Distribuições de abundâncias de espécies

AVANÇOS ANALÍTICOS PARA ENTENDER UM PADRÃO BÁSICO EM ECOLOGIA

Paulo Inácio K. L. Prado

A dominância numérica de poucas espécies nas comunidades biológicas constitui uma das raras leis gerais da ecologia. Muitos modelos quantitativos foram propostos para descrever este padrão, mas persiste a falta de consenso sobre qual seria o mais adequado, ou mesmo se algum o seria. Uma das razões desse "fracasso científico coletivo" é que a multiplicidade de modelos contrasta com uma notável escassez de testes empíricos. Comparações sistemáticas de um certo número de modelos por meio de conjuntos de dados representativos de vários modelos taxonômicos ou de diferentes situações ecológicas são extremamente raras. Nesse sentido, este artigo apresenta uma revisão de dois aspectos analíticos para deduzir e comparar modelos: a teoria da amostragem e seleção de modelos a partir de sua verossimilhança estatística.

Distribution d'abondances d'espèces

PROGRÈS ANALYTIQUES POUR COMPRENDRE UN PATRON EN ÉCOLOGIE

La dominance numérique d'à peine quelques espèces dans les communautés biologiques constitue l'une des rares exemples de loi générale en écologie. Plusieurs modèles quantitatifs ont été proposés pour décrire ce patron; toutefois il n'y a pas encore aujourd'hui une position consensuelle sur quel d'entre eux serait le plus pertinent ou même s'il y en aurait un. L'une des raison de cet "échec scientifique collectif" serait que la multiplicité de modèles théoriques contraste avec un manque remarquable de testes empiriques. Des comparaisons systématiques d'un certain nombre de modèles par des ensembles de données représentatives de modèles taxonomiques divers ou de différentes situations écologiques sont entrêmement rares. Dans ce sens cet article présente une révision de la théorie de l'échantillonage et de la sélection des modèles à partir de leur vraissemblance statistique.

Introdução

A dominância numérica de poucas espécies nas comunidades biológicas é uma das raras leis gerais da ecologia[1]. A maioria das espécies nas comunidades é representada por poucos indivíduos, e também não são muitas as espécies abundantes, resultando em uma distribuição de abundâncias tipicamente côncava, conhecida na literatura como "hollow curve" (figura 1). Há raríssimas exceções a este padrão, o que o torna relevante para a análise de quaisquer outros parâmetros das comunidades. Uma das consequências mais diretas desse fato é que a riqueza de espécies fica determinada pelas espécies raras[2], mas há outras consequências menos óbvias. A concavidade da distribuição de abundância das espécies pode explicar outros padrões importantes nas comunidades, como a inclinação da relação espécie-área[3], bem como a relação positiva entre abundâncias e frequências de ocupação de manchas[4]. Além disso, as distribuições de abundância das espécies (DAE[5]) têm enorme potencial aplicado, uma vez que decifrá-las é compreender as causas da raridade, uma das principais metas da biologia da conservação[6].

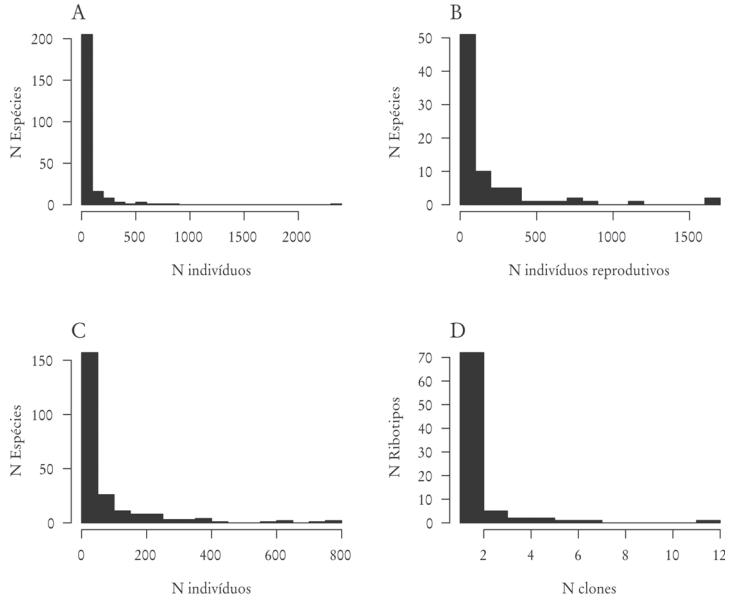

Figura 1: Distribuições de frequência de número de unidades taxonômicas por classes de abundâncias em amostras de (A) Mariposas capturadas com armadilhas luminosa por 4 anos na Estação Experimental de Rothamsted, Inglaterra[7], (B) Aves em estado reprodutivo em Quacker Valley, EUA[8], (C) Plantas lenhosas com DAP > 10 cm em 2005 na parcela permanente de Barro Colorado, Panamá[9], (D) Ribotipos de Archaea em amostra de sedimento do estuário de Santos, Brasil[10].

[1] MCGILL, B. *et al*. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters,* 10:995-1015, 2007.

[2] WHITTAKER, R. Dominance and diversity in land plant communities. *Science*, 147:250-260, 1965.

[3] PRESTON, F. W. Time and space and the variation of species. *Ecology*, 41:611-627, 1960.

[4] GASTON, K. The multiple forms of the interspecific abundance-distribution relationships. *Oikos*, 76:211-220, 1996.

[5] A literatura recente em inglês usa a sigla SAD, de Species Abundance Distribution. Aqui usaremos a forma em português DAE.

[6] HUBBELL, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton: Princeton University Press, 2001.

