ser espeficados por diferenes cores de linhas, como usual.



Fazendo uma analogia grosseira com a ANOVA, o teste é basedo na comparação de distâncias dentro de grupos com as distâncias entre grupos, seja r_b o rank médio de todas as distâncias entre grupos, e r_w o rank médio de todas as distâncias dentro de grupos. A estatística R é então definida por

$$R = \frac{r_b - r_w}{N(N-1)/4}.$$

R positivos (até 1) significam dissimilaridades entre os grupos. A significância unicaudal é calculada por permutação de amostras em grupos, com 9 999 réplicas (pode ser alterado).

Comparações ANOSIM par-a-par entre todos os pares de grupos são fornecidas como um teste post-hoc. Comparações significativas (em $p<0.05$) são mostradas em rosa. A correção opcional de Bonferroni multiplica os valores de p pelo número de comparações. Esta correção é muito conservadora (produz valores elevados de p). A opção sequencial de Bonferroni (*sequential Bonferroni*) não mostra os valores corrigidos de p, mas a significância é decidida com base em Bonferroni sequência *step-down*, o qual tem ligeiramente mais poder do que Bonferroni simples.

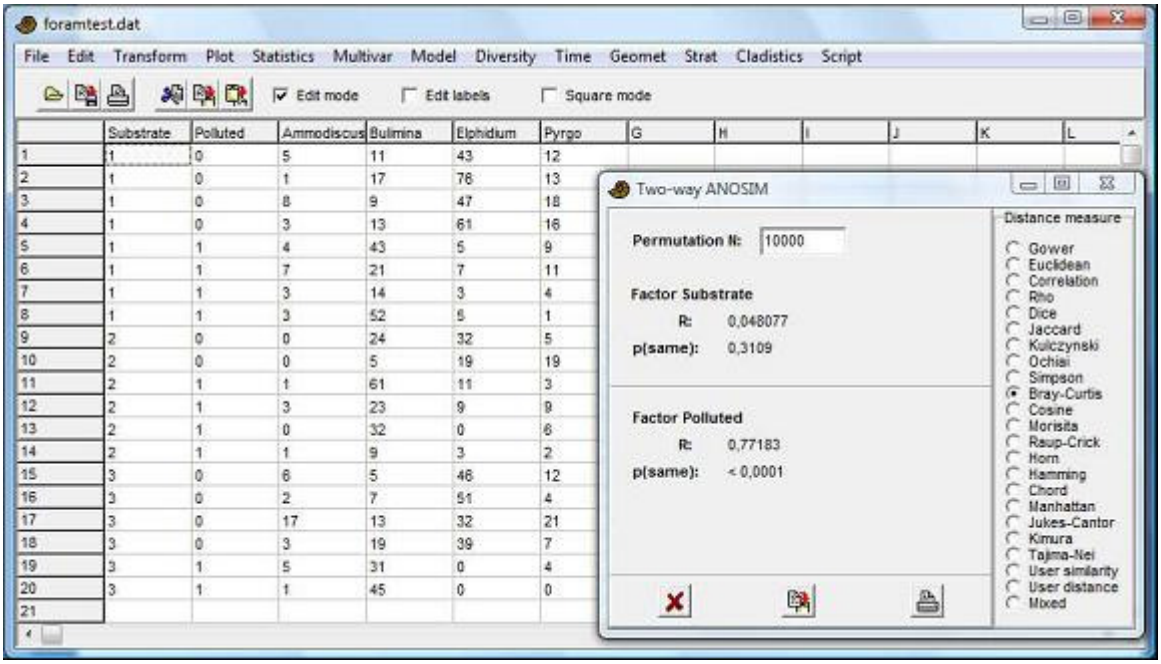
Dados ausentes: suporte por deleção (não para distâncias de Raup-Crick, Rho e definida por usuário).

Referência

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

ANOSIM bifatorial (Two-way ANOSIM)

O ANOSIM bifatorial no Past usa o delineamento cruzado (*crossed design*) (Clarke 1993). Para mais informações, ver ANOSIM unifatorial, mas repare que os grupos (níveis) não são codificados por cores, e sim com números inteiros nas primeiras duas colunas.



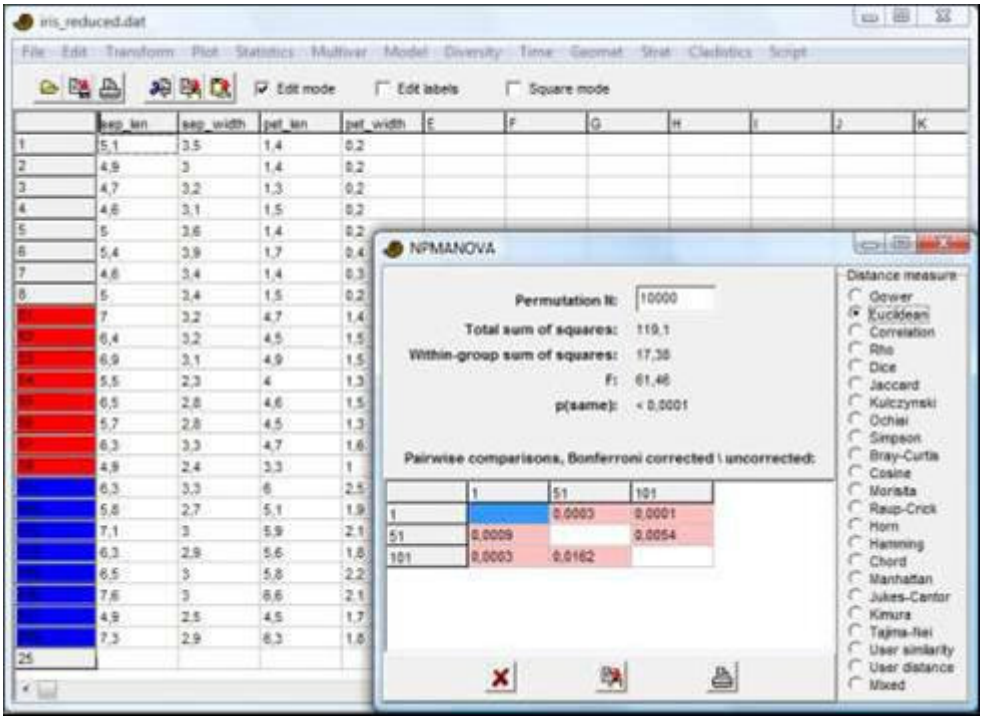
No exemplo acima, a fauna foraminífera é significativamente diferente entre as amostras poluída e não-poluída, mas não é significativa entre os substratos.

Referência

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

NPMANOVA unifatorial (One-way NPMANOVA)

NPMANOVA (MANOVA Não-Paramétrica, também conhecida como PERMANOVA) é um teste não-paramétrico para diferença significativa entre dois ou mais grupos, baseado em qualquer medida de distância (Anderson 2001). NPMANOVA normalmente é usada para dados ecológicos de táxons-em-amostras, onde grupos de amostras precisam ser comparados, mas também pode ser usada como uma MANOVA não-paramétrica geral. Ítens vão em linhas, variáveis em colunas, e os grupos devem ser especificados por cores de linhas, como usual.



NPMANOVA calcula valores de *F* de forma análoga à ANOVA. De fato, para conjuntos de dados univariados e com a medida de distância Euclideana, NPMANOVA é equivalente à ANOVA e dá o mesmo valor de *F*.

A significância é calculada permutando as amostras entre grupos, com 9 999 réplicas (pode ser alterado pelo usuário).

NPMANOVAs par-a-par entre todos os pares de grupos são fornecidas como um teste post-hoc. Comparações significativas (em $p<0.05$) são mostradas em rosa. A correção de Bonferroni mostrada no triângulo superior da matriz multiplica os valores de *p* pelo número de comparações. Esta correção é muito conservadora (produz valores elevados de *p*).

Dados ausentes: suporte por deleção par-a-par.

Referência

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

NPMANOVA bifatorial (Two-way NPMANOVA)

A NPMANOVA bifatorial (Anderson, 2001) no PAST usa o delineamento cruzado (*crossed design*). O delineamento deve ser balanceado, ou seja, cada combinação de níveis deve ter o mesmo número de linhas. Para mais informações, ver NPMANOVA unifatorial, mas repare que grupos (níveis) não são codificados com cores, e sim com números inteiros nas duas primeiras colunas (como para ANOSIM bifatorial).

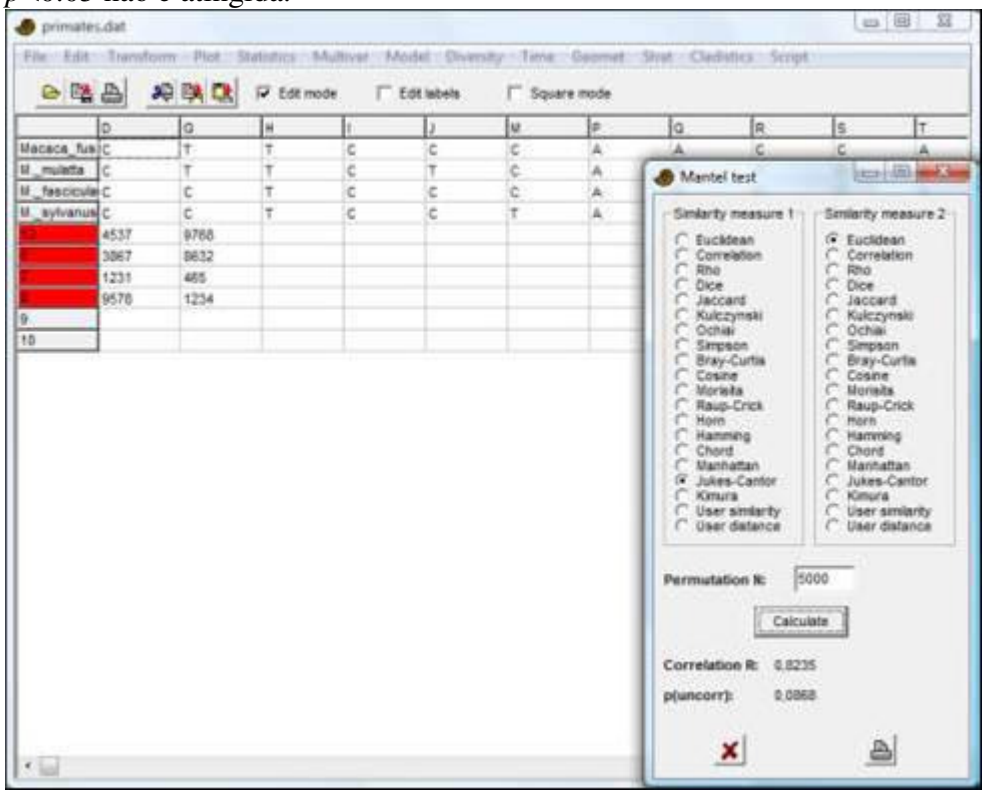
Referência

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

Teste de Mantel (Mantel test) e teste parcial de Mantel (partial Mantel test)

O teste de Mantel (Mantel 1967, Mantel & Valand 1970) é um teste por permutação para correlação entre duas matrizes de distância ou similaridade. No PAST, essas matrizes também podem ser calculadas automaticamente a partir de dois conjuntos de dados originais. A primeira matriz deve ser colocada acima da segunda matriz na planilha, e as linhas devem ser marcadas com duas cores diferentes. As duas matrizes precisam ter o mesmo número de linhas. Caso sejam matrizes de distância ou similaridade, elas também precisam ter o mesmo número de colunas.

No exemplo abaixo, a primeira matriz consiste de dados de sequência para quatro espécies de *Macaca* e a segunda matriz contém suas coordenadas geográficas. Os dois conjuntos de dados parecem estar correlacionados (R=0.82), mas a significância de p<0.05 não é atingida.



O valor de R é simplesmente o coeficiente de correlação parcial de Pearson entre todos os valores das duas matrizes (como as matrizes são simétricas, só é necessário correlacionar os triângulos inferiores). Ele varia de -1 a +1. O teste por permutação compara o R original com o R calculado em e.g. 5000 permutações aleatórias. O valores de p relatado é unicaudal (*one-tailed*).

Teste de Mantel parcial

É possível adicionar uma terceira matriz **C** embaixo das matrizes **A** e **B** como descrito acima. Esta matriz deve ser marcada como acima, e conter o mesmo número de linhas que **A** e **B**. Uma terceira medida de similaridade pode então ser escolhida para a matriz. Caso uma terceira matriz seja incluída, o programa realizará um teste de Mantel parcial

para a correlação entre **A** e **B** controlado por similares dadas em **C** (Legendre & Legendre 1998). Apenas a matriz **A** é permutada, e o valor de *R* é calculado por

$$R(\mathbf{AB} \bullet \mathbf{C}) = \frac{R(\mathbf{AB}) - R(\mathbf{AC})R(\mathbf{BC})}{\sqrt{1 - R(\mathbf{AC})^2}\sqrt{1 - R(\mathbf{BC})^2}}$$

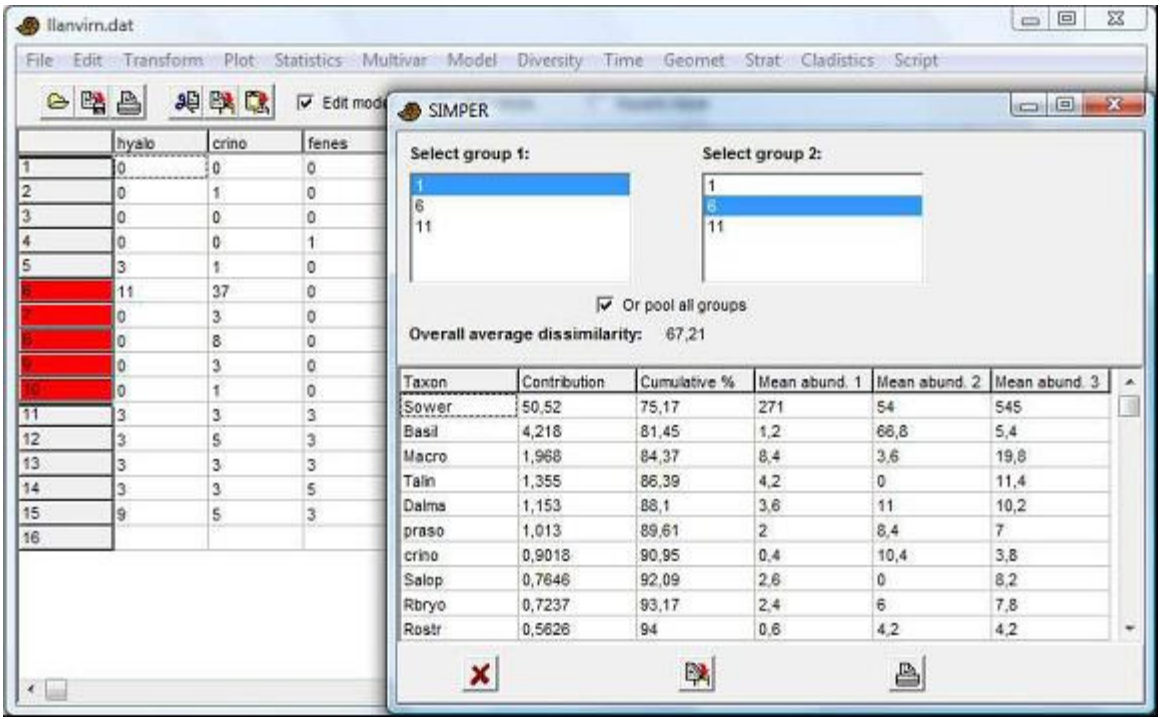
onde $R(\mathbf{AB})$ é o coeficiente de correlação entre **A** e **B**.

Referências

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

O SIMPER (*Similarity Percentage* – Porcentagem de Similaridade) é um método simples para verificar quais táxons são os principais responsáveis por uma diferença observada entre grupos de amostras (Clarke 1993). A significância geral da diferença frequentemente é verificada por meio de ANOSIM. A medida de similaridade de Bray-Curtis (multiplicada por 100) é a mais comumente usada no SIMPER, mas medidas Euclideana, coseno (*cosine*) e chord também podem ser usadas.
Caso mais de dois grupos sejam selecionados, você pode comparar dois grupos (par-a-par) escolhendo na lista de grupos ou você pode agrupar todas as amostras para realizar um único SIMPER geral para grupos múltiplos. Neste último caso, todos os pares possíveis de amostras são comparados usando a medida de Bray-Curtis. A dissimilaridade geral é calculada usando todos os táxons, enquanto as dissimilaridades táxon-específicas são calculadas para cada táxon individualmente.



Amostras vão em linhas, agrupadas por cores, e táxons vão em colunas. Neste exemplo, os três grupos (cada um com cinco amostras) são comparados. Na tabela de saída (*output*), os táxons são ordenados em ordem descendente de contribuição para a diferença entre os grupos. As últimas três colunas mostram a abundância média em cada um dos três grupos.
Dados ausentes: suporte por substituição pela média da coluna.

Referência

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

Calibração a partir de CABFAC (Calibration from CABFAC)

Este módulo reconstrói um (único) parâmetro ambiental a partir de dados de abundância de táxons-em-amostras. O programa também irá pedir um arquivo de função de transferência CABFAC (*CABFAC transfer function file*), como feito previamente por análise fatorial CABFAC (*CABFAC factor analysis*). O conjunto de táxons (colunas) deve ser indêntico na planilha e no arquivo de função de transferência.

Calibração a partir de ótimos (Calibration from optima)

As três primeiras linhas podem ser geradas de dados de abundância e dados ambientais conhecidos (Recente) na opção *"Species packing"* no menu *Model*. A terceira linha (abundância máxima – *peak abundance*) não é usada, e a segunda coluna (tolerância) é usada apenas quando a caixa *"Equal tolerances"* ("Tolerâncias iguais") não é marcada. O algoritmo é um cálculo de média ponderada, com a opção de atribuir pesos por tolerância, de acordo com Braak & van Dam (1989).

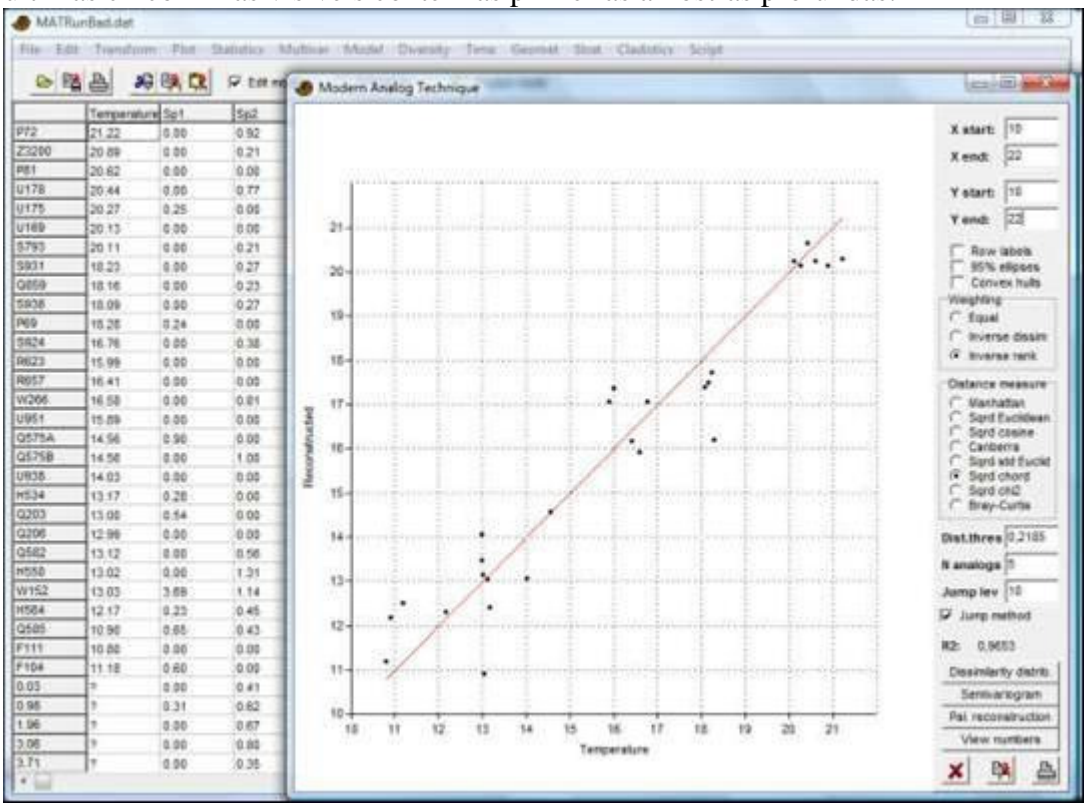
Referência

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.

Técnica de Análogo Moderno (Modern Analog Tecnhique)

A *Modern Analog Technique* funciona encontrando sítios modernos com associações de fauna similar àquelas de amostras profundas (*downcore*). Dados ambientais de sítios modernos são então usados para estimar o ambiente profundo.

A variável ambiental (única), normalmente temperatura, entra nas primeiras colunas, e os táxons nas colunas consecutivas. Todos os sítios modernos, com valores conhecidos para a variável ambiental, vão nas primeiras linhas, seguidos pelas amostras mais profundas (estas devem ter pontos de interrogação na coluna ambiental). No exemplo abaixo, as últimas cinco linhas visíveis contêm as primeiras amostras profundas.



Parâmetros para ajustar:

- Pesos (*Weighting*): Quando uma série de análogos modernos são ligados a uma amostra profunda, os seus valores ambientais podem receber os mesmos pesos, podendo ser inversamente proporcional à distância faunal, ou inversamente proporcional ao rank da distância faunal.
- Medida de distância (*Distance measure*): Uma série de medidas de distância comumente usadas no MAT são disponíveis. "*Squared chord*" tem se tornado a escolha padrão na literatura.

- Limiar de distância (*Distance threshold*): Apenas análogos modernos próximos a este limiar são usados. Um valor-padrão é dado, equivalente ao décimo percentil das distâncias entre todos os pares de amostras nos dados modernos. O histograma de "Distribuição de dissimilaridade" (*Dissimilarity distribution"*) pode ser útil para escolher este limiar.
- *N analogs*: este é o número máximo de análogos modernos usados em cada amostra profunda.
- Método de salto (*Jump method*) (on/off): Para cada amostra profunda, amostras modernas são ordenadas em distâncias ascendentes. Quando a distância aumenta mais do que a porcentagem selecionada, os análogos modernos subsequentes são descartados.

Repare que uma ou mais destas opções podem ser desligadas ao colocar nelas um número grande. Por exemplo, um limiar de distância muito grande nunca será aplicado, então o número de análogos será escolhido apenas pelo valor *"N analogs"* e opcionalmente pelo método de salto.

Validação cruzada (*Cross validation*)
O gráfico de dispersão e o valor de R^2 mostras os resultados de uma validação cruzada deixe-um-de-fora (*jackknife*) aplicada nos dados modernos. A linha $y=x$ é mostrada em vermelho. Isso reflete a "qualidade" do método apenas parcialmente, já que fornece pouca informação sobre a acurácia da estimativa para amostras profundas.

Distribuição de dissimilaridade (*Dissimilarity distribution*)
Um histograma de todas as distâncias no dados nos dados superficiais (modernos) (*core-top*).

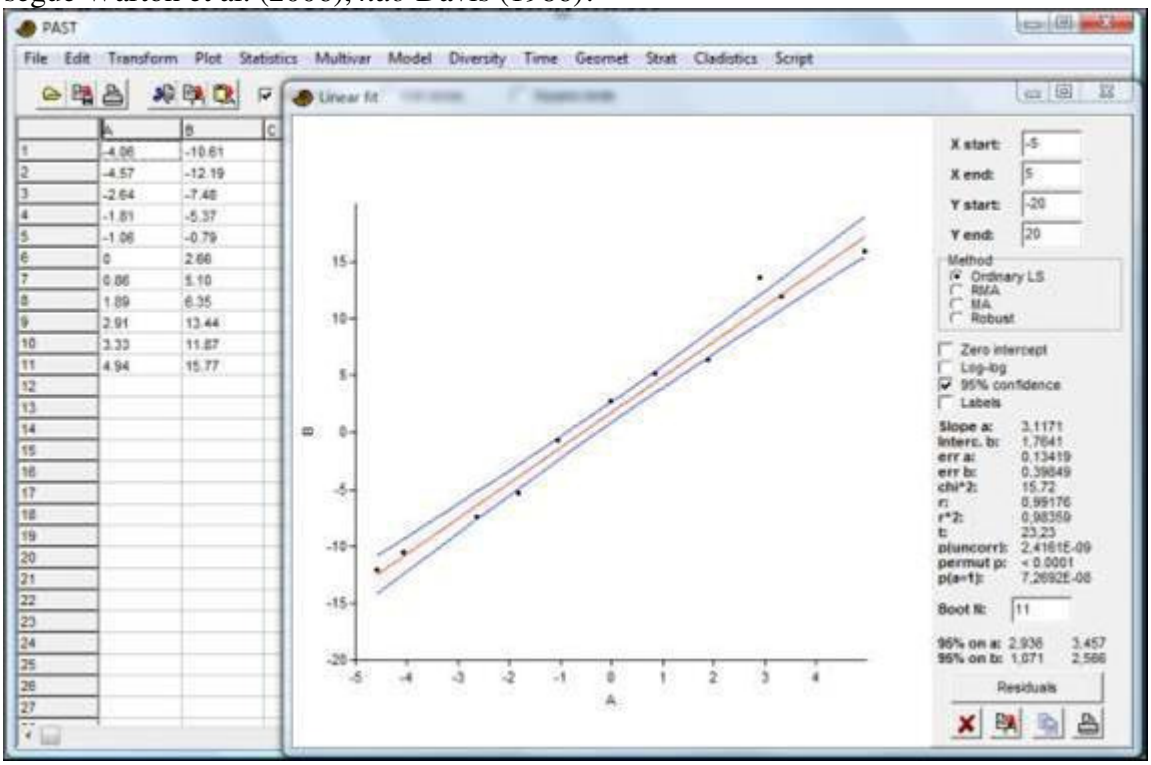
Semivariograma *(Semivariogram)*
Mostra um semivariograma da variâncias na variável ambiental, como função da diferença faunal. Mais de um modelo de semivariograma pode ser ajustado. Este tipo de gráfico é familiar da geoestatística espacial, mas é também útil para MAT porque dá uma boa impressão do grau de "ruído" nos dados de fauna no que diz respeito à predição do ambiental.

Reconstrução paleoambiental *(Pal. reconstruction*)
Reconstrução dos valores paleoambientais usando MAT.

Model menu (Modelagem)

Linear

Se duas colunas são selecionadas, elas representam valores de *x* e *y* respectivamente. Se uma coluna é selecionada, ela representa os valores de *y*, e tomam-se valores de *x* de uma sequência de números positivos (1, 2, …). Uma linha reta $x=ax+b$ é encaixada nos dados. Há quatro algoritmos diferentes disponíveis: quadrados mínimos (*Ordinary Least Squares – OLS)*, eixo maior reduzido (*Reduced Major Axis – RMA*), eixo maior (*Major Axis – MA*) e robusto (*Robust*). A regressão OLS assume que os valores de *x* são fixos e acha a linha que minimiza o quadrado dos erros nos valores de *y*. Use esta opção se seus valores de *x* têm muito pouco erro associado a eles. RMA e MA tentam minimizar os erros tanto em *x* quanto em *y*. O encaixe de RMA/MA e a estimativa do erro padrão segue Warton et al. (2006), *não* Davis (1986)!



O método "Robusto" é um Modelo I (valores fixos de *x*) de regressão avançado, robusto a valores extremos (*outliers*). Ele, às vezes, dá resultados estranhos, mas pode ter muito sucesso em casos de erros "quase" normalmente distribuídos mas com alguns valores muito discrepantes. O algoritmo é "Mínimos Quadrados Aparados" (*"Least Trimmed Squares"*) baseado no código "FastLTS" de Rousseeuw & Driessen (1999). Estimativas paramétricas de erros não são disponíveis, mas o Past fornece intervalos de confiança na inclinação e intercepto por *bootstrap* (cuidado – isso é extremamente lento para conjuntos grandes de dados).

Os valores tanto de *x* quanto de *y* podem ser transformados em log (base 10), encaixando efetivamente os dados a uma função "alométrica" $y=10^b x^a$. Um valor de *a* por volta de 1 indica que um encaixe de linha reta ("isométrico") pode ser mais aplicável.

São fornecidos os valores de *a* e *b*, seus erros, um valor de correlação por qui-quadrado (não para RMA/MA), coeficiente de correlação de Pearson *r*, e a probabilidade de que as duas colunas *não* são correlacionadas. Note que o r^2 é simplesmente o quadrado do coeficiente de Pearson – ele não é ajustado para o método de regressão.
O cálculo dos erros padrão para inclinação e intercepto assume distribuição normal dos resíduos e independência entre as variáveis, e a variância residual. Se estas premissas forem fortemente violadas, é preferível usar o intervalo de confiança de 95% *bootstrap* (2000 réplicas). O número de pontos aleatórios selecionados para cada réplica deve normalmente ser mantido em *N*, mas pode ser reduzido para aplicações especiais.
O teste por permutação para a correlação (r^2) utiliza 10 000 réplicas.

Faixa de confiança (*Confidence band*)

Em regressão OLS (mas não RMA/MA/Robusta), é disponibilizada uma faixa de confiança "Working-Hotelling" de 95% para a linha encaixada (não para os pontos de dados!). O intervalo de confiança é calculado como

$$CI = b + ax \pm t_{0.05/2,n-1}\sqrt{SE_{reg}^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

onde o quadrado da soma os resíduos $SE_{reg}^2 = \sum(y_i - b - ax_i)^2$.

Quando o intercepto é forçado a zero, a faixa de confiança é calculada como

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

Intercepto zero (*Zero intercept*)

Força a linha da regressão por zero. Isso tem implicações também para o cálculo da inclinação e do erro padrão da inclinação. Opção disponível para os quatro métodos.

Resíduos (*Residuals*)

A janela Resíduos (*Residuals*) relata as distâncias de cada ponto até a linha da regressão, nas direções *x* e *y*. Apenas a última é de interesse quando usamos regressão linear ordinária ao invés de RMA ou MA. Os resíduos podem ser copiados de volta à planilha e inspecionados para distribuição normal e independência entre a variável independente e a variância residual (homoscedasticidade).

Teste de Durbin-Watson

O teste de Durbin-Watson para autocorrelação positiva dos resíduos em *y* (violando uma premissa da regressão OLS) é fornecido na janela Resíduos. A estatística do teste varia de zero (autocorrelação positiva total) passando por 2 (sem autocorrelação) até 4 (autocorrelação negativa). Para $n \leq 400$, um valor exato de *p* para ausência de autocorrelação positiva é calculado pelo algoritmo PAN (Farebrother 1980, com correções mais recentes). O teste não é preciso quando usamos a opção Intercepto zero.

Teste de Breush-Pagan

O teste de Breush-Pagan para heteroscedasticidade, ou seja, variância não-estacionária de resíduos (violando uma premissa da regressão OLS), é dado na janela Resíduos. A estatística do teste é $LM=nr^2$, onde *r* é o coeficiente de correlação entre os valores de *x* e

os quadrados dos resíduos. A sua distribuição é assimptótica à de χ^2 com um grau de liberdade. A hipótese nula do teste é homoscedasticidade.

Funções exponenciais (*Exponential functions*)
Para encaixar aos seus dados uma função exponencial $y=e^b e^{ax}$, primeiro transforme em log apenas a sua coluna y (no menu Transform) e depois realize o encaixe de uma linha reta.

Equações RMA

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

Erro padrão de $a = abs(a)\sqrt{\frac{1-r^2}{n-2}}$.

Intercepto $b = \bar{y} - a\bar{x}$.

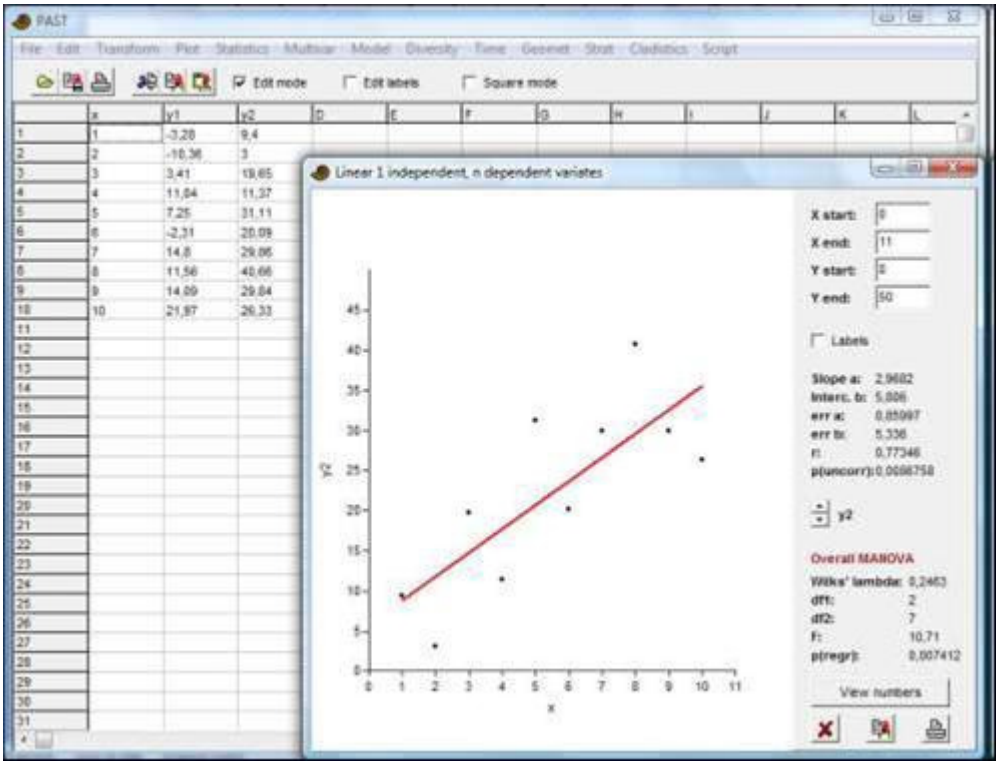
Erro padrão de $b = \frac{s_r^2}{n} + \bar{x}^2 s_a^2$, onde s_r é a estimativa do desvio padrão dos resíduos e s_a é o erro padrão da inclinação.
Para intercepto zero (b=0), coloque $\bar{x}$=0 e $\bar{y}$=0 para o cálculo da inclinação e do seu erro padrão (incluindo o cálculo do r no cálculo do r no erro padrão), e use $n - 1$ ao invés de $n - 2$ graus de liberdade no cálculo do erro padrão

Dados ausentes: suportados por deleção da linha.

Referências
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, uma independente, n dependentes (regressão multivariada) (Linear, onde independent, n dependent (multivariate regression))

Quando você tem uma variável independente e uma série de variáveis dependentes, você pode ajustar separadamente cada variável dependente à variável independente usando regressão linear simples. Este módulo torna o processo mais conveniente ao apresentar um botão de rolagem (*scroll*) que alterna entre as variáveis dependentes.
Este módulo espera duas ou mais colunas de dados mensurados, com a variável independente na primeira coluna e as dependentes nas colunas seguintes.



Adicionalmente, um teste MANOVA global da regressão multivariada é fornecido. A estatística de teste, lambda de Wilks, é calculada como a razão dos determinantes

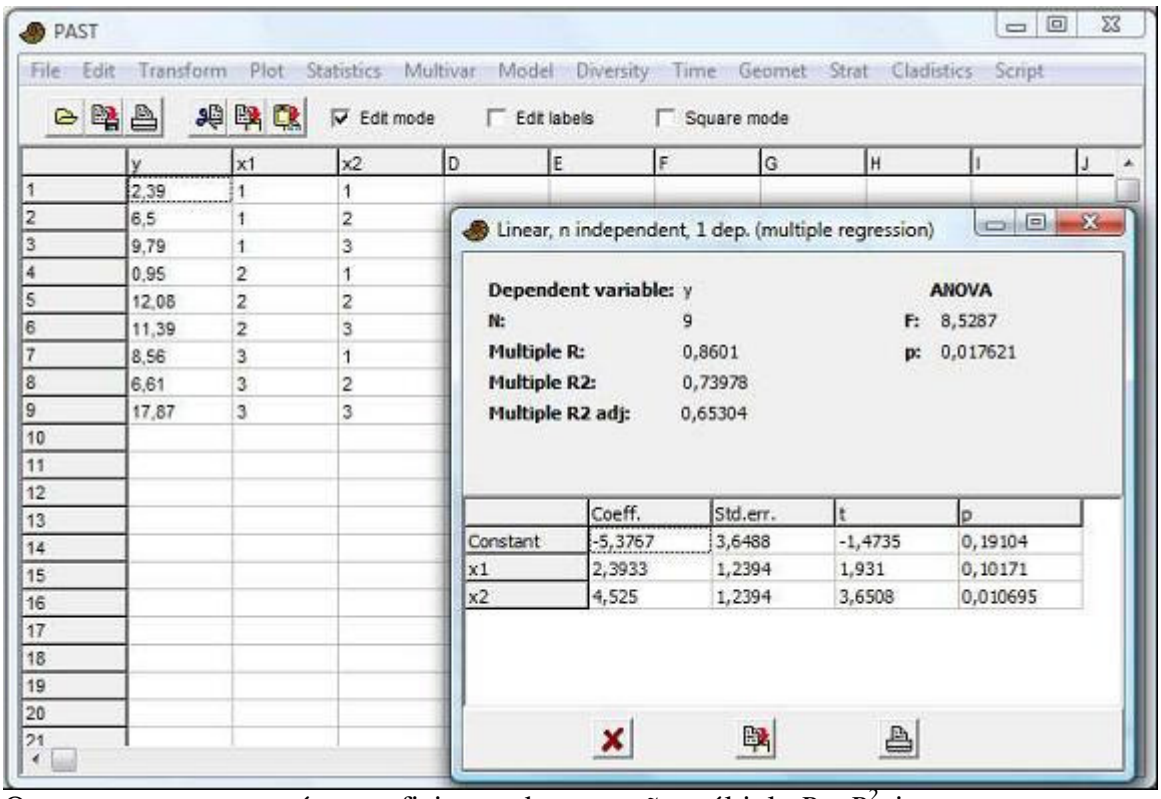
$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|},$$

onde **E** é a soma dos quadrados e produtos de erros (resíduos) (*error (residual) sum of squares and crossproducts*) e **H** é a soma dos quadrados e produtos da hipótese (predições) (*hypothesis (predictions) sum of squares and crossproducts*). A estatística *F* de Rao é calculada a partir do lambda de Wilks e sujeita a um teste *F* unicaudal (veja "Linear, n independentes, n dependentes" abaixo).
Dados ausentes são suportados por substituição pela média da coluna (*column average substitution*).

