OECOLOGIA BRASILIENSIS
Volume II: Tópicos em Tratamento de Dados Biológicos
PERES-NETO, P.R.; VALENTIN, J.L. & FERNANDEZ, F.A.S. (editores), 1995, p.27-55.
Programa de Pós-Graduação em Ecologia - Instituto de Biologia - UFRJ, Rio de Janeiro - RJ.

AGRUPAMENTO E ORDENAÇÃO

VALENTIN, J.L.

Resumo:

Em estudos de ecossistemas, os ecólogos enfrentam um grande número de variáveis bióticas e abióticas, ligadas por uma complexa rede de interrelações. Definir e descrever os padrões estruturais de um ecossistema, bem como formular hipóteses acerca de sua função requer análises multivariadas, como as análises de classificação e ordenação. Diferentes métodos de classificação e ordenação são listados e descritos sumariamente. Alguns comentários e recomendações são formulados quanto à escolha do índice de semelhança e a interpretação de dendrogramas e planos fatoriais.

Abstract:

"Ordination and cluster analysis"

In ecosystem studies, the ecologist has to face a great number of biotic and abiotic variables linked by a complex net of interrelations which rule the structure and the function of this ecosystem. Defining and describing structural patterns of a ecosystem, as well as formulating hypothesis on its function, requires a multivariate analysis of the data, like classification and ordination. This exposition aims a simple description of these technics and comments on some important points for an adequate use. A criterious choice of the similarity index must be done, depending on the nature of data and the objective of the work. Different cluster and ordination methods are listed with comments. Advise on reading dendrogram and factorial plans and interpreting results is introduced.

A medida da semelhança

As medidas de semelhança são grandezas numéricas que quantificam o grau de associação entre um par de objetos ou de descritores. A primeira utilização de uma medida de semelhança em estudos biológicos deve-se a Heincke (1898), mas foi com o índice floral de Jaccard (1908) que assistimos ao surgimento de um número cada vez maior dessas medidas. Muitos estudos sobre medidas de semelhança relacionadas a aplicações ecológicas podem ser consultados numa abundante bibliografia, dentro da qual podemos citar algumas das referências mais importantes que apresentam uma síntese sobre o assunto: Sneath & Sokal (1973), Wolda (1981), Legendre & Legendre (1983), Pielou (1984), Ludwig & Reynolds (1988).

Qual índice escolher?

A resposta a essa pergunta depende da resposta a uma série de outras perguntas:

- quer fazer um estudo comparativo entre amostras (modo Q) ou entre descritores (modo R)?
- os dados são qualitativos binários (presença-ausência), quantitativos merísticos (contagem de organismos), quantitativos contínuos (variável ambiental), semi-quantitativos (códigos de abundância) ou ordinais?
- a tabela é homogênea (= contingência) ou heterogênea (descritores com unidades diferentes)?

Daremos alguns exemplos de índices frequentemente usados em ecologia.

Para os estudos em modo Q. Neste tipo de estudo, que visa a relação entre objetos, os coeficientes mais utilizados na literatura são de similaridade e de distância (ou dissimilaridade).

a) Os índices binários e o problema da dupla-ausência. Os coeficientes de similaridade foram desenvolvidos inicialmente para medidas binárias e posteriormente generalizados a outros tipos de dados com o avanço da informática. Os índices binários mais usados em estudos ecológicos são os de Jaccard $S1 = \frac{a}{a+b+c}$ e de Sorensen $S2 = \frac{2a}{2a+b+c}$

onde a é o número de espécies comuns às duas amostras, b e c sendo o número de espécies ocorrendo em uma ou outra amostra (alternâncias). Esses índices variam entre 0 (nenhuma similaridade entre as duas amostras) e 1 (similaridade completa). Sorensen teria preferência sobre Jaccard quando se pretende valorizar a ocorrência simultânea de duas espécies. Neste dois índices não é considerada a dupla-ausência. Colocamos aqui o importante problema do significado ecológico do valor 0: a ausência de uma espécie na amostra indica realmente que esta espécie não existe no seu universo amostral, ou é simplesmente devida a uma deficiência metodológica

(seletividade do amostrador, amostra pequena demais...) ? É evidente que a ausência de "baleias" nas amostras de plâncton não pode ser levada em conta para comparar essas amostras ! Assim, nos estudos de comunidades é geralmente desaconselhado o uso de coeficientes que incluem a dupla-ausência.

b) Um índice de similaridade quantitativo excluindo a dupla-ausência. O índice do antropólogo Czekanowski, atribuído também ao matemático Steinhaus (Legendre & Legendre 1983) compara, para cada espécie, a menor abundância (W) entre as duas amostras e a média das abundâncias (A e B) nessas amostras: $S3 = \frac{2W}{A+B}$. Este índice, que varia entre 0 e 1, é derivado dos índices binários, pois quando aplicado a dados de presença-ausência, ele é igual ao índice de Sorensen. Ele foi muito utilizado nos estudos fitosociológicos, baseados nas taxas de recobrimento vegetal.

c) O índice de similaridade de Morisita. Este índice, atualmente muito usado, foi proposto por Morisita (1959) para medir a similaridade entre duas comunidades. Ele varia de 0 até um valor máximo próximo de 1. Inicialmente formulado para expressar a similaridade entre amostras de contagem de indivíduos, ele foi posteriormente simplificado tornando-o apropriado também para percentagens e valores de biomassa, recobrimento, produtividade, etc. A literatura considera o índice de Morisita um dos melhores para estudos ecológicos. Sua fórmula, relativamente complexa, é exemplificada em Krebs (1989).

d) Os coeficientes de distância. Eles são preferencialmente aplicados quando se pretende visualizar graficamente a proximidade entre duas amostras, em função da composição específica ou de qualquer outro descritor dessas amostras. Quanto mais próximas foram as amostras, i.e. menor a distância métrica entre os pontos representativos dessas duas amostras, maior será a similaridade entre elas. Um índice de distância corresponde então a uma dissimilaridade. Logo, é possível se transpor uma similaridade S para uma distância D fazendo, por exemplo, D=1-S. Entretanto, para ser realmente uma distância, no sentido métrico do termo, este coeficiente deve respeitar os seguintes axiomas de "metricidade": (1) $D_{A-B} = D_{B-A}$, (2) se A = B então $D_{A-B} = 0$, (3) se $A \neq B$ então $D_{A-B} > 0$ e (4) $D_{A-B} + D_{C-B} \geq D_{A-C}$ (regra do triângulo).

A distância mais conhecida, e perfeitamente métrica, pois baseada no teorema de Pitágoras sobre a hipotenusa do triângulo retângulo, é a Distância Euclidiana: $D_{A-B} = \sqrt{(x_{A1} - x_{B1})^2 + (x_{A2} - x_{B2})^2}$

D_{A-B} é a distância euclidiana entre as amostras *A* e *B*, em função da abundância *x* de duas espécies *1* e *2* (Fig. 3).

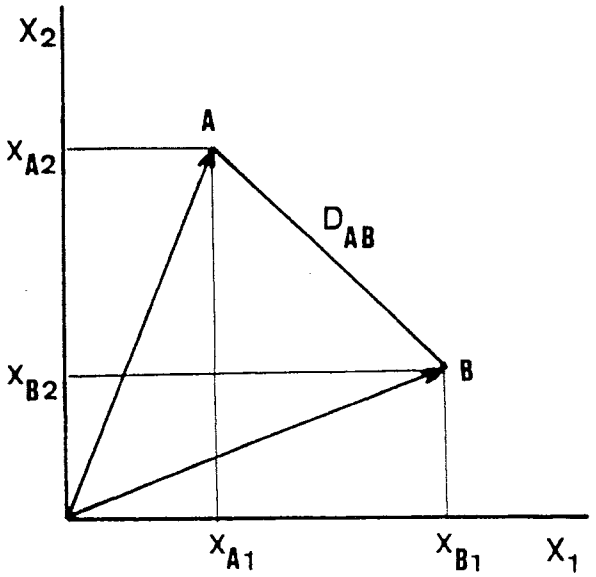

Fig. 3. Representação gráfica da distância euclidiana D_{AB} entre 2 objetos A e B em função dos descritores X1 e X2.