[7] FISHER, R.; CORBET, A. & WILLIAMS, C. The relation between the number of the species and the number of individuals in a random sample from animal population. *Journal of Animal Ecology*, 12:42-58, 1943.

[8] Dados de SAUNDERS, A. Ecology of the birds of Quaker run valley, Allegany state park, New York. *The University of the state of New York,* 1936 transcritos em PRESTON, F. W. The commonness and rarity of species. *Ecology*, 29:254-283, 1948.

[9] CENTER FOR TROPICAL FOREST STUDIES. *Species abundance in Barro Colorado Island*. Sítio na Internet do Smithsonian Tropical Research Institute URL: *http://www.ctfs.si.edu/site/Barro+Colorado+Island/abundance/?size=100* [consultado em dezembro de 2009].

[10] PIZA, F. F.; PRADO, P. I. & MANFIO, G. P. Investigation of bacterial diversity in Brazilian tropical estuarine

sediments reveals high actinobacterial diversity. *Anton Leeuw Int. J. G.*, 86:317-328, 2004.

[11] FISHER, R. *et. al.*. *Op. cit.*

[12] PRESTON, F. W. *Op. cit.*. 1948.

[13] TOKESHI, M. *Species Coexistence* – Ecological and Evolutionary Perspectives. Blackwell Science, 1999. McGILL, B. *et al.*. *Op. cit.*

[14] WHITTAKER, R. *Op. cit.* MAGURRAN, A. *Measuring Biological Diversity*. Oxford: Blackwell, 2004.

[15] McGILL, B. *et al.*. *Op. cit.*

[16] McGILL, B. Strong and weak tests of macroecological theory. *Oikos*, 102:679-685, 2003.

[17] FISHER, R. *et al.*. *Op. cit.*

[18] A contribuição de E. C. Pielou para estes e muitos outros temas de ecologia quantitativa vai muito além da formalização de conceitos. Seus livros combinam prosa elegante e extremo rigor lógico e teórico para mostrar o enorme potencial que há no diálogo entre matemática e biologia. São leitura obrigatória, e sempre inspiradora, para todos os interessados neste diálogo:
PIELOU, E. C. *Mathematical Ecology*. New York: John Wiley and Sons, 1977.
PIELOU, E. C. *Ecological Diversity*. New York: John Wiley and Sons, 1975.
Especificamente quanto à teoria da amostragem aplicada às DAE, uma contribuição essencial, que fornece a lógica usada neste artigo, é
GREEN, J. L. & PLOTKIN, J. B. A statistical theory for sampling species abundances. *Ecology Letters*, 10:1037-1045, 2007.

O uso de um histograma para representar as DAE evoca uma distribuição teórica subjacente. Não por acaso, essas duas ideias foram apresentadas juntas em 1943 por Ronald Fisher e colaboradores[11] – que propuseram a distribuição de série logarítmica de probabilidades para explicar a forma das DAE de amostras de comunidades biológicas com muitas espécies – e em seguida por Frank Preston[12] – que defendeu o uso de uma distribuição log-normal como o modelo mais adequado. A falta de consenso persiste, com o agravante de que pelo menos outros vinte modelos foram propostos[13]. A proliferação de tantos modelos teóricos para explicar um mesmo padrão empírico é um reconhecido embaraço para os ecólogos[14] e foi definido em uma revisão recente sobre o assunto como "fracasso científico coletivo"[15].

Uma das razões desse fracasso é que a abundância de modelos teóricos contrasta com uma notável escassez de testes empíricos. Comparações sistemáticas de vários modelos por meio de conjuntos de dados representativos de vários grupos taxonômicos ou diferentes situações ecológicas são extremamente raras na literatura.[16] Esse tipo de comparação depende de dois desenvolvimentos analíticos, que ainda não estão plenamente incorporados pela ecologia de comunidades: (1) a inclusão do efeito da amostragem nos modelos de DAE, e (2) a definição de um protocolo estatístico para a seleção dos modelos que melhor descrevem cada conjunto de dados. As próximas seções constituem uma apresentação destes dois aspectos, aplicados depois a um exemplo simples.

Teoria da amostragem de DAE

As DAE podem ser descritas com um modelo probabilístico, que tem dois ingredientes: as abundâncias das espécies na comunidade, e o processo de amostragem desta comunidade, o qual resulta nas abundâncias observadas. Tal abordagem foi usada pela primeira vez por Sir Ronald Fisher[17], e depois formalizada e generalizada por Evelyn C. Pielou[18]. Não pretendemos aqui repetir as deduções destes autores, ou manter o seu formalismo matemático, que infelizmente ainda afugentam muitos dos leitores da área de biologia. Nossa escolha foi apresentar a lógica subjacente, com o mínimo de matemática possível. Esperamos, com isso, interessar o leitor o suficiente para prosseguir no estudo do assunto.

Abundâncias na comunidade

Uma descrição probabilística do padrão "J invertido" é propor que uma espécie tomada ao acaso da comunidade apresente chance muito alta de ser pouco abundante, e uma chance pequena de ser muito abundante. Buscamos, portanto, uma função matemática que atribua tais probabilidades aos valores de abundâncias, ou seja, certas *funções de distribuição de probabilidades*[19]. Algumas distribuições com estas propriedades, tradicionalmente usadas para descrever as DAE, são a log-normal, a geométrica, e a série logarítmica.[20]

As abundâncias das espécies são contagens, mas se aceitamos que na comunidade há uma grande quantidade de indivíduos, uma aproximação contínua é razoável e traz algumas conveniências de cálculo. Acrescentando esta premissa, nosso modelo para a distribuição de abundâncias das espécies na comunidade será uma *função de densidade probabilística*, que chamaremos de *f(n)*, sendo *n* o número de indivíduos de uma espécie tomada ao acaso da comunidade. Tal modelo nos dá a probabilidade de que uma espécie na comunidade tenha uma abundância entre dois valores, consistindo na área sob a curva de *f(n)* neste intervalo (figura 2).