Deformações de pontos de referência e deformação EFA (Landmark warps and EFA deformation)
Se a regressão foi feita sobre pontos de referência com ajuste de Procrustes (*Procrustes-fitted landmarks*) ou com coeficientes Elípticos de Fourier (*Elliptic Fourier coefficients*) como as variáveis dependentes, a janela permite visualizar as formas como uma função da variável independentes.

Linear, n independentes, uma dependente (regressão múltipla) (Linear, n independent, one dependente (multiple regression))

Duas ou mais colunas de dados mensurados, com a variável dependente na primeira coluna e as variáveis independentes nas colunas seguintes.



O programa apresentará os coeficientes de regressão múltipla R e R^2, juntamento com o R^2 "ajustado" e um teste global de significância do tipo ANOVA. (*overall ANOVA-type significance test*).

Sendo SSR a soma dos quadrados da regressão, SSE a soma dos quadrados de erro (residual), n o número de pontos e k o número de variáveis independentes, temos que R^2=SSR/SST,

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

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

Os coeficientes (intercepto, e inclinação para cada variável independente) são apresentados juntamente com seus erros padrões estimados e testes t.

Dados ausentes suportados por substituição pela média da coluna (*column average substitution*).

Linear, n independentes, n dependentes (regressão múltipla multivariada) (Linear, n independent, n dependent (multivariate multiple regression)

Requer duas ou mais colunas de dados mensurados, com as variáveis dependentes na(s) primeira(s) coluna(s) e as independentes nas colunas seguintes. O programa irá perguntar o número de variáveis dependentes. A saída (*output*) consiste de quatro partes principais.

MANOVA global (*Overall MANOVA*)

Um teste global da significância da regressão multivariada. A estatística de teste, lambda de Wils, é calculada como a razão de determinantes

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|},$$

onde **E** é a soma dos quadrados e produtos de erros (resíduos) (*error (residual) sum of squares and crossproducts*) e **H** é a soma dos quadrados e produtos da hipótese (predições) (*hypothesis (predictions) sum of squares and crossproducts*).
A estatística *F* de Rao é calculada a partir do lambda de Wilks. Sendo *n* o número de linhas, *p* o número de variáveis dependentes e *q* o número de variáveis independentes, nós temos:

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

$$\tau = \begin{cases} \sqrt{\dfrac{p^2q^2 - 4}{p^2 + q^2 - 5}} & \text{se } p^2 + q^2 - 5 > 0 \\ 1 & \text{caso contr?io} \end{cases}$$

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

O teste *F* tem pq e $m\tau + 1 - pq/2$ graus de liberdade.

Testes nas variáveis independentes
O teste para o efeito global de cada variável independente (sobre todas as variáveis dependentes) é baseado em um desenho similar ao da MANOVA acima, mas comparando os resíduos da regressão com e sem a variável independente em questão.

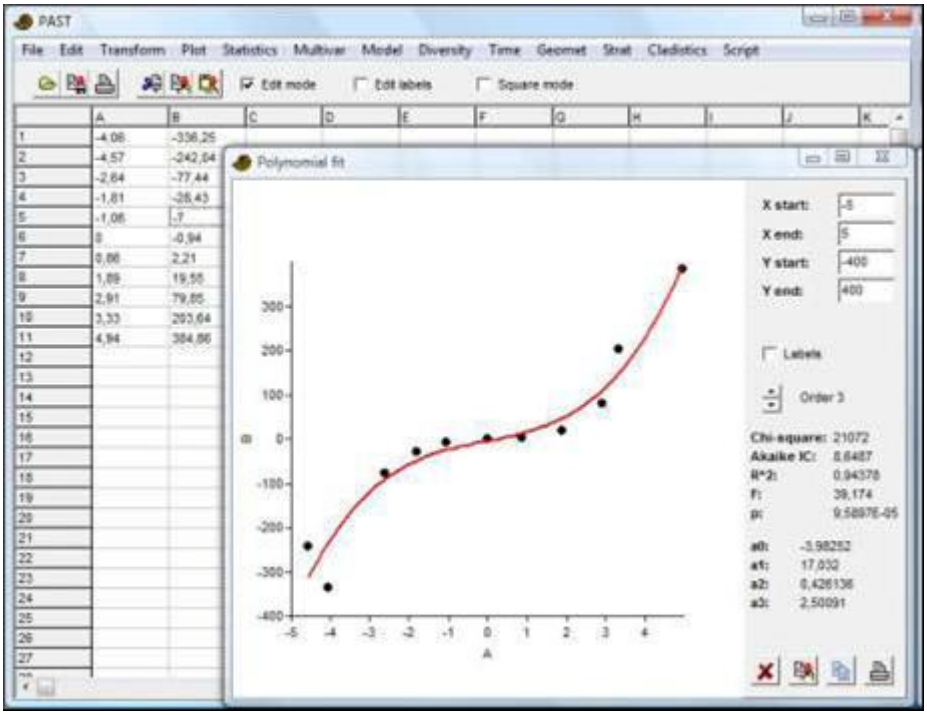
Testes nas variáveis dependentes
Veja "Linear, n independentes, uma dependente" acima para detalhes dos testes ANOVA para o efeito global de todas as variáveis independentes em cada variável dependente.

Coeficientes de regressão e estatísticas
O conjunto completo de coeficientes e suas significâncias para todas as combinações de variáveis dependentes e independentes.

Dados ausentes suportados por substituição pela média da coluna.

Regressão polinomial (Polynomial regression)

Duas colunas devem ser selecionadas (valores de *x* e de *y*). Um polinômio de até quinta ordem é ajustado aos dados. O algoritmo é baseado em um critério de mínimos quadrados e decomposição em valores singulares (*singular value decomposition*) (Press et al. 1992), com padronização de média e variância para melhor estabilidade numérica.



O polinômio é dado por

$$y = a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x^1 + a_o .$$

O valor de qui-quadrado é uma medida do erro de ajuste – valores maiores significam ajuste pior. O Critério de Informação de Akaike (*Akaike Information Criterium - AIC*) tem uma penalidade para o número de termos. O AIC deve ser tão baixo quanto possível para maximizar o ajuste, mas evitar um ajuste exagerado (*overfitting*).

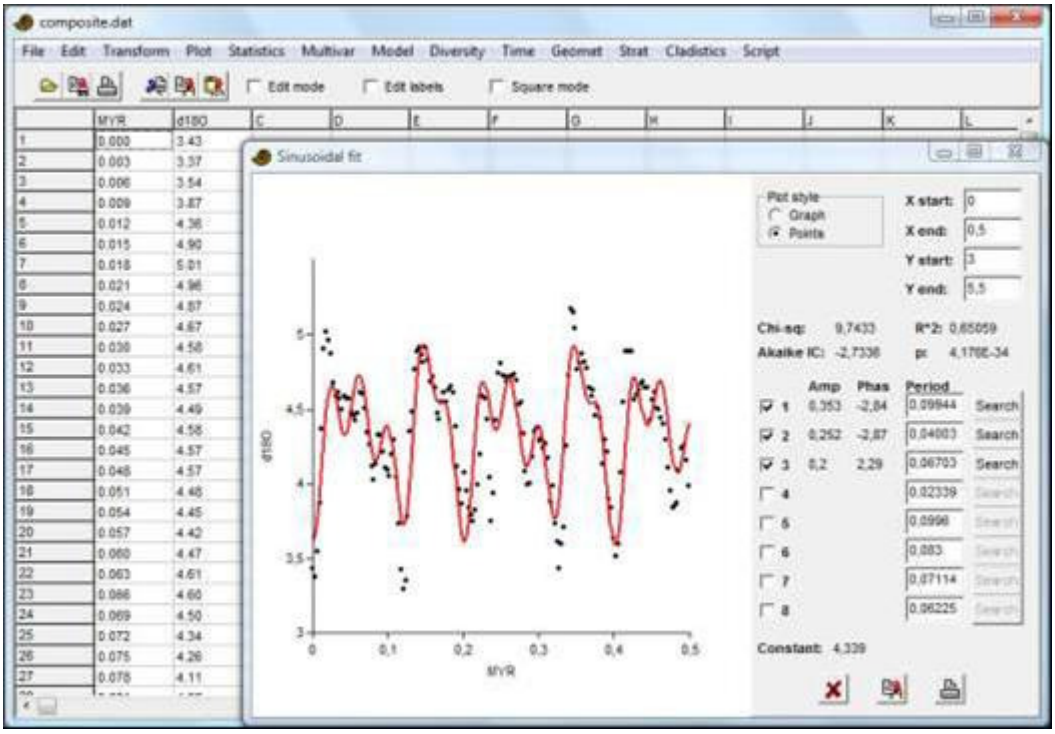
R^2 é o coeficiente de determinação, ou a proporção de variância que é explicada pelo modelo. Finalmente, um valor de p, baseado em um teste F, dá a significância do ajuste.

Referência

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

Regressão sinusoidal (Sinusoidal regression)

Duas colunas devem ser selecionadas (valores de x e de y). Uma soma de até oito sinusóides com períodos especificados pelo usuário, mas com amplitudes e fases desconhecidas, é ajustada aos dados. Isso pode ser útil para modelar periodicidades em séries temporais, como taxas anuais de crescimento ou ciclos climáticos, normalmente em combinação com análise espectral (*spectral analysis*). O algoritmo é baseado em um critério de mínimos quadrados e em decomposição em valores singulares (*singular value decomposition*) (Press et al. 1992). Por definição, os períodos são estabelecidos como sendo a extensão (*range*) dos valores de x e harmônicos (1/2, 1/3, 1/4, 1/5, 1/6, 1/7 e 1/8 do período fundamental). Estes valores podem ser mudados e não precisam estar em proporção harmônica.



O valor de qui-quadrado é uma medida do erro de ajuste – valores maiores significam ajuste pior. O Critério de Informação de Akaike (*Akaike Information Criterium – AIC*) tem uma penalidade para o número de sinusóides (a equação usada assume que os períodos são estimados dos dados). O AIC deve ser o mais baixo possível para maximizar o ajuste, mas evitar um ajuste exagerado (*overfitting*).

R^2 é o coeficiente de determinação, ou a proporção da variância que é explicada pelo modelo. Finalmente, um valor de p, baseado em um teste F, dá a significância do ajuste.

Uma função de "busca" ("*search*") para cada sinusóide irá otimizar a frequência daquele sinusóide (por toda a extensão significativa (*meaningful*) de um período até a frequência de Nyquist), mantendo as frequências de todos os outros sinusóides constantes. O algoritmo é lento, mas muito robusto e é quase garantido que ele encontre o ótimo global.

Para uma análise espectral "cega", encontrando todos os parâmetros e um número ótimo de sinusóides, siga este procedimento: Comece com apenas o primeiro sinusóide selecionado. Aperte "procurar" (*"search"*) para otimizar período, amplitude e fase. Isso vai encontrar o sinusóide mais forte nos dados. Anote o AIC. Adicione (selecione) o segundo sinusóide, e clique o botão de procura para otimizar todos os parâmetros de ambos os sinusóides, exceto o período do primeiro sinusóide. Isso vai encontrar o segundo sinusóide mais forte. Continue até o AIC parar de diminuir.

Não faz sentido (*it is not meaningful*) especificar periodicidades que são menores do que o dobro do espaçamento típico dos pontos de dados.

Cada sinusóide é dado por $y = a*cos(2*pi*(x-x_0) / T - p)$, onde a é a amplitude, T é o período e p é a fase. x_0 é o primeiro (menor) valor de x.

Também há opções para forçar uma série seno ou coseno pura, ou seja, com fases fixas.

Referências

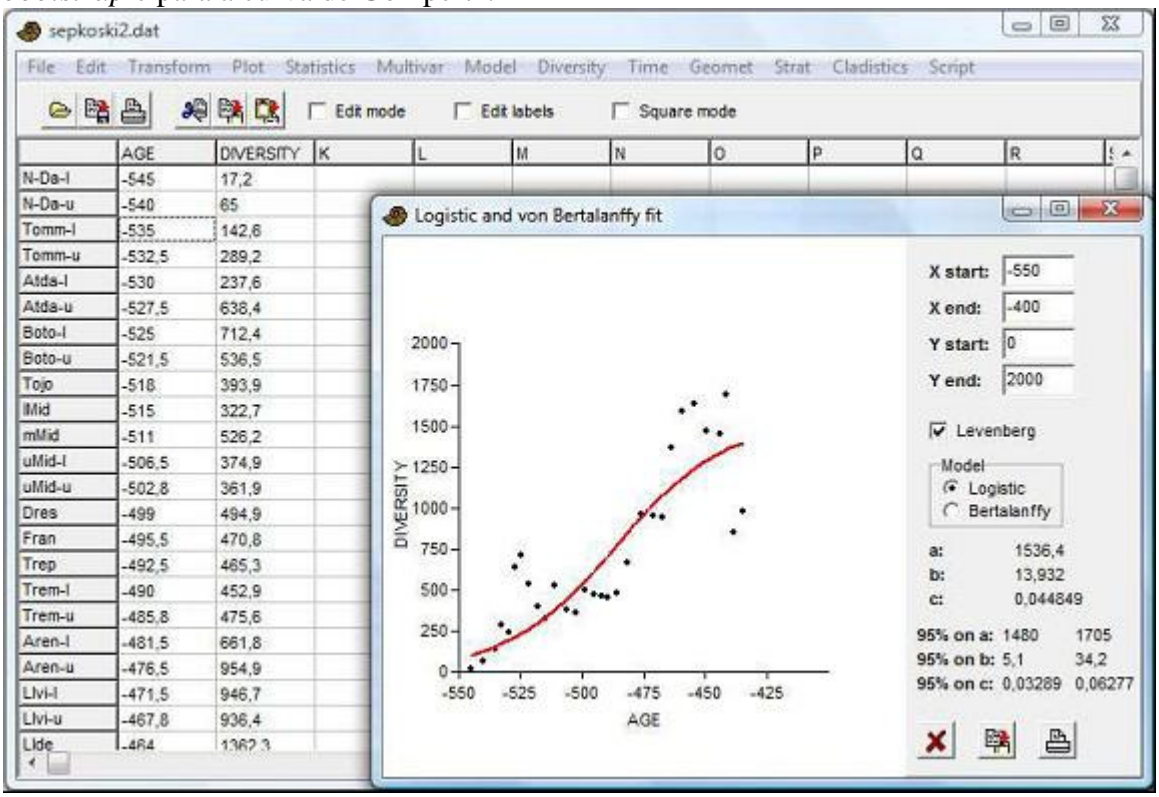
Press, W.H., S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. 1992. Numerical Recipes in C. Cambridge
University Press.

Logistic / Bertalanffy / Michaelis-Menten / Gompertz

Visa ajustar a duas colunas de dados x-y um de três modelos "de saturação" (*"saturation" models*).

A equação *logística* é $y=a/(1+be^{-cx})$. O valor de *a* primeiro é estimado pelo valor máximo de *y*. Os valores de *b* e *c* são então estimados um ajuste de linha reta a um modelo linearizado.

O modelo pode ser melhorado ao usar os valores estimados como um palpite inicial para a otimização de Levenberg-Marquardt (Press et al. 1992). Devido à instabilidade numérica, isso pode falhar com uma mensagem de erro, especialmente durante o *bootstrap* e para a curva de Gompertz.



O intervalo de confiança de 95% é baseado em 2000 réplicas de *bootstrap*.

A opção *von Bertalanffy* usa o mesmo algoritmo que acima mas ajusta a equação $y=a(1-be^{-cx})$. Esta equação é usada para modelar o crescimento de animais multicelulares (em unidade de comprimento ou largura, não volume).

A opção *Michaelis-Menten* ajusta a equação $y=ax(b+x)$. O algoritmo usa estimadores de máxima verossimilhança para a chamada transformação de Eadie-Hofstee (Raaijmakers 1987; Colwell & Coddington 1994). A estimativa normalmente melhora ao usar a otimização de Levenberg.

A opção *Gompertz* ajusta a equação $y=x*exp(b*exp(cx))$. A estimativa inicial é calculada através de regressão em um modelo linearizado.

A equação logística pode ser usada para modelar crescimento com saturação e foi usada por Sepkoski (1984) para descrever a estabilização proposta da diversidade marinha no Paleozóico tardio. Os modelos de crescimento logístico e de von Bertalanffy são descritos por Brown & Rothery (1993). A curva de Michaelis-Menten pode proporcionar ajustes precisos a curvas de rarefação, e pode, portanto (com alguma controvérsia), ser

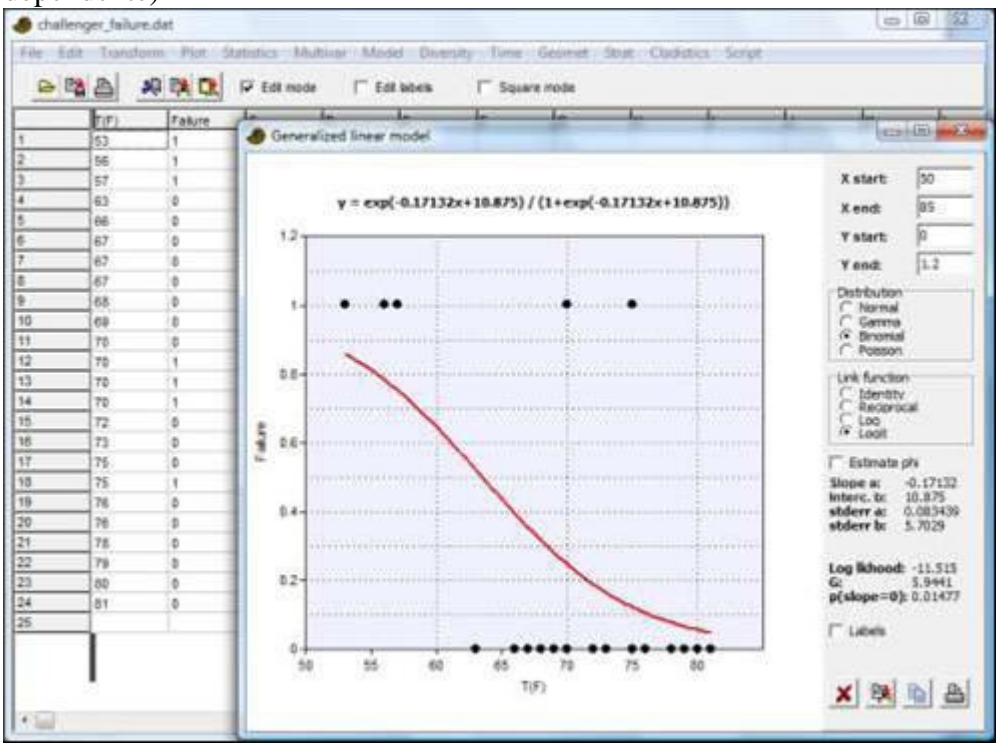
usada para extrapolar estas curvas para estimar a biodiversidade (Colwell & Coddington 1994).
O Critério de Informação de Akaike (*Akaike Information Criterium – AIC*) pode auxiliar na seleção do modelo. Valores mais baixos ao AIC implicam um ajuste melhor ajustado ao número de parâmetros.

Referências

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.

Modelo Linear Generalizado (Generalized Linear Model)

Este módulo calcula uma versão básica do Modelo Linear Generalizado, para uma única variável explanatória. Requer duas colunas de dados (variável independente e dependente)



O GLM (modelo linear generalizado) permite distribuições não-normais e também "transformações" do modelo através de uma função de ligação (*link function*). Algumas combinações particularmente úteis de distribuição e função de ligação são:
Distribuição normal & ligação identidade (Normal distribution and the identity link): É equivalente à regressão linear dos mínimos quadrados

Distribuição normal & ligação recíproca (Normal distribution and the reciprocal link): ajusta a função *y=1/(ax+b)*
Distribuição normal ou gamma e ligação log (Normal or gamma distribution and the log link): ajusta a função *y=exp(ax+b)*
Distribuição binomial (Bernoulli) e ligação logit (Binomial (Bernoulli) distribution and the logit link: Regressão logística para uma variável-resposta binária (ver figura acima).

Detalhes técnicos

O programa utiliza o algoritmo Mínimos Quadrados Repesados Iterativamente (*Iteratively Reweighted Least Squares – IRLS)* para a estimativa de máxima verossimilhança.

O parâmetro de dispersão φ, o qual é usado apenas para a inferência, não para estimativa dos parâmetros, é fixado em φ=1 a não ser que a opção "Estimar fi" *("Estiamte phi"*) seja selecionada; neste caso ele é estimado pelo qui-quadrado de Pearson. Tipicamente assume-se que φ é igual a 1 para as distribuição de Poisson e binomial.

A log-verossimilhança (*log-likelihood*) LL é calculada a partir do desvio *D* por

$$LL = -\frac{D}{2\phi} \ .$$

O desvio é calculado como se segue:

Normal: $D = \sum_i (y_i - \mu_i)^2$

Gamma: $D = 2\sum_i \left[-\ln \frac{y_i}{\mu_i} + \frac{y_i - \mu_i}{\mu_i} \right]$

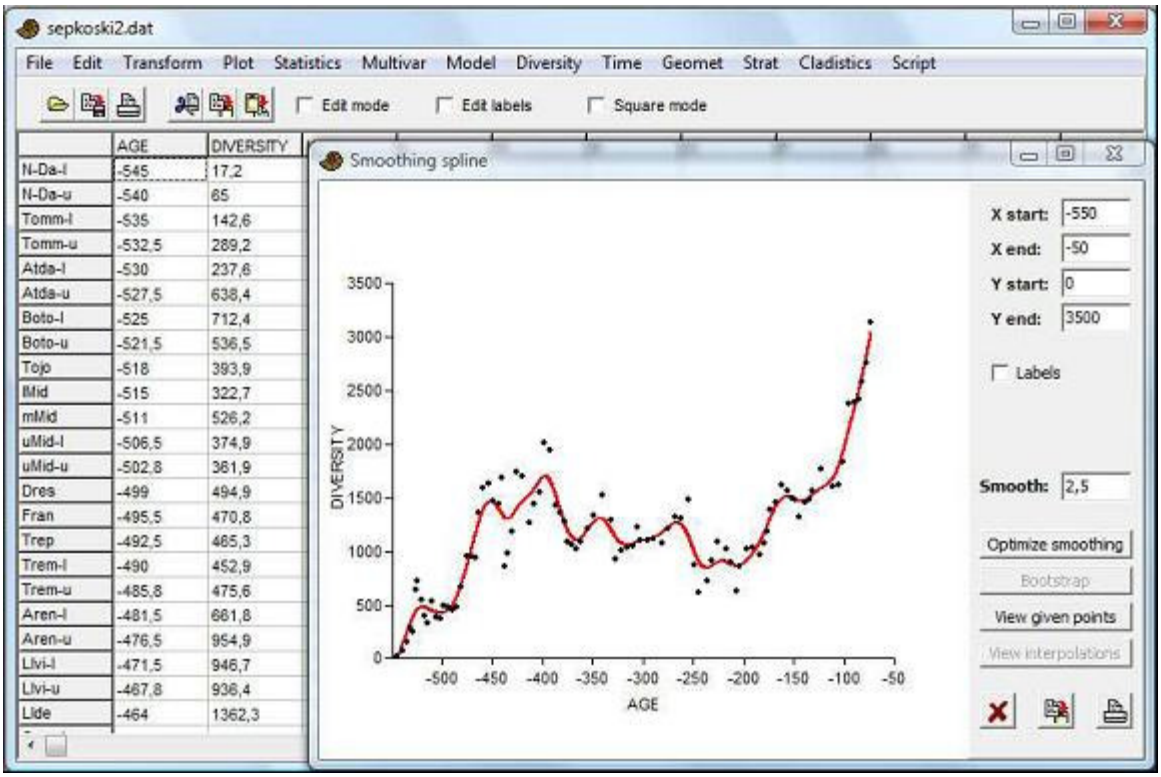
Bernoulli: $D = 2\sum_i \left[y_i \ln \frac{y_i}{\mu_i} + (1 - y_i) \ln \frac{1 - y_i}{1 - \mu_i} \right]$ (o primeiro termo é definido como zero se y_i=0)

Poisson: $D = 2\sum_i \left[y_i \ln \frac{y_i}{\mu_i} - (y_i - \mu_i) \right]$

A estatística *G* é a diferença do *D* de um modelo completo e um modelo GLM adicional onde apenas o intercepto é ajustado. A distribuição de *G* é aproximadamente igual à do qui-quadrado com um grau de liberdade, gerando um valor de significância para a inclinação.

Alisamento polinomial (Smoothing spline)

Duas colunas devem ser selecionadas (valores de X e Y). Os dados são ajustados a uma *smoothing spline* (algo como uma "curva de suavização), que é uma sequências de polinômios de terceira ordem contínuos até a segunda derivada. Uma aplicação típica é a construção de uma curva suave através de um conjunto de dados turbulento (*noisy*). O algoritmo segue de Boor (2001). Saltos bruscos nos dados podem resultar em oscilações na curva, e você também pode obter grandes excursões (*excursions*) em regiões com poucos pontos de dados. Pontos múltiplos em um mesmo valor de *X* são colapsados para um único ponto através do cálculo da média ponderada (*weighted averaging*) e de um desvio padrão combinado.



Uma terceira coluna opcional especifica os desvios padrões dos pontos de dados. Estes são usados para ponderar os dados. Se não forem especificados, são todos fixados em 10% do desvio padrão dos valores de *Y*.

O valor de suavização (*smoothing value*) estabelecido pelo usuário é uma versão normalizada do fator de suavização (*smoothing factor*) de de Boor (1 por padrão). Valores maiores resultam em curvas mais suaves. Um valor de 0 irá começar um segmento da curva em cada ponto. Clicando em "Otimizar suavização" (*"Optimize smoothing"*) irá calcular uma suavização "ótima" por um procedimento de validação cruzada (*crossvalidation procedure*).

"Ver pontos fornecidos" (*"View given points"*) fornece uma tabela dos pontos de *X, Y* e desvio padrão de Y (stdev(Y)), os valores de *Y* correspondentes na curva (ys) e os resíduos. O teste de qui-quadrado em cada ponto pode ser usado para identificar valores extremos (*outliers*). A coluna final sugere um valor de stdev(Y) para ser usado se o valor de *p* está sendo forçado para 0.5.

Uma quarta coluna, opcional (se usada então a terceira coluna também deve ser preenchida com valores de desvio padrão) pode conter um número de valores diferentes das colunas anteriores. Ela contém valores de *X* para serem usados para interpolação entre os pontos de dados. Colunas 5-7, opcionais, contêm limites inferior e superior para os valores de *X* (distribuição retangular) e devios padrões dos valores de *Y* (distribuição normal), a serem usados em simulação por *bootstrap* (Monte Carlo) para fornecer barras de erro para os valores interpolados. Estas funções são incluídas principalmente para calcular idades de limite (*boundary ages*) para a escala de tempo geológica.

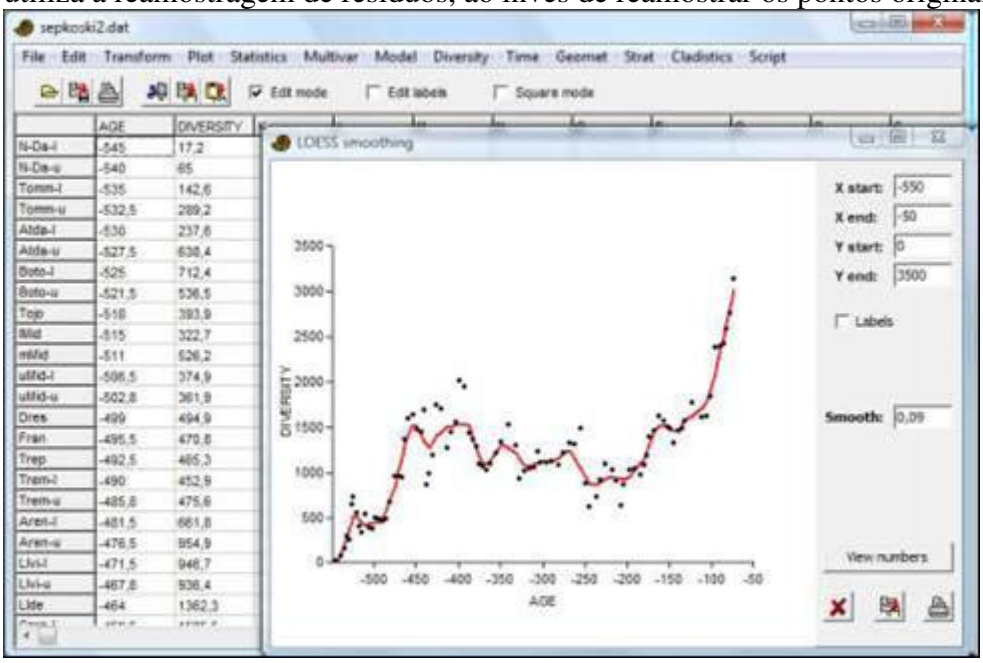
Referência

de Boor, Carl. 2001. A practical guide to splines. Springer.

Alisamento LOESS (LOESS smoothing)

Duas colunas devem ser selecionadas (valores de *x* e *y*). O algoritmo usado é "LOWESS" *(LOcally WEighted Scatterplot Smoothing* – Suavização de gráfico de Dispersão Ponderada Localmente*;* Cleveland 1979, 1981), com os seus valores padrões de parâmetros recomendados (incluindo duas iterações de robustez). Dado um número de pontos *n* e um parâmetro de suavização (*smoothing*) *q* especificado pelo usuário, o programa ajusta os *nq* pontos ao redor de cada ponto para uma linha reta, com uma função de ponderamento que decresce com a distância. O novo ponto suavizado é o valor da função linear ajustada na posição original *x*.

A opção *Bootstrap* irá estimar uma faixa de confiança de 95% para curva, com base em 999 réplicas aleatórias. Para manter a estrutura original da interpolação, o procedimento utiliza a reamostragem de resíduos, ao invés de reamostrar os pontos originais.



LOESS ou smoothing spline?

É quase uma questão de gosto. Compare as curvas acima, para o mesmo conjunto de dados. A *smoothing spline* frequentemente dá uma curva mais agradável esteticamente por causa das suas derivadas contínuas, mas você corre o risco da curva ser exagerada (*overshooting*) perto de curvas abruptas nos dados.

Referências

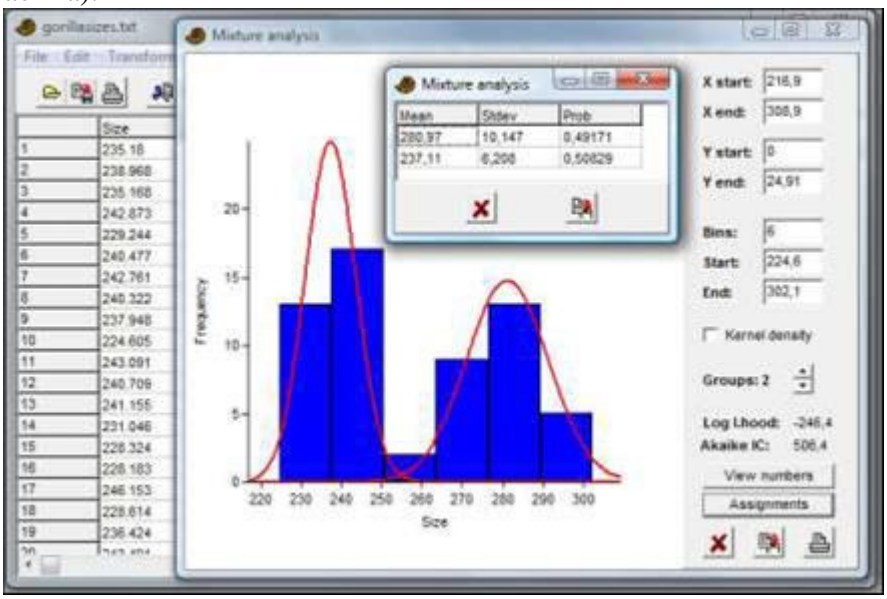
Cleveland, W.S. 1979. Robust locally weighted fitting and smoothing scatterplots. *Journal of the American Statistical Association* 74:829-836.

Cleveland, W.S. 1981. A program for smoothing scatterplots by robust locally weighted fitting. *The American Statistician* 35:54.

Análise de mistura (Mixture analysis)

A análise de mistura é um método de máxima verossimilhança para estimar os parâmetros (médio, desvio padrão e proporção) de duas ou mais distribuições normais univariadas com base em uma amostra univariada agrupada. O programa também pode estimar média e proporção de distribuições normal e de Poisson. Por exemplo, o método

pode ser usado para estudar diferenças entre sexos (dois grupos), ou uma série de espécies, ou classes de tamanho, quando nenhuma informação independente sobre pertencimento a grupos (*group membership*) está disponível.
O programa espera uma coluna de dados univariados, e assume-se que tenham sido tomados de uma mistura de populações com distribuição normal (ou exponencial ou Poisson). No exemplo abaixo, os tamanhos de gorilas machos e fêmeas foram agrupados em uma única amostra. As médias, desvios padrões e proporções das duas amostras originais foram "recuperados" quase perfeitamente (veja "Univariado" (*"Univariate")* acima).



O PAST usa o algoritmo EM (Dempster et al. 1977), o qual pode ficar preso em um ótimo local. O procedimento é então feito automaticamente 20 vezes, cada vez posições iniciais aleatórias novas para as médias. Os valores iniciais para o desvio padrão são estabelecidos em *s/G*, onde *s* é o desvio padrão agrupado e *G* é o número de grupos. Os valores iniciais das proporções são estabelecidos em 1/*G*. Ainda é recomendado que o usuário rode o programa algumas vezes para verificar a estabilidade da solução (soluções "melhores" têm valores menos negativos da log-verossimilhança (*log likelihood values*)).
O Critério de Informação de Akaike (*Akaike Information Criterium – AIC; Akaike 1974*) é calculado com uma correção para amostra pequena:

$$\mathrm{AICc} = 2k - 2\ln L + \frac{2k(k+1)}{n-k-1}$$

onde *k* é o número de parâmetros, *n* é o número de pontos de dados e *L* é a verossimilhança (*likelihood*) do modelo com os dados fornecidos. Um valor mínimo do AIC indica que você escolheu o número de grupos que produz o melhor ajuste sem ajustar demais (*without overfitting*).
É possível atribuir cada um dos pontos de dados a um dos grupos a partir de uma abordagem de máxima verossimilhança. Isto pode ser usado como um método de agrupamento não-hierárquico para dados univariados. O botão "Atribuições" (*"Assignments"*) irá abrir uma janela onde o valor de cada função de probabilidade de densidade (*probability density function*) é dado para cada ponto de dados. Os pontos de dados podem ser atribuídos ao grupo que mostra o maior valor.

Dados ausentes: suporte por deleção.