A distância euclidiana não tem limite superior. Ela aumenta à medida que aumenta o número de descritores. Além disso, ela depende da escala de valores de cada descritor. Este inconveniente pode ser corrigido pela padronização dos dados (dados centrados e reduzidos), e pelo uso da distância euclidiana média **D/*n*** onde *n* é o número de descritores. De maneira geral, a distância euclidiana deve ser evitada para comparar amostras quanto à abundância de espécies, principalmente quando ocorrer um grande número de duplas-ausências.

A Distância de Bray & Curtis (1957) é de uso frequente, por ser disponível na maioria dos pacotes estatísticos. Ela varia entre 0 (similaridade) e 1 (dissimilaridade). Este índice não considera as duplas-ausências e é fortemente influenciado pelas espécies dominantes. As espécies raras acrescentam muito pouco ao seu valor. Seu cálculo é baseado nas diferenças absolutas e nas somas das abundâncias de cada espécie (*i*) nas duas amostras:

$$D_{A-B} = \frac{\Sigma \left| x_{Ai} - x_{Bi} \right|}{\Sigma \left(x_{Ai} + x_{Bi} \right)}$$

Vários autores preferem definir esta medida como "Similaridade", fazendo (1-*D*). Neste caso, o índice de Bray & Curtis equivale ao coeficiente de similaridade de Czekanowski.

Outras medidas de distância podem ser encontradas na literatura, com formulação parecida à de Bray & Curtis, tais como as distâncias de Manhattan e de Camberra.

Para os estudos em modo R. O estudo em modo R de uma matriz de dados ecológicos tem por finalidade definir as relações entre descritores.

a) Descritores métricos. Os descritores métricos são aqueles nos quais é possível aplicar medidas de dependência paramétricas, i.e. que dependem dos parâmetros (média e desvio padrão) da sua distribuição de frequência.

O coeficiente de correlação linear **r** de **Pearson** é um dos mais usados para quantificar a dependência linear entre duas variáveis. Seu uso adequado exige, entretanto, certos cuidados: ele expressa exclusivamente a intensidade da relação linear entre duas variáveis, podendo ser submetido a um teste estatístico para verificar se seu valor é significativamente diferente de zero. Este teste exige a normalidade dos dados. Caso contrário é possível, seja aplicar uma transformação normalizante aos dados, seja utilizar um coeficiente não-paramétrico. O coeficiente **r** só pode ser aplicado em modo **R** (associação entre descritores), exceto para tabela homogêneas, do tipo contingência (dados de contagem ou de frequência), quando é então possível calcular **r** entre objetos (modo **Q**).

b) Descritores ordenados não-métricos. O coeficiente de correlação ρ (rhô) de Spearman é chamado "não-paramétrico" por ser aplicável em descritores não-métricos, cujas medidas são ordinais (postos) e, consequentemente, a distribuição de frequência não depende de média nem desvio padrão.

Este coeficiente pode ser também aplicado a descritores métricos, após transformação dos dados em "postos", mas com menos eficiência que o **r** de Pearson. Entretanto, no caso de descritores métricos em relação não linear, o ρ de Spearman seria mais eficiente (a "eficiência" de um coeficiente é a sua capacidade de detectar mais facilmente um relação entre descritores, i.e. rejeitar a hipótese nula de independência).

Da mesma maneira que o **r** de Pearson, o ρ de Spearman varia também entre -1 e +1, sendo o valor 0 a ausência de relação monótona. O cálculo de ρ é bastante simples, sua significância podendo ser testada da mesma maneira que o **r** de Pearson. Para comparar dois descritores com pequeno número de amostras (<10) existe uma tabela de significância de ρ.

A fórmula do coeficiente de correlação ρ de Spearman escreve-se:

$$\rho = 1 - \frac{6 \sum_{j=1}^{n} d_j^2}{(n^3 - n)}$$

onde d é a diferença entre os postos de cada amostra nos dois descritores e n o número de amostras. No caso, bastante frequente, de amostras com mesmo posto, atribui-se a cada um o valor médio dos postos. Se a quantidade de empates for muito elevado, deve ser aplicado um fator de correção no cálculo de ρ.

Como para o **r** de Pearson, é desaconselhado utilizar ρ em modo **Q,** pois neste modo a noção de posto não faz sentido quando os descritores são variáveis ambientais com unidades e escalas diferentes.

Em tabelas de contagem de organismos ocorre frequentemente um grande número de espécies raras. O valor real do posto de cada uma delas no ecossistema é bastante incerto e impreciso e, consequentemente, o cálculo de ρ entre duas amostras seria fortemente influênciado pelas espécies menos abundantes, geralmente mal amostradas.

O coeficiente τ (tau) de Kendall é um outro coeficiente de correlação de posto com aplicação idêntica ao ρ de Spearman. Ele é descrito em vários manuais de estatística não paramétrica.

Agrupamento

Há uma tendência normal do Ecólogo em procurar agrupar amostras de mesmas características bióticas ou abióticas, ou associar espécies em comunidades, de acordo com o objetivo do seu trabalho, visando com isso descrever, da maneira mais clara e sintética possível, a estrutura de um ecossistema, determinando a composição e a extensão das suas unidades funcionais.

Agrupar objetos consiste em reconhecer entre eles um grau de similaridade suficiente para reuni-los num mesmo conjunto. Os métodos ecológicos de agrupamento devem poder destacar os grupos de objetos similares entre si, deixando do lado os pontos intermediários que permaneçam geralmente entre os grupos quando a amostragem é suficientemente extensa. A não ser que o meio físico seja fortemente descontínuo e que a amostragem tenha sido realizada de cada lado de um forte gradiente, o ecólogo terá geralmente dificuldade em definir nitidamente grupos de amostras ou de espécies, em virtude do conceito de *continuum* que caracteriza os ecossistemas.

As etapas de uma análise de agrupamento (*Cluster Analysis* em inglês) são as seguintes:

1. coleta dos dados, que serão reunidos numa tabela com m colunas (descritores) e n linhas (objetos).
2. escolha do modo de análise: modo **Q** (agrupamento de objetos) ou modo **R** (agrupamento de descritores), de acordo com o objetivo do trabalho.
3. escolha do coeficiente de associação (similaridade, distância, dependência).

4. escolha do método de agrupamento, que depende de critérios baseados no menor grau de distorção (maior coeficiente cofenético), mas que na realidade, é geralmente definida pela sua disponibilidade nos pacotes estatísticos.
5. elaboração e interpretação do dendrograma.

Os métodos de agrupamento

A escolha do método de agrupamento é tão crítica quanto a escolha do coeficiente de associação. Dele dependerá a correta classificação de uma amostra dentro de um ou outro grupo já formado. Os métodos de agrupamento foram desenvolvidos a partir de modelos e dados diversos. Sneath & Sokal (1973) apresentam a seguinte classificação, dividindo os métodos em:

- **sequenciais** (os objetos são reunidos um após o outro, respeitando uma determinada sequência de operações), ou **simultâneos** (o agrupamento é realizado numa única etapa. É o caso da Ordenação que veremos a seguir).
- **aglomerativos** (os objetos inicialmente isolados são progressivamente reunidos em grupos sucessivos até formar um único grupo) ou **divisivos** (inicia-se com um único grupo o qual, em função de determinados critérios, é dividido em subgrupos, para chegar no final aos objetos individuais. É o processo aplicado nas chaves de identificação em taxonomia).
- **monotéticos** (baseados num único descritor a cada vez ou **politéticos** (baseados em vários descritores).
- **hierárquicos** (os elementos-objetos de um grupo tornam-se elementos do grupo superior, constituindo assim uma série hierarquizada) ou **não-hierárquicos** (procuram maximizar a homogeneidade intra-grupo, sem considerar a hierarquia entre grupos, tais como os métodos de ordenação e de otimização da matriz de associação).
- **probabilísticos**: recomendados para o agrupamento de espécies, eles são porém pouco usados em razão da complexidade dos cálculos e da necessidade de muito espaço de memória em computador. Com eles deve ser empregado o índice de similaridade probabilístico de Goodall (cf. Legendre & Legendre 1983). Os métodos de agrupamento probabilísticos permitem, para um conjunto de elementos, definir todos os grupos cuja matriz de associação intra-grupo tem uma certa probabilidade de ser homogênea. A vantagem destes métodos é de poder estabelecer grupos de maneira objetiva, baseando-se em probabilidade.