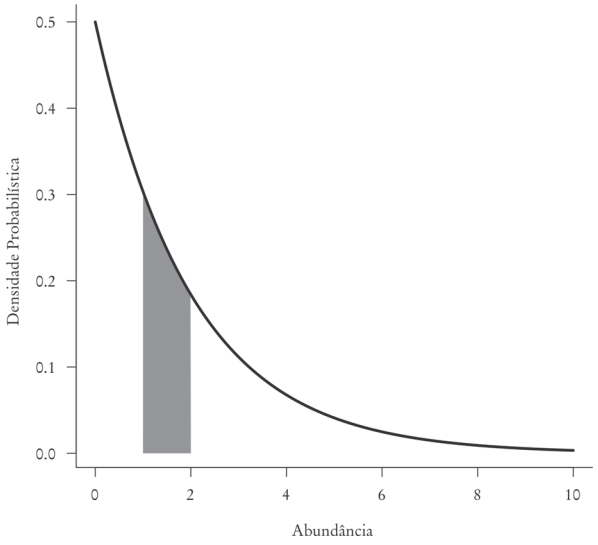

Figura 2: Gráfico da distribuição exponencial, uma das funções de densidade probabilística usadas para descrever as distribuições de abundâncias de espécies numa comunidade. Esta função tem um único parâmetro, em geral denominado pela letra grega lambda (λ). Em cinza está a área sob a curva entre os valores de abundância um e dois. Tal área corresponde à probabilidade de que uma espécie na comunidade esteja neste intervalo de abundâncias. Neste exemplo o valor do parâmetro é λ = 0,5 e a probabilidade assinalada é de p = 0,24.

[19] Para contagens, como é o caso de número de indivíduos, estas funções atribuem uma probabilidade a cada valor e são chamadas distribuições discretas. Para variáveis contínuas, como biomassa, o que temos é uma função de densidade probabilística, que atribui uma probabilidade ao fato de a observação estar em um certo intervalo de valores. Uma ótima introdução a modelos de distribuição de probabilidades pode ser encontrada nos manuais: OTTO, S. P. & DAY, T. *A biologist's guide to mathematical modeling in ecology and evolution*. Princeton: Princeton University Press, 2007. BOLKER, B. *Ecological Models and Data*. Princeton: Princeton University Press, 2008.

[20] Uma introdução muito simples a estes e outros modelos de DAE é dada no capítulo 2 de MAGURRAN, A. E. *Measuring biological diversity*. *Op. cit.* Para uma apresentação dos detalhes matemáticos destes modelos, veja PIELOU, E. C. *Mathematical Ecology*. *Op. cit.* e *Ecological diversity*. *Op. cit.*, bem como MAY, R. M. Patterns of Species Abundance and Diversity. *In*: CODY, M. L. & DIAMOND, J. M. (ed.). *Ecology and Evolution of Communities*. Harvard: Harvard University Press, 1975. p. 81-120.

Função de amostragem

Normalmente não tomamos espécies ao acaso de uma comunidade, e sim indivíduos. Se a abundância de uma espécie na comunidade é *n* e uma fração *a* dos indivíduos da comunidade é amostrada ao acaso, o número esperado de indivíduos desta espécie na amostra é o produto *a* x *n*. No entanto, o acaso fará com que algumas amostras tenham menos ou mais indivíduos do que este valor esperado, mas valores muito diferentes do esperado devem ser menos prováveis do que os valores próximos. Temos aqui outra fonte de variação das abundâncias na amostra, que permaneceria mesmo que as abundâncias na comunidade fossem fixas (figura 3). Para descrever a variação, usamos uma outra função, que nos dá a probabilidade de que a abundância *X* de uma espécie na amostra seja um certo valor *x*, *dada uma certa abundância n* na comunidade:

$$g(x \mid n) = P(X = x \mid n) \quad (1)$$

Combinando os dois modelos

A probabilidade de que uma espécie tenha certa abundância *n* na comunidade e uma certa abundância *x* na amostra (dado *n*) é o produto das probabilidades destes dois eventos independentes:

$$P(N = n, X = x) = g(x \mid n) \times f(n) \quad (2)$$

O modelo nos indica que cada valor de abundância na amostra pode ocorrer para os diferentes valores de abundância na comunidade. Somando a probabilidade de cada uma dessas combinações[21], temos o nosso modelo final, que nos dá a probabilidade de ocorrência de certo número de indivíduos na amostra. Trata-se, então, de somar o produto acima para todos os valores de *n*. Como a função *f(n)* é contínua, a soma torna-se uma integral:

$$\int_0^\infty f(n)\, g(x \mid n)\, dn \quad (3)$$

A solução dessa integral nos fornece a distribuição de probabilidades das abundâncias de cada espécie na amostra, dadas as distribuições que descrevem as abundâncias na comunidade e o processo amostral. Conhecendo propriedades de diferentes distribuições, podemos combiná-las para construir

[21] Como cada espécie só pode ter uma abundância na comunidade, estas combinações são eventos excludentes, portanto a soma de suas probabilidades nos dá a probabilidade do evento composto de uma certa abundância na amostra.

modelos apropriados para cada situação. Por exemplo, se os indivíduos distribuem-se aleatoriamente pelas unidades amostrais, uma distribuição de Poisson é um modelo adequado para descrever a amostragem. Mas se há agregação dos indivíduos, como é comum em parcelas de amostragem de vegetação, outras distribuições devem ser usadas.[22]

[22] GREEN, J. L. & PLOTKIN, J. B. *Op. cit.* propõem o uso da distribuição binomial negativa para estes casos. Um tratamento mais aprofundado é feito por ALONSO, D.; OSTLING, A. & ETIENNE, R. S. The implicit assumption of symmetry and the species abundance distribution. *Ecology Letters*, 11:93-105, 2008.