Referências

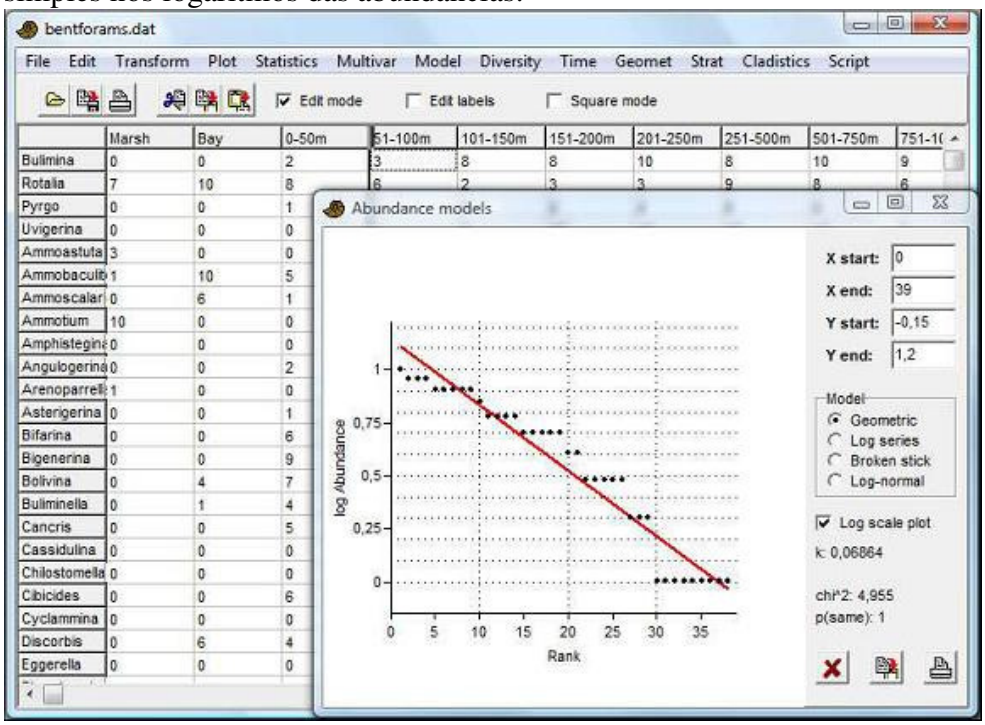
Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716-723.

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.

Modelos de abundância (Abundance models)

Este módulo pode ser usado para plotar abundância de táxons em ordem ordinal decrescente de *rank* em uma escala linear ou logarítmica (gráfico de Whittaker), ou o número de espécies em classes de abundância de oitavas (como mostrado ao ajustar à distribuição log-normal). Os táxons vão nas linhas. Também pode ser usado para ajustar os dados a um dos quatro modelos padrão de abundância:

- Geométrico, onde a segunda espécie mais abundante deve ter uma contagem de táxon k<1 vezes a da mais abundante, a 3ª mais abundante uma conta de táxon igual a ka vezes a 2ª mais abundante etc, para um *k* constante. Sendo n_ia contagem do *i*-ésimo táxon mais abundante, temos $n_i = n_1 k^{i-1}$. Isto resultará em uma linha reta descendente no gráfico de Whittaker. O ajuste é feito por regressão linear simples nos logaritmos das abundâncias.



- Log-series, com dois parâmetros *alfa* e *x*. O algoritmo usado no ajuste é de Krebs (1989). O número de espécies com *n* indivíduos (esta esquação não se traduz diretamente para a representação gráfica de Whittaker):

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

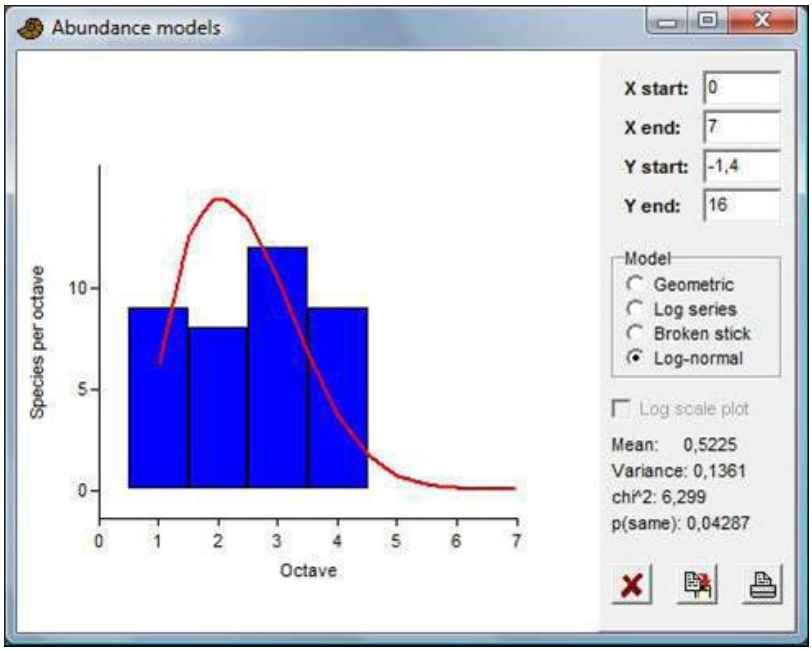
- *Broken stick* (MacArthur 1957). Não há parâmetros livres para serem ajustados a este modelo. Sendo S_{tot} o número total de espécies e n_{tot} o número total de

indivíduos:

$$n_i = \frac{n_{tot}}{S_{tot}} \sum_{j=0}^{S_{tot}-i} \frac{1}{S_{tot} - j}.$$

- Log-normal. O algoritmo de ajuste é de Krebs (1989). O logaritmo (base 10) da média e variância ajustadas são dados. As *oitavas (octaves)* referem-se a classes de abundância da potência de 2:

Oitava	Abundância
1	1
2	2-3
3	4-7
4	8-15
5	16-31
6	32-63
7	64-127
...	...



Um valor de significância baseado em qui-quadrado é fornecido para cada um destes modelos, mas o poder do teste não é o mesmo para os quatro modelos e os valores de significância, portanto não devem ser comparados. É importante, como sempre, lembrar que um valor elevado de *p* não pode ser tomado como implicando um bom ajuste. Um valor baixo não implica que o ajuste é ruim. Note também que os testes de qui-quadrado no PAST parecem não corresponder com alguns outros programas, possivelmente porque o PAST usa contagens ao invés dos valores log-transformados dos gráficos de Whittaker.

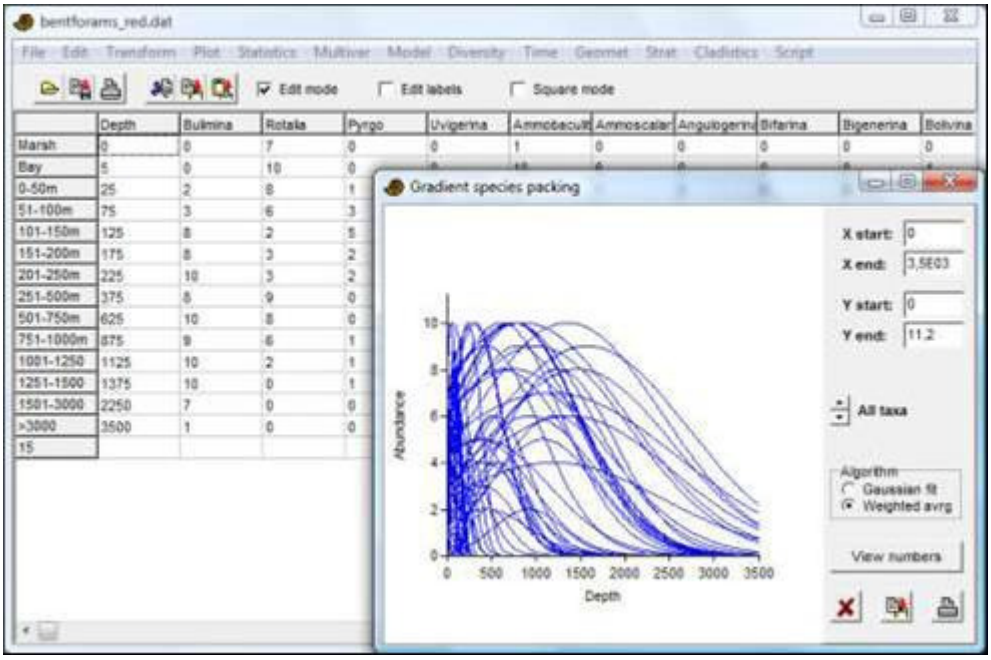
Referências

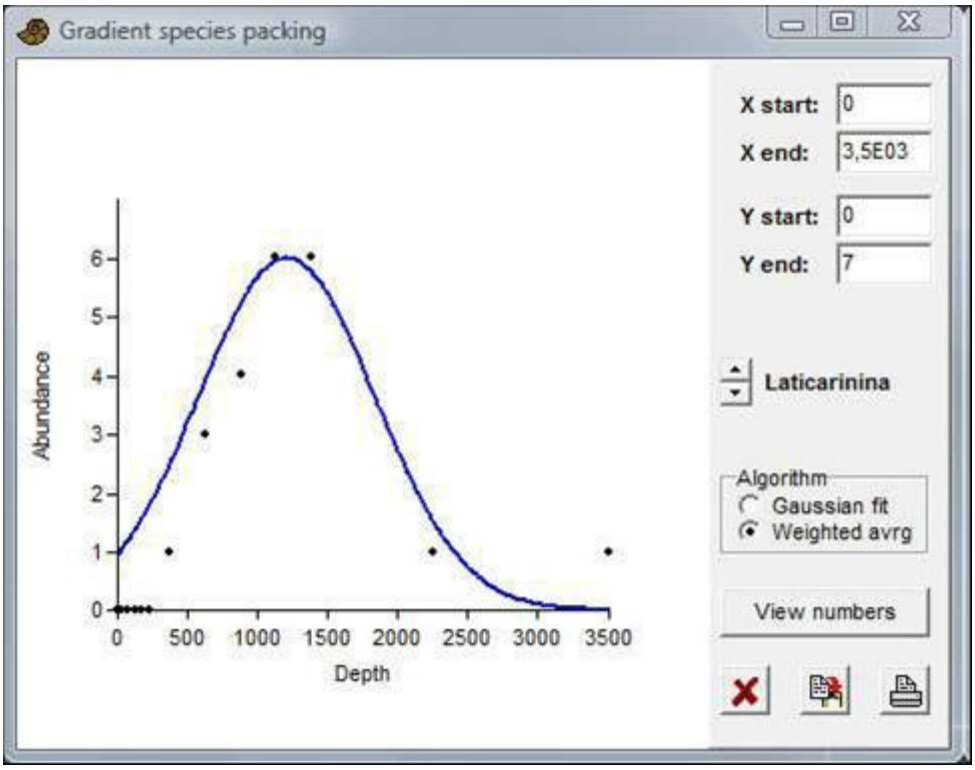
Krebs, C.J. 1989. Ecological Methodology. Harper & Row, New York.

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

Empacotamento de espécies (Gaussiano) (Species packing (Gaussian))

Este módulo ajusta modelos de resposta Gaussianos às abundâncias de espécies ao longo de um gradiente, para uma ou mais espécies. Os parâmetros ajustados são: ótimo (média), tolerância (desvio padrão) e máximo.
Uma coluna de medidas ambientais nas amostras (e.g. temperatura), e uma ou mais colunas de dados de abundância (táxons em colunas).
O algoritmo é baseado em média ponderada de acordo com ter Braak & von Dam (1989).



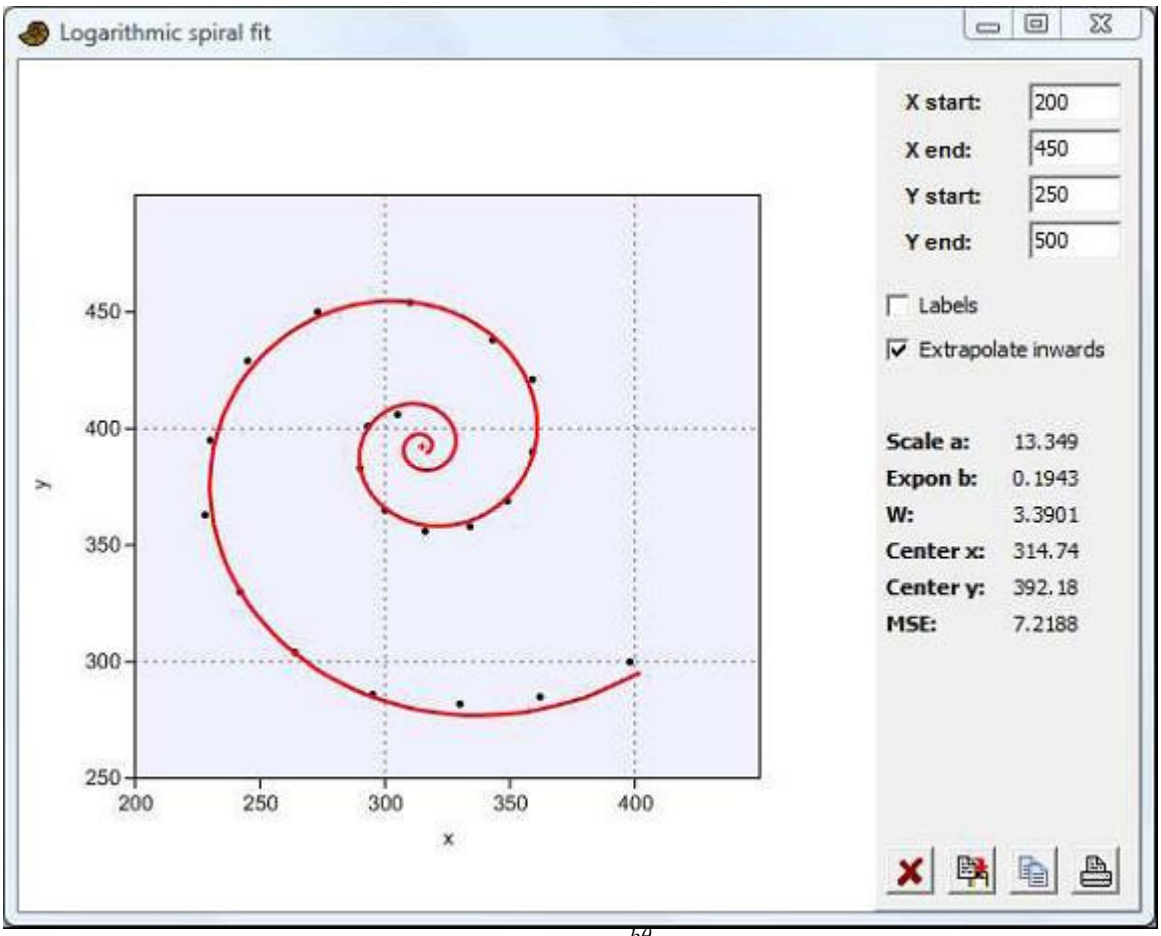


Referência

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.

Espiral logarítmica (Logarithmic spiral)

Ajusta um conjunto de pontos no plano a uma espiral logarítmica. Útil para caracterizar e.g. conchas de moluscos, dentes, garras e chifres. Requer duas colunas de coordenadas (x e y). Os pontos devem ser dados na sequência, para dentro ou para fora. São aceitas espirais voltadas tanto para direita quanto para esquerda.



A espiral ajustada em coordenadas polares: $r=ae^{b\theta}$. A escala a e o expoente b são fornecidos, juntamente com o ponto central estimado, marcado com uma cruz vermelha. A taxa de expansão da espiral W (fator de incremento no raio por giro) é calculada a partir do b como $W = e^{2\pi b}$.
A posição central é estimada por otimização não-linear e a própria espiral por linearização e regressão.

Diversity menu (Diversidade)

Índices de diversidade (Diversity indices)

Essas estatísticas se aplicam a dados de associação, onde o número de indivíduos é tabulado nas linhas (táxons) e possivelmente em mais de uma coluna (associações). As estatísticas disponíveis são as seguintes, calculadas para cada associação:

- Número de táxons (S)
- Número total de indivíduos (n)

- Dominância = 1 – Índice de Simpson. Varia de 0 (todos os táxons presentes em iguais quantidades) a 1 (um táxon domina completamente a comunidade).

  $$D = \sum_i \left(\frac{n_i}{n} \right)^2$$ onde n_i é o número de indivíduos do táxon *i*.
- Índices de Simpson 1 – D. Mede a "equitabilidade" ("*evenness*") da comunidade, de 0 a 1. Preste atenção na confusão existente na literatura – a dominância e o índice de Simpson são frequentemente trocados!
- Índice de Shannon (entropia). Um índice de diversidade que leva em conta não só o número de táxons, mas também o número de indivíduos. Varia de 0 para comunidades com um único táxon até valores elevados para comunidades com muitos táxons, cada um com alguns indivíduos.

  $$H = -\sum_i \frac{n_i}{n} \ln \frac{n_i}{n}$$
- Índice de equitabilidade de Buzas e Gibson: e^H/S
- Índice de Brillouin:

  $$HB = \frac{\ln(n!) - \sum_i \ln(n_i!)}{n}$$
- Índice de riqueza de Menhinick: $\frac{S}{\sqrt{n}}$
- Índice de riqueza de Margalef: $(S-1) / \ln (n)$
- Equitabilidade. Índice de diversidade de Shannon dividido pelo logaritmo do número de táxons. Esta medida representa e equitabilidade com a qual os indivíduos se distribuem entre os táxons presentes.
- Alfa de Fisher (Fisher's alpha) – um índice de diversidade, definido implicitamente pela fórmula $S = a*\ln(1+n/a)$, onde *S* é o número de táxons, *n* é o número de indivíduos, e *a* é o alfa de Fisher
- Dominância de Berger-Parker: simplesmente o número de indivíduos do táxon dominante em relação ao *n*.

Muitos desses índices são explicados em Harper (1999).

Intervalos de confiança aproximados para todos estes índices podem ser calculados por um procedimento de *bootstrap*. São produzidas 1000 amostras aleatórias (200 antes da versão 0.87b), cada uma com o mesmo número total de indivíduos que na amostra original. As amostras aleatórias são retiradas do conjunto de dados total (agrupando todas as colunas). Para cada indivíduo da amostra aleatória, o táxon é escolhido de acordo com as abundâncias agrupadas (*pooled abundances*) originais. Um intervalo de confiança de 95% é então calculado. Repare que a diversidade das réplicas frequentemente será menor, e nunca maior, que a diversidade da amostra total agrupada (*pooled diversity*).

Como estes intervalos de confiança são calculados em relação ao conjunto de dados agrupado (*pooled data set*), eles não representam intervalos de confiança das amostras individuais. São úteis principalmente para identificar amostras nas quais um dado índice de diversidade está fora do intervalo de confiança. Comparação por

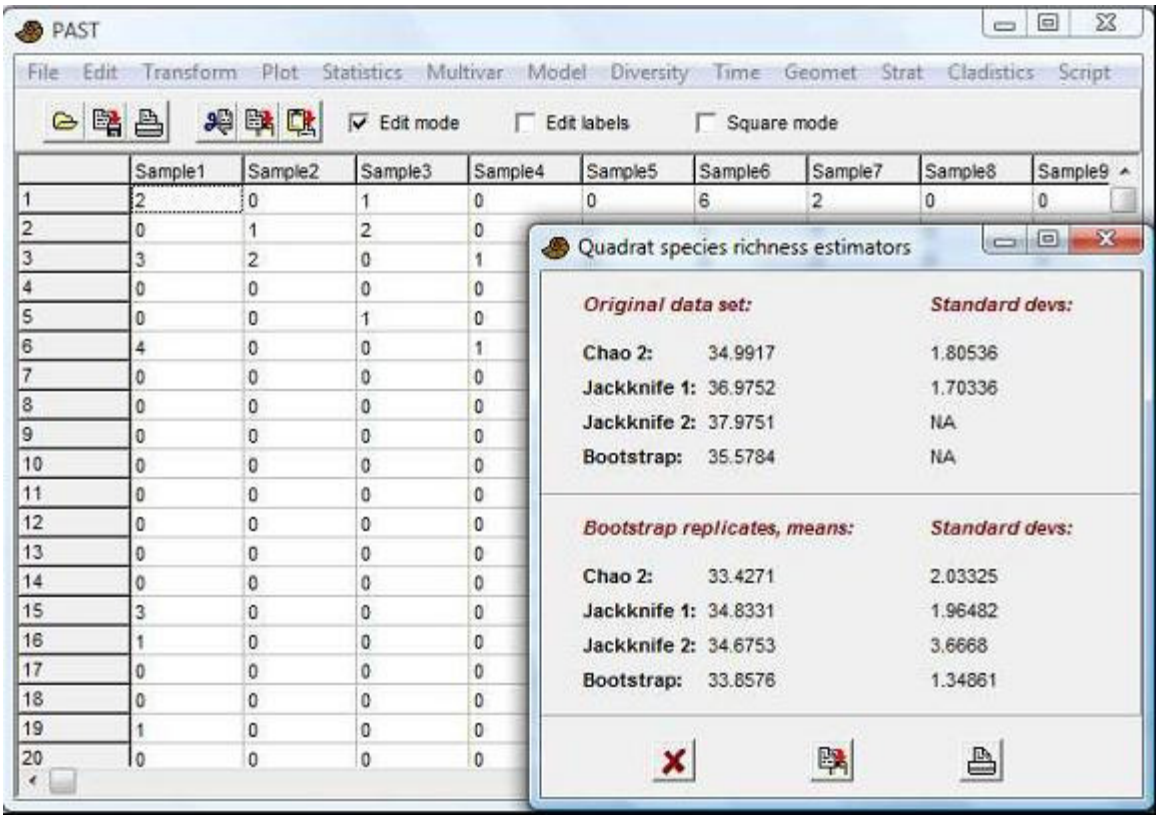
bootstrap dos índices de diversidade de duas amostras é fornecida no módulo *Compare diversities* (Comparar diversidades).

Referência

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

Riqueza quadrática ou por parcela (Quadrat richness)

Requer duas ou mais colunas, cada um com dados de presença/ausência (1/0) de diferentes táxons descendo as linhas (abundância positiva é tratada como presença).
São incluídos no PAST quatro estimadores não-paramétricos de riqueza de espécies: Chao 2, *jackknife* de primeira e segunda ordem, e *bootstrap*. Todos eles requerem dados de presença/ausência em duas ou mais parcelas de tamanhos iguais amostradas. Colwell & Coddington (1994) revisaram estes estimadores e encontraram que Chao2 e *jackknife* de segunda ordem apresentam melhor performance.
O *output* (resultado fornecido) do Past é dividido em dois painéis. Primeiro, os estimadores de riqueza e seus desvios-padrões analíticos (appenas para Chao2 e *Jackknife* 1) são calculados do conjunto de amostras. A seguir os estimadores são calculados a partir de 1000 reamostragens aleatórias das amostras com reposição (*bootstrap*), e suas médias e desvios-padrões são relatados. Em outras palavras, os desvios-padrões relatados aqui são estimados por *bootstrap*, não baseados em equações analíticas.



Chao2

O estimador Chao2 (Chao 1987) é calculado como no EstimateS, versão 8.2.0 (Colwell 2009), com correção de viés:

$$\hat{S}_{Chao2} = S_{obs} + \left(\frac{m-1}{m}\right)\frac{Q_1(Q_1-1)}{2(Q_2+1)}$$

onde S_{obs} é o número total observado de espécies, m é o número de amostras, Q_1 é o número de ocorrências únicas (espécies que ocorrem em precisamente uma amostra) e Q_2 é o número de duplicatas (espécies que ocorrem em precisamente duas amostras).
Se $Q_1>0$ e $Q_2>0$, a variância é estimada por

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

Se $Q_1>0$ mas $Q_2=0$:

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

Se $Q_1=0$:

$$\text{vâr}(\hat{S}_{Chao2}) = S_{obs}e^{-M/S_{obs}}\left(1-e^{-M/S_{obs}}\right),$$

onde M é o número total de ocorrências de todas as espécies em todas as amostras.

Jackknife 1
Jackknife de primeira ordem (Burnham & Overton 1978, 1979; Heltsche & Forrester 1983):

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

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

onde f_j é o número de amostras que contêm j espécies únicas

Jackknife 2
Jackknife de segunda ordem (Smith & van Belle 1984):

$$\hat{S}_{jack2} = S_{obs} + \frac{Q_1(2m-3)}{m} - \frac{Q_2(m-2)^2}{m(m-1)}.$$

Nenhuma estimativa analítica da variância é disponível.

Bootstrap
Estimador por *bootstrap* (Smith & van Belle 1984):

$$\hat{S}_{boot} = S_{obs} + \sum_{k=1}^{S_{obs}} (1 - p_k)^m ,$$

onde p_k é a proporção de amostras contendo k espécies. Nenhuma estimativa analítica da variância é disponível.

Referências

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.

Diversidade beta (Beta diversity)

Duas ou mais linhas (amostras) com dados de presença/ausência (0/1), com os táxons em colunas.
O módulo diversidade beta do Past pode ser usado para qualquer número de amostras (não é limitado a apenas duas amostras). As oito medidas disponíveis são descritas por Koleff et al. (2003):

Past	**Koleff et al.**	**Equação**	**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: número total de espécies; $\overline{\alpha}$: número médio de espécies; *N*: número de amostras; *g(H)*: ganho total de espécies ao longo do gradiente (amostras ordenadas ao longo das colunas); *l(H)*: perda total de espécies; e_i: número de amostras que contêm a espécie *i*; *T*: número total de ocorrências.

Referências

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.

Distinção taxonômica (Taxonomic distinctness)

Uma ou mais colunas, cada uma contendo contagens de indivíduos de diferentes táxons descendo as linhas. Além disso, as linhas da esquerda precisam conter nomes de gêneros/famílias etc (Ver abaixo).
Diversidade taxonômica e distinção taxonômica são definidas por Clarke & Warwick (1998), incluindo intervalos de confiança calculados de 200 réplicas aleatórias retiradas do conjunto de dados agrupado (*pooled dataset*) (todas as colunas). Note que a "lista global" de Clarke & Warwick não é inserida diretamente, mas é calculada internamente pelo agrupamento (somatória) das amostras fornecidas.
Estes índices dependem de informação taxônomica não só a níveis de espécies, mas também acima dele. Esta informação é inserida da seguinte forma: Nomes de espécies vão na coluna de nomes (coluna fixa da extrema esquerda), nomes de gêneros na coluna 1, família na coluna 2, etc (é claro que você pode substituir por outros níveis taxonômicos, contanto que eles estejam em ordem ascendente). Contagem de espécies é colocada nas colunas seguintes. O programa irá perguntar qual é o número de colunas contendo informação taxonômica acima do nível de espécie.
Para dados de presença-ausência, diversidade e distinção taxonômica serão válidas, mas iguais uma à outra.

A distinção taxonômica em uma amostra é dada por (repare que existem outras formas equivalentes):

$$\Delta = \frac{\sum\sum_{i<j} w_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j + \sum_i x_i (x_i - 1)/2},$$

onde w_{ij} são pesos de modo que w_{ij}=0 se *i* e *j* são da mesma espécie, w_{ij}=1 se eles são do mesmo gênero, etc. Os *x* são abundâncias.
Distinção taxonômica:

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

Referência

Clarke, K.R. & Warwick, R.M. 1998. A taxonomic distinctness index and its statistical properties.
Journal of Applied Ecology 35:523-531.

Rarefação individual

Para comparar a diversidade taxonômica entre amostras de diferentes tamanhos. Requer uma ou mais colunas de contagem de indivíduos de diferentes táxons (cada coluna precisa ter o mesmo número de valores). Ao comparar amostras: amostras devem ser taxonomicamente similares, obtidas com amostragem padronizada e amostradas em "hábitats" similares.
Dada uma ou mais colunas de dados de abundância para um número de táxons, este módulo estima quantos táxons você esperaria encontrar em uma amostra com um número total menor de indivíduos. Usando análise de rarefação na sua amostra *maior*, você pode verificar o número de táxons esperados em qualquer amostra de tamanho menor (incluindo o tamanho da sua *menor* amostra). O algoritmo foi retirado de Krebs (1989), usando uma função log Gamma para o cálculo dos termos combinatórios. Um exemplo de aplicação para paleontologia pode ser encontrado em Adrain et al. (2000).
Seja *N* o número total de indivíduos em uma amostra, *s* o número total de espécies, e *Ni* o número de indivíduos da espécie *i*. O número esperado de espécies $E(S_n)$ em uma amostra de tamanho *n* e a sua variância $V(S_n)$ são dadas por

$$E(S_n) = \sum_{i=1}^{s} \left[1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right]$$

$$V(S_n) = \sum_{i=1}^{s} \left[\frac{\binom{N-N_i}{n}}{\binom{N}{n}} \left(1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right) \right]$$

$$+ 2\sum_{j=2}^{s} \sum_{i=1}^{j-1} \left[\frac{\binom{N-N_i-N_j}{n}}{\binom{N}{n}} - \frac{\binom{N-N_i}{n}\binom{N-N_j}{n}}{\binom{N}{n}\binom{N}{n}} \right]$$

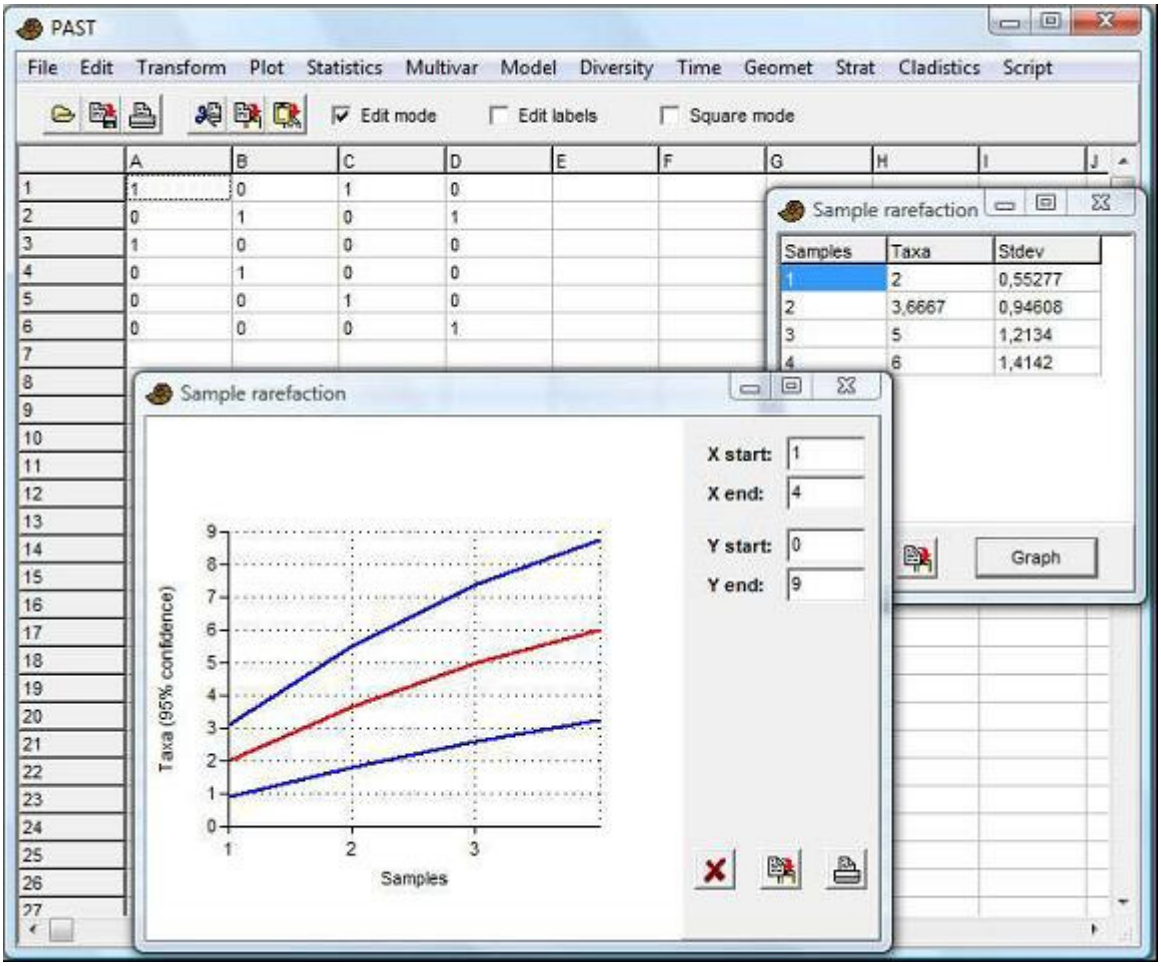
Erros padrões (raízes quadradas das variâncias) são fornecidos pelo programa. No gráfico, estes erros padrões são convertidos em intervalos de confiança de 95%.

Referências

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.

Rarefação por amostra (Sample rarefaction) (Mao tau)

A rarefação por amostra requer uma matriz de dados de presença-ausência (abundâncias tratadas como presenças), com táxons em colunas e amostras em linhas. Rarefação baseada em amostras (também conhecida como curva de acumulação de espécies) é aplicável quando uma certa quantidade de amostras é disponível, a partir das quais a riqueza de espécies é estimada como função do número de amostras. PAST implementa a solução analítica conhecida por "Mao tau', com desvio padrão. No gráfico, os erros padrões são convertidos em intervalos de confiança de 95%.



Ver Colwell et al. (2004) para detalhes.

Com H amostras e S_{obs} o número total de espécies observadas, sejam s_j o número de espécies encontradas em j amostras, de modo que s_1 é o número de espécies encontrado em exatamente uma amostra, etc. O número total de espécies esperadas em $h \leq H$ amostras é então

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

Os coeficientes combinatoriais α são

$$\alpha_{jh} = \begin{cases} \dfrac{(H-h)!(H-j)!}{(H-h-j)!H!} & para \quad j+h \leq H \\ 0 & para \quad j+h > H \end{cases}$$

Estes coeficientes são calculados por meio de uma função log Gamma. O estimador da variância é

$$\tilde{\sigma}^2 = \sum_{j=1}^{H} (1-\alpha_{jh})^2 s_j - \frac{\tilde{\tau}^2(h)}{\tilde{S}}$$

onde $\tilde{S}$ é um estimador para a riqueza total (desconhecida) de espécies. Seguindo Colwell et al. (2004), um estimador do tipo Chao2 é usado. Para $s_2 > 0$,

$$\tilde{S} = S_{obs} + \frac{(H-1)s_1^2}{2Hs_2}.$$

Para $s_2 = 0$,

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

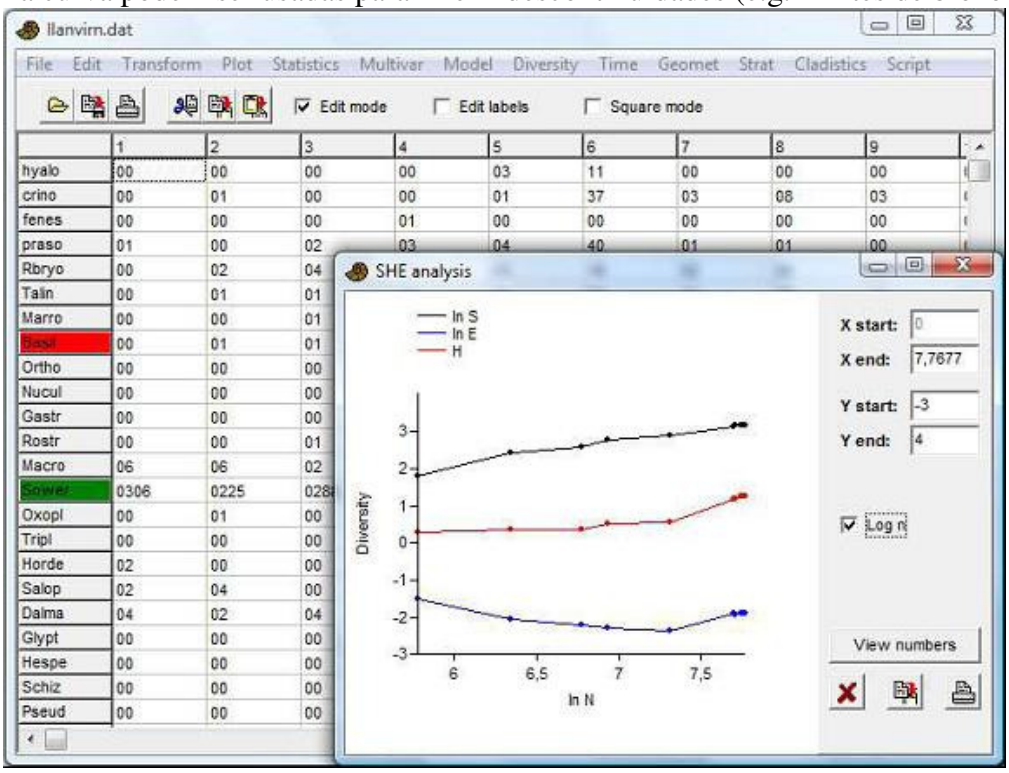
Para modelar e extrapolar a curva usando uma equação de Michaelis-Mentem, use o botão Copiar Dados (*Copy Data*), cole numa nova planilha do Past, e use o módulo para encaixe de funções (*fitting module*) no menu *Model* (Modelar).

Referência

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

Análise SHE (SHE analysis)

A análise SHE (Hayek & Buzas 1997, Buzas & Hayek 1998) requer uma matriz de dados inteiros de abundância (contagens), com táxons em linhas e amostras em colunas. O programa calcula: log abundância de espécies (ln *S*), índices de Shannon (*H*) e log equitabilidade (evenness) (ln *E* = *H* – ln *S*) para a primeira amostra. Então a segunda amostra é adicionada à primeira e o processo continua. Os perfis cumulativos de SHE resultantes podem ser intereprétados ecologicamente. Se as amostras são retiradas não de uma população homogênea, mas de um gradiente ou de uma seção estratigráfica, quebras na curva podem ser usadas para inferir descontinuidades (e.g. limites de biozonas).



Referências

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.

Comparar diversidades (Compare diversities)

Espera duas colunas de dados de abundância, com táxons descendo as linhas. Este módulo calcula um número de índices de diversidade para duas amostras e então compara as diversidades por meio de dois procedimentos diferentes de aleatorização, como segue.