Nesta curta exposição mencionaremos a seguir, somente alguns dos métodos mais usados em Ecologia, que são os métodos de agrupamento politéticos não-probabilísticos, por aglomeração sequencial hierárquica. O critério básico da fusão entre um objeto e um grupo ou entre dois grupos, é sempre o mesmo: serão reunidos

os grupos que tem maior similaridade entre eles. O problema é: como calcular esta similaridade? O método de cálculo depende do método de aglomeração escolhido.

Método por ligação simples. Este método, também chamado de salto mínimo e de mais próximo vizinho, é de concepção simples, podendo ser realizado sem ajuda do computador. O dendrograma é montado a partir dos pares de objetos mais similares (os de menor distância), e em seguida, os objetos ou grupos já formados vão se reunir em função de similaridades decrescentes (ou de distâncias crescentes). O exemplo a seguir ilustra o método de aglomeração por ligação simples, a partir de uma matriz de distâncias entre 5 amostras (Tabela 2):

Tabela 2. Matriz de distâncias entre as amostras.

Objeto	(1)	(2)	(3)	(4)	(5)
(1)	0	0,40	0,10	0,40	0,37
(2)		0	0,48	0,36	0,20
(3)			0	0,42	0,40
(4)				0	0,18
(5)					0

O dendrograma será montado da maneira seguinte (Fig.4):

- no eixo horizontal (ou vertical, tanto faz), serão posicionadas as amostras. O intervalo entre elas é arbitrário, sem valor métrico.
- no eixo vertical ou (horizontal) são plotados os valores de distância, iniciando por 0.
- procura-se na matriz a menor distância (ou a maior similaridade): é a distância de valor 0,10, entre as amostras (1) e (3) que serão reunidas no dendrograma na altura do valor 0,10. É assim formado um primeiro núcleo, chamado (6).
- a segunda menor distância da matriz é entre as amostras (4) e (5) que devem ser posicionadas no eixo horizontal e reunidas no valor 0,18, constituindo um segundo núcleo chamado (7).
- a próxima distância é 0,20, entre a amostra (5) que já pertence ao grupo (7), e a amostra (2). A amostra (2) deve ser então reunida no grupo (7) ao nível da distância 0,20, formando o grupo (8).
- a próxima distância é 0,36 entre (2) e (4). Mas como (4) pertence ao grupo (7), já ligado a (2), passamos à próxima distância, que é 0,37 entre a amostra (1) pertencendo ao grupo (6) e a amostra (5) pertencendo ao grupo (8). Ficam assim ligados ao nível 0,37 de distância os grupos (6) e (8). O dendrograma ficou completo, com um grupo único (9) aglomerando todas as amostras.

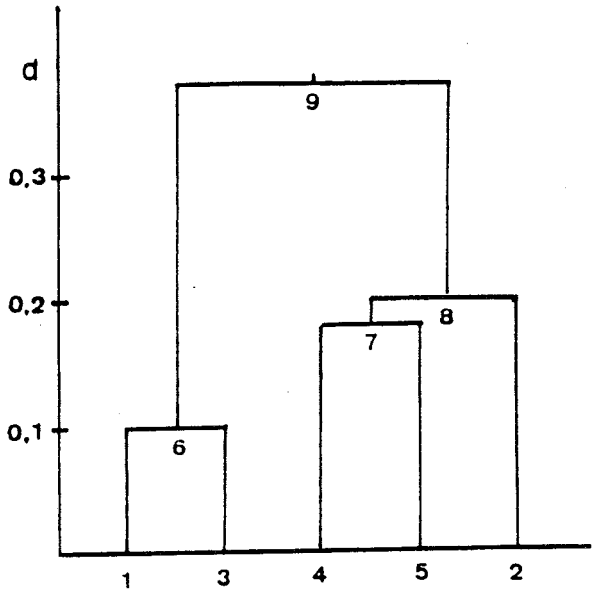

Fig. 4 Dendrograma elaborado a partir da matriz de distâncias en-tre 5 amostras (Tabela 2).

Em razão da sua simplicidade, este método apresenta grande desvantagem. O fato de reunir um objeto ao elemento "mais próximo" do grupo já formado, faz com que objetos intermediários entre os grupos sejam rapidamente aglomerados a esses. Ocorre então um encadeamento de objetos que dificulta a separação dos grupos. Nos estudos ecológicos onde as amostras de características intermediárias são geralmente numerosas, este método deve ser evitado.

Método por ligação completa. Este método, também chamado de "método de aglomeração pelo diâmetro", e de "vizinho mais distante", é o oposto do anterior. A fusão de dois grupos depende do par de objetos mais distantes. Em outros termos, um elemento fusionará a um grupo unicamente se for ligado a todos os elementos deste grupo. Com isso à medida que os grupos crescem, é cada mais difícil incluir elementos a esses grupos. O resultado é um dendrograma "dilatado", onde os grupos são facilmente evidenciados, mas onde a maior parte das amostras intermediárias permanece isolada. O método é recomendado em ecologia, quando o objetivo é descobrir fortes descontinuidades.

A Fig. 5 representa de maneira esquemática a diferença entre os métodos por "ligações simples" (I) e pelo "diâmetro" (II). Sejam dois grupos já formados, (A) composto de amostras (a), e (B) composto de amostras (b), e uma amostra isolada X. A que grupo irá se juntar esta amostra X, tal como posicionada no espaço? A resposta depende do método escolhido:

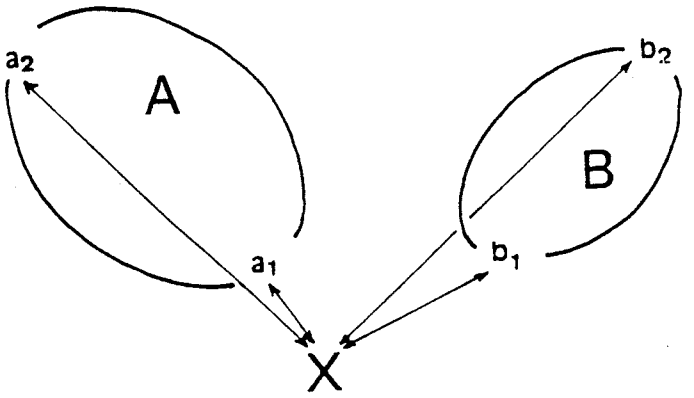

Fig.5. Atribuição de um objeto X a um dos dois grupos A e B de acordo com o método de aglomeração, ligações simples e ligações completas.

Pelo método (I), a amostra X será incluída no grupo A, em razão da sua proximidade com a amostra (a1) deste grupo, comparativamente com a amostra (b1) do grupo B. Pelo método (II), a amostra X será atribuída ao grupo B em razão de sua menor distância com a amostra (b2) deste grupo, comparativamente com a amostra (a2) do grupo A. As amostras (a2) e(b2) sendo as mais afastadas dos respectivos grupos, e situadas na extremidade do diâmetro, daí o nome do método.

Entre os métodos por ligações simples e por ligações completas, existem outros métodos que utilizam a média como critério de aglomeração, e que mencionaremos à seguir. São os métodos pela associação média e pelos pesos proporcionais.

Método pela associação média. Este método conhecido em inglês pelo nome de "arithmetic average clustering" ou UPGMA (Sneath & Sokal, 1973), calcula a média aritmética da similaridade (ou da distância) entre o objeto que se quer incluir num grupo e cada objeto desse grupo. O objeto é atribuído ao grupo com o qual ele tem a maior similaridade média (ou menor distância média) com todos os objetos.

Método dos pesos proporcionais. Em Ecologia acontece frequentemente que grupos de amostras, oriundas por exemplo de regiões distintas, sejam de tamanhos diferentes (uma região mais amostrada que outra). Para evitar que essa diferença de esforço amostral venha interferir no cálculo da associação média (método anterior), Sokal & Michener (1958) sugeriram a aglomeração por pesos proporcionais (Weighted Clustering, ou WPGMA). O método consiste em atribuir um peso igual a dois ramos do dendrograma que estão para fusionar. Para isso, no cálculo da associação média, cada similaridade (ou distância) é multiplicada por dois coeficientes (um para cada objeto), e a associação média é calculada fazendo a soma

ponderada dos diferentes pares de objetos dos dois grupos a fusionar (cf. Legendre & Legendre 1983, para um exemplo numérico). Este método é considerado um dos melhores por Davis (1973).