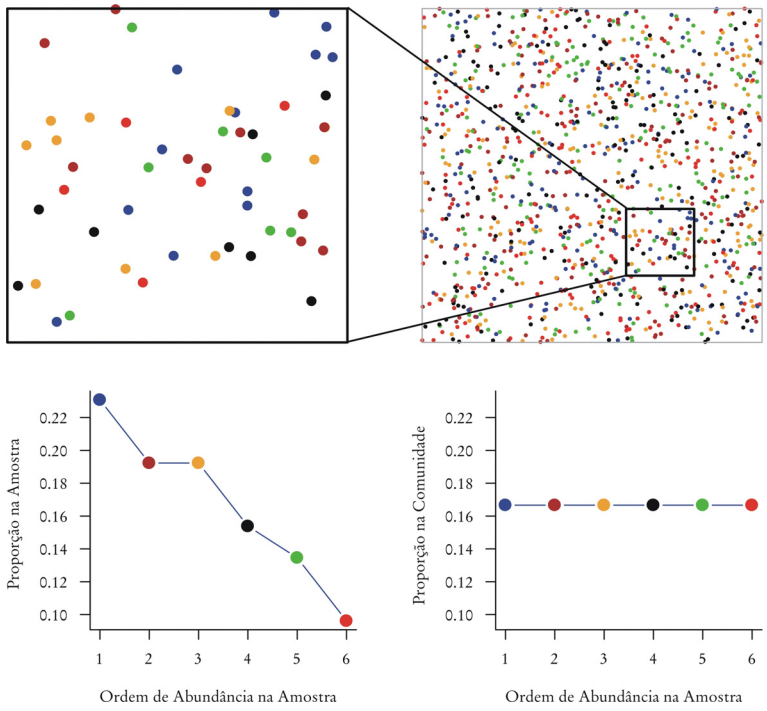

Figura 3: Efeito da amostragem sobre a abundância das espécies. Neste exemplo hipotético, uma amostra ao acaso é tomada de uma comunidade com seis espécies, todas com a mesma abundância, e com os indivíduos distribuídos ao acaso no espaço. Acima, um mapa da posição dos indivíduos, representados por pontos com cores diferentes para cada espécie. À direita está o mapa para toda a comunidade, e à esquerda está ampliada a área amostrada. Abaixo temos a abundância relativa de cada espécie na amostra e na comunidade, ordenadas pelos valores de abundância na amostra. Este exemplo ilustra que a variação ao acaso do número de indivíduos de cada espécie incluídos na amostra cria diferenças na distribuição de abundâncias observada, independente das abundâncias na comunidade.

Em casos simples, utilizamos as ferramentas do cálculo analítico para solucionar a integral apresentada anteriormente. Por exemplo, a combinação de uma distribuição exponencial para descrever as abundâncias das espécies na

comunidade com uma distribuição Poisson para descrever o processo amostral, resulta na distribuição Poisson-exponencial truncada[23]:

$$P(x) = \frac{\lambda a^{x-1}}{(\lambda + a)^x} \quad (4)$$

onde $P(x)$ é a probabilidade de uma espécie ter x indivíduos na amostra, a é a fração da comunidade amostrada, e λ é o parâmetro da exponencial negativa, que pode ser interpretado como um índice de dominância na comunidade.

Assim, a teoria da amostragem nos permite expressar as abundâncias numa amostra como uma função de parâmetros da comunidade. A primeira aplicação desta ideia poderosa é de Sir Ronald Fisher que, tomando uma distribuição gama para descrever a DAE da comunidade e uma distribuição Poisson para a amostragem, demonstrou que a distribuição de probabilidades das abundâncias na amostra seguiria uma distribuição binomial negativa. Fisher fez então um dos parâmetros da binomial negativa tender a zero para deduzir o modelo de série logarítmica[24], um dos modelos de DAE mais conhecidos e utilizados. A série logarítmica tem um único parâmetro, que expressa a diversidade de espécies na comunidade de onde foi tomada a amostra, hoje conhecido como alfa de Fisher.

Nem toda combinação de modelos de DAE da comunidade e de processos amostrais resulta em integral com solução analítica conhecida, porém, mesmo nestes casos, é possível obter aproximações numéricas muito boas. Os recursos computacionais que temos hoje tornam tanto as soluções analíticas como as numéricas acessíveis a todos que tenham um conhecimento básico de matemática e estatística. No apêndice, apresento um exemplo trabalhado, para enfatizar essa afirmação e encorajar os leitores.

Ajuste dos modelos: o método da máxima verossimilhança

Os modelos de DAE utilizados aqui possuem um ou mais parâmetros, que determinam como as probabilidades se distribuem pelos valores de abundância (figura 4). Podemos pensar nos parâmetros como "botões" que giramos para mudar a forma da curva de valores esperados pelo modelo[25]. Portanto, depois de escolher um modelo, precisamos descobrir os valores dos parâmetros que resultam no melhor ajuste às abundâncias que observamos em nossa

[23] Esta equação é truncada porque ignora as espécies não incluídas na amostra. Veja o apêndice para este e outros detalhes da dedução.