Bootstrap

As duas amostras A e B são agrupadas. 1000 pares aleatórios de amostras (A, B) são então retirados deste grupo, com o nesmo número de indivíduos que nas duas amostras originais. Para cada par replicado, são calculados os índices de diverisdade $div(A_i)$ e $div(B_i)$. O número de vezes que $|div(A_i) – div(B_i)|$ é maior ou igual que $|div(A) – div(B)|$ indica a probabilidade que a diferença observada possa ter ocorrido por amostragem aleatória de uma população parental (*parent population*) como estimada pela amostra agrupada.

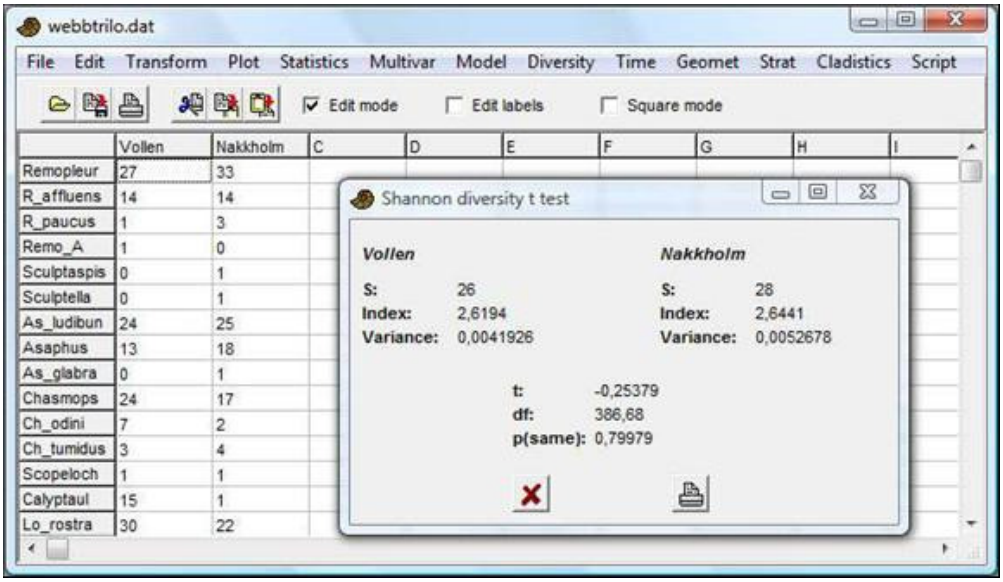
Então, um pequeno valor de probabilidade *p(same)* (*p(igual)* ou *p(mesmo)*) indica uma diferença significativa no índice de diversidade entre as duas colunas.

Permutação

São geradas 1000 matrizes aleatórias com duas colunas (amostras), cada uma com o mesmo número de linhas e total de colunas que na matriz de dados original. O valor de *p* (*p value*) é calculado como no teste por bootstrap.

Teste t de diversidade (Diversity t test)

Comparação dos índices de diversidade de Shannon de duas amostras, por meio de um teste *t* descrito, e.g., por Hutcheson (1970), Poole (1974), Magurran (1988). Esse teste é uma alternativa ao teste por aleatorização disponível no módulo *Comparar diversidadades (Compare diversities)*. Requer duas colunas de dados de abundância com táxons descendo as linhas.



O índice de Shannon aqui inclui uma correção de viés e pode diferir levemente das estimativas não corrigidas calculadas em outros módulos do PAST, ao menos para amostras pequenas. Com p_i a proporção (0-1) do táxon i, S o número de táxons e N o número de indivíduos, o estimador do índice é

$H' = -\sum_{i=1}^{S} p_i \ln p_i - \frac{S-1}{2N}$ (note que o segundo termo está incorreto em Magurran 1988).

O estimador da variância é

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

A estatística t é dada por

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

Os graus de liberdade para o teste t são

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

Referências

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.

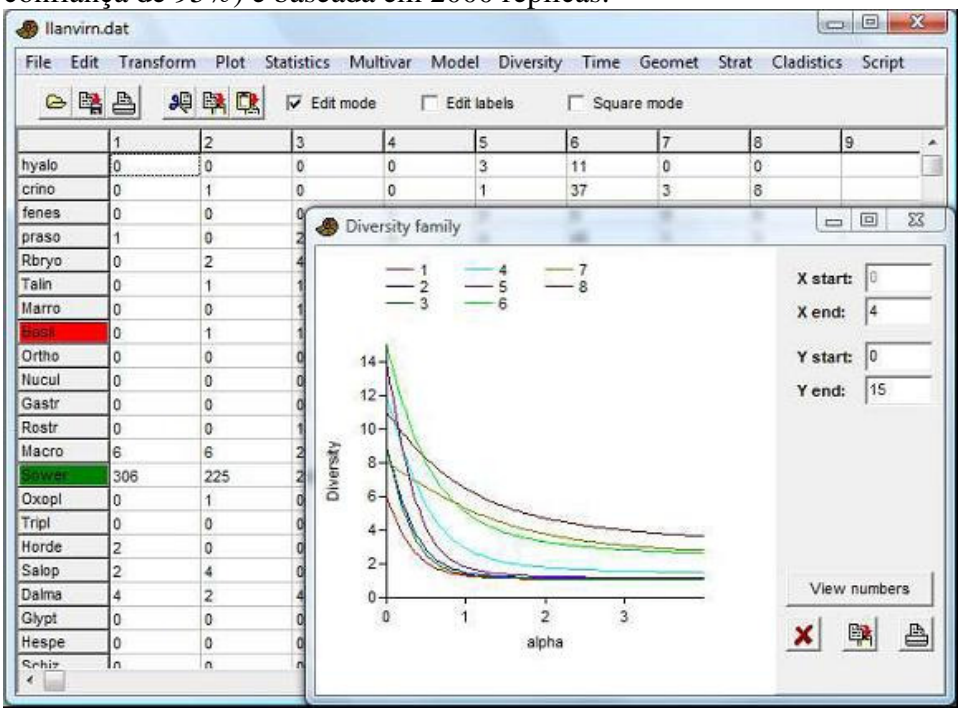
Perfis de diversidade (Diversity profiles)

Este módulo requer uma ou mais colunas de dados de abundância com táxons descendo as linhas. O principal objetivo é comparar a diversidade em uma série de amostras.
A validade de comparar diversidades entre amostras pode ser criticada por causa da escolha arbitrária do índice de diversidade. Uma amostra pode, por exemplo, conter um grande número de táxons, emquanto outra pode ter um índice de Shannon elevado. Uma série de índices de diversidade podem ser comparados para certificar que a ordem da diversidade é robusta. Um jeito formal de fazer isso é definir uma família de índices de diversidade que dependem de um único parâmetro (Tothmeresz 1995).
O PAST usa a exponencial do chamado índice de Renyi, a qual de um parâmetro α. Para $\alpha = 0$ esta função dá o número total de espécies. $\alpha = 1$ (no limite) dá um índice

proporcional ao índice de Shannon, enquanto α=2 dá um índice que se comporta como o índice de Simpson.

$$\exp(H_\alpha) = \exp\left(\frac{1}{1-\alpha}\ln\sum_{i=1}^{S} p_i^\alpha\right)$$

O programa pode plotar uma série de perfis ao mesmo tempo. Se os perfis se cruzam, as diversidades não são comparáveis. A opção de *bootstrap* (fornecendo um intervalo de confiança de 95%) é baseada em 2000 réplicas.



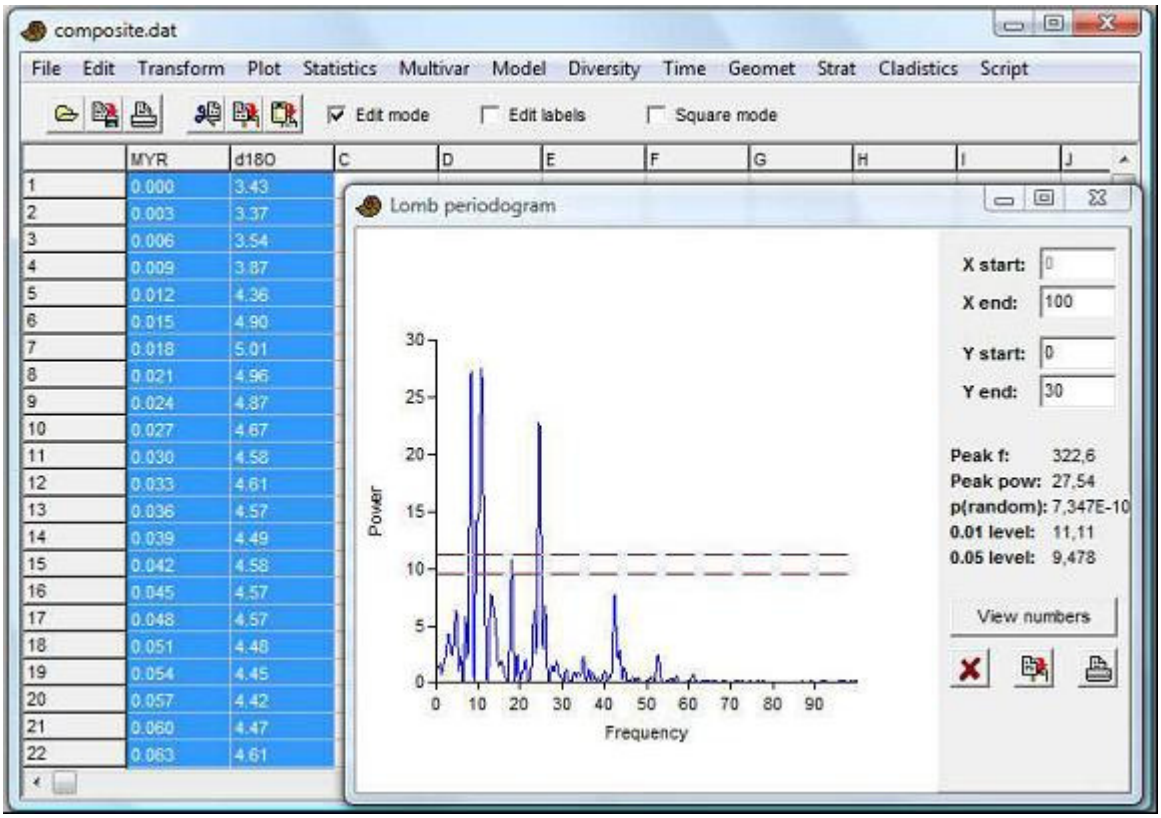
Referência

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

Time series menu (Séries temporais)

Análise espectral (Spectral analysis)

Como dados paleontológicos são frequentemente amostrados de forma desigual *(unevenly sampled)*, métodos baseados em Fourier podem ser difíceis de usar. Por isso o PAST utiliza o algoritmo do periodograma de Lomb para dados amostrados de forma desigual (Press et al. 1992), com valores do tempo dados na primeira coluna e os valores dependentes na segunda coluna. Se apenas uma coluna é selecionada, assume-se um espaçamento igual de uma unidade entre os pontos de dados. O periodograma de Lomb deve então dar resultados similares ao FFT. Os dados são automaticamente destendenciados antes da análise



O eixo da frequência é em unidades de *1/(unidade do x)*. Se, por exemplo, seus valores de *x* estão em milhões de anos, uma frequência de 0.1 corresponde a um período de 10 milhões de anos. O eixo de potência (*power axis*) é em unidades proporcionais ao quadrado das amplitudes das sinusóides presentes nos dados. Note também que o eixo da frequência se estende a valores muito altos. Se seus dados foram amostrados regularmente *(evenly sampled)*, a parte superior do espectro é uma imagem-espelho da metade superior e é de pouca serventia. Se algumas regiões são amostradas de forma menos espaçada *(closely sampled)*, o algoritmo pode ser capaz de encontrar informação útil até mesmo acima do ponto médio (frequência de Nyquist).

O pico mais alto do espectro é apresentado com a sua frequência e seu valor de potência (*power value*), juntamente com a probabilidade de que o pico poderia ocorrer de dados

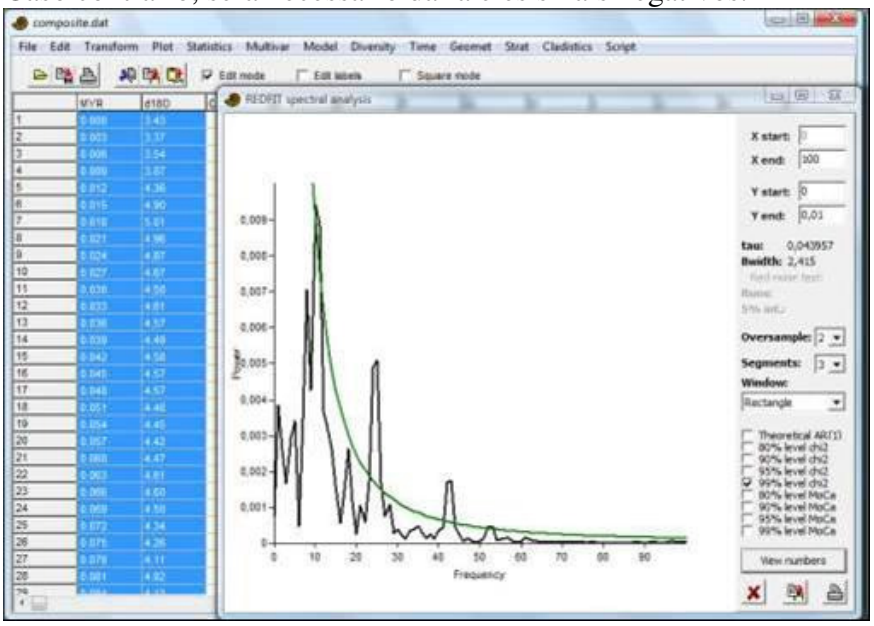
aleatórios. Os níveis de significância de 0.01 e 0.05 ("linhas de barulho branco" – "white noise lines") são mostradas como linhadas tracejadas vermelhas.
O exemplo acima mostra uma análise espectral de um isótopo de oxigênico forâmico (*foram oxygen isotope*) de 1 Ma até Recente, com um espaçamento regular de 0.003 Ma (3 ka). Há periodicidades em frequência de por volta de 9 (pico dividido – *split peak*), 25 e 43 Ma^{-1}, correspondentes a períodos de 111 ka, 40 ka e 23 ka – com claro forçamento orbital (*clearly orbital forcing*).

Referência

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

Análise espectral REDFIT (REDFIT spectral analysis)

Este módulo é uma implementação do procedimento REDFIT de Schulz & Mudelsee (2002). Uma versão mais avançada do periodograma simples de Lomb descrito acima. REDFIT incluir uma opção para "médias de segmentos sobrepostos de Welch" (*Welch overlapped segment averaging*), a qual implica em dividir a série temporal em um número de segmentos com 50% de sobreposição e usar a média dos seus espectros. Isso reduz o barulho (*noise*), mas também reduz a resolução espectral. Adicionalmente, a série temporal é encaixada a um de barulho vermelho AR(1) (*AR(1) red noise model*), o qual é normalmente uma hipótese nula mais apropriada do que o modelo de barulho branco (*white noise model*) descrito acima. As "linhas de falso alarme" (*"false-alarm lines"*) fornecidas são baseadas tanto em aproximações paramétricas (qui-quadrado) quanto em Monte Carlo (usando 1000 realizações aleatórias de um processo AR(1)).
Os dados devem ser inseridos na forma de duas colunas com valores de tempos e de dados, ou uma coluna com valores de dados igualmente espaçados. Os dados são destendenciados automaticamente. O encaixe do modelo AR(1) implica que os dados devem a direção temporal correta (em contraste ao espectrograma simples onde a direção temporal é arbitrária). Espera-se que os valores do tempo sejam eras antes do presente. Caso contrário, será necessário dar a eles sinais negativos.



O valor da superamostragem de frequência (*frequency oversampling value*) contra o número de pontos ao longo do eixo da frequência (mas ter mais pontos não aumenta a resolução da frequência!). Aumentando o número de segmentos, vai reduzir o barulho, mas também reduzirá a resolução. A função de janela (*window function*) influencia o *trade-off* entre resolução espectral e atenuação dos lobos laterais (*attenuation of side lobes*).
O valor (médio) do tau é a escala temporal característica (o parâmetro do modelo AR). A largura de banda (*bandwidth*) é a resolução espectral, dada como a largura entre os -6dB pontos.
O encaixe a um modelo AR(1) pode ser verificado pelo valor de corridas (*runs value*) e seu intervalo de aceitação de 5%. Este teste é disponível apenas com o Monte Carlo ligado, superamostragem (*oversampling*) = 1, segmentos = 1, janela (*window*) = retangular (*rectangular*). Em adição a um conjunto fixo de níveis de falso alarme (90%, 90%, 95% e 99%), o programa também fornece o nível "crítico" de falso alarme (False-al) que depende do comprimento do segmento (Thomson 1990).
Importante: por causa do longo tempo de cálculo, a simulação Monte Carlo não é executada automaticamente, e os níveis de falso-alarme por Monte Carlo, portanto, não são disponíveis. Quando a opção Monte Carlo é ativada, o espectro fornecido pode mudar levemente porque os resultados do Monte Carlo são, então, usados para calcular uma versão com viés corrigido (*"bias-corrected*") (veja Schulz e Mudelsee 2002).

Referências

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.

Análise espectral de afunilamento múltiplo (Multitaper spectral analysis)

Na análise espectral tradicional, os dados são frequentemente "janelados" (*"windowed*") (multiplicados por uma função em forma de sino) para reduzir vazamento espectral (*spectral leakage*). No método de afunilamento múltiplo, algumas funções de janela diferentes (ortogonais) são aplicadas e os resultados são combinados. O espectro resultante tem baixo vazamento, baixa variância, e retém a informação contida no começo e no fim da série temporal. Adicionalmente, testes estatísticos podem ser favorecidos pelas múltiplas estimativas espectrais. Uma possível desvantagem é a resolução espectral reduzida.
O método de afunilamento múltiplo requer dados espaçados regularmente, fornecidos em uma coluna.
A implementação no Past é baseada no código de Lees & Park (1995). O espectro de afunilamento múltiplo pode ser comparado com um periodograma simples (FFT com uma janela coseno de 10%) e um periodograma suavizado (*smoothed*). O número de afunilamentos (*tapers*) (NWIN) pode ser ajustado em 3, 4 ou 5, para diferentes balanços (*tradeoffs*) entre resolução e redução da variância. O "produto tempo-largura de banda" ("*time-bandwidth product"*) *p* é fixado em 3.0.
O teste F para singnificância da periodicidade segue Lees & Park (1995). Os níveis de significância 0.01 e 0.05 são mostrados como linhas horizontais, baseadas em 2 e 2*NWIN-2 graus de liberdade.

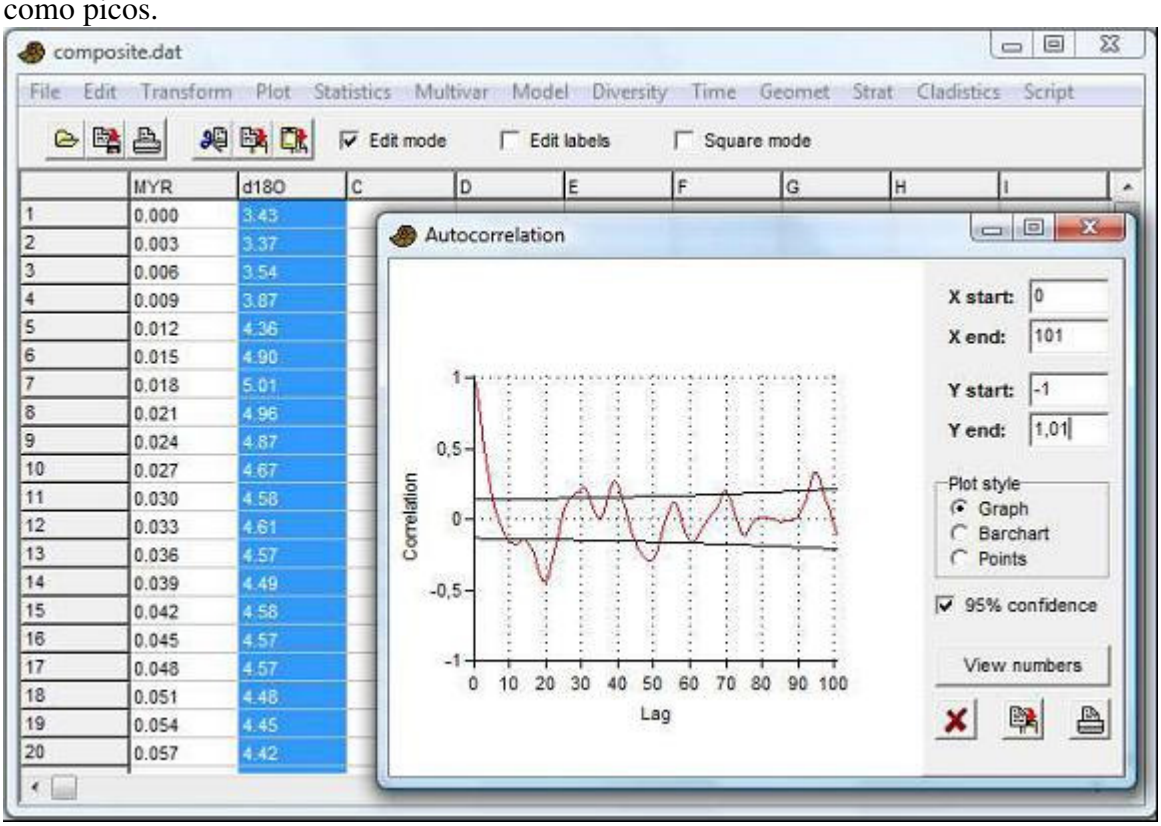
Os dados são zero-acolchoados (*zero-padded*) até a menor potência de 2 que seja maior que o comprimento da sequência. Isto é necessário para produzir os resultados de teste fornecidos por Lees & Park (1995).

Referência

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

Autocorrelação (Autocorrelation)

A autocorrelação (Davis 1986) é feita em duas colunas de dados temporais/estratigráficos *amostrados regularmente*. Tempo de atraso (*lag times*) τ de até *n/2*, onde *n* é o número total de valores no vetor, são mostrados ao longo do eixo x (apenas tempos de atraso positivos – a função de autocorrelação é simétrica em torno de zero). Uma autocorrelação predominantemente igual a zero significa dados aleatórios – periodicidades aparecem como picos.



A opção "intervalo de confiança 95%" (*"95 percent confidence interval"*) desenhará linhas em

$$\pm 1.76\sqrt{\frac{1}{n-\tau+3}}$$

segundo Davis (1986). Este é o intervalo de confiança para pontos aleatórios e independentes (barulho branco). Há duas considerações: Barulho branco é um modelo não realístico, e o intervalo de confiança só é rigorosamente válido em cada atraso *individual* (problema dos testes múltiplos).

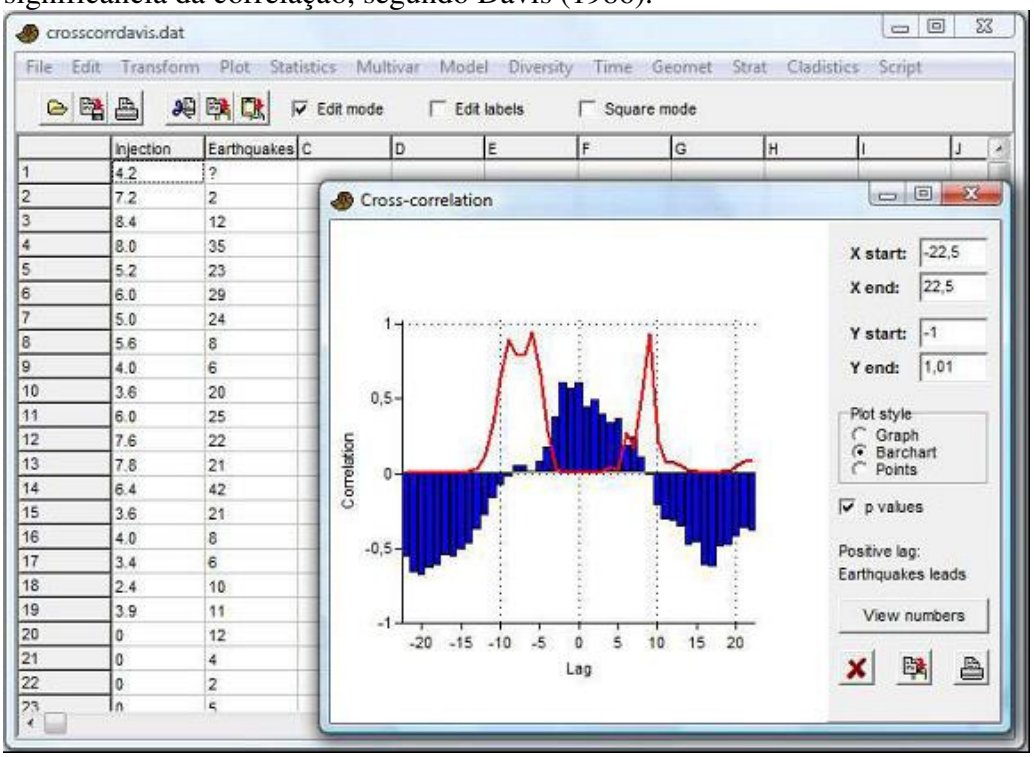
Há suporte para dados ausentes.

Referência

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

Correlação cruzada (Cross-correlation)

A correlação cruzada (Davis 1986) é feita em duas colunas de dados temporais/estratigráficos *amostrados regularmente*. O eixo x mostra o deslocamento da segunda coluna em relação à primeira, o eixo y é a correlação entre as duas séries temporais para um dado deslocamento. A opção "p valores' (*"p values"*) desenhará a significância da correlação, segundo Davis (1986).



Para duas séries temporais **x** e **y**, o valor da correlação cruzada em um tempo de atraso (*lag time*) *m* é

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

As somatórias e os valores médios são tomados apenas nas partes das sequências que se sobrepõem para um dado tempo de atraso.

A equação mostra que para atrasos positivos, **x** é comparado com um **y** que foi atrasado em *m* amostras. Uma alta correlação em atrasos positivos então significa que características de **y** estão guiando, enquanto **x** fica para trás. Um lembrete disso é dado pelo programa.

Um valor de *p* para um dado *m* é dado por um teste *t* com *n-2* graus de liberdade, sendo *n* o número de amostras que se sobrepõem:

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

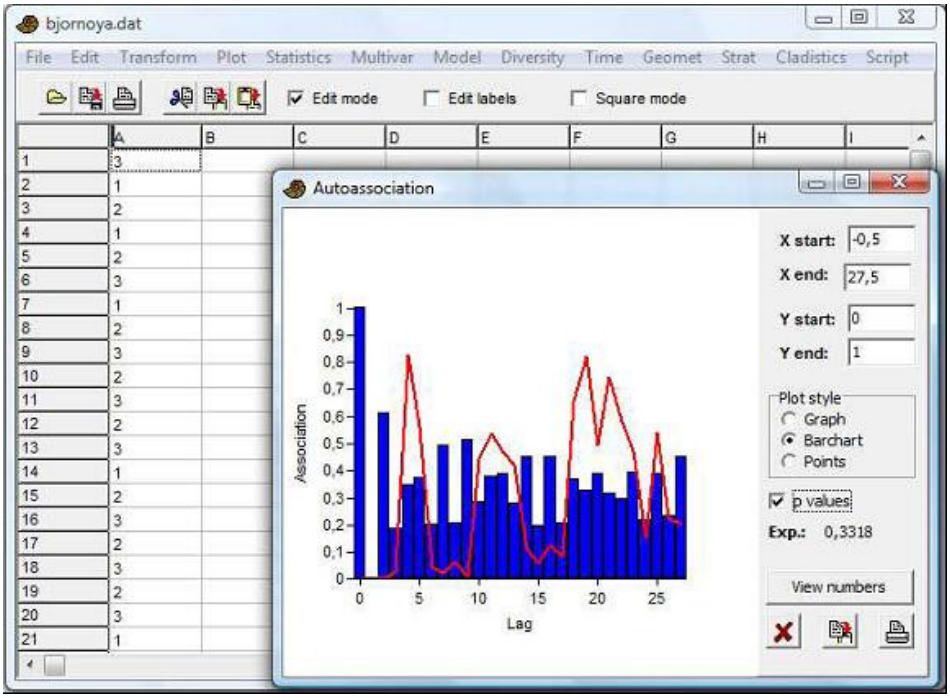
É importante notar que este teste diz respeito *a um m em particular*. Plotar *p* em função de todos os *m* traz a questão de testes múltiplos – valores de *p* menores que 0.05 são esperados para 5% dos tempos de atraso mesmo em conjuntos de dados totalmente aleatórios (não correlacionados).
No exemplo acima, os dados de "terremotos" (*"earthquakes"*) parecem se atrasar em relação aos dados de "injeção" (*"injection"*) com um atraso de 0-2 amostras (neste caso, meses), onde os valores da correlação são maiores. Os valores de *p* (curva vermelha) indicam a significância nestes atrasos. Curiosamente, parece haver significância para a correlação negativa em atrasos positivos e negativos grandes.
Há suporte para dados ausentes.

Referência

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

Autoassociação (Autoassociation)

A autoassociação é análoga à autocorrelação, mas para uma sequência de dados binários ou nominais codificados como números inteiros



Para cada atraso (*lag*), o valor da autoassociação é simplesmente a razão entre o número de posições de mesmo valor (*matching position*) e o número total de posições que foram comparadas. O valor de autoassociação esperado (0.3318 no exemplo acima) para uma sequência aleatória é (Davis 1986)

$$P = \frac{\sum_{k=1}^{m} X_k^2 - n}{n^2 - n}$$

onde *n* é o número total de posições, *m* é o número de estados distintos (3 no exemplo acima), e X_k é o número de observações com o estado *k*.
Para valores de atraso diferentes de zero, um valor de *P* é computado apenas pelas posições com sobreposição, e o número esperado de correspondências é dado por *E=nP*.

Isso é comparado ao número observado de correspondência *O* para produzir um χ^2 com 1 grau de liberdade:

$$\chi^2 = \frac{(O - E - 1/2)^2}{E} + \frac{(O' - E' - 1/2)^2}{E'}$$

com *O'=n-O* e *E'=n(1-P)* os valores observados e esperados de não-correspondências (*mismatches*).

A questão de testes múltiplos surge para o conjunto de valores *p*.

O teste acima não é rigorosamente válido para sequências de "transição" nas quais repetições não são permitidas (a sequência no exemplo acima é desse tipo). Neste caso, selecionar a opção "sem repetições" (*"No repetitions"*). Os valos de *p* serão então computados por um teste exato, onde todas as possíveis permutações sem repetição são computadas e a autoassociação é comparada com os valores originais. Este teste demora muito tempo para rodar para *n>30*, e a opção não está disponível para *n>40*.

Há suporte para dados ausentes.

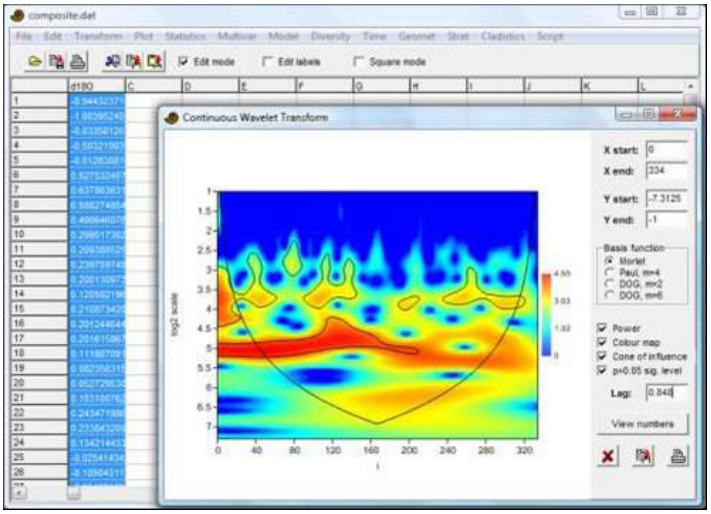
Referência

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

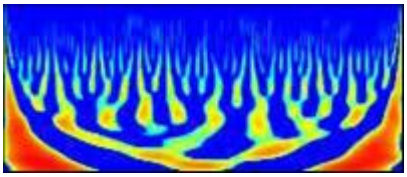
Wavelet (Wavelet transform)

Inspeção de séries temporais em diferentes escalas. Requer uma coluna de dados ordinais ou contínuos com espaçamento regular entre os pontos.

A transformação *wavelet* contínua (*continuous wavelet transform – CWT*) é uma forma de análise em que os dados podem ser inspecionados simultaneamente em escalas pequena, intermediária e grande. Pode ser útil para detectar periodicidades em diferentes comprimentos de onde, auto-similaridade (*self-similarity*) e outras características. O eixo vertical no gráfico é o logaritmo (base 2) da escala de tamanho, com o sinal observado a uma escala de apenas dois pontos consecutivos no topo e a uma escala de um quarto de toda a sequência na base. O topo da figura assim representa uma visão detalhada, de granulação fina (*fine-grained view*), enquanto a base representa uma visão geral suave de tendências mais longas. O poder do sinal (ou, mais corretamente, o quadrado da força de correlação com o *wavelet* gerador (*mother wavelet*) daquela escala) é mostrada como uma escala de cinza ou em cor.



A forma do *wavelet* gerador pode ser estabelecida para Morlet (número de *wavelet* (*wavelet number*) = 6), Paul (4ª ordem) ou DOG (Derivado Do Gaussiano (*Derivative Of Gaussian*), 2ª e 4ª derivadas). O *wavelet* de Morlet normalmente tem o melhor desempenho.
O exemplo acima é baseada em um registro de isótopos de oxigênio de foraminíferos (*foram oxygen isotope*) de 1 Ma até Recente, com um espaçamento regular de 0.003 Ma (3 ka). Uma faixa pode ser vista a uma escala de aproximadamente 2^5 = 32 amostras, ou por volta de 100 ka. Uma faixa mais fraca por volta de $2^{3.7}$=13 amostras corresponde a uma escala de 40 ka. Isto são periodicidades orbitais (*orbital periodicities*). Em contraste com a análise espectral "geral", o escalograma torna visíveis as mudanças de força e frequência ao longo do tempo.
O assim chamando "cone de influência" (*"cone of influence*") pode ser plotado para mostrar a região em que efeitos de fronteira (*boundary effects*) estão presentes.
O algoritmo é baseado na convolução rápida do sinal com o *wavelet* em diferentes escalas, usando o FFT.
Teste de significância: O nível de significância correspondente a p=0.05 pode ser plotado como um contorno (teste qui-quadrado, de acordo com Torrence & Compo 1998). O valor de "atraso" (*"Lag"*), como fornecido pelo usuário, especifica a hipótese nula. Atraso=0 especifica um modelo de barulho branco. Valores de 0 < Atraso < 1 especificam um modelo de barulho vermelho com o dado coeficiente MA(1) de autocorrelação. Este pode ser estimado usando o módulo ARMA no menu *Time* (especificar zero termos AR (*AR terms*) e um termo MA (*MA term*), note que os valores de MA são dados com sinal negativo).
Se a função "Potência" (*"Power"*) for desmarcada, o programa irá mostrar apenas a parte real do escalograma (sem elevar ao quadrado). Isso mostra o sinal no domínio tempo, filtrado em diferentes escalas:



Na janela "Ver números" (*"View numbers"*), cada linha mostra uma escala, com o número da amostra (posição) ao longo das colunas.
A transformação *wavelet* foi usada por Prokoph et al. (2000) para ilustrar ciclos em curvas de diversidade em foraminíferos planctônicos. O código no Past é baseado em Torrence & Compo (1998).

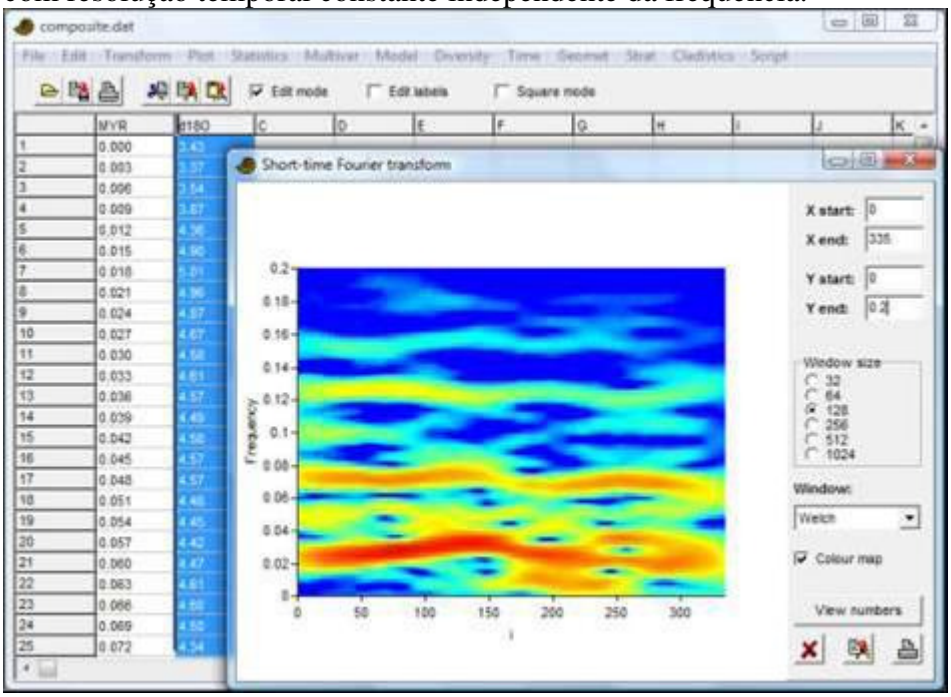
Referências

Prokoph, A., A.D. Fowler & R.T. Patterson. 2000. Evidence for periodicity and nonlinearity in a highresolution 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.