Método pela variância mínima. Neste método, também chamado de Método de Wards (Romesburgh 1984), um grupo será reunido a um outro se essa reunião proporcionar o menor aumento da variância intragrupo. A variância intragrupo será calculada para todas as alternativas de aglomeração, escolhendo a que proporciona a menor variância. O mesmo procedimento é aplicado a todos os passos da análise. O método é altamente eficiente, mas de cálculo mais demorado.

Modelo geral de agrupamento. Lance & Williams (1966, in Legendre & Legendre 1983) propuseram um modelo geral incluindo os diversos métodos de aglomeração. Este modelo oferece a vantagem de poder ser traduzido na forma de um programa único de computador, permitindo passar de um método aglomerativo a um outro fazendo simplesmente variar três parâmetros (α, β, γ) que determinam a estratégia de agrupamento. Assim, a distância $D_{(g,h)}$ entre um objeto g e um grupo h formado de 2 subgrupos j e m, seria igual a:

$$D_{(g,h)} = \alpha_j D_{(j,g)} + \alpha_m D_{(m,g)} + \beta D_{(j,m)} - \gamma \left| D_{(j,g)} - D_{(m,g)} \right|$$

A Tabela 3 dá os valores dos parâmetros da equação geral de Lance & Williams (1966) para alguns modos de agrupamento.

Tabela 3. Valores dos parâmetros da equação de Lance & Williams para 3 métodos de aglomeração (onde ω_j e ω_m são os respectivos números de objetos nos subgrupos j e m).

Métodos/Parâmetros	α_j	α_m	β	γ
Ligação simples	0.5	0.5	0	-0.5
Ligação completa	0.5	0.5	0	0.5
Associação média	$\frac{\omega_j}{\omega_j + \omega_m}$	$\frac{\omega_m}{\omega_j + \omega_m}$	0	0

Maiores explicações sobre esses e outros métodos de aglomeração podem ser obtidas em abundante bibliografia, dentre as quais Sneath & Sokal (1973), Orloci (1978), Gauch (1982), Legendre & Legendre (1983), Pielou (1984), Romesburgh (1984), Ludwig & Reynolds (1988) e Krebs (1989).

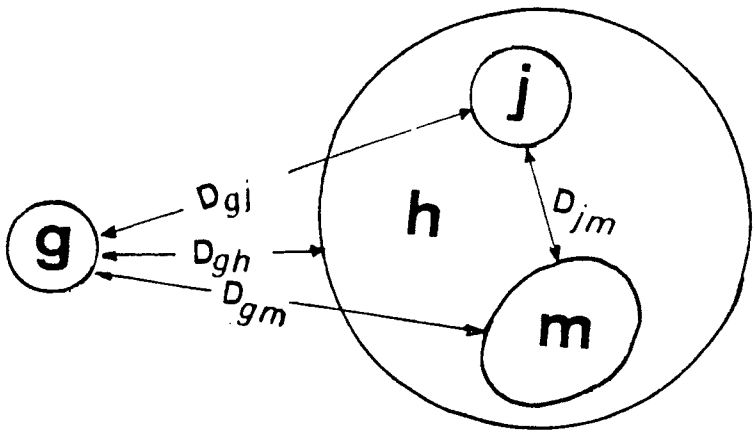

Fig. 6. Cálculo da distância D_{gh} entre um objeto g e um grupo h formado por dois subgrupos j e m, pelo método geral de Lance & Williams (1966).

Qual método escolher?

Um método é melhor do que um outro quando o dendrograma fornece uma imagem menos distorcida da realidade. É possível avaliar o grau de deformação provocado pela construção do dendrograma, calculando o chamado "coeficiente de correlação cofenético". É o coeficiente ***r*** de Pearson, calculado entre os índices de similaridade da matriz original e os índices reconstituídos a partir do dendrograma (chamados valores cofenéticos). Quanto maior ***r***, menor será a distorção. Obviamente há sempre um certo grau de distorção, ***r*** nunca será igual a 1, mas a literatura considera aceitável um coeficiente cofenético superior a 0.8. Certos pacotes estatísticos oferecem a possibilidade de calcular este coeficiente, o que facilita muito a escolhe do melhor método de aglomeração. Em geral, além do método de Wards (variância mínima) unanimemente considerado como o mais eficiente, embora não apresentando o melhor coeficiente cofenético, os métodos de agrupamento baseados nas similaridades ou distâncias médias, ponderadas ou não, são os mais recomendados.

A interpretação do dendrograma.

O passo final, numa análise de agrupamento, é a interpretação do dendrograma, i.e. a identificação dos grupos de espécies ou de objetos. À exceção dos métodos probabilísticos, pouco usados pela complexidade do algoritmo de cálculo, há uma grande parte de subjetividade na decisão de destacar e interpretar os grupos que poderiam, aparentemente, constituir uma realidade ecológica.

Algumas dicas para, por exemplo, uma análise de agrupamento em modo Q (entre objetos):

- escrever, no próprio dendrograma, em frente de cada amostra (que geralmente aparece sob forma de um número na saída do computador), as suas características: local e época de coleta, espécies dominantes, diversidade, características físico-químicas, etc, enfim tudo que poderia revelar os aspectos comuns entre as amostras de um mesmo grupo, e as diferenças com as amostras de outros grupos.
- começar a "ler" o dendrograma de cima para baixo, isto é, dos baixos valores de similaridade para os maiores. Assim, deverão ser interpretados em primeiro lugar os "grandes grupos", geralmente pouco numerosos (dois ou três apenas). Eles representam a macro-estrutura do ecossistema, ligada ao efeito dos principais fatores ambientais: fortes gradientes, discontinuidades espaciais ou temporais, etc. Seria vão tentar explicar os grupos menores, sem ter conseguido formular antes uma hipótese plausível sobre os grandes.

Ordenação

O que é ordenação?

A ordenação é, para o ecólogo, uma operação muito familiar. Ordenar amostras em função de um critério, por exemplo o número de indivíduos de uma espécie ***Sp1***, consiste em posicionar essas amostras ao longo de um eixo representativo da escala de abundância dessa espécie. Assim na Fig. 7a, as amostras A e C, próximas uma da outra, tem uma forte similaridade em razão das suas fracas abundâncias em ***Sp1***. Da mesma maneira, pela sua riqueza em ***Sp1***, as amostras D, E e B, são similares entre si e dissimilares com A e C.

A ordenação em função de dois critérios, por exemplo a abundância em indivíduos de duas espécies ***Sp1*** e ***Sp2***, posiciona as amostras num plano de acordo com suas coordenadas (abundâncias) nos eixos ***Sp1*** e ***Sp2***. Assim, na Fig. 7b verificamos que as amostras A e C tem pouco indivíduos das duas espécies, ao contrário das amostras D, E e B mais ricas em ambas as espécies, e consequentemente, distantes das amostras A e C. Podemos com isso observar que as amostras se distribuem ao longo de um eixo $\lambda 1$, que sintetiza os dois critérios de abundância. Este eixo representa a tendência de maior dispersão das amostras. Uma segunda tendência de dispersão é representada pelo eixo $\lambda 2$, perpendicular ao primeiro. Os eixos $\lambda 1$ e $\lambda 2$ são os eixos principais da elipse dentro da qual se posicionam as amostras. Eles representam a totalidade da variância dos dados, cada um contribuindo para uma determinada fração.