[24] FISHER, R. *et al.*. *Op. cit.*, PIELOU, E. C. *Op. cit.*, 1977.

[25] A analogia com botões é usada em programas de ensino de estatística, como os disponíveis no portal *Statistics Online Computational Resource* da Universidade da Califórnia (*http://www.socr.ucla.edu/htmls/SOCR_Distributions.html*). Nesses programas, há de fato botões de controle para o usuário mudar os valores dos parâmetros e visualizar as mudanças resultantes na curva de distribuição de probabilidades.

amostra. Há muitas maneiras de se obter este ajuste, mas o método que apresentarei tem sólida base teórica e muitas propriedades extremamente convenientes. Um de seus méritos é que ele parte de uma ideia muito simples: para cada conjunto de dados, há valores dos parâmetros que são mais plausíveis do que outros. Mais ainda, entre todos os valores possíveis do parâmetro há um que é o de maior plausibilidade, por isso chamado estimativa de máxima verossimilhança. No exemplo a seguir mostro como obtemos essas estimativas e porque constituem bons palpites para os valores reais dos parâmetros da comunidade.

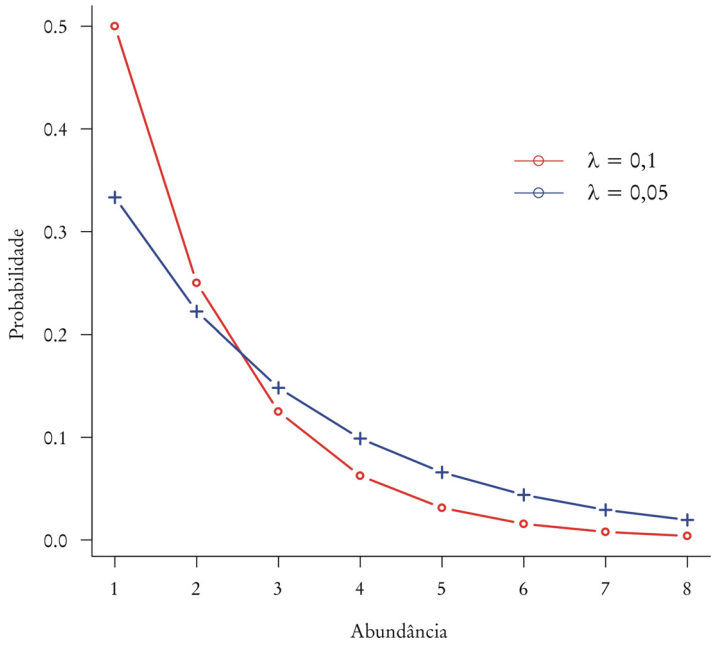

Figura 4: Efeito do valor do parâmetro sobre um modelo de distribuição de abundância de espécies, a Poisson-exponencial truncada (equação). O modelo atribui probabilidades a cada valor de abundância das espécies presentes numa amostra Poisson tomada de uma comunidade com distribuição exponencial (figura 3). Este modelo possui um único parâmetro, lambda (λ), preservado da distribuição exponencial. Valores mais altos de λ tornam abundâncias baixas mais prováveis, e a curva decai mais rapidamente, o que permite interpretar este parâmetro como um índice de dominância.

Vamos supor que amostramos 10% dos indivíduos de uma comunidade e observamos sete espécies, com as seguintes abundâncias: {8 6 6 4 3 1 1}. Com o modelo de distribuição Poisson-exponencial (equação 4), a probabilidade de que uma espécie tenha oito indivíduos amostrados é de

$$L(\lambda) = \frac{\lambda 0,1^{8-1}}{(\lambda + 0,1)^8} = \frac{10^{-7}\lambda}{(\lambda + 0.1)^8} \quad (5)$$

Note que agora temos uma função que nos retorna o valor de probabilidade atribuído pelo modelo a uma observação fixa, para cada valor do parâmetro. A chamamos *função de verossimilhança*, para distingui-la das funções de distribuição de probabilidades. Como seu nome indica, as distribuições de probabilidades nos retornam as probabilidades atribuídas a cada observação possível, conforme valores fixos dos parâmetros (figura 2, por exemplo). Funções de verossimilhança são úteis quando temos observações, mas não conhecemos os parâmetros. Distribuições de probabilidade servem para prever observações, se conhecemos os parâmetros.

A função de verossimilhança possui um máximo (figura 5A), o que nos fornece um critério simples para escolher o valor do parâmetro: será aquele que atribuir a maior probabilidade ao dado de abundância que temos. Este valor é a *estimativa de máxima verossimilhança* do parâmetro. Tal critério é uma aplicação da Lei da Verossimilhança: se temos vários modelos possíveis (no caso, funções com diferentes valores dos parâmetros) e cada um atribui uma probabilidade diferente ao que observamos, o modelo mais plausível é aquele que atribui a maior probabilidade à nossa observação.

A Lei da Verossimilhança tem um forte apelo intuitivo, o que é demonstrado por um famoso experimento mental.[26] Imagine uma urna de onde tiramos bolas por sorteio. Todas as bolas apresentam a mesma chance de serem sorteadas, e são devolvidas à urna em seguida. Há duas hipóteses propostas: (H1) a urna contém apenas bolas pretas e (H2) metade das bolas da urna são pretas, e metade são brancas. Você sorteou três bolas, e todas eram pretas. Se temos que escolher entre essas hipóteses, ficamos com H1, pois ela explica melhor o que aconteceu. Podemos comprovar a intuição calculando as probabilidades que cada hipótese atribui à observação: ela é de um para H1, e de apenas $0,5^3 = 0,125$ para H2. Dada a observação, H1 é mais plausível que H2, pois atribui maior probabilidade ao que ocorreu. Podemos dizer, então, que obter três bolas pretas em três sorteios tem valor de evidência a favor de H1, em comparação com H2, pois torna a primeira hipótese $1/0,125 = 8$ vezes mais plausível que a segunda.