Transformação de Fourier de tempos curtos (Short-time Fourier transform)

Análise espectral usa a transformação de Fourier (*Fourier transform* – FFT), mas divide o sinal em uma sequência de janelas que se sobrepõe, que são analisadas

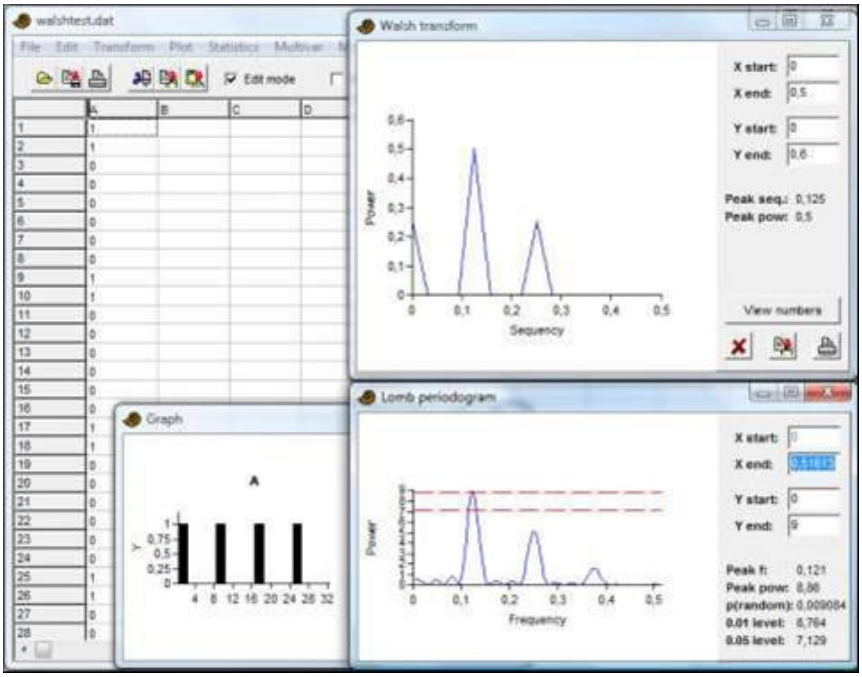
individualmente. Isso permite desenvolvimento do espectro no tempo, contrastando com a análise global fornecida por outros módulos de análise espectral. Posição da amostra é mostrada no eixo *x*, frequência (em períodos por amostra) no eixo *y*, e poder em uma escala logarítmica por uma escala de cor ou escala-de-cinza.
A Transformação de Fourier de Tempos curtos (*Short-time Fourier transform – STFT*) pode ser comparada com análise de *wavelet*, mas com uma escala linear de frequência e com resolução temporal constante independente da frequência.



O tamanho de janela (*window size*) controla a troca entre resolução em tempo e frequência; janelas pequenas dão boa resolução para tempo mais baixa resolução para frequência. Janelas são acolchoadas em zero (*zero-padded*) por um fator de oito para dar uma aparência mais suave do diagrama ao longo do eixo de frequências. As funções de janela (*window functions*) (*Rectangle, Welch, Hanning, Blackman-Harris*, afunilamento múltiplo (*multiple taper*) com 3, 4 ou 5 *tapers*) dão diferentes trade-offs entre resolução de frequência e rejeição de faixas laterais (*sideband rejection*).

Transformação de Walsh (Walsh transform)

A transformação de Walsh é um tipo de análise espectral (para encontrar periodicidades) de dados binários ou ordinais. Assume espaçamento uniforme entre os pontos de dados e espera uma coluna de dados binários (0/1) ou ordinais (inteiros).



Os métodos comuns de análise espectral talvez não sejam ótimos para dados binários, já que eles decompõem as séries temporais em sinusóides, e não em "ondas quadradas". A transformação de Walsh pode então ser uma escolha melhor, usando como base funções que se alternam entre -1 e +1. Estas funções têm "frequências" variáveis (número de transições dividido por dois), conhecidas como *sequências*. No PAST, cada par de funções básicas pares ("cal") e ímpares ("sal") é combinado em uma potência usando cal^2+sal^2, produzindo um "espectro de potências" que é comparável com o periodograma de Lomb.

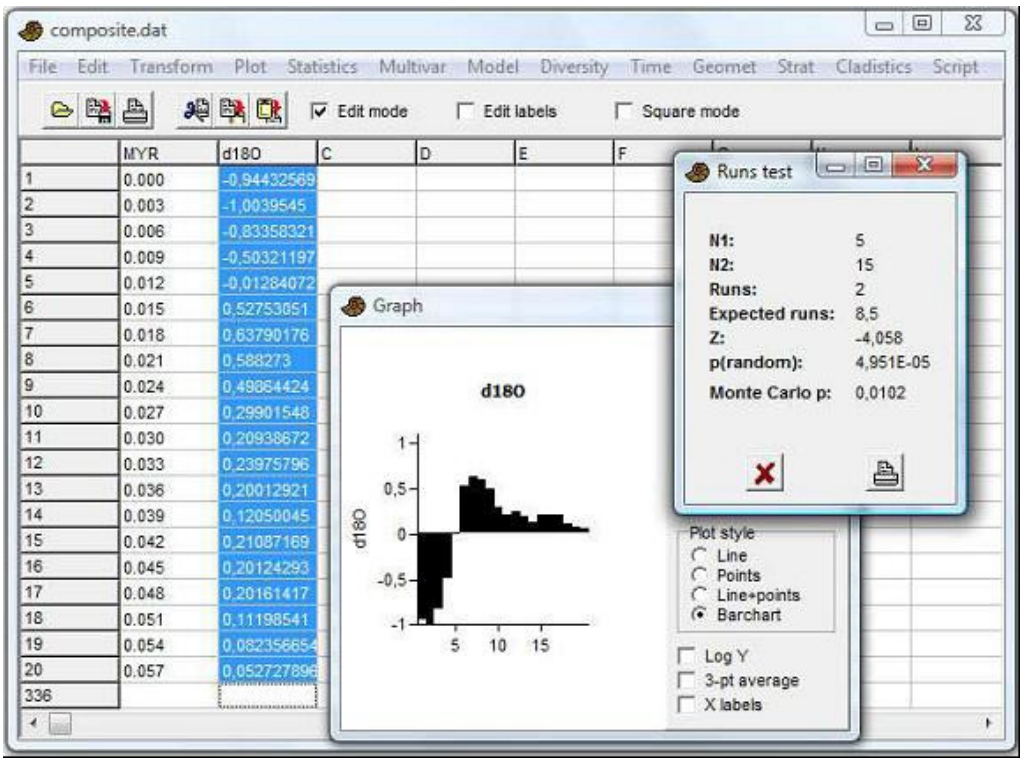
No exemplo acima, compare o periodograma de Walsh (topo) ao periodograma de Lomb (embaixo). O conjunto de dados tem 0.125 períodos por amostra. Ambas as análises mostram harmônicos.

A transformação de Walsh é ligeiramente exótica comparada com a transformação de Fourier, e os resultados devem ser interpretados com cautela. Por exemplo, os efeitos do *duty cycle* (porcentagem de 1s contra a porcentagem de zeros) são um tanto difíceis de entender.

No PAST, os valores de dados são pré-processados multiplicado por dois e subtraindo 1, trazendo os valores binários 0/1 para dentro da amplitude -1/+1, ótima para a transformação de Walsh. Os dados são zero-acolchoados (*zero-padded*) à potência mais próxima de 2 se necessário, como é requerido pelo método.

Runs test ("teste de séries")

O *runs test* ("teste de séries") é um teste não-paramétrico para aleatoriedade em uma sequência de valores como um série temporal. Não-aleatoriedade pode incluir efeitos como autocorrelação, tendência e periodicidade. O módulo requer uma coluna de dados, que são convertidos internamente para 0 ($x \leq 0$) ou 1 ($x > 0$).



O teste é baseado na dicotomia entre dois valores ($x \leq 0$ ou $x > 0$). Ele conta o número de séries (*runs*) (grupos de valores consecutivos iguais) e compara este número a um valor teórico. O *runs test* pode portando ser usado diretamente em sequências de dados binários. Também há opções por "series em torno da média" (*"runs about the mean"*) (o valor médio é subtraído dos dados antes do teste), e "séries para cima e para baixo" (*"runs up and down"*) (são tomadas as diferenças entre um valor e o próximo antes do teste).

Sendo n o número total de pontos de dados, n_1 o número de pontos ≤ 0 e n_2 o número de pontos >0, o número esperado de séries em uma sequência aleatória e a variância são

$$E(R) = \frac{n + 2n_1 n_2}{n},$$

$$Var(R) = \frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}.$$

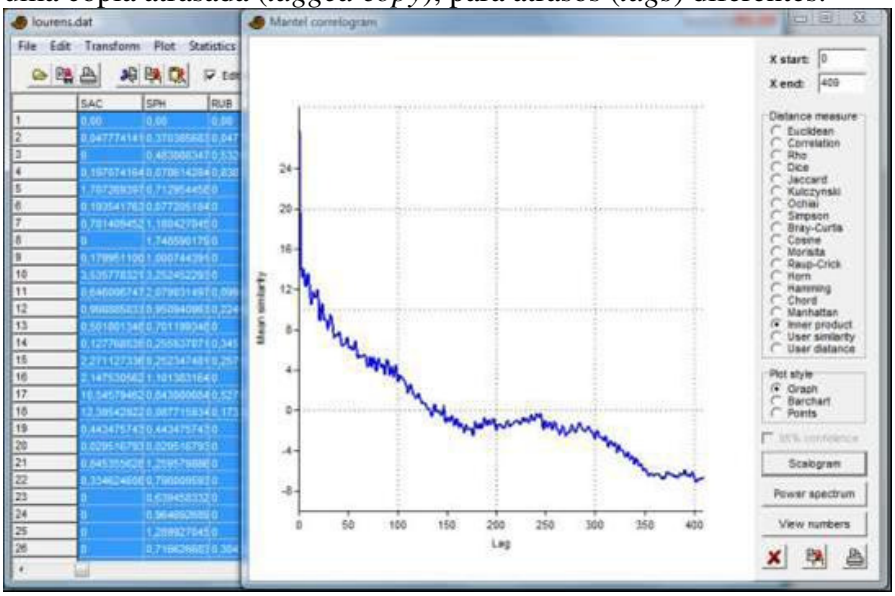
Sendo R o número observado de séries, uma estatística z pode ser escrita como

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

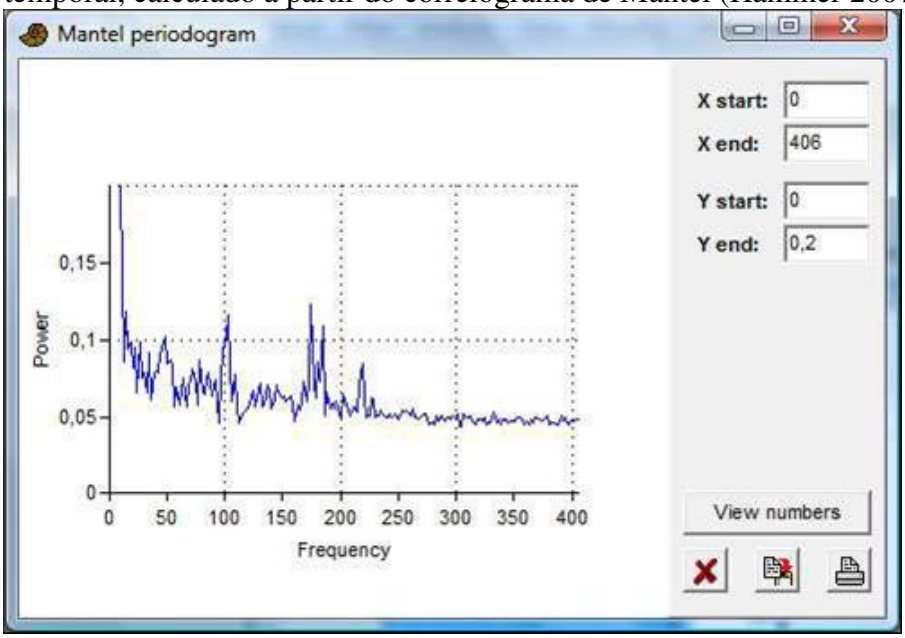
O valor de p bicaudal resultante não é preciso para n<20. Sendo assim, também é incluído um procedimento Monte Carlo baseado em 10 000 réplicas aleatórias usando n, n_1 e n_2.

Correlograma (e periodograma) de Mantel (Mantel correlogram (and periodogram)

Este módulo espera uma série de linhas de dados multivariados, uma linha por amostra. Assume-se que as amostras estejam distribuídas regularmente no tempo.
O correlograma de Mantel (e.g. Legendre & Legendre 1988) é uma extensão multivariada da autocorrelação e é baseado em qualquer medida de similaridade ou distância. O correlograma de Mantel no PAST mostra a similaridade média entre a série temporal e uma cópia atrasada (*lagged copy*), para atrasos (*lags*) diferentes.

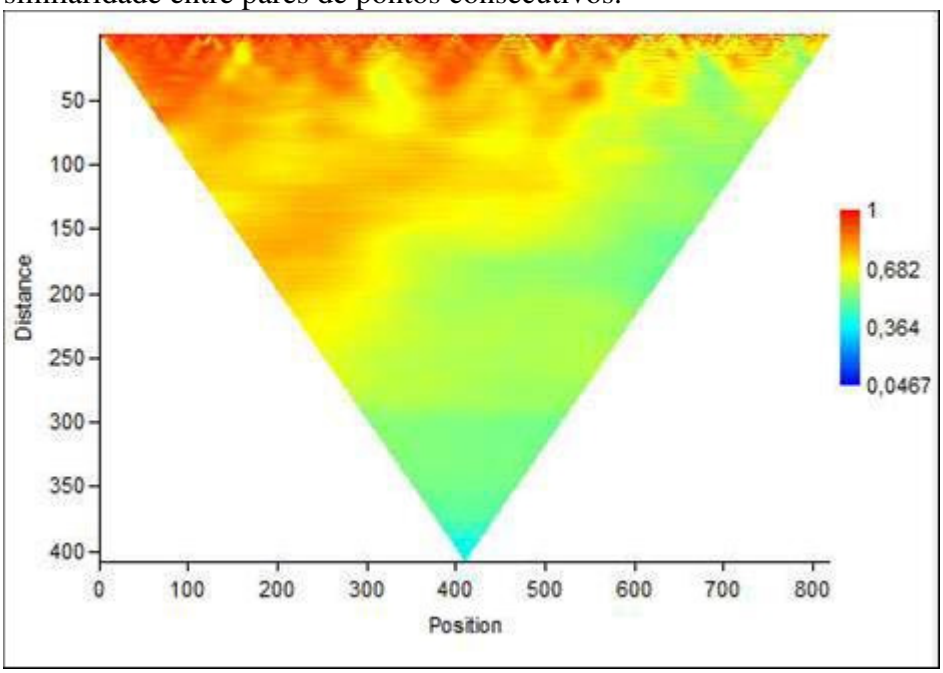


O periodograma de Mantel é um espectro das potências (*power spectrum*) da série temporal, calculado a partir do correlograma de Mantel (Hammer 2007).



O escalograma de Mantel (*Mantel scalogram*) é um gráfico experimental das similaridades entre todos os pares de pontos ao longo da série temporal. O ápice do

triângulo é a similaridade entre o primeiro e o último ponto. A base do triângulo mostra similaridade entre pares de pontos consecutivos.



Referências

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 (e análise de intervenção) (ARMA (and intervention analysis))

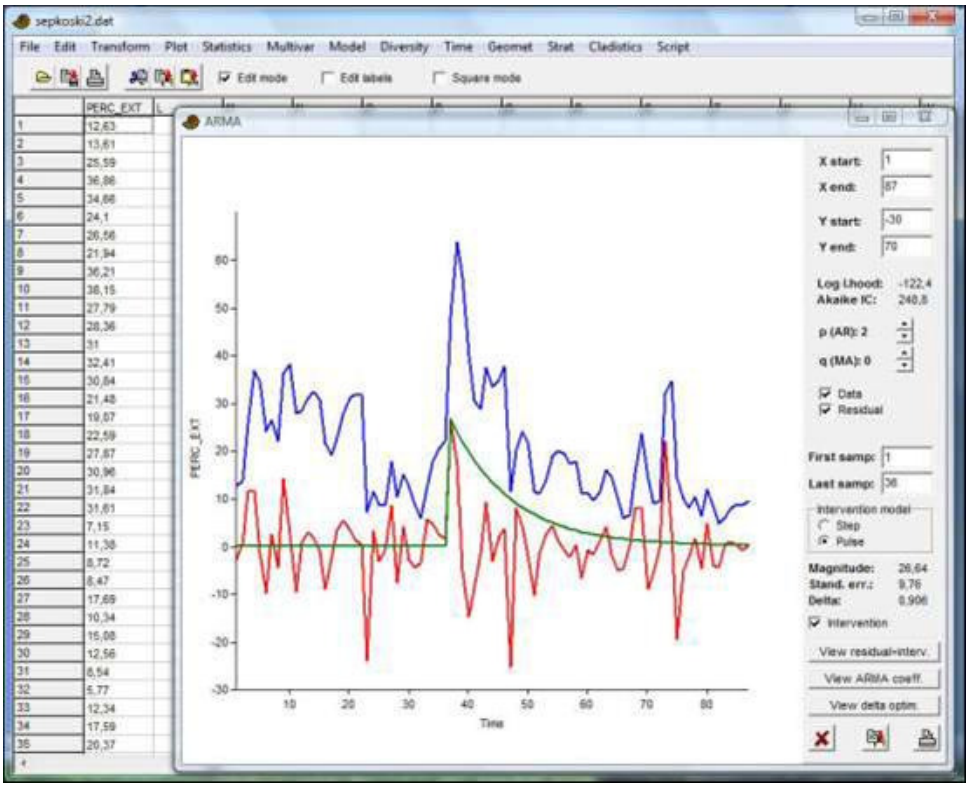
Análise e remoção de correlações seriais (*serial correlations*) na série temporal, e análise do impacto de um distúrbio externo ("intervenção") em um dado ponto no tempo. Séries temporais estacionárias, exceto para uma única intervenção. Uma coluna de dados com espaçamento regular.

Este módulo poderoso, mas um tanto complicado, implementa análise ARMA de máxima verossimilhança e uma versão mínima da análise de intervenção de Box-Jenkins (e.g. para investigar como uma mudança climática pode afetar a biodiversidade).

Por padrão, uma análise ARMA simples sem intervenção é calculada. O usuário seleciona o número de termos AR (auto-regressivos (*autoregressive*)) e MA (média móvel (*movem average*)) que serão incluídos na equação de diferença do ARMA. A log-verossimilhança (*log-likelihood*) e o Critério de Informação de Akaike (*Akaike Information Criterium – AIC*) são fornecidos. Selecione o número de termos que minimiza o critério de Akaike, mas leve em conta que os termos AR são mais "poderosos" do que os termos MA. Dois termos AR podem modelar uma periodicidade, por exemplo.

O principal objetivo da análise ARMA é remover correlações seriais, que caso contrário causariam problemas para ajuste de modelos e estatística. O resíduo deve ser inspecionado para sinais de autocorrelação, por exemplo copiado o resíduo da janela de saída numérica de volta à planilha e usando o módulo de autocorrelação. Repare que para

muitos conjuntos de dados paleontológicos com dados esparsos efeitos que confundem, uma análise ARMA adequada (e, portanto, análise de intervenção) será impossível.
O programa é baseado no algoritmo de verissimilhança de Melard (1984), combinado com otimização multivariada não-linear usando busca por simplex (*nonlinear multivariate optimization using simplex search*).
A análise de intervenção prossegue assim: Primeiro, faça uma análise ARMA apenas nas amostras que precedem a intervenção. Para isso, digite o número da última amostra pré-intervenção na caixa "última amostra" (*"last samp"*). Também é possível fazer a análise ARMA apenas nas amostras que se seguem à intervenção, ao digitar a primeira amostra pós-intervenção na casa "primeira amostra" (*"first samp"*), mas isso não é recomendado por causa do distúrbio pós-intervenção. Também selecione a caixa "Intervenção" (*"Intervention"*) para ver o modelo de intervenção otimizado.
A análise segue Box e Tiao (1975) ao assumir uma "função indicadora" (*"indicator function"*) u(i) que é ou um passo unitário (*unit step*) ou um pulso unitário (*unit pulse*), como escolhido pelo usuário. A função indicadora é transformada por um processo AR(1) com um parâmetro delta e então escalonada (*scaled*) por um magnitude (note que a magnitude dada no PAST é o coeficiente na função indicadora transformada: primeiro faça y(i)=delta*y(i-1)+u(i), então reajuste a escala de y pela magnitude). O algoritmo é baseado na transformação ARMA da sequência completa, então uma transformação ARMA correspondente de y, e finalmente regressão linear para encontrar a magnitude. O parâmetro delta é otimizado por busca exaustiva entre [0,1].
Para impactos pequenos em dados com ruído, o delta pode parar em um sub-ótimo. Tente as opções tanto de passo (*step*) quanto de pulso (*pulse*) e veja qual dá o menor erro padrão na magnitude. Também inspecione os dados de "otimização do delta" (*"delta optimization"*), onde o erro padrão da estimativa é plotado como função de delta, para ver se o valor otimizado pode ser instável.
O modelo de Box-Jenkins pode modelar mudanças abruptas e permanentes (função passo (*step*) com delta=0, ou pulso com delta=1), abruptas e não-permanentes (pulso com delta<1), ou graduais e permanentes (passo com delta<0).
Tome cuidado com o erro padrão da magnitude – ele frequentemente será subestimado, especialmente se o modelo ARMA não se ajusta bem. Por esta razão, um valor de *p* deliberadamente não é calculado (Murtaugh 2002).



O conjunto de dados do exemplo (curva azul) é a curva de Sepkoski para a taxa de extinção percentual em nível de gênero, interpolada para produzir um espaçamento regular de ca. 5 milhões de anos. O pico maior é a extinção no limite entre o Permiano e o Triássico. O usuário especificou um modelo ARMA(2,0). O resíduo é plotado em vermelho. O usuário especificou que os parâmetro do ARMA devem ser calculados para os pontos antes da extinção P-T no tempo 37 e uma intervenção do tipo pulso (*pulse-type intervention*). A análise parece indicar uma constante temporal (delta) elevada para a intervenção, com o efeito durando até o Jurássico.

Referências

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.

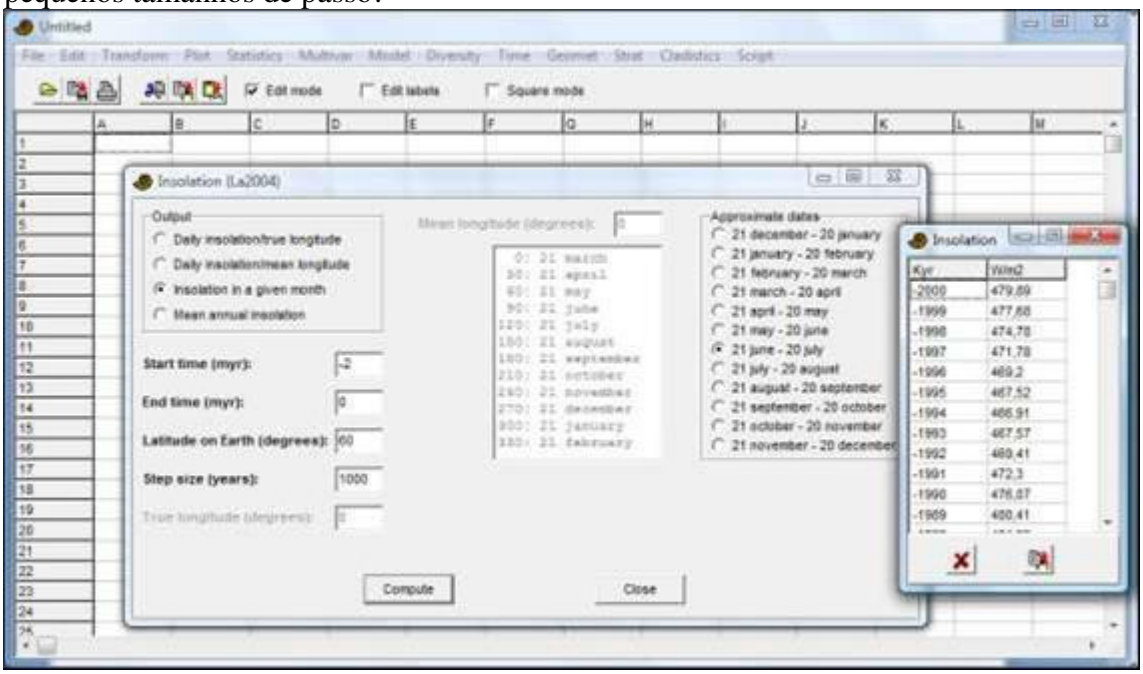
Murtaugh, P.A. 2002. On rejection rates of paired intervention analysis. *Ecology* 83:1752-1761.

Modelo de insolação (forçamento solar) (Insolation (solar forcing) model)

Este módulo calcula a insolação solar em qualquer latitude e em qualquer tempo de 100 Ma até o Recente (os resultados são menos precisos antes de 50 Ma). O cálculo pode ser feito para uma longitude orbital "verdadeira", longitude orbital "média" (correspondente a uma certa data do ano), com a média de um certo mês em cada ano, ou integrada para um ano inteiro.

A implementação no PAST é portada do código de Laskar et al. (2004), por cortesia deste autores. Por favor, cite Laskar et al. (2004) em qualquer publicação.

É necessário especificar um arquivo de dados contendo parâmetros orbitais. Baixe o arquivo http://www.imcce.fr/Equipes/ASD/insola/earth/La2004 e o coloque em qualquer lugar no seu computador. O PAST irá perguntar a localização do arquivo na primeira vez que você fizer o cálculo.
A quantidade de dados pode se tornar excessiva para períodos longos de tempo e pequenos tamanhos de passo!

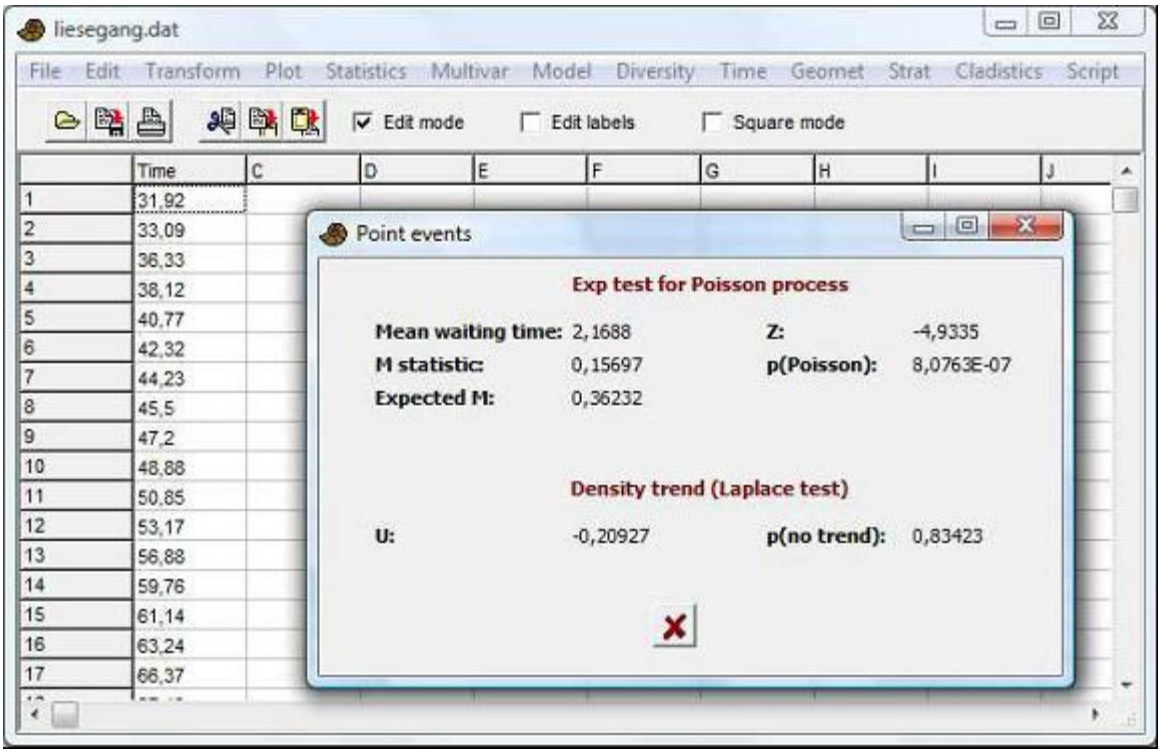


Referência

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.

Eventos pontuais (Point events)

Espera uma coluna contendo tempos de eventos (e.g. terremotos ou divergências de clado) ou posições ao longo de uma linha (e.g. transecto). Os tempos não precisam estar em ordem crescente.



Teste exp para processo de Poisson

O teste exp (Prahl 1999) para um processo estacionário de Poisson (eventos aleatórios e independentes) é baseado no conjunto de n tempos de espera Δt_i entre eventos sucessivos na sequência ordenada. A estatística de teste é:

$$M = \frac{1}{n} \sum_{\Delta t_i < T} \left(1 - \frac{\Delta t_i}{T} \right)$$

onde T é o tempo de espera médio. M irá tender a zero para uma sequência espaçada regularmente (superdispersa – *overdispersed*) e a 1 para uma sequência altamente agrupada. Para a hipótese nula de um processo de Poisson, M tem distribuição assimptoticamente normal com média $1/e - \alpha/n$ e um desvio padrão $\beta/\sqrt{n}$, onde α=0.189 e β=0.2427. Esta é a base para o teste z fornecido.

Resumindo, se p<0.05 a sequência não é Poisson. Você pode então inspecionar a estatística M; se ela for menor do que o valor esperado, isso indica regularidade, se for maior, indica agrupamento.

Tendência de densidade (teste de Laplace)

O teste "de Laplace" para uma tendência na densidade (intensidade) é descrito por Cox & Lewis (1978). Ele é baseado na estatística de teste

$$U = \frac{\bar{t} - \frac{L}{2}}{L\sqrt{\frac{1}{12n}}}$$

onde $\bar{t}$ é o tempo médio de evento, n é o número de eventos e L é o comprimento do intervalo. L é estimado como o tempo do primeiro evento ao último, mais o tempo médio

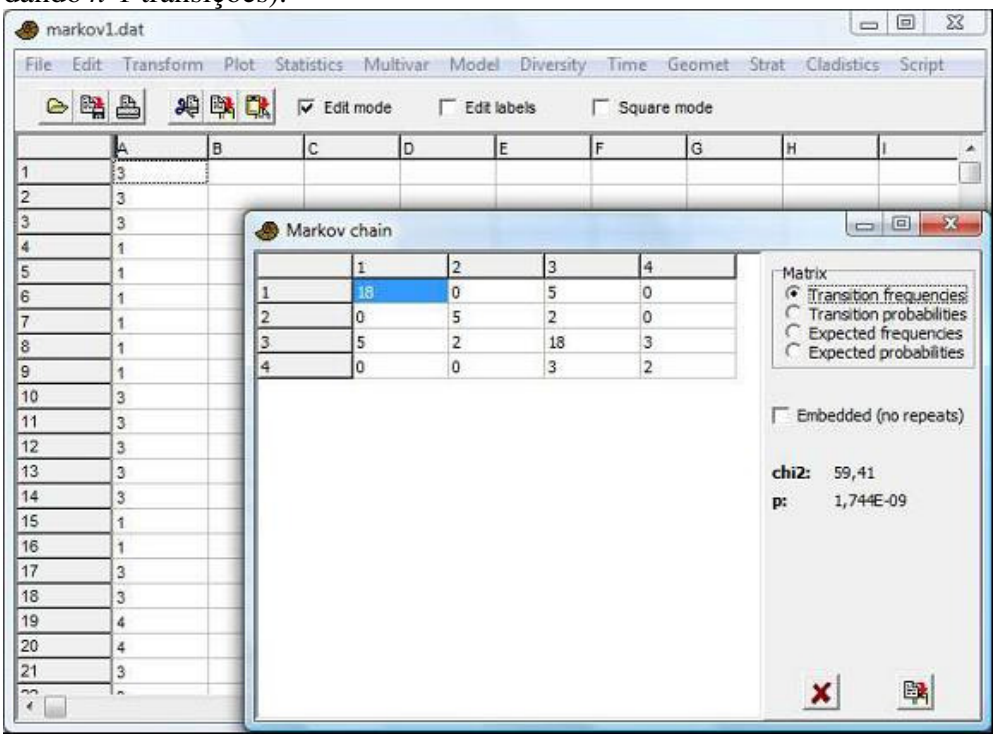
de espera. Na hipótese nula de intensidade constante, *U* tem distribuição aproximadamente normal com média zero e variância um. Esta é a base para o valor de *p* que é fornecido.
Se $p<0.05$, um *U* positivo indica uma tendência de aumento na densidade (redução nos tempos de espera), enquanto um *U* negativo indica uma tendência decrescente. Repare que se uma tendência é detectada por este teste, a sequência não é estacionária e as premissas do teste exp acima são violadas.

Referências

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.

Cadeia de Markov (Markov chain)

Este módulo requer uma única de coluna contendo uma sequência de dados nominais codificados como números. Por exemplo, uma sequência estratigráfica onde 1 significa calcário, 2 significa xisto e 3 significa areia. Uma matriz de transição contendo contagens ou proporções (probabilidades) de transições de estado é mostrada. Os estados originais ("de") estão nas linhas e os estados finais ("para") estão nas colunas.
Também é possível especificar mais de uma coluna, cada uma contendo uma ou mais transições de estado (dois números para uma transição, *n* números para uma sequência dando *n*-1 transições).



O teste de qui-quadrado relata a probabilidade de que os dados foram tomados de um sistema com proporções aleatórias de transições (i.e. sem transições preferenciais). As transições com frequências anômalas podem ser identificadas comparando as matrizes de transição observada e esperada.

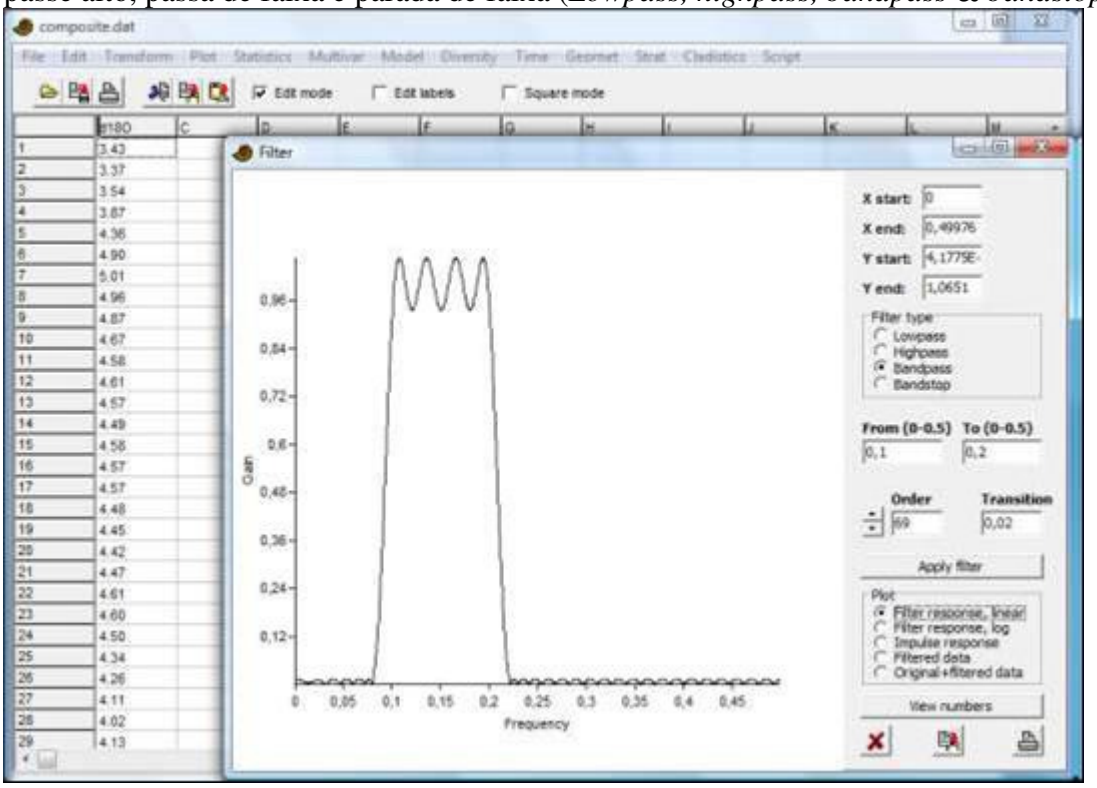
A opção “Incorporada (sem repetições)” (*“Embedded (no repeats)”*) deve ser selecionada se os dados foram coletados de tal modo que transições para o mesmo estado não são possíveis (pontos de dados só são coletados quando há uma mudança). A matriz de transição então terá zeros na diagonal.
Os algoritmos, incluindo um algoritmo iterativo para cadeias de Markov incorporadas, seguem Davis (1986).