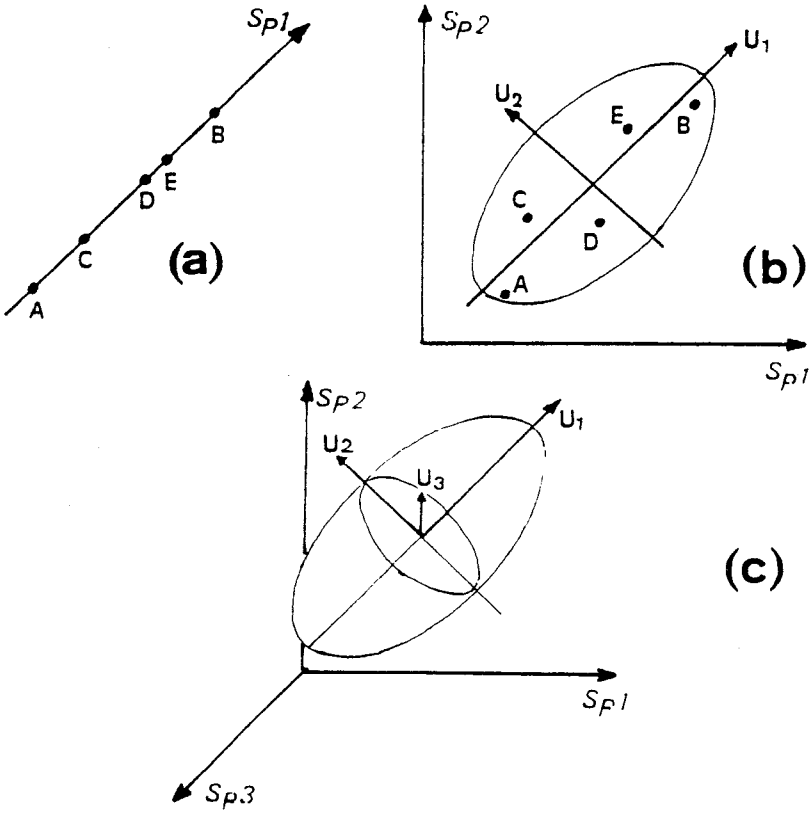

Fig. 7. Ordenação de 5 amostras (A-E) em função de 1 espécie (a), de 2 espécies (b) e de 3 espécies (c).

Com três critérios de ordenação (abundância em 3 espécies ***Sp1, Sp2*** e ***Sp3***), os pontos-amostras se posicionam dentro de um elipsoide, com projeções ao longo dos 3 eixos principais deste elipsoide, **λ1, λ2** e **λ3**, representativos da totalidade da variância dos dados. Cada um desses eixos, de comprimento decrescente, contribui para uma fração cada vez menor dessa variância (Fig. 7c).

Em Ecologia, as amostras são geralmente ordenadas em função de um grande número de critérios (*m* espécies), e se posicionam dentro de um espaço de *m* dimensões (hiperespaço), com projeções ao longo dos *m* eixos de um hiperelipsoide (as suas abundâncias nas *m* espécies). Obviamente, a representação gráfica é impossível, mas graças ao cálculo matricial é possível projetar as amostras num plano, por exemplo o plano formado pelos dois primeiros eixos, e ter assim uma imagem simplificada desta estrutura multidimensional.

Em vista disso, podemos definir a Ordenação como conjunto de técnicas pelas quais objetos são posicionados em relação a um ou mais eixos, de tal maneira que suas posições relativas aos eixos e entre eles, proporcionem o máximo de informação sobre suas semelhanças ecológicas. Essas técnicas, também chamadas de "técnicas fatoriais", visam definir esses eixos de dispersão como fatores ambientais responsáveis pelo determinismo dessa estrutura. Em suma, o princípio da ordenação consiste em simplificar, condensar e representar sinteticamente vastos conjuntos de dados, na esperança que as inter-relações ecológicas emerjam.

Os diversos métodos de ordenação.

Os princípios dos métodos de ordenação são antigos, mas seu desenvolvimento e sua diversificação são relativamente recentes, e ligados à difusão das possibilidades oferecidas pela informática.

A análise fatorial "stricto sensu ". Embora não sendo um método de ordenação propriamente dito, esta análise deve ser citada como pioneira das técnicas fatoriais. Inventada e aperfeiçoada por Spearman (1904) e Thurstone (1947), ela foi aplicada aos estudos da psicologia. Assim, as notas obtidas em teste psicológicos poderiam ser "explicadas" a partir de um pequeno número de fatores inerentes ao ser humano, tais como a inteligência e a memória, por exemplo. Em razão das suas limitações, e por não ser puramente descritivo, este método não é utilizado em Ecologia.

A análise de componentes principais. A análise de componentes principais (PCA) foi, talvez, o método de ordenação mais usado em ecologia, em razão principalmente da sua disponibilidade nos programas de computador. A sua primeira aplicação em Ecologia foi de Goodall (1954), a partir de um desenvolvimento da técnica por Pearson (1901, in Ludwig & Reynolds 1988). A PCA estabelece, a partir de uma matriz de semelhança (correlações, variâncias-covariâncias ou até mesmo de similaridades), um conjunto de eixos (ou componentes ou fatores) perpendiculares. Cada componente corresponde a um autovetor dessa matriz. Assim, a partir de uma matriz de correlação entre *m* variáveis, serão calculados *m* autovetores (ou eixos fatoriais) de comprimento $\lambda 1, \lambda 2, \ldots \lambda m$ decrescente em função da sua contribuição à variância total dos dados. Esses comprimentos correspondem aos *m* autovalores (ou raízes latentes) da matriz. Deste modo, o primeiro eixo da PCA, sobre o qual serão ordenadas as amostras, representará a maior parte da variação dos dados. O resultado disso é um sistema reduzido de coordenadas, proporcionando informações sobre as semelhanças ecológicas das amostras.

Em razão da sua importância, maiores comentários serão formulados ulteriormente sobre as etapas de cálculo e a interpretação dos resultados dessa análise.

A análise fatorial de correspondência. A análise fatorial de correspondência (AFC) foi desenvolvida separadamente por diversos autores. Primeiramente descrita para a análise de tabelas de contingência por Fisher (1940) sob o nome de "contingency table analysis" e Benzecri (1969), ela foi aplicada à Ecologia para análise de tabelas Espécies x Amostras por Hatheway (1971) sob o nome de "RQ analysis", Hill (1973, 1974) ("correspondence analysis"), Orloci (1978) ("reciprocal averaging") e outros, mas seu uso foi generalizado por Benzecri (1973) ("analyse factorielle des correspondences").

O ponto forte dessa análise é que as ordenações das espécies e das amostras são obtidas simultaneamente, permitindo que o ecólogo examine as relações entre amostras e espécies a partir de uma única análise. A AFC utiliza a mesma abordagem que a análise em componentes principais, pelo cálculo dos autovetores e autovalores de uma matriz de variâncias-covariâncias com a diferença que cada dado é previamente transformado em probabilidade (dividido pela soma total dos dados) e, em seguida, duplamente centrado pelas probabilidades marginais da linha e da coluna correspondentes. Com essas transformações, o cálculo das variâncias-covariâncias corresponde à distância do χ^2, e possibilita uma perfeita correspondência entre as linhas e as colunas, permitindo assim analisar indiferentemente em modo Q ou em modo R. Os autovetores das matrizes entre linhas e entre colunas são os mesmos, podendo então ser escolhida a menor dimensão da tabela para fazer a análise. Por exemplo, se tivermos 50 espécies coletadas em apenas 3 amostras, poderemos realizar a AFC sobre a matriz de 3 x 3, e projetar no mesmo plano as 50 espécies e as 3 amostras para uma interpretação conjunta. Uma outra peculiaridade deste método reside na possibilidade de analisar qualquer tipo de dados quantitativos e positivos, qualitativos (binários) ou semi-quantitativos (códigos de abundância), desde que eles sejam homogêneos. Não há exigência quanto à normalidade das distribuições, podendo ser incluídas, por exemplo, espécies raras.

A análise fatorial dos postos. Trata-se de uma análise de componentes principais aplicada a uma matriz de coeficientes de correlação não-paramétrica (ρ de Spearman, por exemplo). Esta técnica é escolhida no lugar da PCA quando os dados são ordinais (postos) ou quando as distribuições de frequências das variáveis não seguem uma distribuição normal, mesmo após transformação. Os procedimentos de cálculos são idênticos aos da PCA.

A análise em coordenadas principais. Descrita por Gower (1966), essa técnica corresponde a uma PCA em modo Q, a partir de uma matriz de distâncias euclidianas entre amostras. A análise em coordenadas principais preserva, em espaço de dimensões reduzidas, as distâncias entre objetos caracterizadas por descritores de qualquer tipo. O inconveniente do método é a impossibilidade de interpretar os eixos a partir da projeção dos descritores, já que eles não participam do cálculo desses eixos.