O exemplo da urna também nos mostra como calcular verossimilhanças quando temos mais de uma observação. Se as observações são independentes, basta multiplicar as probabili-

[26] ROYALL, R. *Statistical Evidence*: A Likelihood Paradigm. London: Chapman & Hall, 2000.

dades atribuídas a cada uma pelo modelo. Os modelos de distribuição de abundâncias de espécies assumem esta premissa de independência[27], com a qual podemos calcular a função de verossimilhança dos modelos de DAE, como no caso de nossa amostra fictícia, para o modelo Poisson-exponencial:

$$L(\lambda) = \frac{10^{-7}\lambda}{(\lambda+0.1)^8} \times 2\frac{10^{-5}\lambda}{(\lambda+0.1)^6} \times \frac{10^{-3}\lambda}{(\lambda+0.1)^4} \times \frac{10^{-2}\lambda}{(\lambda+0.1)^3} \times 2\frac{\lambda}{(\lambda+0.1)} \quad (6)$$

Pelo fato de produtos de probabilidades como este tenderem rapidamente a números muito pequenos, em geral usamos o seu logaritmo. Outra vantagem desta transformação é que o logaritmo de um produto é uma soma, o que é bem mais fácil de manipular e interpretar. Com isto temos a *função de log-verossimilhança*. Encontrando o máximo dessa função obtemos a estimativa de máxima verossimilhança dos parâmetros do modelo, para nossa amostra (figura 5B). Em casos simples, as funções de log-verossimilhança e seu máximo podem ser deduzidos analiticamente. Em geral, no entanto, é necessário usar aproximações numéricas, com a ajuda de programas de matemática e estatística.[28]

[27] Portanto assumimos que as abundâncias das espécies são realizações independentes da mesma distribuição de probabilidades, ou seja, são variáveis aleatórias independentes e identicamente distribuídas (iid). Esta premissa é avaliada criticamente por autores como ALONSO, D. *et. al.*. *Op. cit.* e ENGEN, S. Comments on two different approaches to the analysis of species frequency data. *Biometrics, International Biometric Society*, 33:205-213, 1977.

[28] BOLKER, B. *Ecological models... Op. cit.* faz uma excelente introdução a todos estes procedimentos.

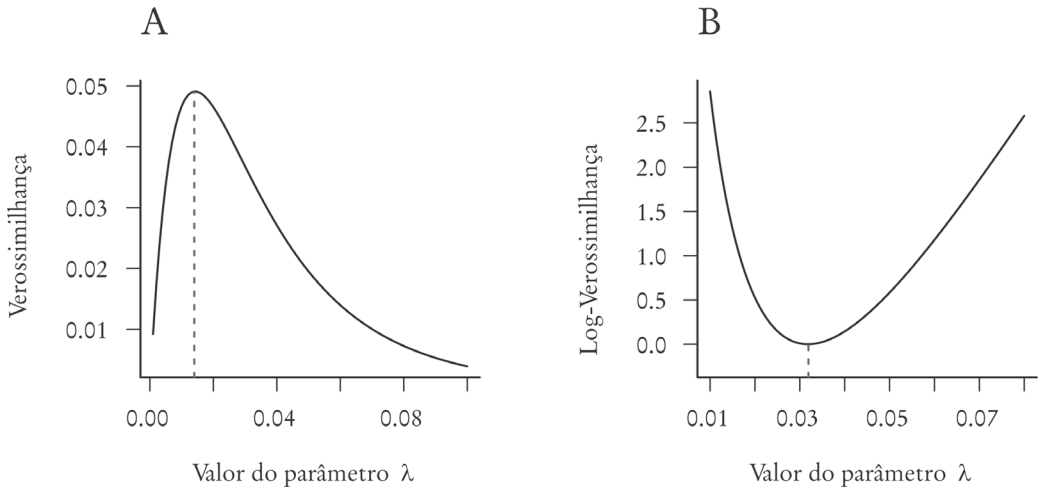

Figura 5: Funções de verossimilhança da distribuição Poisson-exponencial truncada (equação 4) (A) Função de verossimilhança dada à observação de uma espécie com abundância de oito indivíduos na amostra. O valor do parâmetro λ, que resulta no maior valor da função, é a estimativa de máxima verossimilhança deste parâmetro, para esta observação, e está indicado pela linha pontilhada. (B) Função de log-verossimilhança negativa, dada à observação de sete espécies na amostra, com abundâncias de {8 6 6 4 3 1 1}. O máximo da função de verossimilhança corresponde ao mínimo do negativo de seu logaritmo. O valor do parâmetro λ, que resulta neste mínimo, é a estimativa de máxima verossimilhança de λ para esta amostra, e está indicado pela linha pontilhada.

Seleção de Modelos

Agora que descrevemos como modelos de DAE podem ser deduzidos e ajustados, necessitamos de um critério para selecionar entre vários modelos candidatos[29]. Cada modelo tem seu valor máximo de verossimilhança, que é usado para estimar seus parâmetros. É este mesmo valor que usamos para comparar modelos diferentes. Novamente lançamos mão da Lei da Verossimilhança: preferimos o modelo que atribuir a maior probabilidade aos dados observados, ou seja, aquele que tiver o maior valor da função de verossimilhança, ou de log-verossimilhança.