Referência

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

Filtrar (Filter)

Filtrar os dados, de modo a deixar certas faixas de frequência de fora, pode ser útil, em análise de séries temporais, para suavizar (*smooth*) uma curva, remover variação lenta ou enfatizar certas periodicidades (e.g ciclos de Milaknovitch). Espera uma coluna de dados com espaçamento regular. O Past usa filtros FIR, que foram desenhados usando o algoritmo de Parks-McClellan. Os seguintes tipos de filtro são disponíveis: Passe baixo, passe alto, passa de faixa e parada de faixa (*Lowpass, highpass, bandpass & bandstop*).



Parâmetros do filtro

Algum esforço é necessário para delinear o melhor filtro. As frequências são especificadas na faixa de 0-0.5, i.e. T_0/T onde T_0 é o intervalo de amostragem (não especificado para o computador) e T é o período requerido. Por exemplo, se o seu intervalo de amostragem é de 1000 anos, uma frequência correspondente a um período de 23000 anos é especificada como 1000/23000=0,043.

Depois de definir o tipo de filtro, você deve escolher uma largura de transição (*transition*) (ou deixar o valor padrão de 0,02). Reduzir a largura da transição vai produzir um filtro mais preciso (*sharper*) ao custo de ondulações maiores ("ondas" na resposta da frequência).
Repare que os valores nos campos do texto não são atualizados até que você pressione Enter. Além disso, se uma combinação inválida for colocada (e.g. faixa de transição cruzando 0 ou 0.5, ou limite superior menor que o limite inferior) o programa irá reiniciar alguns valores para evitar erros. Portanto, é necessário inserir os números em uma ordem para que o filtro sempre seja válido.
Os tipos de filtro são os seguintes:

1. **Passe baixo (*lowpass*):** A frequência *De (From)* é forçada a zero. As frequências até a frequência *Até (Up)* passam pelo filtro. As frequências de *Até+Transição* até 0.5 são bloqueadas.
2. **Passe alto (*highpass*):** A frequência *Para* é forçada para 0.5. Frequências acima da frequência *De* passam pelo filtro. Frequências de 0 até *De-Transição* são bloqueadas.
3. **Passe de faixa (*bandpass*):** Frequências de *De* até *Até* passam pelo filtro. Frequências abaixo de *De-Transição* e acima de *Até+Transição* são bloqueadas.
4. **Parada de faixa (*bandstop*):** Frequências de *De* até *Até* são bloqueadas. Frequências de 0 até *De-Transição* e de *Até+Transição* até 0.5 passam pelo filtro.

Ordem do filtro (*Filter order*)
A ordem do filtro deve ser grande o suficiente para dar um filtro aceitavelmente preciso com poucas ondulações. No entanto, um filtro de ordem *n* vai dar resultados menos acurados nas *n/2* primeiras e últimas amostras da série temporal, o que coloca um limite prático na ordem do filtro para séries pequenas.
O algoritmo de Parks-McClennan nem sempre irá convergir. Isso dá uma resposta de frequência obviamente incorreta, e uma tentativa de aplicar este filtro aos dados dá uma mensagem de aviso. Tente mudar a ordem do filtro (normalmente aumentando-a) para resolver este problema.

Suavizadores simples (Simple smoothers)

Um conjunto de suavizadores simples para uma única coluna de dados espaçados regularmente.
Há suporte para dados ausentes.
Média móvel *(Moving average)*
Uma média móvel simples, centrada, de *n* pontos (*n* deve ser ímpar). Seu uso é comum, mas tem propriedades indesejáveis como uma resposta de frequência (*frequency response*) não-monotônica.
Gaussiana *(Gaussian)*
Média móvel ponderada usando um Kernel Gaussiano com desvio padrão de ¼ do tamanho da janela (de *n* pontos). Este é provavelmente, de modo geral, o melhor método do módulo.
Mediana móvel *(Moving median)*

Similar à média móvel, mas usa a mediana ao invés da média. Este método é mais robusto em relação a valores extremos (*outliers*).

AR1 (Exponencial) ***(AR1 (Exponential))***

Filtro recursivo (autoregressivo), $y_i = \alpha y_{i-1} + (1-\alpha)x_i$ com α sendo um coeficiente de alisamento de 0 até 1. Isso corresponde ao cálculo de médias ponderadas com pesos que decaem exponencialmente. Dá um atraso de fase e também um transitório (*transient*) no começo da série. Incluído para deixar o módulo mais completo.

Conversão de data/tempo (Date/time conversion)

Ferramenta para converter datas e/ou tempos em uma variável contínua para análise. O algoritmo espera uma ou duas colunas, cada uma contendo datas ou tempos. Se ambas são fornecidas, o tempo é adicionado à data para dar o valor final do tempo.

Datas podem ser fornecidas no formato Ano/Mês/Dia ou Dia/Mês/Ano. Anos precisam de todos os dígitos (um ano inserido como 11 significa 11 d.C., não 2011). Há suporte apenas para datas do calendário Gregoriano. Anos bissextos são levados em conta.

Tempo pode ser fornecido como Horas:Minutos ou como Horas:Minutos:Segundos (segundos podem incluir decimais).

A unidade de saída pode ser anos (usando o calendário médio Gregoriano de 365.2425 dias), dias (de 86400 segundos), horas, minutos ou segundos.

O tempo inicial (tempo zero) pode ser o menor tempo fornecido, o começo do primeiro dia, o começo do primeiro ano, ano 0 (repare a convenção "astronômica" onde o ano antes do ano 1 é ano 0), ou o começo do primeiro dia Juliano (meio-dia, ano -4712).

O programa opera com tempo simples (UT), definido em relação à rotação da Terra e com um número fixo de segundos por dia (86400).

Se os dados de entrada consistem de valores separados por espaço, como "2011/12/24 18:00:00.00", você pode ter que usar a função "Importar arquivo de texto" (*"Import text file"*) para ler os dados de modo que as datas e os tempos sejam separados em colunas distintas.

O cálculo do dia Juliano (usado para encontrar o número de dias entre duas datas) segue Meeus (1991):

se *mês* <= 2 começar *ano* := *ano*-1; *mês* := *mês* +12; fim
A = base[4](*ano*/100);
$B = 2 - A$ + base(A/4);
JD = base (365.25(*ano*+4716)) + base(30.6001(*mês*+1) + *dia* + B – 1524.5;

Referência

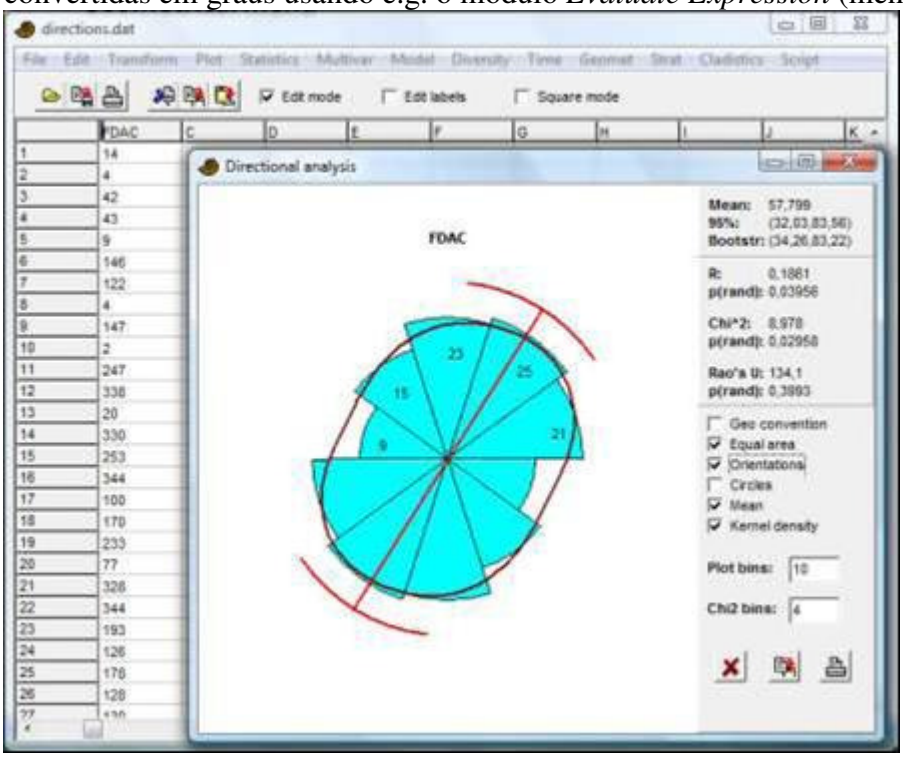
Meeus, J. 1991. *Astronomical algorithms*. Willmann-Bell, Richmond.

[4] Traduzi *floor* como *base*.

Geometrical menu

Direções – uma amostra (Directions – one sample)

Este módulo plota um diagrama de rosa (*rose diagram*), ou histograma polar, de direções. Usado para plotar espécimens orientados a correntes, orientação de caminhos, orientação de elementos morfológicos (e.g. linhas de terraceamento), etc.
Uma coluna de dados direcionais (0-360) ou orientacionais (0-180), em graus, é esperada. Dados direcionais ou periódicos em outras formas (radianos, 0-24 horas, etc) devem ser convertidas em graus usando e.g. o módulo *Evaluate Expression* (menu *Transform*).



Por padrão, a convenção "matemática" de ângulos anti-horários a partir do leste é escolhida. Caso você queria usar a convenção "geográfica" de ângulos em sentido horário a partir do norte, marque a caixa *Geo. convention*.
Você também pode escolher se terá abundâncias proporcionais ao raio do diagrama de rosa ou proporcionais à área (*equal area*).
A opção "Densidade Kernel" (*"Kernel density"*) plota uma estimativa circular da densidade por Kernel.

Estatística descritiva

O ângulo médio (*mean angle*) leva em conta a circularidade:

$$\bar{\theta} = \tan^{-1} \frac{\sum \sin \theta_i}{\sum \cos \theta_i} \text{ (levado ao quadrante certo)}$$

O intervalo de confiança de 95% da média é estimado de acordo com Fisher (1983). Ele assume distribuição normal circular, e não é muito preciso para variâncias muito grandes (intervalo de confiança maior do que 45 graus) ou tamanhos amostrais pequenos. O

intervalo de confiança de 95% das médias por *bootstrap* utiliza 5000 réplicas de *bootstrap*. O gráfico usa o intervalo de confiança por *bootstrap*.
O parâmetro de concentração κ é estimado por aproximação iterativa à solução da equação

$$\frac{I_1(\kappa)}{I_0(\kappa)} = \overline{R}$$

onde I_0 e I_1 são funções imaginárias de Bessel de ordens 0 e 1, estimadas de acordo com Press et al. (1992), e o R definido abaixo (ver e.g. Mardia 1972).

Teste de Rayleigh para distribuição uniforme
O valor de *R* (comprimento médio resultante – *mean resultant length*) é dado por

$$\overline{R} = \sqrt{\left(\sum_{i=1}^{n} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n} \sin\theta_i\right)^2} \Big/ n .$$

O *R* é então testado em relação a uma distribuição aleatória por meio do teste de Rayleigh para dados direcionais (Davis 1986). Repare que este procedimento assume dados distribuídos de forma uniforme ou unimodal (von Mises) – o teste não é apropriado para, por exemplo, dados bimodais. Os valores de *p* são calculados usando uma aproximação dada por Mardia (1972):

$$K = n\overline{R}^2$$

$$p = e^{-K}\left(1 + \frac{2K - K^2}{rn} - \frac{24K - 132K^2 + 76K^3 - 9K^4}{288n^2}\right)$$

Teste de espaçamento de Rao (*Rao's spacing test*) para distribuição uniforme
O teste de espaçamento de Rao (Batschelet 1981) para distribuição uniforme tem a estatística de teste

$$U = \frac{1}{2}\sum_{i=1}^{n} |T_i - \lambda| ,$$

onde λ=360°/*n*. $T_i = \theta_{i+1} - \theta_i$ para $i < n, T_n = 360^\circ - \theta_n + \theta_1$. Esse teste é não-paramétrico, e não assume, e.g., distribuição de vone Mises. O valor de *p* é estimado por interpolação linear a partir das tabelas de probabilidade publicadas por Russel & Levitin (1995).
Um teste de qui-quadrado para distribuição uniforme também é disponível, com o número de grupos definido pelo usuário (igual a 4 por padrão).

Teste U^2 de Watson para qualidade-de-ajuste (*goodness-of-fit*) da distribuição de von Mises
Seja *f* a distribuição de von Mises para os parâmetros estimados de ângulo médio e concentração:

$$f(\theta; \overline{\theta}, \kappa) = \frac{e^{\phi\cos(\theta - \overline{\theta})}}{2\pi I_0(\kappa)} .$$

A estatística do teste (e.g. Lockhart & Stevens 1985) é

$$U^2 = \sum\left(z_i - \frac{2i-1}{2n}\right)^2 - n\left(\overline{z} - \frac{1}{2}\right)^2 + \frac{1}{12n}$$

onde

$$z_i = \int_0^{\bar{\theta}_i} f(\theta;\bar{\theta},\kappa)d\theta ,$$

estimado por intergração numérica. Valores críticos para a estatística de teste são obtidos por interpolação linear da Tabela 1 de Lockhart & Stevens (1985). São aceitavelmente precisos para n>=20.

Dados axiais (*Axial data*)

A opção *"Orientations"* ("Orientações") permite análise de orientações lineares (axiais) (0-180 graus). Os testes de Rayleigh e Watson são então feitos sobre os ângulos dobrados (o truqe é descrito por Davis 1986); o teste de qui-quadrado usa quatro grupos de 0 a 180 graus; os diagramas de rosa espelham o histograma ao redor da origem.

Referências

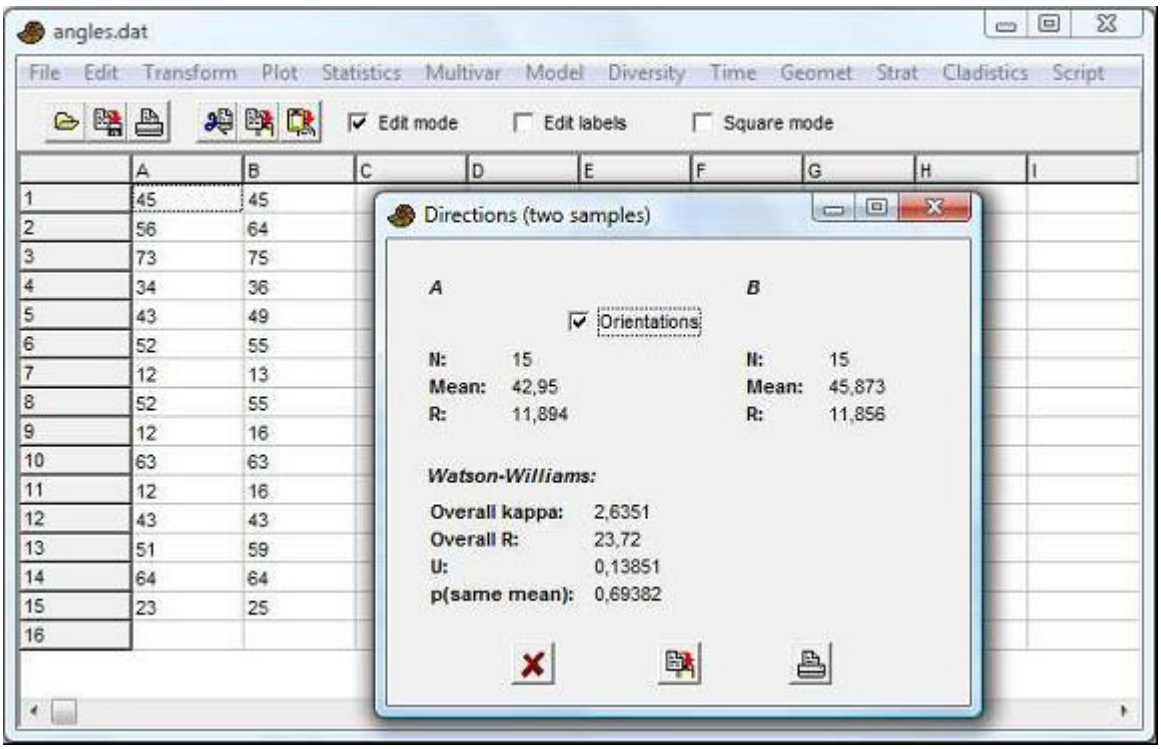
Batschelet, E. 1981. Circular statistics in biology. Academic Press.
Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley & Sons.
Fisher, N.I. 1983. Comment on "A Method for Estimating the Standard Deviation of Wind Directions". *Journal of Applied Meteorology* 22:1971.
Lockhart, R.A. & 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.

Direções – duas amostras (Directions – two samples)

Teste de Watson-Williams

O teste de Watson-William para ângulo médio igual em duas amostras é um teste paramétrico, assumindo distribuição de von Mises, mas é razoavelmente robusto. O módulo espera duas colunas de dados direcionais (0-360) ou orientacionais (0-180) em graus.

O parâmetro de concentração κ deve ser maior do que 1.0 para testes precisos. Adicionalmente, o teste assume variâncias angulares (valores de R) similares.



As duas amostras **φ** e **θ** têm n_1 e n_2 valores. O espalhamento de Rayleigh (*Rayleigh's spread*) R é calculado para cada amostra e para a amostra combinada:

$$R_1 = \sqrt{\left(\sum_{i=1}^{n_1} \cos\phi_i\right)^2 + \left(\sum_{i=1}^{n_1} \sin\phi_i\right)^2}$$

$$R_2 = \sqrt{\left(\sum_{i=1}^{n_2} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n_2} \sin\theta_i\right)^2}$$

$$R = \sqrt{\left(\sum_{i=1}^{n_1} \cos\phi_i + \sum_{i=1}^{n_2} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n_1} \sin\phi_i + \sum_{i=1}^{n_2} \sin\theta_i\right)^2}$$

A estatística de teste U é calculada por

$$U = (n-2)\frac{R_1 + R_2 - R}{n - (R_1 + R_2)}$$

A significância é calculada inicialmente corrigindo o U de acordo com Mardia (1972a):

$$U = \begin{cases} \dfrac{U}{1 - \dfrac{\kappa^2}{8} + \dfrac{1}{n\kappa^2}} & R/n < 0.45 \\ \left(1 + \dfrac{3}{8\kappa}\right)U & R/n < 0.95 \end{cases}$$

onde $n=n_1+n_2$. O valor de p é então dado pela distribuição F com 1 e n-2 graus de liberdade. O parâmetro de concentração combinada (*combined concentration parameter*) κ é de máxima-verossimilhança, calculada como descrito em "Direções – uma amostra" acima.

Teste de Mardia-Watson-Wheeler

Esse teste não-paramétrico para igualdade de distribuição é calculado de acordo com Mardia (1972b).

$$W = 2\left(\frac{C_1^2 + S_2^2}{n_1} + \frac{C_2^2 + S_2^2}{n_2}\right)$$

onde, para a primeira amostra,

$$C_1 = \sum_{i=1}^{n_1} \cos(2\pi r_{1i} / N), \qquad S_1 = \sum_{i=1}^{n_1} \sin(2\pi r_{1i} / N)$$

e de modo similar para a segunda amostra (N=n_1+n_2). Os r_{1i} são os ranks dos valores da primeira amostra dentro da amostra agrupada.

Para N>14, W tem distribuição aproximada de qui-quadrado com 2 graus de liberdade.

Referências

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.

Correlações circulares (Circular correlations)

Teste de correlação entre duas variáveis direcionais ou orientacionais. Assume um número "grande" de observações. Requer duas colunas de dados direcionais (0-360) ou orientacionais (0-180) em graus.

O módulo usa o procedimento de correlação circular o teste de significância paramétrico de Jammalamadaka & Sengupta (2001).

O coeficiente de correlação circular r entre os vetores de ângulos **α** e **β** é

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

onde as médias angulares são calculadas como descrito antes. A estatística de teste T é calculada como

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

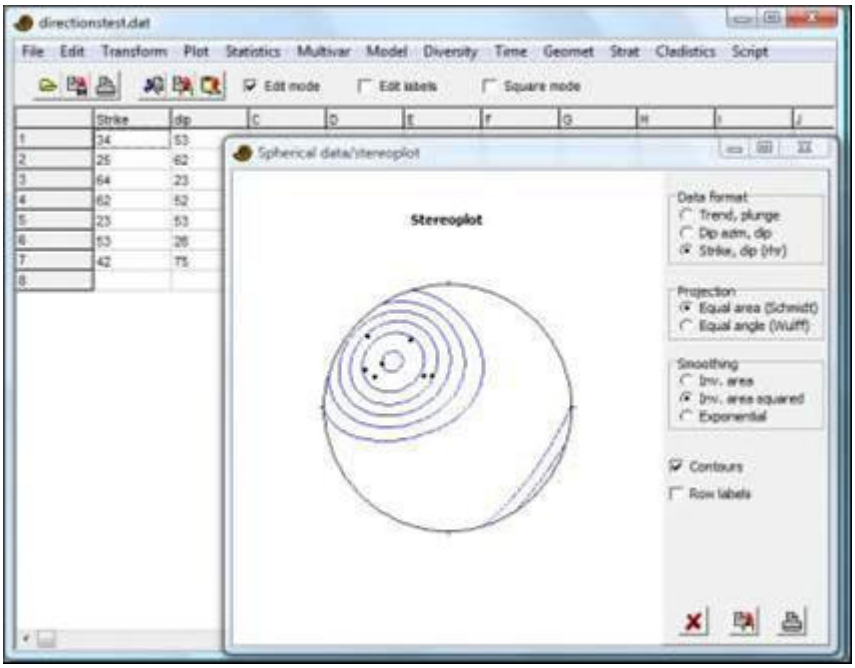
Para n grandes, essa estatística tem distribuição assimptoticamente normal com média 0 e variância 1 na hipótese nula de correlação nula, constituindo a base para o cálculo do p.

Referência

Jammalamadaka, S. R. & A. Sengupta. 2001. Topics in circular statistics. World Scientific.

Esférico – uma amostra (Spherical – one sample)

Este módulo faz gráficos estéreos ("stereo") de dados esféricos axiais (e.g. medidas *strike-dip* em geologia estrutural). Estatísticas esféricas poderão ser adicionadas em versões futuras.



Três formatos de dados podem ser usados, todos usando a convenção geográfica de ângulo (ângulos, sentido horário a partir do norte):

- Tendência (*trend* – azimute) e imersão (*plunge* – ângulo para baixo a partir da horizontal) para dados axiais
- Azimute da imersão e ângulo da imersão (para baixo a partir da horizontal) para planos. O eixo (*pole* – vetor normal) do plano é plotado.
- Golpe (*strike*) e imersão (*dip*) para planos, usando a convenção da regra da mão direita com a impressão para baixo e para a direita do golpe. O eixo do plano é plotado.

O contorno da densidade é baseado em um algoritmo modificado do método de Kamb, por Vollmer (1995). Tanto projeções de área igual (Schmidt) quanto de ângulo igual (Wulff) são disponíveis. Projeções são para o hemisfério inferior. Estimativas de densidade podem usar área inversa, área inversa elevada ao quadrado, ou lei exponencial, resultados em graus maiores de alisamento (*smoothing*).

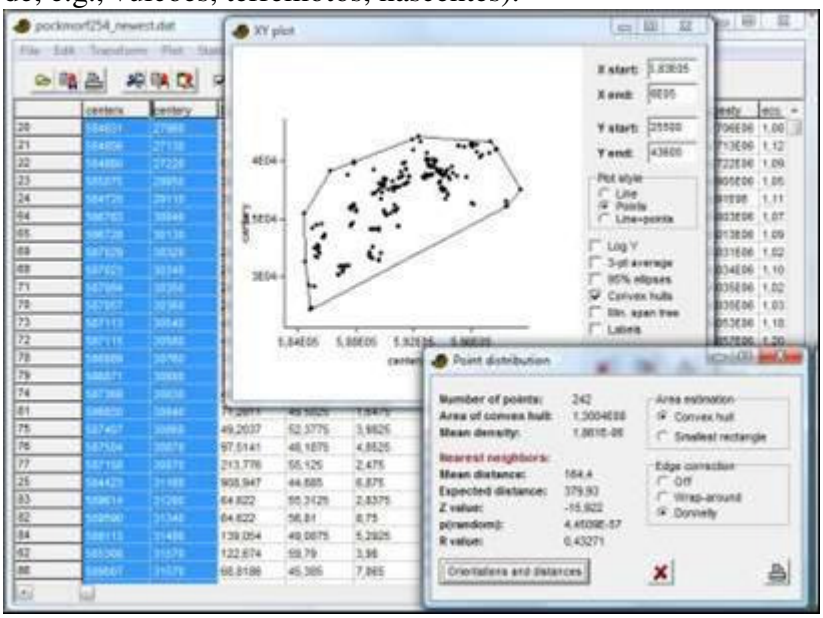
Referência

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

Análise de vizinho mais próximo do padrão de pontos (Nearest neighbour point pattern analysis)

Este módulo testa o agrupamento *(clustering*) ou superdispersão (*overdispersion*) de pontos fornecidos por valores bi-dimensionais de coordenadas. O procedimento assume que todos os elementos são pequenos em comparação com suas distâncias, que o domínio é predominantemente convexo, e n>50. Duas colunas de posições x/y são necessárias.

Aplicações deste módulo incluem ecologia espacial (braquiópodos são agrupados in-situ?), morfologia (tubérculos de trilobitas são superdispersos?) e geologia (distribuição de, e.g., vulcões, terremotos, nascentes).



O cálculo das estatística de distribuição de pontos usando o vizinho mais próximo é de acordo com Davis (1986), com modificações. A área é estimada usando ou o menor retângulo que envolve todos os pontos ou o casco convexo (*convex hull*), que é o menor polígono convexo que envolve todos os pontos. Ambos são inapropriados para pontos em domínios muito côncavos. Dois métodos de ajuste diferentes para efeitos de borda (*edge effects*) são disponíveis- *wrap-around* ("torus") e correção de Donnelly. Detecção de borda *wrap-around* só é apropriada em domínios retangulares.

A hipótese nula é um processo aleatório de Poisson, dando uma distribuição exponencial modificada de vizinho mais próximo (ver abaixo) com média

$$\mu = \frac{\sqrt{A/n}}{2}$$

onde A é a área e n é o número de pontos.

A probabilidade de que a distribuição é Poisson é fornecida, juntamente com o valor de R:

$$R = \frac{\bar{d}}{\mu} = \frac{2\bar{d}}{\sqrt{A.n}}$$

onde $\bar{d}$ é a distância média observada entre vizinhos mais próximos. Pontos agrupados dão $R<1$, padrões de Poisson dão $R\sim1$, enquanto pontos superdispersos são $R>1$.

A distribuição esperada (teórica) sob a hipótese nula é plotada como uma curva contínua junto com o histograma das distâncias observadas. A função de probabilidade de densidade esperada em função da distância r é

$$g(r) = 2\rho\pi r \exp(-\rho\pi r^2)$$

onde $\rho = n/A$ é a densidade de pontos (Clark & Evans 1954).

As orientações (0-180 graus) e comprimentos das linhas entre os vizinhos mais próximos também são incluídas. As orientações podem ser sujeitas a análise direcional para

verificar se os pontos estão organizados ao longo de linhas (ver Hammer 2009 para métodos mais avançados).

Referências

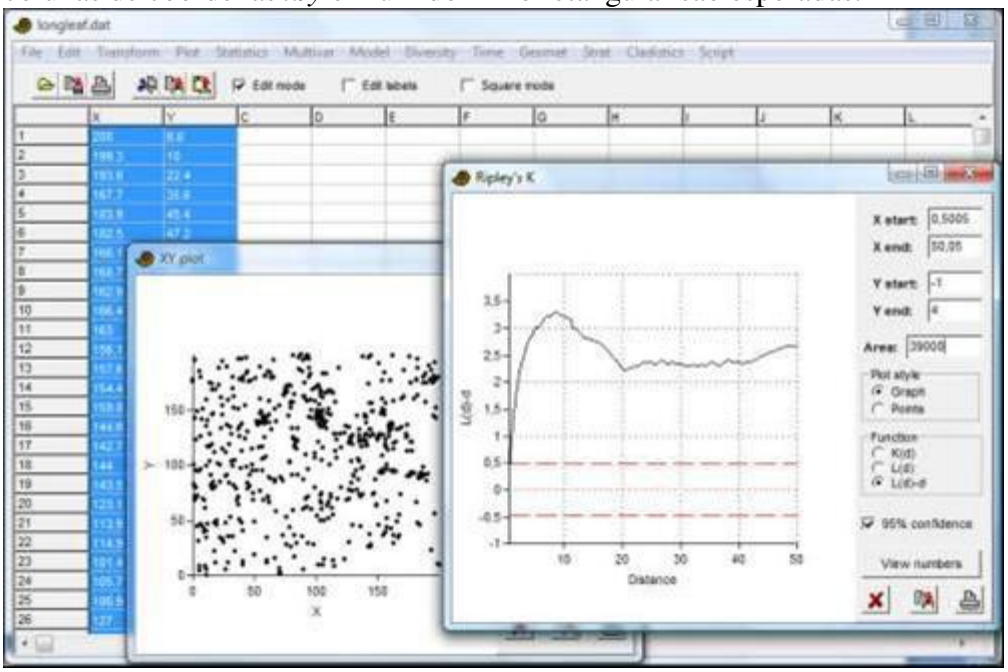
Clark, P.J. & Evans, F.C. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* 35:445-453.

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.

Análise do padrão de pontos pelo K de Ripley (Ripley's K point pattern analysis)

O *K* de Ripley (Ripley 1979) é a densidade média de pontos como função da distância de cada ponto. É útil quando características do padrão de pontos mudam com escala, e.g. superdispersão em pequenas escalas, mas com agrupamento em grandes escalas. Duas colunas de coordenas *x*/*y* em um domínio retangular são esperadas.



Defina a intensidade estimada do padrão de pontos, com *n* pontos em uma área *A*, como $\lambda=n/A$. A distância entre os pontos *i* e *j* é d_{ij}. A estimativa do *K* de Ripley, como função de distância, é então calculada como

$$K(d) = \frac{1}{\lambda n} \sum_{i=1}^{n} \sum_{j \neq 1} I(d_{ij} \leq d),$$

onde a função indicadora *I* é um se o argumento é verdadeiro, zero caso contrário.

A normalização de *K* é tal que para aleatoriedade espacial completa (*complete spatial randomness* – CSR), espera-se que *K(d)* aumente como área de círculos, i.e. $K(d)=\pi d^2$. A função *L(d)* é uma transformação correspondente de *K(d)*:

$$L(d) = \sqrt{\frac{K(d)}{\pi}}$$

Para CSR, $L(d)=d$, e $L(d)-d=0$. Um intervalo de confiança de 95% para CSR é estimada usando 1000 simulações Monte Carlo dentro do retângulo que delimita a área (versões anteriores usaram a aproximação $1.42\sqrt{A}/n$).
A correção de Ripley para bordas (*Ripley's edge correction*) é incluída, dando pesos a contagens dependendo da proporção do círculo que está dentro do domínio retangular.
O exemplo acima mostra localizações de árvores em uma floresta. $L(d)-d$ fica acima do intervalo de 95% para CSR, indicando agrupamento. Adicionalmente, as interações espaciais parecem ser mais proeminentes em uma escala de aproximadamente 10 m, acima da qual a curva fica plana de um modo esperado para CSR.

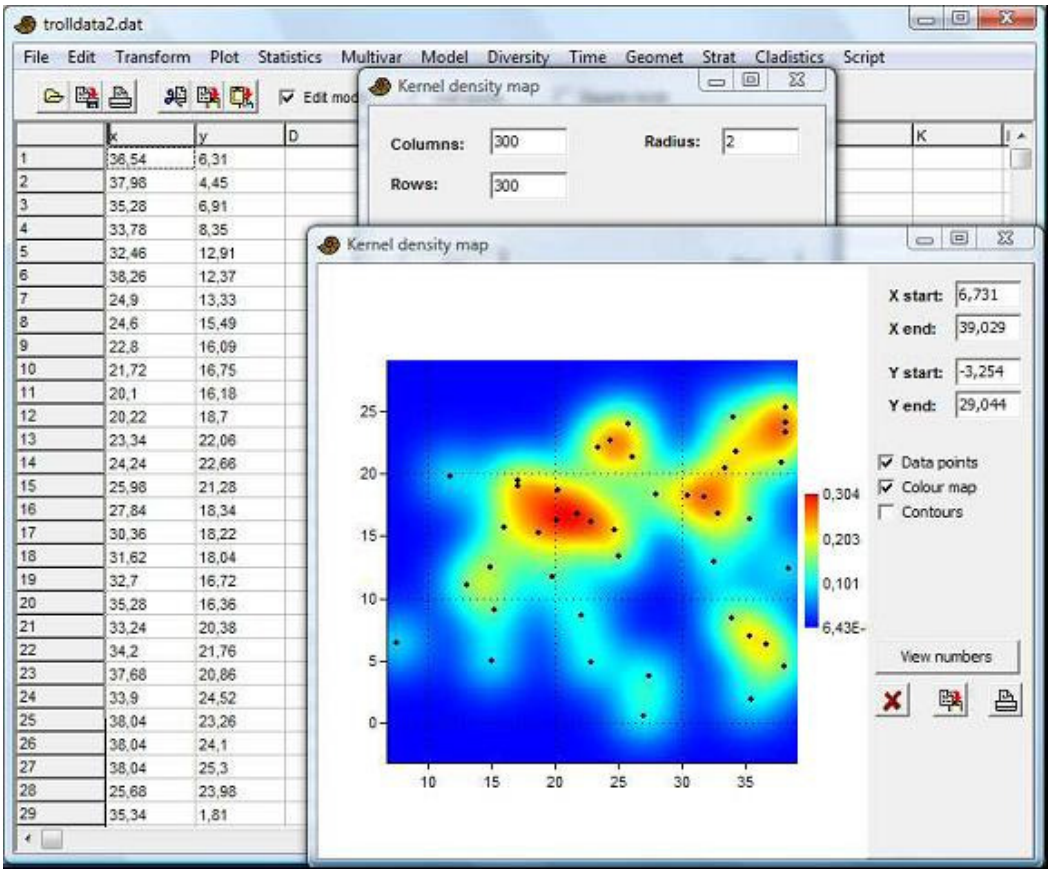
Área
Para que o K de Ripley seja calculado corretamente, a área deve ser conhecida. Na primeira rodada, a área é calculada usando o menor retângulo que engloba a área, mas isso pode super ou subestimar a área real. A área pode ser ajustada pelo usuário. Uma área superestimada normalmente irá aparecer como uma forte tendência linear geral com inclinação positiva para $L(d)-d$.

Dimensão fractal (*Fractal dimension*)
A dimensão fractal (caso exista alguma) pode ser estimada pela inclinação linear assimptótica em um gráfico log-log de $K(d)$. Para CSR, a inclinação log-log deve ser 2.0. Fractais devem ter inclinações menores do que 2.

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

Densidade Kernel (Kernel density)

Cria um mapa suave da densidade de pontos em 2D. Duas colunas de dados *x*/*y* em um domínio retangular são esperadas. O usuário pode especificar o tamanho da grade (número de linhas e colunas). O valor "*Radius*" ("Raio") estabelece a escala *r* do Kernel. Automaticamente não há uma seleção de raio "ótimo", de modo que este valor deve ser definido pelo usuário dependendo da escala de interesse.