As técnicas acima citadas visam a análise descritiva da estrutura de um conjunto de dados, i.e. a ordenação das amostras e dos descritores em gradientes dentro de um *continuum*, ou em subconjuntos. Para interpretar essa estrutura, i.e. formular hipóteses sobre os fatores responsáveis, é preciso relacionar essa estrutura com os descritores identificados como potencialmente explicativos pelas teorias ecológicas. As técnicas a seguir são destinadas a esta finalidade.

A análise das correlações canônicas. O método das correlações canônicas é uma generalização da correlação múltipla. Ele tem por finalidade achar a correlação máxima entre combinações lineares de dois conjuntos de descritores. Teoricamente, esse método deveria constituir uma boa ferramenta para ordenar e analisar matrizes duplas de dados formados por descritores biológicos e físicos, por exemplo, ou interpretar as componentes principais de uma PCA a partir de um conjunto de descritores. Entretanto, em virtude da complexidade dos cálculos, da dificuldade de interpretação e, sobretudo. da exigência de linearidade nas relações entre variáveis, os autores aconselham proceder em duas etapas, a partir de uma PCA: (a) ordenar os descritores biológicos, (b) interpretar essa ordenação com projeção de variáveis ambientais nos planos fatoriais.

A análise discriminante. Este tipo de análise é destinado a interpretar grupos de objetos, definidos a priori pelos métodos de agrupamento e ordenação. A técnica não consiste em estabelecer grupos já que eles são previamente conhecidos, mas em interpretá-los a partir de variáveis ambientais. Os dados são assim apresentados (Fig. 8).

A discriminação dos grupos de espécies I e II é geralmente causada por um combinação de diversas variáveis ambientais. Como mostrado na figura 9, nenhuma das duas espécies permite discriminar perfeitamente os dois grupos de amostras, mas sim uma combinação das duas, chamada de "função discriminante" (*d*).

A análise discriminante é aplicada nos seguintes problemas:

- atribuir uma amostra isolada a um ou outro grupo, conhecendo suas características ecológicas (valores dos descritores) e a função discriminante,
- calcular uma distância D2 entre dois grupos de amostras, chamada "distância generalizada de Mahalanobis", e verificar se ela é significativa,
- determinar a percentagem explicativa de cada variável ambiental na separação de dois grupos de amostras.

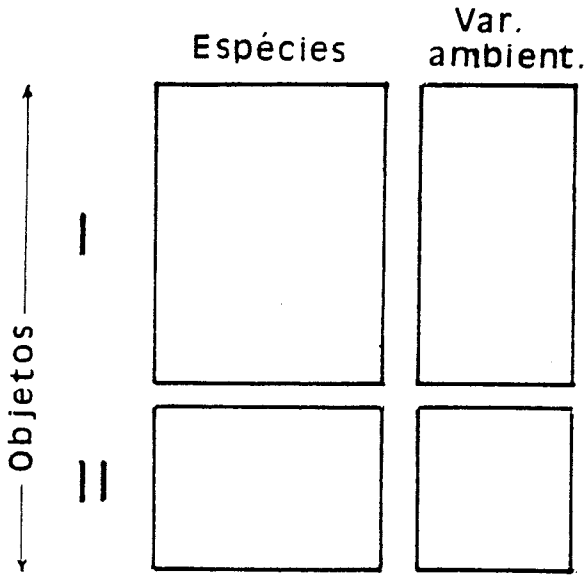

Fig. 8. Organização das matrizes de dados para uma análise discri-minante simples.

A análise discriminante é dita "simples" (ADS) quando ela é aplicada a somente 2 grupos, neste caso, uma única função é calculada para separar esses 2 grupos, ou "múltipla" (ADM) para mais de 2 grupos, e neste caso, são calculadas funções discriminantes múltiplas, chamadas também de "variáveis canônicas".

As análises de ordenação canônica. A ordenação canônica é um conjunto de técnicas visando relacionar a composição em espécies de amostras com as variáveis ambientais. É uma combinação de ordenação (PCA e AFC) e de regressão múltipla. A interpretação é realizada com ajuda de dados externos, calculando o coeficiente de correlação entre variáveis ambientais e os eixos fatoriais. Esses eixos são determinados por uma combinação linear das variáveis ambientais. O programa CANOCO ("Canonical Community Ordination", Ter Braak, 1988), desenvolve essas técnicas. Ele é uma extensão do programa DECORANA ("Detrended Correspondence Analysis", Hill, 1979), cujo objetivo é eliminar a dependência quadrática (efeito de arco) entre o segundo e o primeiro eixo fatorial.

As etapas de uma ordenação

Tomando por base a Análise de Componentes Principais, as etapas de cálculos são:

Preparação dos dados. A tabela de dados ecológicos é elaborada com *m* variáveis (descritores) e *n* objetos (amostras). A PCA deve ser realizada em

modo R. Consequentemente, é preciso definir perfeitamente, a priori, o que corresponde aos "descritores" e aos "objetos". Deve ser verificada a necessidade de uma transformação normalizante dos dados, exigida na PCA, que utiliza o coeficiente *r* de Pearson como índice de semelhança. Não há necessidade de uma padronização dos dados, já que ela está incluída no próprio cálculo de *r* (os dados são centrados e reduzidos).

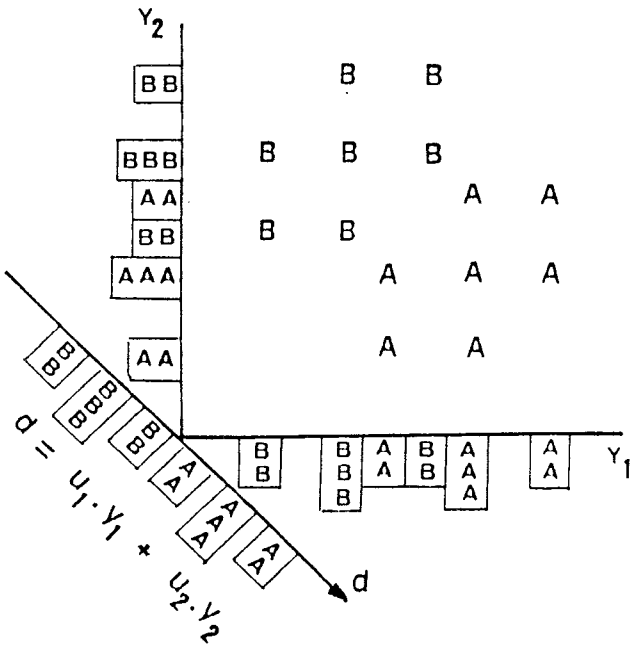

Fig. 9. Representação gráfica de uma análise discriminante: os dois grupos de amostras A e B não podem ser separados apenas pela abundância em S1 ou em S2, mas pelo eixo discriminante *d* cuja função é $d = \cos(45).S1 + \cos(45).S2$ (Legendre & Legendre, 1983).

Cálculos propriamente ditos. Devem ser realizadas as seguintes etapas de cálculos:

- cálculos da matriz de coeficientes de correlação linear (ou variâncias-covariâncias) entre descritores.
- definição dos eixos fatoriais (componentes principais), em direção e comprimento, pelo cálculo dos autovetores e autovalores da matriz de correlação. Trata-se da resolução da equação característica $\left(S - \lambda_h I\right)U_h = 0$, onde S é a matriz de correlação, U_h é o autovetor (eixo fatorial h), λ_h o autovalor deste eixo, e I uma matriz unidade.

- cálculo das coordenadas das amostras ("scores") e das variáveis ("factor loading") no novo sistema de eixos.
- cálculo das contribuições das amostras e das variáveis na construção dos eixos. Em PCA, essas contribuições são iguais às coordenadas dos pontos sobre o eixo, o que dispensa o seu cálculo. Na análise fatorial de correspondências (AFC), entretanto, as contribuições levam em conta o "peso" (total das frequências) das linhas e das colunas, e devem ser calculadas para a interpretação dos eixos.