Portanto, ajuste e seleção de modelos são indissociáveis. Escolhemos modelos candidatos, identificamos os parâmetros de cada um que resultam no melhor ajuste, e então comparamos o melhor ajuste de cada modelo. Uma boa analogia do processo todo é a da compra de sapatos: primeiro separamos os modelos que nos agradam, em seguida experimentamos e selecionamos o número que melhor nos serve de cada modelo, para então decidir qual modelo comprar[30].

Podemos usar a máxima verossimilhança dos modelos para compará-los, ou alguma função derivada dela. O *critério de informação de Akaike (AIC)* é hoje a métrica mais utilizada para este fim. Tendo em vista a log-verossimilhança máxima de um modelo (L) e seu número de parâmetros (k), o AIC é dado por:

$$AIC = 2k - 2L \qquad (7)$$

O AIC estima a distância relativa de cada modelo ao "verdadeiro", que não precisa ser conhecido e nem estar entre os competidores. De qualquer forma, o melhor entre os modelos comparados tende a apresentar a menor distância a este modelo hipotético[31]. Assim, o modelo mais plausível terá o menor valor de AIC. Para expressar quão piores são os demais modelos, calculamos a diferença de seu AIC em relação ao do melhor modelo:

$$\Delta AIC = AIC - min(AIC) \qquad (8)$$

Um regra canônica é que modelos com $\Delta AIC \leq 2$ são igualmente plausíveis. Como muitas outras convenções, sua principal vantagem é evitar a polêmica. Em nosso experimento mental, essa decisão equivale aproximadamente a aceitar que todas as bolas na urna são pretas, se três ou mais sorteios só retornaram bolas pretas. Se há motivos para maior rigor, valores de corte mais altos podem ser usados, tendo em

[29] A introdução básica sobre seleção de modelos em ecologia, na qual esta seção se baseia, é: BURNHAM, K. P. & ANDERSON, D. R. *Model Selection and Multimodel Inference* – A Practical-Theoretic Approach. New York: Springer-Verlag, 2002.

[30] Gonçalo Ferraz (comunicação pessoal).

[31] A dedução do AIC demonstra que o modelo mais próximo do verdadeiro entre os comparados terá, em média, valores de AIC menores, se repetimos a seleção para várias amostras. Como se trata de uma esperança estatística, no entanto, é possível que um modelo pior tenha menor AIC em algumas das amostras. Este erro diminui com o aumento do tamanho amostral. Veja BURNHAM, K. P. & ANDERSON, D. R. *Model Selection and Multimodel Inference... Op. cit.*

mente que o AIC está em escala de logaritmos naturais. O valor convencionado considera equivalentes modelos até $e^2 = 7{,}4$ vezes menos plausíveis que o de menor AIC. Aumentando este critério para $\Delta AIC \leq 3$, seríamos muito rigorosos, pois mesmo modelos que sejam $e^3 = 20$ vezes menos plausíveis seriam considerados equivalentes.

Uma outra maneira de expressar a distância entre os modelos é o peso de evidência ou peso de Akaike (wAIC), calculado para cada um dos modelos:

$$wAIC_i = \frac{e^{-\Delta_i/2}}{\sum_{j=1}^{n} e^{-\Delta_j/2}} \qquad (9)$$

Onde Δ são os ΔAIC, e os índices indicam cada um dos n modelos competidores. Os pesos de evidência dos modelos comparados somam um, o que nos dá uma escala simples de qualidade relativa. Além disso, o peso de evidência estima a probabilidade de cada modelo ser selecionado, se os dados forem reamostrados com reposição e uma nova seleção for feita.[32]

[32] BURNHAM, K. P. & ANDERSON, D. R. *Model Selection and Multimodel Inference... Op. cit.*

O resultado da seleção de modelos é apresentado em uma tabela com os valores de AIC, ΔAIC e wAIC. Em nosso exemplo fictício, os modelos Poisson-exponencial e Série Logarítmica são descrições igualmente plausíveis dos dados (tabela 1). Empates como este são comuns quando as amostras são pequenas, indicando corretamente que os dados não trazem informação suficiente para identificar conclusivamente o melhor modelo. À medida que aumentamos a amostra, a chance de identificarmos o melhor modelo aumenta.[33]

[33] BURNHAM, K. P. & ANDERSON, D. R. *Model Selection and Multimodel Inference... Op. cit.*

ROYLE, J. & DORAZIO, R. *Hierarchical modeling and inference in ecology*. London: Academic Press, 2008.

Tabela 1: Comparação do ajuste de três modelos teóricos de DAE a abundâncias de espécies em uma amostra fictícia. Os modelos são comparados com o Critério de Informação de Akaike (AIC), que é uma função da log-verossimilhança e do número de parâmetros de cada modelo. O modelo mais plausível é o de menor AIC. Modelos com uma diferença de AIC igual ou menor que 2 são considerados igualmente plausíveis. O $\Delta AICc$ indica esta diferença em relação ao modelo mais plausível, diagnosticando modelos equivalentes. O peso de evidência (wAIC) expressa a plausibilidade de cada modelo em uma escala que soma um. Neste exemplo usamos o AIC corrigido para pequenas amostras (AICc, Burham & Anderson).