A estimativa de densidade é baseada em uma de quatro funções Kernel, com parâmetro de raio *r*. Sendo $d_i = \sqrt{(x-x_i)^2 + (y-y_i)^2}$:

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

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

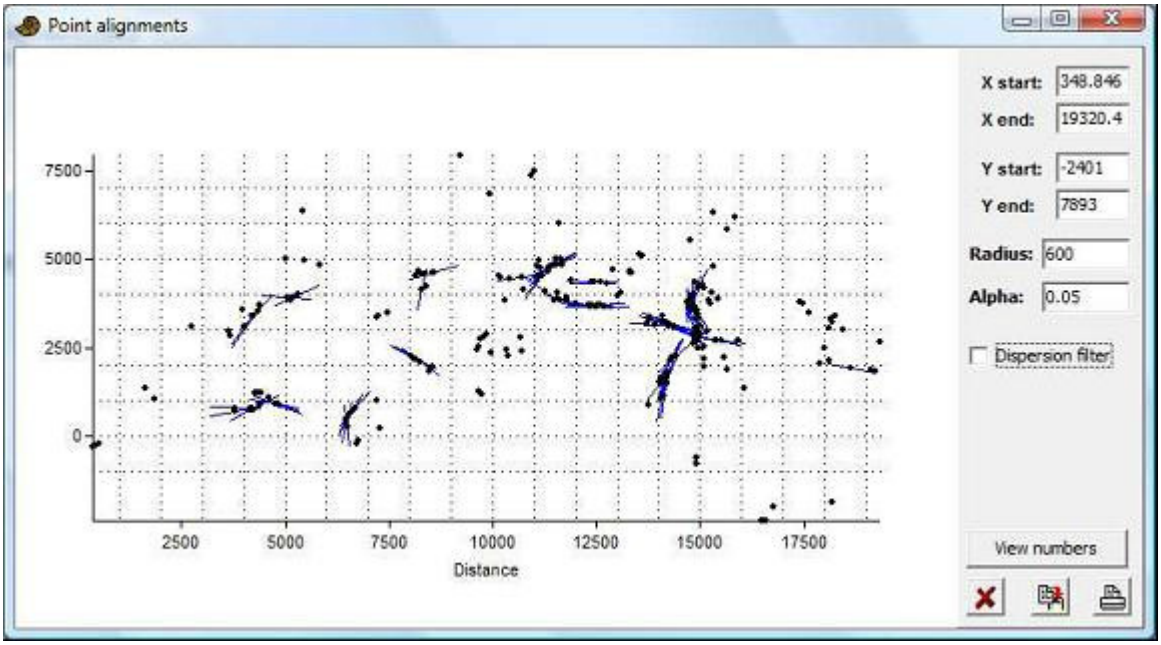
Triangular: $$f(x,y) = \frac{2}{\pi r^2} \sum_i \begin{cases} 1 - \frac{d_i}{r} & d_i \le r \\ 0 & d_i > r \end{cases}$$

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

Este escalonamento fornece uma estimativa do número estimado de pontos por área, não uma densidade de probabilidade. Os Kernels gaussiano e parabolóide (quadrático) normalmente têm melhor desempenho. O Kernel uniforme resulta em gráficos muito pouco suaves.

Alinhamento de pontos (Point alignments)

Detecção de alinhamentos lineares em um padrão de pontos 2D, usando o método dos setores contínuos (*continuous sector method* – Hammer 2009). Aplicações típicas são em geologia e geografia, para estudar a distribuição de terremotos, vulcões, fontes etc, associadas com falhas ou outras estruturas lineares.



O parâmetro *Radius* (raio) estabelece a escala da análise. No exemplo acima, alinhamentos com comprimento de 1200 m (o dobro do raio) são detectados.
Alpha estabelece o nível de significância para o teste de Rayleigh usado por este procedimento. Repare que esta é uma significância ponto-a-ponto, não corrigida para testes múltiplos de todos os pontos.
O filtro de dispersão (*Dispersion filter*) desativa alinhamentos com distribuição desigual de pontos ao longo da linha.
View number (Ver números) lista as posições de alinhamentos e suas orientações, que então podem ser sujeitas à estatística circular se necessário (módulo Direções).

Referência

Hammer, Ø. 2009. New methods for the statistical detection of point alignments. *Computers & Geosciences* 35:659-666.

Autocorrelação espacial – I de Moran (Spatial autocorrelation – Moran's I)

Autocorrelação espacial no Past requer três colunas, contendo coordenadas *x* e *y* e valores correspondentes de dados *z* para uma série de pontos. A estatística de correlação *I* de Moran é então calculada dentro de cada uma de uma série de classes de distância (classes ou *bins*), indo de distâncias pequenas a distâncias grandes.
O valor crítico unicaudal para $p<0.05$ pode ser plotado para cada classe. Valores de *I* de Moran que excedam o valor crítico podem ser considerados significativos, mas ajuste de Bonferroni ou algum outro ajuste para testes múltiplos deve ser considerado por causa da existência de várias classes.

O cálculo é de acordo com Legendre & Legendre (1998). Para cada classe de distância *d*, calcule

$$I(d) = \frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(z_h - \bar{z})(z_i - \bar{z})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Aqui, n é o número total de pontos, W é o número de pares de pontos com distâncias entre eles dentro da classe de distância, e w_{hi} é uma função de ponderamento (*weight function*) tal que w_{hi}=1 se os pontos h e i estão dentro da classe de distância e w_{hi}=0 caso contrário (delta de Kronecker). Repare que esta equação está incorreta em algumas publicações.

Para o nível crítico unicaudal $I_{0.05}$, calcule

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

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

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

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

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

Aqui, w_{i+} e w_{+i} são somatórias de linhas e de colunas. O fator de correção $k_{0.05}$ é ajustado em $\sqrt{10 \cdot 0.05} = 0.707$ se $4(n - \sqrt{n}) < W \leq 4(2n - 3\sqrt{n} + 1)$, caso contrário $k_{0.05}$=1.

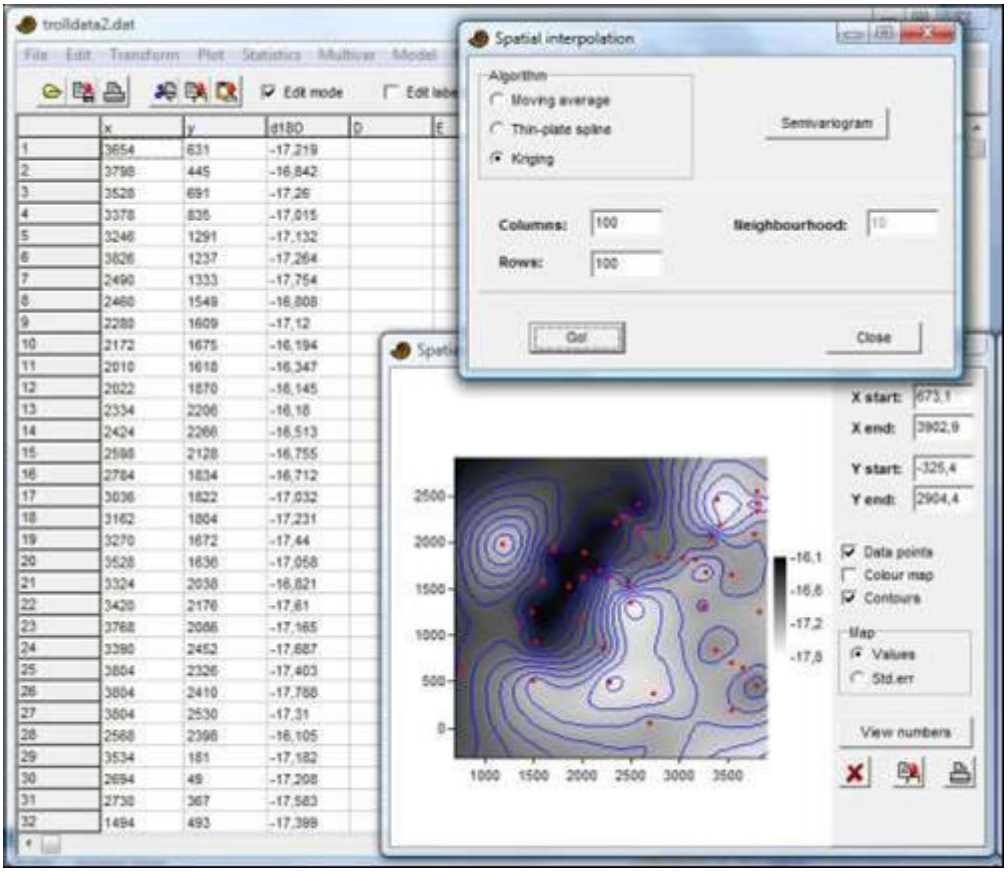
Referência

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

Gridagem – interpolação espacial (Gridding – spatial interpolation)

"Gridagem" (*"Gridding"*) é a operação de interpolação espacial que produz uma grade (*grid*) regular a partir de pontos de dados 2D espalhados. Três colunas com posição (*x*,*y*) e valores correspondentes são necessárias.

A gridagem permite produzir um mapa mostrando uma estimativa espacial contínua de alguma variável tal como abundância de fósseis ou espessura de uma unidade de rocha, com base em pontos de dados espalhados. O usuário pode especificar o tamanho da grade (número de linhas e de colunas). A cobertura espacial do mapa é gerada automaticamente como um quadrado cobrindo os pontos de dados. Ao fazer o gráfico, isso pode ser reduzido a um casco convexo (*convex hull*) dos pontos.



Uma superfície linear de mínimos-quadrados (tendência) é automaticamente ajustada aos dados, removida antes da gridagem e finalmente adicionada novalmente. Isso é útil principalmente para a modelagem de semivariograma (*semivariogram modelling*) e para o método de krigagem (*kriging*).

Validação cruzada (Cross validation): Esta opção irá remover um ponto de dados por vez e re-calcular a superfície com base nos pontos remanescentes (*"jackknife"*). As diferenças entre os valores originais e os valores obtidos por validação cruzada indica a acurácia da precisão do modelo de superfície. Estas diferenças são relatadas para cada ponto, junto com o erro quadrado médio (*mean squared error – MSE*) calculado para todos os pontos.

Quatro algoritmos de interpolação são disponíveis:

Ponderamento pelo inverso da distância (Inverse distance weighting)
O valor no nó da grade é apena a média dos *N* pontos mais próximos, como especificado pelo usuário (o padrão é usar todos os pontos de dados). Os pontos são ponderados em uma proporção inversa à distância. Este algoritmo é rápido, mas nem sempre dará resultados bons (suaves). Um artefato típico é o "alvo" (*"bull's eyes"*) em volta dos pontos de dados. Uma vantagem é que os valores interpolados nunca irão exceder a amplitude (*range*) dos pontos de dados. Estabelecendo *N*=1, o algoritmo fica reduzido ao *método do vizinho mais próximo (nearest-neighbour method)*, que estabelece o valor em um nó da grade igual ao valor do ponto de dados mais próximo.

Alisamento polinomial de placa fina (Thin-plate spline)
Interpolador que dá a máxima suavidade. Pode produzir valores elevados ou baixos demais na presença de curvaturas abruptas na superfície. É um método radial com função radial básica (*radial basis function*) $\varphi = r \ln r$.

Multiquadrático
Função radial básica $\varphi = r$. Bastante usado para modelagem de terreno.

Krigagem (Kriging)
É necessário que o usuário estabeleça um modelo para o semivariograma, escolhendo um dos quatro modelos comuns e parâmetros correspondentes para ajustar as semivariâncias empíricas (a soma dos quadrados residuais – *residual sum of squares* – deve ser a menor possível. O semivariograma é calculado dentro de cada um de um número de classes (*bins*). Usando a opção histograma, escolha o número de *bins* tal que cada *bin* (com a possível exceção dos da extrema direita) contenha pelo menos 30 distâncias.
O parâmetro *nugget* é uma constante adicionada ao modelo. Ele implica uma variância diferente de zero na distância zero, e, portanto, permitirá que a superfície não passe exatamente pelos pontos de dados. O parâmetro *range* controla a extensão da curva ao longo do eixo das distâncias. Nas equações abaixo, o valor de distância normalizado h representa *distância/range*. O *scale* (escala) controla a extensão da curva ao longo do eixo da variância.

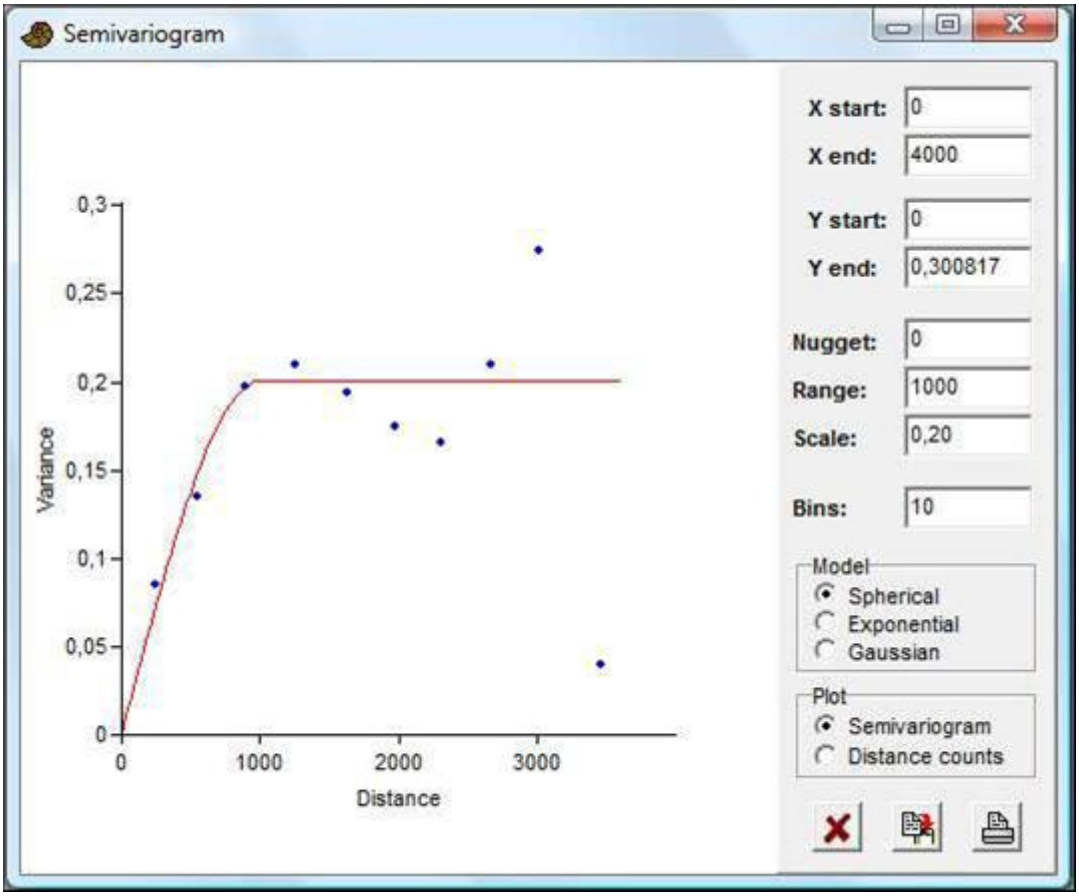
Esférico (*Spherical*): $$\gamma(h) = \begin{cases} nugget + scale\left(\frac{3h}{2} - \frac{1}{2}h^3\right) & h < 1 \\ nugget + scale & h \geq 1 \end{cases}$$

Exponencial (*Exponential*): $$\gamma(h) = nugget + scale(1 - e^{-h})$$

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

Cúbico (*Cubic*): $$\gamma(h) = \begin{cases} nugget + scale(7h^2 - 8.75h^3 + 3.5h^5 - 0.75h^7) & h < 1 \\ nugget + scale & h \geq 1 \end{cases}$$

O botão *"Optimize all"* ("Otimizar todos") irá selecionar o modelo e os parâmetros que dão a menor soma de quadrados dos resíduos do semivariograma. Isso pode não ser o que você quer: por exemplo, você pode querer usar um modelo específico ou ter um *nugget* igual a zero para garantir uma interpolação exata. Para isso será necessário ajustar os valores manualmente.
O procedimento de krigagem também fornece uma estimativa dos erros padrão ao longo do mapa (para isso, o modelo de semivariograma deve ter boa acurácia). Krigagem no PAST não funciona com semivariância anisotrópica.
Aviso: Krigagem é um processo lento, não tente caso você tenha mais de aproximadamente 1000 pontos de dados em uma grade 100x100.



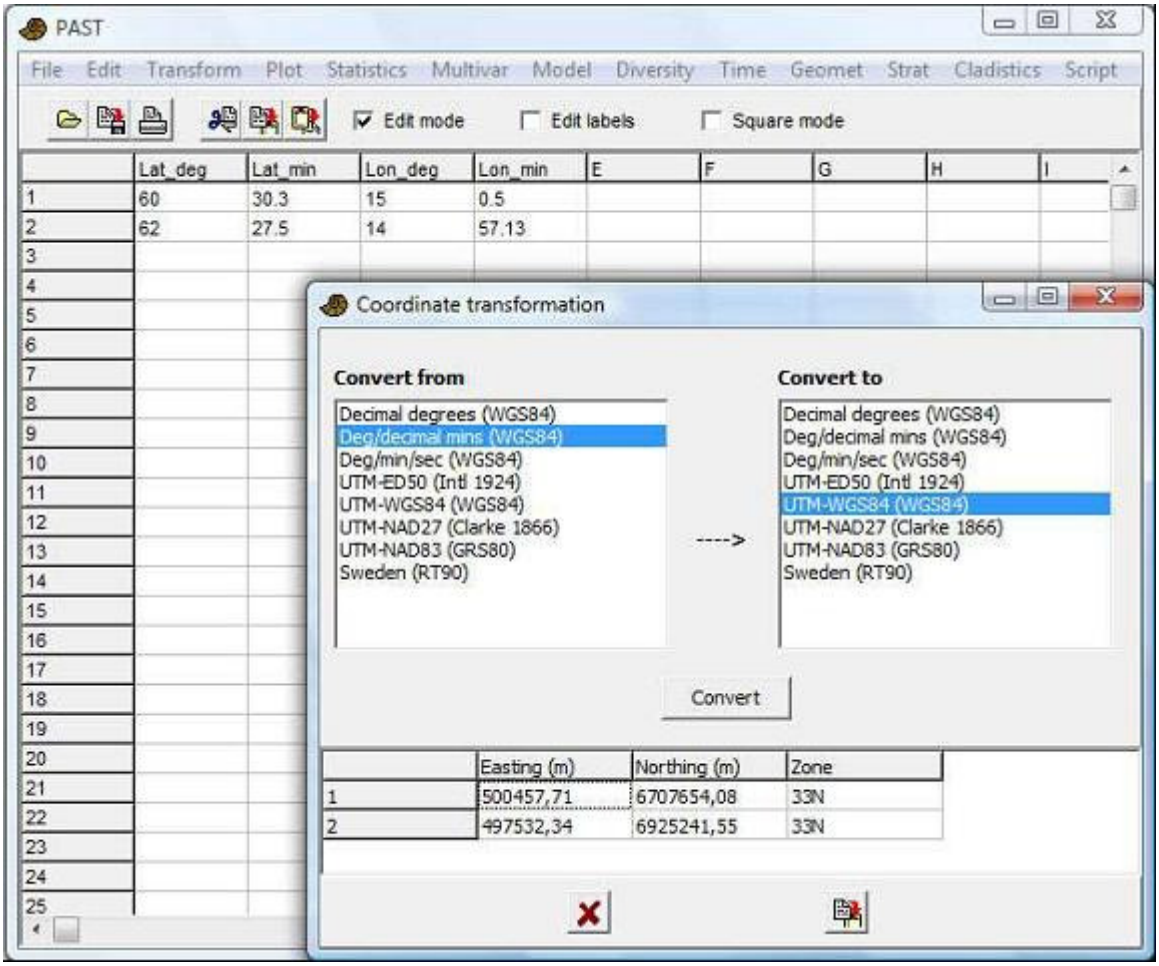
Veja e.g. Davis (1986) ou Smith et al. (2009) para mais informação sobre krigagem.

Referências

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.

Transformação de coordenadas (Coordinate transformation)

Conversão entre coordenadas geográficas em diferentes grades (*grids*) e datums. O número de colunas de entrada (*input*) depende dos tipos de dados, como descrito abaixo.



Graus decimais (*Decimal degrees* – WGS84)
Duas colunas: Latitude e Longitude, em graus decimais (60.5 é 60 graus, 30 minutos). Valores negativos para o sul do equador e a oeste de Greenwich. Referenciado ao datum WGS84.

Graus, minutos decimais (*Deg/ decimal mins* – WGS84)
Quatro colunas: Graus de latitude, minutos decimais (40.5 é 40 minutos, 30 segundos), graus de longitude, minutos decimais. Referenciado ao datum WGS84.

Graus/minutos/segundos (*Deg/min/sec* – WGS84)
Seis colunas: graus de latitude, minutos, segundos, graus de longitude, minutos, segundos. Referenciado ao datum WGS84.

UTM-ED50 (Intl 1924)
Três colunas: Leste (*Easting*) (metros), norte (*northing*) (metros), e zona. Use números de zonas negativos para o hemisfério sul. O tratamento das zonas UTM leva em conta as situações especiais de Svalbard e do oeste da Noruega. Referenciado ao datum europeu ED50 em Potsdam.

UTM-WGS84 (WGS84)
Três colunas: Leste (metros), norte (metros) e zona. Referenciado ao datum WGS84.

UTM-NAD27 (Clarke 1866)
Três colunas: Leste (metros), norte (metros) e zona. Referenciado ao datum NAD27. Conversão para/de este formato é ligeiramente imprecisa (5-6 metros).

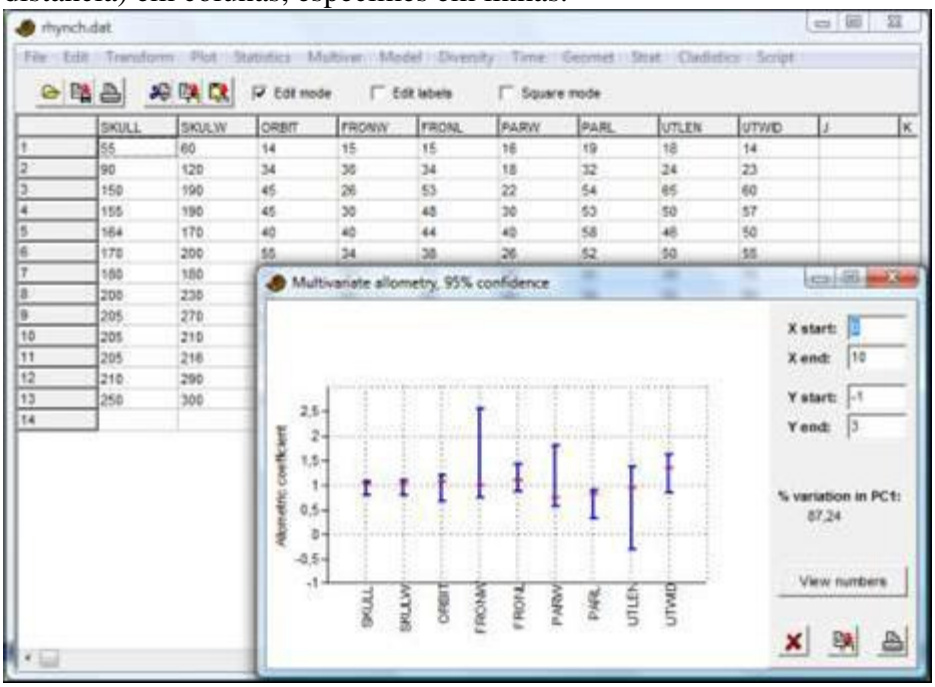
UTM-NAD83 (GRS80)
Três colunas: Leste (metros), norte (metros) e zona. Referenciado ao datum NAD83 (praticamente idêntico ao WGS84).

Sweden (RT90)
Duas colunas: Leste (metros) e norte (metros).

As transformações são baseadas em código gentilmente fornecido por I. Scollar.

Alometria multivariada (Multivariate allometry)

Este módulo é usado para investigar a alometria de um conjunto multivariado de dados morfométricos. Espera-se um conjunto multivariado de dados com variáveis (medidas de distância) em colunas, espécimes em linhas.



Este método para a investigação de alometria em um conjunto multivariado de dados é baseado em Jolicoeur (1963) com extensões por Kowalewski et al. (1997). Os dados são (automaticamente) transformados em log e sujeitos a uma PCA. O primeiro componente principal (PC1) é então considerado como eixo de tamanho (*size axis*) (isso só é válido caso a variação explicada pela PCA seja grande, digamos mais de 80%). O coeficiente alométrico de cada variável original é estimado dividindo o peso (*loading*) daquela variável no PC1 pelo peso médio de todas as variáveis no PC1.

Intervalos de confiança de 95% para os coeficientes alométricos são estimados por *bootstrap* dos espécimes. 2000 réplicas de *bootstrap* são feitas.
Dados ausentes: suporte por substituição pela média da coluna.

Referências

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.

Forma de Fourier – 2D (Fourier shape – 2D)

Análise do contorno da forma de fósseis (2D). Forma apresentável em coordenadas polares, número suficiente de pontos digitalizados para capturar as características. Coordenadas *x*/*y* digitalizadas ao redor de um contorno. Espécimes em linhas, coordenadas de valores alternantes de *x* e *y* em colunas (veja Encaixe de Procrustes – *Procrustes fitting* no menu Transform).
Aceita coordanadas X-Y digitalizadas ao redor de um contorno. Mais de uma forma (linha) pode ser analisada simultaneamente. Os pontos não precisam ser uniformemente espaçados. A forma deve poder ser expressa como uma função única de coordenadas polares, ou seja, qualquer linha reta irradiando do centro da forma deve cruzar o contorno uma única vez.
O algoritmo é de acordo com Davis (1986). A origem do sistema de coordenadas polares é encontrada por aproximação numérica do centróide. 128 pontos são então produzidos em incrementos angulares uniformes ao redor do contorno por interpolação linear. O centróide é então recalculado e os raios são normalizados (de modo que o tamanho é removido da análise). Os componentes seno e coseno (*sine* e *cosine*) são dados para os vinte primeiros harmônicos, mas repare que apenas *N*/2 harmônicos são "válidos", onde *N* é o número de pontos digitalizados. Os coeficientes podem ser copiados para a planilha principal para análises subsequentes (e.g. por PCA).
O janela "Ver forma" (*"Shape view"*) permite uma visualização gráficas da(s) aproximação(ões) de Fourier.

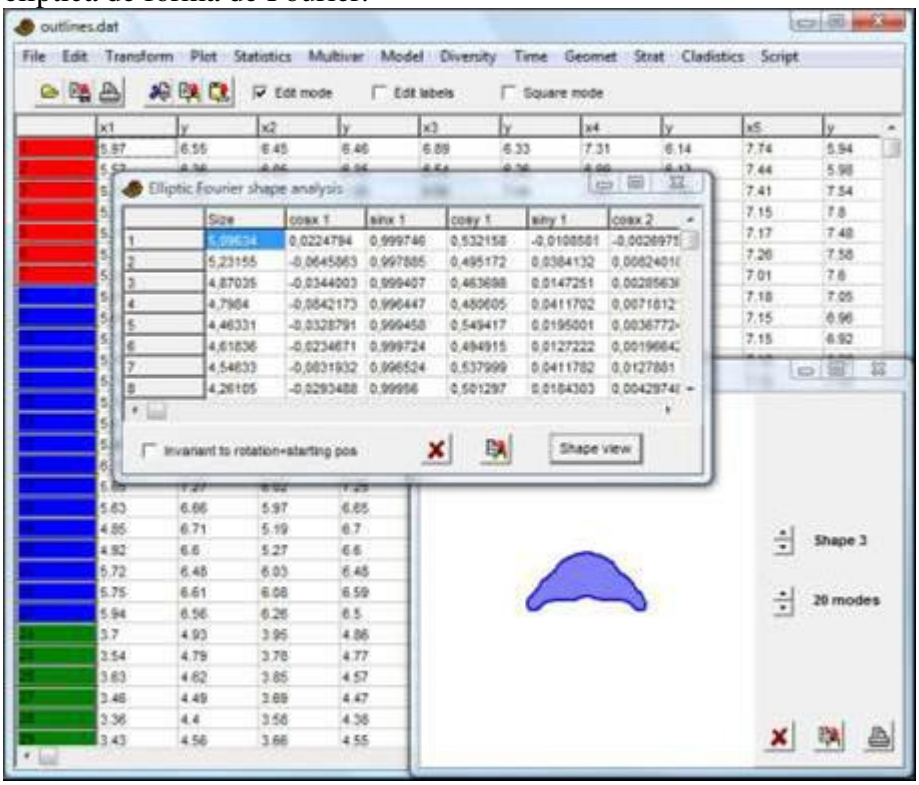
Referência

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

Análise elíptica de forma de Fourier (Elliptic Fourier shape analysis)

Requer coordenadas digitalizadas *x*/*y* ao redor de contornos. Espécimes em linhas, coordenadas de valores alternantes de *x* e *y* em colunas. A análise elíptica de forma de Fourier é superior à análise de forma simples de Fourier em diversos aspectos. Uma vantagem é que o algoritmo pode lidar com formas complicadas que podem não ser expressas como uma função única de coordenadas polares. Formas elípticas de Fourier é atualmente um método padrão para análise de contorno. O algoritmo usado no PAST é descrito por Fearson et al. (1985).
Componentes coseno (*cosine*) e seno (*sine*) de incrementos *x* e *y* ao longo contorno para os primeiros 30 harmônicos são fornecidos, mas apenas os primeiros *N*/2 harmônicos

deveriam ser usados, sendo *N* o número de pontos digitalizados. Tamanho e translação posicional (*positional translation*) são removidos por normalização e não entram nos coeficientes. O tamanho (antes da normalização) é fornecido na primeira coluna. A normalização opcional para rotação ou ponto inicial (*starting point*), que segue Fearson et al., às vezes inverte formas. Isso deve ser verificado com a opção "Ver forma" (*"Shape view"*) – pode ser necessário remover estes espécimes.
Os coeficientes podem ser copiados para a planilha principal para análises subsequentes, como PCA e análise de discriminantes. Os módulos PCA e regressão linear (1 independente, *n* dependentes) contêm funções para mostrar os contornos de formas correspondentes a determinados escores de PCA ou valores da variável independente.
A janela "Ver forma" (*"Shape view"*) permite visualizar graficamente a aproximação elíptica de forma de Fourier.



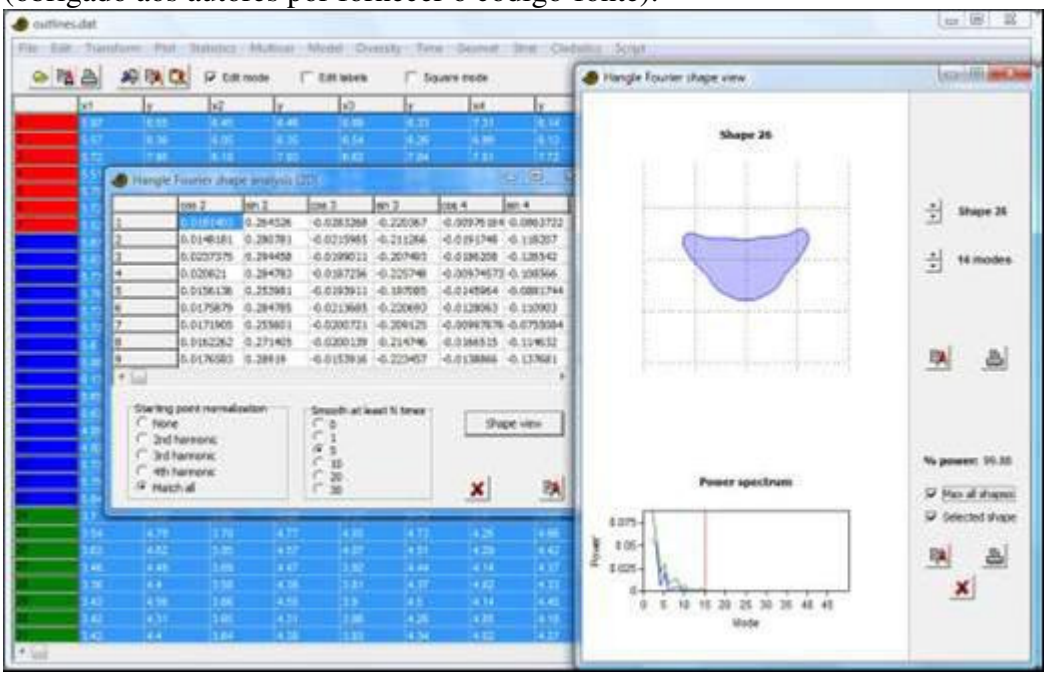
Referência

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

Análise Hangle de forma de Fourier (Hangle Fourier shape analysis)

Requer coordenadas *x*/*y* digitalizadas ao redor de contornos. Espécimes em linhas, coordenadas de valores alternantes de *x* e *y* em colunas.
O método "Hangle" para análise de contornos fechados, proposto por Haines & Crampton (2000), é um concorrente da Análise Elíptica de Fourier. Hangle tem algumas vantagens em relação à AEF, sendo a mais importante o fato de menos coeficientes serem necessários para capturar o contorno com a precisão desejada. Isso é importante para

testes estatísticos (e.g. MANOVA) e para análise de discriminantes. A implementação no Past é baseada nos pacotes Hangle/Hmatch/Htree/Hshape de Haines & Crampton (obrigado aos autores por fornecer o código-fonte).



O *output* consiste de 46 coeficientes de Fourier, que são os coeficientes coseno e seno dos 24 harmônicos (modos), começando no harmônico de número 2. Cope estes números de volta à planilha do Past para análises de formula multivariadas subsequentes.

Normalização do ponto inicial (Starting point normalization)
Normalmente deve ser deixado em "todos condizentes" (*"match all"*), com o método "Hmatch" ou (talvez mais indicado) "Htree" para alinhar todos os contornos. Uma alternativa é selecionar o harmônico 2.-4., o que irá mudar a fase de cada contorno de acordo com o modo selecionado (ver Haines & Crampton 2000).

Suavização (Smoothing)
Aumentar o parâmetro de *smoothing* pode reduzir ruído de alta frequência, mas ao mesmo tempo pode haver perda de informações de alta frequência que podem ser importantes para a descrição da forma.

Ver forma (Shape view)
Use esta função para inspecionar as formas reconstruídas a partir dos coeficientes de Fourier. Verifique se a rotina de *matching* não rotacionou incorretamente alguma forma. Além disso, use esta função para selecionar o número de modos necessário para capturar a forma. No exemplo acima, o número de modos foi deixado em 14, o que captura 99.88% do poder integrado total (quadrado da amplitude) (*total integrated power – amplitude squared*) da forma selecionada. O número de modos é mostrado pela linha vermelha no espectro de poder – tenha certeza de que as principais características do espectro estejam à esquerda desta linha para todas as formas.

Nota: Reconstrução de forma por PCA, regressão e CVA (como para AEF) ainda não foi implementada para Hangle.

Referência

Haines, A.J. & J.S. Crampton. 2000. Improvements to the method of Fourier shape analysis as applied in morphometric studies. *Palaeontology* 43:765-783

Análise de autoforma (Eigenshape analysis)

Coordenadas digitalizadas *x*/*y* ao redor de uma forma. Espécimes em linhas, coordenadas de valores alternados de *x* e *y* em colunas (veja Encaixe de Procrustes no menu Transform).

Autoformas são componentes principais de contornos. O gráfico de dispersão (*scatter plot*) dos contornos no espaço de componentes principais é mostrado, e combinações lineares das próprias autoformas podem ser visualizadas.

A implementação no Past é baseada parcialmente em MacLeod (1999). Ela encontra o número ótimo de pontos espaçados uniformemente ao redor do contorno por meio de uma busca iterativa, de modo que os pontos originais não precisam ser espaçados uniformemente. A autoanálise (*eigenanalysis*) é baseada na matriz de covariância dos incrementos de giro angular não-normalizados ao longo dos contornos. O algoritmo não assume uma curva fechada, e os pontos extremos, portanto, não precisam coincidir nas formas reconstruídas. Autoanálise com registro de pontos de referência (*landmark-registered eigenanalysis*) não é incluída. Todos os contornos devem começar no "mesmo" ponto.

Referência

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

Polinômios de placa fina e deformações (Thin-plate splines and warps)

Coordenadas digitalizadas *x*/*y* de pontos de referência (*landmarks*). Espécimes em linhas, coordenadas de valores alternados de *x* e *y* em colunas. Padronização de Procrustes é recomendada.

O primeiro espécime (primeira linha) é usado como referência, como uma grade quadrada associada. As deformações de todos os espécimes em relação a este espécime podem ser visualizadas. Você também pode usar a forma média como referência.

A opção "Fatores de expansão" (*"Expansion factors"*) irá mostrar o fator de expansão (ou contração) de área ao redor de cada ponto de referência com números amarelos, indicando o grau de crescimento local. Isso é calculado usando a Jacobiana da deformação. Além disso, as expansões são codificadas por cores para todos os elementos da grade, com verde para expansão e púrpura para contração.

Em cada ponto de referência, as principais tensões (*strains*) podem ser mostradas, com a tensão principal mostrada em preto e as tensões maiores mostradas em marrom. Estes vetores indicam esticamento direcional.

Uma descrição de grades de transformação polinomial de placa fina é feita por Dryden & Mardia (1998).

Deformações parciais (Partial warps)

A partir da janela do polinômio de placa fina, você pode escolher a visualização de deformações parciais iara uma deformação polinomial particular. A primeira deformação parcial irá representar alguma deformação de larga escala na grade, enquanto deformações de ordens maiores normalmente serão relacionadas e deformações mais locais. Os componentes *affine* da deformação (também conhecidos como deformação de ordem zero – *zeroth warp*) representam translação linear, escalonamento, rotação e cisalhamento (*shearing*). Na versão atual do PAST não é possível ver as deformações principais.
Ao colocar valores maiores que zero no fator de amplitude, a configuração original dos pontos de referência e uma grade serão deformadas progressivamente de acordo com a deformação parcial escolhida.