Interpretação dos resultados. Interpretar uma PCA consiste em tentar definir o que representa cada eixo em termos de fator ecológico, responsável pela ordenação das amostras. Na medida em que a importância dos eixos, i.e. a sua participação à variância total (autovalor), vai diminuindo, a sua interpretação torna-se cada vez mais difícil. A interpretação de um eixo deve ser baseada nas coordenadas dos descritores (variáveis, espécies, etc.) neste eixo, a partir dos quais foi elaborada a matriz de correlação que deu origem aos autovetores. As coordenadas das amostras vem ajudar, em seguida, nessa interpretação:

- uma proximidade maior ou menor entre dois pontos-variáveis no plano traduz uma maior ou menor correlação entre essas variáveis, principalmente quando elas são afastadas do centro do plano.
- a coordenada de um ponto-variável sobre um eixo fatorial é igual ao coeficiente de correlação entre esta variável e o eixo. Uma variável é considerada significativamente ligada a um eixo, e por consequência suscetível de ser utilizada para a interpretação deste eixo, quando a sua distância d ao centro do plano é $d = \sqrt{\frac{2}{m}}$, onde m = número de variáveis (Legendre & Legendre 1983).
- a proximidade entre dois pontos-amostra traduz uma certa similaridade entre essas duas amostras, em termos de composição em espécies por exemplo.

Validade dos resultados.

Ao termo de uma análise fatorial, o pesquisador enfrenta uma série de dúvidas na hora de interpretar os resultados gerados pelo computador.

A significância dos eixos. Quantos fatores extraídos pela análise são estatisticamente significativos, i.e. representam uma variância fatorial e não residual? Em outros termos, a partir de que eixo devemos parar a interpretação? Não existe uma resposta perfeitamente clara a esta pergunta. Um método indireto consiste em comparar os autovalores com os que seriam obtidos se os dados fossem todos aleatórios (sem influência fatorial) (Paes & Blinder, neste volume). Os eixos cujos autovalores

ultrapassam os obtidos pelos dados aleatórios podem ser considerados como representando uma certa variância fatorial e suscetíveis de interpretação ecológica. Este tipo de teste levou à seguinte constatação: numa tabela de dados ao acaso, o decaimento dos autovalores é progressivo. Conseqüentemente, aconselha-se parar a interpretação a partir do eixo cujo autovalor é anormalmente inferior ao anterior (Fig. 10).

Geralmente, além dos dois ou três primeiros eixos, a interpretação torna-se mais difícil. Benzecri (1973) considera válida o que ele chama de "prova semântica": um eixo significativo pode não ser interpretável, por causa por exemplo de uma insuficiência de informações sobre o ambiente, mas um eixo interpretado tem grande chance de ser significativo.

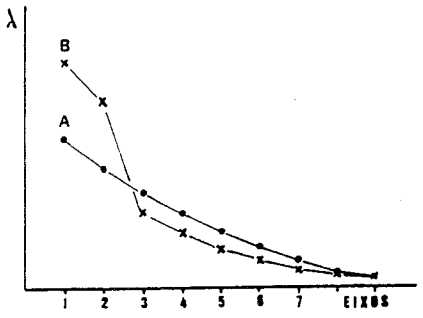

Fig. 10. Decaimento dos autovalores λ em duas situações: (A) dados aleatórios, (B) dados com eixos I e II significativos.

Na maioria dos trabalhos, os autores param a interpretação a partir do segundo eixo. É preciso insistir na importância de tentar prosseguir com a análise dos demais eixos, que podem revelar estruturas subjacentes invisíveis a "olho nu". É neste aspecto que reside toda a força das técnicas fatoriais. Para isso, é necessário entretanto dispor:

- de dados em quantidade suficiente: essas técnicas são destinadas a grandes conjuntos de dados.
- de dados de qualidade, sem vícios de amostragem, e adequadamente coletados de acordo com os objetivos do trabalho.
- de informações exaustivas sobre o meio ambiente e a ecologia dos organismos.

A estabilidade dos resultados. Após ter evidenciado uma estrutura, o ecólogo pode se perguntar se essa estrutura que caracteriza o ecossistema, é estável, i.e. representativa de uma realidade e não ligada a algum artefato dos próprios dados: valores aberrantes, influência de variáveis não conformes às exigências do método

(normalidade, excesso de valores zero...), influência das unidades, da transformação ou codificação aplicadas aos dados. Sempre que for necessário deverá ser refeita a análise eliminando os dados aberrantes ou duvidosos para verificar esta estabilidade. Isso leva, às vezes, a um trabalho repetitivo de computação, indispensável para tirar conclusões seguras sobre a estrutura do sistema estudado e dos fatores que a regem.

Apresentação dos resultados.

Raramente todos os resultados de uma análise fatorial merecem ser publicados. Isso depende essencialmente da contribuição que eles podem dar à interpretação dos dados e à elaboração das conclusões do trabalho. A parte publicável pode ser ínfima comparativamente ao esforço dispensado.

É recomendado fornecer ao leitor as figuras dos planos fatoriais que melhor ilustram sua interpretação. Essas figuras são o melhor veículo da comunicação entre pesquisadores, ao mesmo título que uma fotografia ou um mapa. Mas, para isso, há de ter o maior cuidado com a clareza de apresentação do plano (cf apresentação gráfica).

As informações que devem previamente constar na publicação são:

- as dimensões da matriz de dados: número de variáveis e observações.
- a natureza dos dados e as transformações eventuais.
- a listagem das variáveis com, eventualmente, indicações dos seus valores de média, frequência, dispersão, o código ou a abreviatura utilizada na figura. Dizer se são variáveis "ativas" ou "ilustrativas" (cf a seguir).
- a necessidade de análises preliminares para testar a estabilidade e se foi preciso eliminar certas variáveis ou observações.

Variáveis e observações suplementares. Alem das variáveis utilizadas para o cálculo dos autovetores (variáveis chamadas "ativas"), podemos eventualmente dispor de outras variáveis, as quais, por motivos diversos (e.g., espécies raras, não normalidade) não puderam ser introduzidas na análise. Essas variáveis, chamadas "ilustrativas, suplementares ou passivas", podem ser projetadas nos planos fatoriais, pelo cálculo das suas coordenadas nos eixos, e contribuir enormemente à interpretação desses eixos. Do mesmo modo, podemos utilizar observações suplementares. Essas variáveis ou observações suplementares devem constar de arquivos separados a serem chamados pelo programa de computador quando necessário. Poucos programas, entretanto, possuem essa opção de cálculo, que é de grande utilidade para a interpretação dos resultados.

Apresentação gráfica dos planos fatoriais. É com a projeção gráfica dos pontos Variáveis e Observações no primeiro plano (formado pelos eixos I e II) que deve iniciar a interpretação. É neste plano que deve ser possível explicar a maior parte da variabilidade dos dados, e descrever as grandes linhas da sua estrutura. Os planos sucessivos (por exemplo, I-III, II-III, III-IV etc.) deverão ser projetados para

descrever estruturas mais finas, e tentar identificar fatores de menor importância, mas que somente este tipo de análise poderia colocar em evidência.

Certos procedimentos elementares tornam mais clara a leitura dos planos. Além de indicar obrigatoriamente, na extremidade de cada eixo, seu número e sua percentagem explicativa da variância, podemos:

- destacar por simbologia diferente, os pontos variáveis e observações de mesmas características ecológicas (mesmo período do ano, região, regime alimentar, sexo, etc.), podendo, inclusive com ajuda dos resultados da análise de agrupamento, delimitar no plano os grupos assim identificados.
- escrever no plano fatorial, informações concisas que sintetizem a interpretação.
- traçar eventualmente um eixo diagonal que ilustre melhor a interpretação dada ao plano fatorial; um eixo, mesmo que significativo, pode não ter uma definição clara.

É preciso lembrar, por fim, que a interpretação de um eixo deve ter um caracter "residual" em relação aos anteriores já interpretados. Não deverá ser dada uma mesma definição a dois ou mais eixos, já que eles são matematicamente independentes (ortogonais). Entretanto, se dois eixos sucessivos tem autovalores iguais ou muito próximos, é, neste caso, preferível tentar interpretar o plano globalmente, sem definir especificamente o papel de cada um no plano. A Fig. 11 pode servir de exemplo de apresentação de um plano fatorial. O estudo visa descrever a estrutura de comunidades planctônicas, numa região costeira marinha onde ocorre alternância de massas de água de origem tropical e profunda.

A sua interpretação poderia ser assim formulada:

- **Eixo I** (52% da variância explicada). Ele é positivamente ligada as espécies Sp1, Sp3, Sp7 e Sp9, as quais ocorrem com maior abundância nas amostras (1), (2), (3) e (4), contribuindo para uma elevada diversidade específica, em águas relativamente quentes (coordenadas positivas das variáveis Temperatura e Diversidade). Em oposição, projetam-se negativamente no eixo I, as espécies Sp2 e Sp5, as amostras (8), (12), (13), (14) e (15), e os valores de Nitratos. Essa primeira componente principal sintetiza o efeito preponderante das condições hidrológicas (alternância de água profunda fria e rica em nutrientes e água tropical quente) sobre a distribuição do plâncton.
- **Eixo II** (18%). Ele é formado pelas contribuições positiva da espécie Sp8, abundante na amostra (9), e negativa da espécie Sp6 abundante nas amostras (10) e (11). Em razão da projeção positiva da variável Salinidade sobre este eixo, podemos interpretá-lo como sendo um fator de influência continental. A espécie Sp6 deve ser de afinidade dulcícola, preponderante nos pontos de coleta próximos aos efluentes continentais

(rio, despejos, etc.). Além disso, as coordenadas negativas das variáveis NH4 e Clorofila revelam um ambiente rico em amônia (poluição orgânica) favorável ao crescimento do fitoplâncton. As amostras (5), (6) e (7), no centro do plano, apresentam características bióticas e abióticas intermediárias.

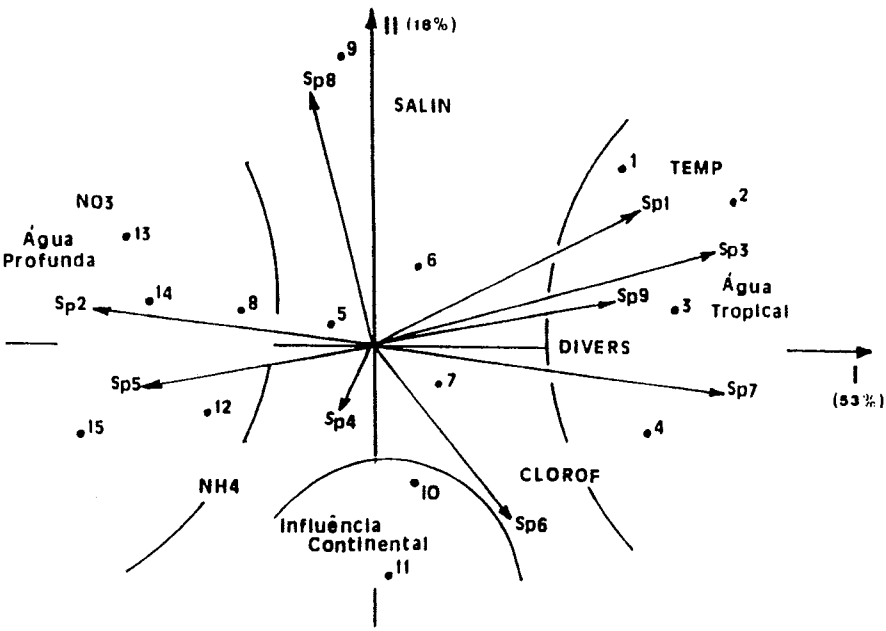

Fig. 11. Exemplo de apresentação de um plano fatorial numa Análise de Componentes Principais. Foram projetados, os vetores espécies (Sp1, Sp2...), os pontos amostras (1), (2),..., as variáveis suplementares (TEMP =temperatura, DIVERS = diversidade específica, CLOROF = teor de clorofila "a", SALIN = salinidade, NO3 e NH4 = teores em nitrato e amônia.

Embora sucinto, este exemplo de análise fatorial revela o poder de síntese dessa técnica, para descrever de maneira integrada a estrutura de dados multidimensio-nais, e em particular a complexa rede de interações, que caracterizam os ecossistemas.

Agradecimentos

Trabalho realizado com apoio financeiro do Centro Nacional de Desenvolvimento Científco e Tecnológico (CNPq) e da Fundação Universitária José Bonifácio (FUJB).

Referências

BRAY, R.J. & J.T. CURTIS. 1957. An ordination of the upland forests communities of southern Winsconsin. *Ecological Monography*, 27: 325-349.

BENZECRI, J.P. 1969. Statistical analysis as a tool to make patterns emerge from data. pp. 35-60, *In*: S. Watanabe, (ed.), Methodologies of pattern recognition. Academic Press, New York.

BENZECRI, J.P. 1973 L'analyse des données. Tome I: La taxinomie, Tome II: L'analyse des correspondences. Dunod, Paris.

DAVIS, J.C. 1973. Statistics and Data Analysis in Geology. J.Wiley & Sons, Inc., New York.

FISHER, R.A. 1940. The precision of discriminant functions. *Ann Eugen. Lond.*, **10**:422-429.

GAUCH, H.G. 1982. Multivariate Analysis in Community Ecology. Cambridge Univ. Press, N.York.

GOODALL, D.W. 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. *Australian Journal of Botany*, **2**:304-324.

GOWER, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**:325-338.

HATHEWAY, W.H. 1971. Contingency-table analysis of rain forest vegetation. pp. 271-313. *In*: G.P. Patil, E.C. Pielou & W.E. Waters (eds.), Statistical Ecology, Vol.3: Many species populations, ecosystems and systems analysis. Pennsylvania State University Press, London.

HILL, M.O. 1973. Reciprocal averaging: an eigenvector method of ordination. *Journal of Ecology*, **61**:237-249.

HILL, M.O. 1974. Correspondence analysis: a neglected multivariate analysis. *Applied Statistics*, **23**:340-354.

HILL, M.O. 1979. DECORANA - A FORTRAN program for Detrended Correspondence Analysis and Reciprocal Averaging. Ithaca, N. Y.: Cornell University.

JACCARD, P. 1908. Nouvelles recherches sur la distribution florale. *Bulletin Society Sciences Naturale*, **44**:223-270.

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

LANCE, G.N. & W.T. WILLIAMS. 1966. A generalized sorting strategy for computer classifications. *Nature (London)*, **212**:218.

LEGENDRE, L. & P. LEGENDRE. 1983. Numerical Ecology. Elsevier, New York.

LUDWIG, J.A. & J.F. REYNOLDS. 1988. Statistical Ecology. A primer on Methods and Computing. J.Wiley & Sons, Inc., New York.

MORISITA, M. 1959. Measuring of interspecific association and similarity between communities. *Memoirs Faculty Kyushy University*, **Series E3**: 65-80.

ORLOCI, L. 1978. Multivariate Analysis in Vegetation Research (2nd. Ed.). W. Junk, The Hague.

PEARSON, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine,* **Sixth Series 2**: 559-572.

PIELOU, E.C. 1984. The Interpretation of Ecological Data. Wiley, Newe York

ROMESBURGH, H.C. 1984. Cluster Analysis for Researchers. Lifetime Learning Publications, Belmont.

SNEATH, P.H.A. & R.R. SOKAL. 1973. Numerical Taxonomy. Freeman, San Francisco.

SOKAL, R.R. & C.D. MICHENER. 1958. A statistical method for evaluating systematic relationship. *University of Kansas Science Bulletin*, **38**:1409-1438.

SPEARMAN, C. 1904. "General intelligence", objectively determined and measured. *American Journal of Psychology*, **15**:201-293.

TER BRAAK, C.J.F. 1988. CANOCO-A FORTRAN program for canonical community ordination. Microcomputer Power, Ithaca.

THURSTONE, L.L. 1947. Multiple-factor Analysis - A development and expansion of the vectors of mind. Chicago Univ. Press, Illinois.

WOLDA, H. 1981. Similarities indices, sample size, and diversity. *Oecologia*, **50**:296-302.

Endereço

JEAN LOUIS VALENTIN
Departamento de Biologia Marinha, Universidade Federal do Rio de Janeiro
CEP: 21949-900 - Cidade Universitária, Rio de Janeiro - RJ