Escores das deformações parciais (Partial warp scores)
A partir da janela do polinômio de placa fina, você também pode ver os escores de deformação parcial para todos os espécimes. Cada escore de deformação parcial tem dois componentes (*x* e *y*), e os escores são, portanto, apresentados em gráficos de dispersão.
Referência
Dryden, I.L. & K.V. Mardia 1998. Statistical Shape Analysis. Wiley.

Deformações relativas (Relative warps)

Ordenação de um conjunto de formas. Coordenadas digitalizadas *x*/*y* de pontos de referência. Espécimes em linhas, coordenadas de valores alternados de *x* e *y* em colunas. Recomenda-se usar padronização de Procrustes.
As deformações relativas podem ser vistas como os componentes principais de uma série de transformações de placa fina (*thin-plate transformations*) a partir da forma média para cada forma estudada. Esta análise fornece uma alternativa para PCA feita diretamente nos pontos de referência (ver *Shape PCA* acima).
O parâmetro *alpha* pode ser ajustado em um de três valores:

- *alpha*=-1 enfatiza variação de pequena escala.
- *alpha*=0 é PCA aplicada diretamente nos pontos de referência, e é equivalente a *Shape PCA* (ver acima) mas *sem a inclusão de um componente affine (uniforme).*
- *alpha*=1 enfatiza variação de larga escala.

As deformações relativas são ordenadas de acordo com a importância, e a primeira e segunda deformações normalmente são as mais informativas. Repare que os valores de porcentagem dos autovalores são relativos à parte total não-*affine* da transformação – parte *affine* não é incluída (veja *Shape PCA* para deformações relativas com inclusão do componente *affine*).
As deformações relativas são visualizadas com grades de transformação de placa fina. Ao aumentar o diminuir o fator de amplitude a partir de zero, a configuração original de pontos de referência e a grade sofrerá deformações progressivas de acordo com a deformação selecionada.
Os escores das deformações relativas de pares de deformações relativas consecutivas são mostrados em gráficos de dispersão, e todos os escores podem ser mostrados em uma matriz numérica.

O algoritmo para o cálculo das deformações relativas é tirado de Dryden & Mardia (1998).

Referência

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

Tamanho a partir de pontos de referência – 2D ou 3D (Size from landmarks – 2D or 3D)

Coordenadas digitalizadas *x*/*y* ou *x*/*y*/*z* de pontos de referência. Espécimes em linhas, coordenadas com valores alternados de *x* e *y* (e *z* para 3D) em colunas. Não devem ser ajustadas por Procrustes ou normalizadas para tamanho!

Calcula o tamanho do centróide para cada espécime (norma Euclideana das distâncias de todos os pontos de referência até o centróide).

Os valores na coluna *"Normalized"* são tamanhos de centróide divididos pela raiz quadrada do número de pontos de referência – pode ser útil para comparar espécimes com diferentes quantidades de pontos de referência.

Normalizar tamanho – Normalize size

A opção *"Normalize size"* no menu Transform permite a remoção do tamanho ao dividir os valores das coordenadas pelo tamanho do centróide de cada espécime. Para dados 2D também podem ser usadas coordenadas de Procrustes, que são também normalizadas em relação ao tamanho.

Veja Dryden & Mardi (1998), p. 23-26.

Referência

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

Distância a partir de pontos de referência – 2D ou 3D (Distance from landmarks – 2D or 3D)

Coordenadas digitalizadas *x*/*y* ou *x*/y/*z* de pontos de referência. Espécimes em linhas, coordenadas com valores alternados de *x* e *y* (e *z* para 3D) em colunas. Podem ou não ser ajustados por Procustes ou normlaizados para tamanho.

Calcula as distâncias Euclideanas entre dois pontos de referência definidos para um ou muitos espécimes. Você precisa escolher estes pontos de referência – estes são nomeados de acordo com a primeira coluna do ponto de referência (valor de *x*).

Todas as distâncias a partir de pontos de referência – EDMA (All distances from landmarks – EDMA)

Coordenadas digitalizadas *x*/*y* ou *x*/*y*/*z* de pontos de referência. Espécimes em linhas,m coordenadas com valores alterados de *x* e *y* (e *z* para 3D) em colunas. Podem ou não ser ajustadas por Procrustes ou normalizadas para tamanho.

A função irá substituir os dados de pontos de referência por uma matriz de dados composta por distâncias entre todos os pares de pontos de referência, com um espécime por linha. O número de pares é *N(N-1)/2* para *N* pontos de referência. A transformação irá permitir análise multivariada de dados de distância, que não são sensíveis à rotação ou translação dos espécimes originais, de modo que o ajuste de Procrustes não é indispensável antes desta análise. O uso de dados de distância também permite a

transformação em logaritmo, assim como análise ou ajuste da equação alométrica para pares de distância.
Dados ausentes: suporte por substituição pela média da coluna.

Ligação de pontos de referência (Landmark linking)

Esta função do menu Geomet permite a escolha de qualquer par de pontos de referência a ser ligado com linhas nos gráficos morfométricos (polinômios de placa fina, deformações parciais e relativas, etc), para melhorar a leitura. Os pontos de referência devem estar presentes na planilha principal antes que as ligações possam ser definidas.
Pares de pontos de referência são selecionados ou desmarcados clicando na matriz simétrica. O conjunto de ligações também pode ser salvo em um arquivo de texto. Repare que há pouca checagem de erros neste módulo.

Strat menu

Associações unitárias (Unitary associations)

Análise de Associações Unitárias (*Unitary Associations analysis* – Guex 1991) é um método de correlação bioestratigráfica (veja Angiolini & Bucher 1999 para uma aplicação típica). O *input* de dados consiste de uma matriz de presença/ausência com amostras em linhas e táxons em colunas. Amostras pertencentes à mesma seção (localidade) devem ser marcadas com a mesma cora e ordenadas estratigraficamente dentro de cada seção, de tal modo que a amostra mais profunda seja colocada na última linha da seção. Cores podem ser reutilizadas em conjuntos de dados com um número grande de seções.

Descrição geral do método

O método de Associações Unitárias é lógico, mas um tanto complicado, sendo composto por uma série de passos. Para detalhes, veja Guex (1991). A implementação no PAST inclui grande parte dos aspectos do programa original, chamado BioGraph (Savary & Guex 1999), e graças a uma colaboração frutífera com Jean Guex o módulo no Past também inclui uma série de opções e melhorias que não são encontradas na versão atual daquele programa.

A idéia básica é gerar uma série de zonas de assembleia (similares às "zonas Oppel") ótimas, no sentido de que elas dão a máxima resolução estratigráfica com o mínimo de contradições de superposição (*superpositional contradictions*). Um exemplo de uma contradição assim seria uma seção contando a espécie A acima da espécie B, enquanto a assembléia 1 (que contém a espécie A) fica baixo da assembléia 2 (que contém a espécie B). PAST (e BioGraph) fazem a análise pelos seguintes passos:

1. **Horizontes residuais máximos (*Residual maximal horizons*)**
   O método assume a premissa de *range-through*, o que significa que se considera que os táxons estiveram presentes em todos os níveis entre a primeira e a última aparição em cada seção. A seguir, qualquer amostra com um conjunto de táxons que esteja contido dentro de outra amostra é descartada. As amostras restantes são chamadas de *horizontes residuais máximos*. A idéia por trás do descarte de dados é que o táxon ausente na amostra descartada pode simplesmente não ter sido encontrado mesmo que ele tenha existido originalmente. Assim, ausências não são tão informativas quanto as presenças.

2. **Sobreposição e co-ocorrência de táxons**
   A seguir, as relações de superposição entre todos os pares (A,B) de táxons são investigados: A abaixo de B, B abaixo de A, A junto de B, ou desconhecido. Caso A ocorra abaixo de B em uma localidade e B ocorra abaixo de A em outra localidade, eles são considerados co-ocorrentes apesar de nunca terem sido encontrados, de fato, juntos.
   As sobreposições e co-ocorrências de táxons podem ser vistos no *gráfico bioestratigráfico*. Neste gráfico, táxons são codificados numericamente. Co-ocorrências entre pares de táxons são mostrados com linhas azuis contínuas.

Sobreposições são mostradas como linhas vermelhas tracejadas, com traços longo para o táxon que ocorre acima e traços curtos para o táxon que ocorre embaixo. Alguns táxons podem ocorrer nos chamados *sub-gráficos proibidos* (*forbidden sub-graphs*), o que indica inconsistências nas suas relações de sobreposição. Dois de uma série de tipos de grafos como esses podem ser plotados no PAST: *ciclos* C_n *(*C_n *cycles)*, que são ciclos de sobreposição (A->B->C->A), e *circuitos* S_3 *(*S_3 *circuits)*, que são inconsistência do tipo "A co-ocorrendo com B, C acima de A, e C abaixo de B". Interpretações de grafos proibidos são sugeridas por Guex (1991).

3. Cliques máximos (*Maximal cliques*)

*Cliques máximos*são grupos de táxons co-ocorrentes que não estão contidos dentro de outro grupo de táxons co-ocorrentes. Os cliques máximos são candidatos ao *status* de associações unitárias, mas sofrerão processamento adicional subsequente. No PAST, cliques máximos recebem um número e também são nomeados por um horizonte máximo no conjunto de dados originais que seja idêntico ao, ou contido no (marcado com um asterisco) clique máximo.

4. Sobreposição de cliques máximos

As relações de sobreposição entre cliques máximos são decididas por inspeção de relações de sobreposição entre os seus táxons constituintes, como calculado no passo 2. Contradições (alguns táxons no clique A ocorrem abaixo de alguns táxons do clique B, e vice-versa) são resolvidas por um "voto majoritário". As contradições entre cliques podem ser visualizadas no PAST.

As sobreposições e co-ocorrências de cliques podem ser vistas no *grafo de cliques máximos (maximam clique graph)*. Neste gráfico, cliques são codificados por números. Co-ocorrências entre pares de cliques são mostradas como linhas azuis contínuas. Sobreposições são mostradas como linhas vermelhas tracejadas, com traços longos do clique que ocorre acima e traços curtos do clique que ocorre abaixo. Além disso, ciclos entre cliques máximos (ver abaixo) podem ser visualizados como linhas verdes.

5. Ciclos resolventes (*Resolving cycles*)

Pode acontecer de os cliques máximos serem ordenados em ciclos: A abaixo de B, que é abaixo de C, que é novamente abaixo de A. Isso é claramente contraditório. O "elo mais fraco" (relação de sobreposição que recebe suporte do menor número de táxons) nestes ciclos é destruído.

6. Redução para um caminho único

Neste estágio, idealmente devemos ter um único caminho (cadeia) de relações de sobreposição entre cliques máximos, do topo ao fundo. No entanto, frequentemente isso não acontece, por exemplo, quando A e B ficam abaixo de C, que fica abaixo de D, ou se temos caminhos isolados sem relações (A abaixo de B e C abaixo de D). Para produzir um único caminho, é necessário unir cliques de acordo com regras especiais.

7. Pós-processamento dos cliques máximos

Finalmente, uma série de manipulações menores são feitas para "polir" o resultado: Geração da propriedade de "uns consecutivos" (*"consecutive ones"*), reinserção de co-ocorrências e sobreposições virtuais residuais, e compactação para remover quaisquer cliques não-máximos que tenham sido gerados. Detalhes sobre estes procedimentos podem ser encontrados em Guex (1991). Finalmente, agora nós temos as Associações Unitárias, que podem ser visualizadas no PAST.
As associações unitárias têm associado a elas um índice de similaridade de uma AU para a próxima, conhecido por D:
$D_i = |AU_i - AU_{i-1}| / |AU_i| + |AU_{i-1} - AU_i| / |AU_{i-1}|$

8. Correlação usando Associações Unitárias
As amostras originais são agora correlacionadas por meio das associações unitárias. Uma amostra pode conter táxons que a coloquem unicamente em uma associação unitária, ou ela pode não ter táxons-chave que a diferenciariam entre duas ou mais associações unitárias. Neste último caso, só é fornecida uma extensão das possíveis associações unitárias. Estas correlações podem ser visualizadas no PAST.

9. Matriz de reprodutibilidade (*Reproducibility matrix*)
Algumas associações unitárias podem ser identificadas em apenas uma ou poucas seções, e neste caso pode ser considerada a possibilidade de unir associações unitárias para melhorar a reprodutibilidade geográfica (ver abaixo). A matriz de reprodutibilidade deve ser inspecionada para identificar associações unitárias como essas. A AU que só é identificada unicamente em uma seção é mostrada como um quadrado preto, enquanto as extensões de AUs (como dadas na lista de correlações) são mostradas em cinza.

10. Grafo de reprodutiblidade (*Reproducibility graph*) e junções sugeridas de AUs (biozonação)
O gráfico de reprodutibilidade (Gk' em Guex 1991) mostra a sobreposição das associações unitárias que são de fato observadas nas seções. O PAST irá reduzir internamente este grafo a um único caminho máximo (Guex 1991, seção 5.6.3), e neste processo também pode juntar algumas AUs. Estas junções são mostradas como linhas vermelhas no grafo de reprodutibilidade. A sequência de AUs únicas e juntadas pode ser vista como uma biozonação sugerida.

Funcionalidade especial
A implementação do método das Associações Unitárias no PAST inclui uma série de opções e funções que ainda não foram descritas na literatura. Para questões sobre estas, favor nos contatar.

Referências
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.

Ranqueamento-Escalonamento (Ranking-Scaling)

Ranqueamento-Escalonamento (Agterberg & Gradstein 1999) é um método de bioestratigrafia quantitativa baseado em *eventos* em uma série de poços (*wells*) ou seções (*sections*). O *input* de dados consiste de poços em linhas, com um poço por linha, e eventos (e.g. FADs e/ou LADs – datums de primeiro e último aparecimento) em colunas. Os valores na matriz são profundidades de cada evento em cada poço, aumentando para cima (você pode querer usar valores negativos para conseguir isso). Ausências são codificadas por zero. Caso apenas a ordem dos eventos seja conhecida, esta pode ser codificada como números inteiros crescentes (*ranks*, com possíveis números repetidos (*ties*) para eventos co-ocorrentes) dentro de cada poço.
A implementação do ranqueamento-escalonamento no PAST não é abrangente, e usuários avançados podem querer usar os programas RASC e CASC de Agterberg e Gradstein.

Visão geral do método

O método de Ranqueamento-Escalonamento é feito em dois passos:

1. Ranquamento
O primeiro passo do Ranqueamento-Escalonamento é produzir uma ordem única e abrangente dos eventos, mesmo que os dados contenham contradições (evento A acima de B em um poço, mas B acima de A em outro) ou ciclos mais compridos (A acima de B acima de C acima de A). Isso é feito por um "voto majoritário", contando o número de vezes que cada evento ocorre acima, abaixo ou junto de todos os outros. Tecnicamente, isso é efeito por *Presorting* (Pré-ordenamento) seguido pelo Método Modificado de Hay (*Modified Hay Method*) (Agterberg & Gradstein 1999).

2. Escalonamento
A análise bioestratigráfica pode acabar no ranqueamento, mas informações adicionais podem ser adquiridas estimando as distâncias estratigráficas entre eventos consecutivos. Isso é feito contando o número de relações de sobreposição observadas (A acima ou abaixo de B) entre cada par (A, B) de eventos consecutivos. Um baixo número de contradições implica uma distância grande.
Algumas distâncias calculadas podem aparecer como negativas, iniciando que a ordem dada no passo de ranqueamento não foi ótima. Caso isso aconteça, os eventos são reordenados e as distâncias são recalculadas para certificar que haja apenas distâncias positivas entre eventos.

RASC no PAST

Parâmetros

- Limiar de poços (*Well threshold*): O número mínimo de poços em que o evento deve ocorrer para ser incluído na análise.
- Limiar de pares (*Pair threshold*): O número mínimo de vezes que uma relação entre eventos A e B deve ser observada em sequência para que o par (A,B) seja incluído no passo de ranqueamento

- Limiar de escalonamento (*Scaling threshold*): Limiar de pares para o passo de escalonamento
- Tolerância (*Tolerance*): usado no passo de ranqueamento (ver Agterberg & Gradstein)

Ranqueamento
É fornecida a ordem dos eventos depois do passo de ranqueamento, com o primeiro evento no fundo da lista.

Escalonamento
É fornecida a ordem dos eventos depois do passo de escalonamento, com o primeiro evento aparecendo no fundo da lista. Para uma explicação de todas as colunas, ver Agterberg & Gradstein (1999).

Distribuição de eventos (*Event distribution*)
Um gráfico mostrando o número de eventos em cada poço, com os poços ordenados de acordo com o número de eventos.

Gráficos de dispersão (*Scattergrams*)
Para cada poço, a profundidade de cada evento no poço é plotada em relação à sequência ótima (depois do escalonamento). Idealmente, os eventos devem ser colocados em uma sequência ascendente.

Dendrograma
Gráfico das distâncias entre eventos na sequência escalonada, incluindo um dendrograma que pode auxiliar na zonação.

Referência
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 (Otimização Restrita)

Tabela de profunidades/níveis, com poços/seções em linhas e pares de eventos em colunas: FADs (*First Appearance Datums* – Datums do Primeiro Aparecimento) em colunas ímpares e LADs (*Last Appearance Datums* – Datum do Último Aparecimento) em colunas pares. Eventos faltantes são codificados por zero.
O PAST inclui uma versão simples da Otimização Restrita (*Constrained Optimization* – Kemple et al. 1989). Tanto FAD quanto LAD de cada táxon devem ser especificados em colunas alternadas. Usando o assim chamado Arrefecimento Simulado (*Simulated Annealing*), o programa procura por uma sequência global (composta) de eventos que implique na menor possível quantidade total do aumento de extensão (*range extension*) (penalidade) nos poços/seções individuais. Os parâmetros do procedimento de otimização inclui uma temperatura inicial de arrefecimento, o número de passos de resfriamento, a razão de resfriamento (porcentagem, menor que 100), e o número de testes (*trials*) por passo. Para explicação e recomendações, ver Kemple et al. (1989).

A janela de *output* inclui uma história de otimização com a temperatura e a penalidade em função do passo de resfriamento, a solução global composta e as extensões que ela implica para cada seção individual.
A implementação de CONOP no PAST é baseada código de otimização em FORTRAN fornecido por Sadler e Kemple.

Referência

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.

Ordenação de Eventos de Aparecimento (Appearance Event Ordination

Ordenação de Eventos de Aparecimento (Alroy 1994, 2000) é um método de seriação e correlação bioestratigráfca. O *input* de dados é no mesmo formato que para Associações Unitárias, consistindo de uma matriz de presen/ca/ausência co amostras em linhas e táxons em colunas. Amostras pertencendo à mesma seção (localidade) devem ser marcadas com a mesma cor, e ordenadas estratigraficamente dentro de cada seção de modo que a amostra mais profunda seja colocada na linha de baixo. Cores podem ser reutilizadas em conjuntos de dados com um grande número de seções.
A implementação no PAST é baseada em código fornecido por John Alroy. Ele inclui OEA de máxima verossimilhança (*Maximum Likelihood AEO*) (Alroy 2000)

Referências

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.

Curva de diversidade (Diversity curve)

Matriz de abundância ou presença/ausência com amostras em linhas (amostra mais profunda na última linha) e táxons em colunas.
Encontrado no menu "Strat", esta ferramenta simples permite plotar curvas de diversidade a partir de dados de ocorrência na coluna estratigráfica. Repare que as amotras precisam estar em ordem estratigráfica, com a amostra menos profunda (mais jovem) na linha de cima. Datas são sujeitas à premissa *ranghe-through* (ausências entre o primeiro e o último aparecimento são tratadas como presenças). Originações (*originations*) e extinções são em números absolutos, não em porcentagem.
A opção de "Correção de ponto final" ("*Endpoint correction*") atribui à primeira ou última ocorrência (FAD ou LAD) em uma amostra o peso de 0.5 ao invés de 1 naquela amostra. Um ponto que seja ao mesmo tempo FAD e LAD (singleton) na amostra conta como 0.33. Veja Hammer & Harper (2006).

Referência

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

Intervalos de confiança de extensão (Range confidence intervals)

Estimativa de intervalos de confiança para o primeiro e último aparecimento ou para a extensão total, para um táxon. Assume distribuição aleatória de horizontes fossilíferos ao longo da coluna estratigráfica ou ao longo do tempo. Requer amostragem contínua de seções.

Assumindo uma distribuição aleatória (Poisson) de horizontes fossilíferos, intervalos de confiança para a extensão estratigráfica de um táxon podem ser calculados a partir do datum (nível) do primeiro aparecimento, datum do último aparecimento e do número total de horizontes em que este táxon é encontrado (Strauss & Sadler 1989, Marshall 1990).

Nenhum dado precisa ser inserido na planilha. O programa irá perguntar pelo número de horizontes em que o táxon é encontrado, e os níveis ou datas da primeira e da última aparição. Se necessário, use valores negativos para certificar que o datum do último aparecimento tenha um valor numérico mais elevado do que o datum do primeiro aparecimento. Intervalos de confiança de 80%, 95% e 99% são calculados para o FAD (datum do primeiro aparecimento) isolado, para o LAD (datum do último aparecimento) isolado e para a extensão total. O valor de α é o comprimento do intervalo de confiança dividido pelo comprimento da extensão observada.

Para o caso de um único ponto final (*endpoint*):

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

onde C_1 é o intervalo de confiança e H é o número de horizontes fossilíferos.

Para o caso de pontos terminais juntos (*joint endpoint*) (extensão total), α é encontrado por solução iterativa da equação

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

Leve em consideração que a premissa de distribuição uniforme será violada em muitas situações reais.

Referências

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.

Intervalos de confiança da extensão livres de distribuição (Distribution-free range confidence intervals)

Estimativa de intervalos de confiança para o primeiro e último aparecimento. Não assume correlação entre posição estratigráfica e tamanho da lacuna (*gap size*). Requer amostragem contínua de seções. Espera uma coluna por táxon, com níveis ou datas de todos os horizontes onde o táxon é encontrado.

O programa fornece os limites superior e inferior dos comprimentos dos intervalos de confiança, usando uma probabilidade de 95% de confiança para níveis de confiança de 50, 80 e 95 porcento. Valores que não podem ser calculados são marcados com um asterisco (ver Marshall 1994).

Referência

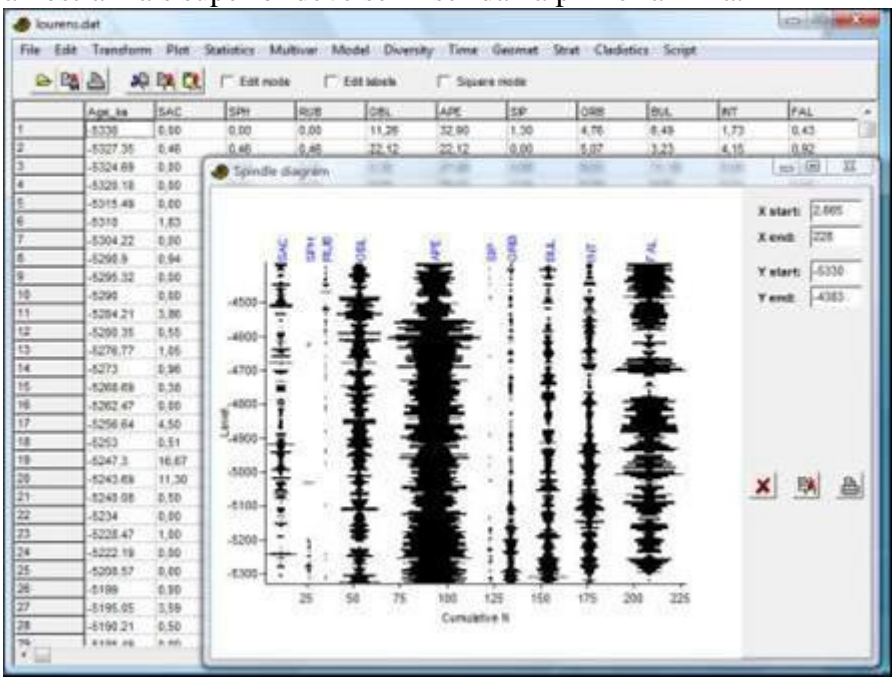
Marshall, C.R. 1994. Confidence intervals on stratigraphic ranges: partial relaxation of the assumption of randomly distributed fossil horizons. *Paleobiology* 20:459-469.

Diagrama de carretel (Spindle diagram)

Tipo padrão de gráfico usado na paleontologia para ilustrar a abundância de táxons fósseis ao longo de uma seção ou um núcleo estratigráfico. Amostras são colocadas em linhas, táxons em colunas. O programa irá perguntar se a primeira coluna contém níveis estratigráficos (e.g. metros ou anos).

Caso os níveis estratigráficos sejam fornecidos, cada caixa será desenhado de um dado nível até a próxima linha. Assim, uma última linha adicional (*dummy*) deve ser fornecida, com um nível estratigráfico final mas com zero para todos os táxons. Caso os níveis sejam apresentados em profundidades ou idades, números negativos devem ser usados para certificar que a figura esteja orientada de modo correto.

Se os níveis estratigráficos não forem fornecidos, todas as caixas terão a mesma altura. A amostra mais superior deve ser inserida na primeira linha.



Cladistics

Análise de parcimônia (Parsimony analysis)

Aviso: o pacote Cladistics nos PAST é totalmente operacional, mas não tem uma funcionalidade abrangente. O algoritmo eurístico parece não ter um desempenho tão com como em alguns outros programas (isso está sendo investigado). O pacote cladístico do PAST é adequado para ensino e para exploração inicial dos dados, mas para trabalho mais "sério" recomendamos um programa especializado, como o PAUP.
Análise semi-objetiva das relações entre táxons a partir de evidência morfológica ou genética.
Matriz de caracteres com táxons em linhas, grupo externo (*outgroup*) na primeira linha.
Para o cálculo de índices de congruência estratigráfica (*stratigraphic congruence indicies*), os datums de primeiro e último aparecimento devem ser dados nas duas primeiras colunas.
Algoritmos são de Kitching et al. (1998).
Estados de caracteres devem ser codificados usando números inteiros de 0 a 255 ou letras c, a, g, t, u (maiúsculas ou minúsculas). O primeiro táxon é tratado como grupo externo e será colocado na base da árvore.
Valores ausentes são codificados por ponto de interrogação (?) ou por -1. Repare que o PAST não colapsa ramos com comprimento zero. Por causa disso, valores ausentes podem levar a uma proliferação *ad nauseam* de árvores igualmente curtas, muitas das quais serão na verdade equivalente.
Quatro algoritmos são disponíveis para encontrar as árvores mais curtas:

Branch-and-bound ("ramificar e unir")
O algoritmo *branch-and-bound* garante encontrar todas as árvores mais curtas. O número total de árvores mais curtas é relatado, mas um máximo de 10000 árvores são salvas. O algoritmo *branch-and-bound* pode consumir muito tempo para conjuntos de dados com mais de 12 táxons.
Exaustivo (*Exhaustive*)
O algoritmo exaustivo avalia todas as árvores possíveis. Como o algoritmo *branch-and-bound*, ele irá necessariamente encontrar todas as árvores mais curtas, mas ele é muito lento. Para 12 táxons, mais de 600 milhões de árvores são avaliadas! A única vantagem que tem em relação ao *branch-and-bound* é a plotagem da distribuição de comprimentos de árvores. Este histograma pode indicar a "qualidade" da sua matriz, no sentido de que deveria haver uma cauda para a esquerda, de tal modo que poucas árovres curtas são "isoladas" da massa maior de árvores mais compridas (mas veja Kitching et al. 1998 para comentários críticos sobre isso). Para mais de 8 táxons, o histograma é baseado em um subconjunto de comprimentos de árvores e pode não ser preciso.

Heurístico, intercâmbio de vizinhos mais próximos (*Heuristic, nearest neighbour interchange*)
Este algoritmo heurístico adiciona táxons sequencialmente na ordem em que eles aparecem na matriz ao ramo em que isso produzirá o menor aumento no comprimento da

árvore. Depois da adição de cada táxon, todas as árvores vizinhas mais próximas são permutadas na tentativa de encontrar uma árvore ainda mais curta.
Como todas as buscas heurísticas, este algoritmo é muito mais rápido do que os algoritmos acima e pode ser usado para quantidades grandes de táxons, mas não há garantia de que ele encontre todas ou alguma das árvores mais parcimoniosas. Para reduzir a probabilidade de acabar em um sub-ótimo local mínimo, um número de *reordenamentos (reorderings)* pode ser especificado. Para cada reordenamento, a ordem de entrada dos táxons será permutada aleatoriamente e será feita uma nova busca heurística.
Repare: Por causa da reordenação aleatória, as árvores encontradas pelas buscas heurísticas normalmente serão diferentes em cada rodada. Para reproduzir exatamente uma busca, você precisa começar o modo de parcimônia novamente do menu, usando o mesmo valor para "semente aleatória" (*"Random seed"*). Isso irá reiniciar o gerador de números aleatórios para o valor da semente.

Heurístico, corte e retransplante de subárvores (*Heuristic, subtree pruning and regrafting)*
Este algoritmo (SPR) é similar ao acima (NNI), mas com um esquema mais elaborado de permutação de ramos: Uma subárvore é cortada da árvore e replantada em todos os outros ramos da árvore na tentativa de achar uma árvore mais curta. Isso é feito depois da adição de cada táxon e para todas as subárvores possíveis. Apesar de mais lento que o NNI, SPR frequentemente irá encontrar árvores mais curtas.

Heurístico, bissecção e reconexão de árvores (*Heuristic, tree bisection and reconnection)*
Este algoritmo (TBR) é similar ao acima (SPR), mas com um esquema ainda mais complexo de permuta de ramos. A árvore é dividida em duas partes, e estas são reconectadas por todos os pares de ramos possíveis para encontrar uma árvore mais curta. Isso é feito depois da adição de cada táxon e para todas as divisões possíveis da árvore. TBR frequentemente irá encontrar árvores mais curtas do que SPR e NNI ao custo de um maior tempo de cálculo.

Critérios de otimização de carateres (*Character optimization criteria*)
Três algoritmos diferentes para de otimização são disponíveis:

Wagner
Caracteres são reversíveis e ordenados, significando que 0->2 custa mais do que 0->1, mas tem o mesmo custo que 2->0.

Fitch
Caracteres são reversíveis e não-ordenados, significando que todas as mudanças têm o mesmo custo. Isso é o critério com o menor número de premissas, e, portanto, normalmente é preferível.

Dollo

Caracteres são ordenados, mas a aquisição de um estado de caráter (de um valor mais baixo para um mais alto) pode acontecer uma única vez. Toda homoplasia é representada por reversões (*reversals*) secundárias. Assim, 0->1 pode acontecer uma única vez, normalmente relativamente próximo à base da árvore, mas 1->0 pode acontecer qualquer número de vezes árvore acima. (Essa definição foi debatida na lista de emails do PAST, especialmente quanto à necessidade de ordenação dos caracteres Dollo).

Bootstrap

Bootstrap é feito quando o valor de "Réplicas bootstrap" (*"Bootstrap replicates"*) é colocado em um valor diferente de zero. O número especificado de réplicas (tipicamente 100 ou até 1000) da sua matriz de caracteres é feito, cada um com caracteres recebendo pesos arbitrários. Uma réplica fornece suporte ao grupo se o grupo existe na árvore de consenso majoritário (*majoritary rule consensus tree*) das árvores mais curtas feitas pela réplica.

Aviso: Especificar 1000 réplicas por *bootstrap* claramente resulta em um tempo de cálculo 1000 vezes maior do que sem bootstrap! Busca exaustiva com bootstrap não é realística e não é permitida.

Plotagem de cladograma (*Cladogram plotting*)

Todas as árvores mais curtas (mais parcimoniosas) podem ser visualizadas, até um máximo de 10000 árvores. Caso tenha sido feito bootstrap, um valor de bootstrap é dado na raiz da subárvore que especifica cada grupo.

Estados de caracteres podem ser plotados na árvore, como selecionado pelo botão *"Character"*. Esta reconstrução de caracteres só é única na ausência de homoplasia. No caso de homoplasia, mudanças de caracteres são colocadas o mais próximo possível da raiz, favorecendo aquisição em um único tempo com reversão subsequente de um estado de caráter ao invés de mais de uma aquisição independentes (conhecido como *transformação acelerada – accelerated transformation*).

A opção "Filograma" (*"Phylogram"*) permite plotar árvores onde o comprimento das linhas verticais (juntando clados) é proporcional ao comprimento dos ramos.

Índice de consistência (*Consistency index*)

O índice de consistência por caractere (*per-character consistency index* – ci) é definido por m/s, onde m é o menor número possível de mudanças de caracteres (passos) em qualquer árvore e s é o número de passos de fato observado na árvore atual. Este índice, portanto, varia de 1 (sem homoplasia) e desce até zero (muita homoplasia). O índice de consistência de assembléia (*ensemble consistency index* – CI) é um índice similar somado para o conjunto de caracteres.

Índice de retenção (*Retention index*)

O índice de retenção por caractere (ri) é definido como $(g-s)/(g-m)$, onde m e s são como definidos para o índice de consistência e g é o número máximo de passos para o caractere em qualquer cladograma (Farris 1989). O índice de retenção mede a sinapomorfia da árvore e varia de 0 a 1.

Repare que não versão atual o índice de retenção só é calculado corretamente quando a otimização de Fitch é usada.

Árvore de consenso (*Consensus tree*)
A árvore de consenso de todas as árvores mais curtas (mais parcimoniosas) também pode ser vista. Duas regras de consenso são implementadas: Estrito (*Strict* – grupos suportados por todas as árvores) e majoritário (*majority* – grupos devem ser suportando por mais de 50% das árvores).

Suporte de Bremer (índice de decaimento) (*Bremer support – decay index*)
O suporte de Bremer para um clado é o número extra de passos que são necessários para construir uma árvore (consistente com os caracteres) sem aquele clado. Existem razões para dar preferência a este índice ao invés do valor de bootstrap. O PAST não calcula diretamente o suporte de Bremer, mas para conjuntos de dados menores isso pode ser feito "manualmente" da seguinte maneira:

- Faça uma análise de parcimônia por busca exaustiva ou *branch-and-bound*. Anote os clados e o comprimento *N* da(s) árvore(s) mais curta(s) (por exemplo 42). Caso haja mais de uma árvore mais curta, olhe a árvore de consenso estrito. Caldos que não são mais encontrados na árvore de consenso têm um valor de suporte de Bremer igual a 0.
- Na caixa para "Árvore mais longa mantida" (*"Longest tree kept"*), coloque o número *N*+1 (43 no nosso exemplo) e faça uma nova busca.
- Clados adicionais que não são mais encontrados na árvore de consenso estrito têm um valor de suporte de Bremer igual a 1.
- Para "Árvore mais longa mantida", coloque o número *N*+2 (44) e faça uma nova busca. Clados que agora desaparecem da árvore de consenso têm um valor de suporte de Bremer igual a 2.
- Continue até que todos os clados tenham desaparecido.

Índice de congruência estratigráfica (*Stratigraphic congruence indices*)
Para calcular índices de congruência estratigráfica, as duas primeiras colunas na matriz de dados devem conter os datums de primeiro e último aparecimento, respectivamente, para cada táxon. Estes datums devem ser fornecidos de tal modo que idade mais jovem (ou o nível estratigráfico mais alto) tenha o maior valor numérico. Pode ser necessário usar valores negativos para conseguir isso (e.g. 400 milhões de anos antes do presente é codificado como -400.0). A boxa "*FADs/LADs in first columns*" na caixa de diálogo *Parsimony* deve ser marcada.
O *Índice de Congruência Estratigráfica (SCI)* de Huelsenbeck (1994) é definido como a proporção de nós estratigraficamente consistentes no cladograma, e varia de 0 a 1. Um nó é estratigraficamente consistente quando a primeira ocorrência mais antiga acima do nó tem a mesma idade ou é mais jovem do que a primeira ocorrência no seu táxon (nó) irmão.
O *Índice de Completude Relativa (Relative Completeness Index – RCI)* de Benton & Storrs (1994) é definido como (1-MIG/SRL)x100%, onde MIG (*Minimum Iplied Gap* – Lacuna Mínima Implícita) é a soma das durações de de extensões-fantasma (*ghost ranges*) e SRL é a soma das durações das extensões observadas. O RCI pode ser negativo, mas normalmente varia de 0 a 100.

A *Razão de Excesso de Lacunas (Gap Excess Ratio – GER)* de Wills (1999) é definida por $1-(MIG-G_{min})/G_{max}-G_{min})$ onde G_{min} é a menor somatória possível de extensões-fantasma em qualquer árvore (ou seja, a somatória das distâncias entre FADs consecutivos) e G_{max} é a maior somatória possível (ou seja, a somatória das distâncias do primeiro FAD a todos os outros FADs).
Estes índices são submetidos a um teste de permutação, onde todas as datas são redistribuídas aleatoriamente 1000 vezes entre os diferentes táxons. A proporção de permutações onde o índice recalculado excede o índice original é fornecida. Se pequena (e.g. $p<0.05$), isso indica um desvio estaticamente significativo da hipótese nula de não haver congruências entre o cladograma e a estratigrafia (em outras palavras, a congruência é significativa). As probabilidades de permutação de RCI e GER são iguais para qualquer conjunto de permutações, já que são baseadas no mesmo valor de MIG.

Referências

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.