Modelo	N Parâmetros	Log-Verossimilhança	AICc	$\Delta AICc$	wAICc
Poisson-exponencial	1	-16,03	34,9	0,0	0,67
Série Logarítmica	1	-17,03	36,9	2,0	0,24
Poisson-lognormal	2	-15,89	38,8	3,9	0,09

Conclusões

Mostrei como é possível construir modelos probabilísticos para descrever as distribuições de abundâncias das espécies em uma amostra. Os modelos combinam processos biológicos – que governam a variação das abundâncias das espécies nas comunidades – com a variação dessas abundâncias causada pela maneira como foi tomada a amostra. Além disso, dois níveis estabelecem uma relação explícita entre abundâncias, constituindo, portanto, modelos hierárquicos.[34] Uma vez criados os modelos, a abordagem da verossimilhança e a seleção de modelos fornecem um protocolo geral para compará-los. Todo o procedimento se resume a cinco passos:

1. Defina os modelos de DAE da comunidade.
2. Defina o modelo para representar a amostragem.
3. Deduza os modelos compostos (equação 3), ou obtenha uma aproximação numérica.
4. Ajuste cada modelo aos dados com máxima verossimilhança.
5. Compare os modelos com o AIC.

Mas, por que nos preocuparmos com as DAE? As distribuições de abundâncias das espécies são atributo básico das comunidades, e em ecologia um dos poucos exemplos de padrão universal, que são as curvas côncavas. Se aceitamos tal padrão como princípio sintético que pode ser expresso por meio de modelos quantitativos, podemos usá-lo para deduzir novos enunciados, como regularidades e diferenças entre as comunidades. O uso de leis empíricas tem justificativas epistemológicas e, na prática, já promoveu importantes avanços teóricos em diversas ciências, incluindo a biologia[35]. Foi esta a motivação para o desenvolvimento e a aplicação de modelos de DAE, muito bem resumida por Pielou[36]:

> *If it should turn out that one single form of probability distribution with a small number of parameters (say two or three) fitted the data from the majority of observed communities, with only the parameter values varying from one community to the another, interesting relationships might be discovered between the values of the parameters and the types of community they describe.*

As abordagens de dedução, ajuste e comparação de modelos aqui apresentadas ampliaram as possibilidades de implementação dessa ideia simples, trazendo-a de volta ao centro das atenções da ecologia de comunidades. Contribuições neste campo possuem alto potencial de impacto para esta área de conhecimento básico, bem como para quaisquer de suas aplicações que envolvam monitoramento da diversidade biológica.

[34] ROYLE, J. & DORAZIO, R. *Hierarchical modeling and inference in ecology*. *Op. cit.*, p. 16-17 definem modelos hierárquicos como uma distribuição de probabilidades compostas de outras duas: (1) a distribuição que descreve como os dados foram obtidos, condicionada à (2) distribuição que descreve a dinâmica de um processo ecológico.

[35] EL-HANI, C. N. Generalizações ecológicas. *Oecologia Brasiliensis*, 10:17-68, 2006.

[36] PIELOU, E. C. *Mathematical Ecology*. *Op. cit.*. p. 269.

Agradecimentos aos colegas João Luis Ferreira Batista, Camila Mandai, Paulo de Marco Jr., Alexandre Adalardo, Gonçalo Ferraz, Charbel El-Hani, Pedro Rocha, Thomas Lewinsohn, André Belloto; aos companheiros do Laboratório MP3, Glauco Machado, Pérsio Santos e Paulo Guimarães Jr.; e ao CNPq pela Bolsa de Produtividade em Pesquisa.

Paulo Inácio K. L. Prado é graduado em Biologia e professor do Departamento de Ecologia do Instituto de Biociências da Universidade de São Paulo.
prado@ib.usp.br

APÊNDICE

Um exemplo da aplicação da Teoria de Amostragem

Omiti detalhes matemáticos no texto principal, como uma estratégia para tornar a lógica da teoria da amostragem acessível a um público mais amplo, e facilitar o posterior entendimento desses detalhes. Não há estatística sem matemática. Se fui bem sucedido em despertar seu interesse pelo assunto, o próximo passo é adquirir conhecimento matemático suficiente para compreender a dedução dos modelos. Não é muito: bastam algumas noções de cálculo numérico e teoria de probabilidades. Aprofundando-se um pouco mais você será capaz de deduzir seus próprios modelos. Uma excelente introdução a esta matemática operacional para biólogos é o livro de Sarah Otto e Troy Day[37]. Outra ferramenta muito útil são programas para manipulação de expressões simbólicas, como derivadas, integrais ou operações algébricas. As passagens do exemplo neste apêndice foram conferidas com o programa MAXIMA[38], sob a interface gráfica WXMAXIMA[39], que são *softwares* livres sob a licença GNU[40], e estão disponíveis gratuitamente na internet. Por fim, mas não por último, busque sempre o apoio de profissionais com sólida formação matemática, seja pelo aprendizado formal, seja por colaboração informal[41].

Neste exemplo vamos deduzir a distribuição Poisson-exponencial truncada (equação 4). A função abaixo é um modelo para descrever a probabilidade de que uma espécie tenha certa abundância numa amostra, dado que uma fração *a* da comunidade seja amostrada. O primeiro passo é definir a distribuição de probabilidades que descreve as abundâncias na comunidade. Aceitando que na comunidade há um número muito grande de indivíduos, podemos usar uma função contínua, o que irá facilitar os cálculos. Vamos então supor que a distribuição das abundâncias na comunidade (n) segue uma exponencial (figura 2):

$$f(n) = \lambda e^{-n\lambda} \qquad \text{(A1)}$$

Se a amostragem foi feita de maneira que cada indivíduo é incluído na amostra independentemente, a distribuição de Poisson descreve a probabilidade de cada abundância (x) numa amostra, dada uma taxa média de captura ou inclusão (θ):

$$g(x) = \frac{\theta^x e^{-\theta}}{x!} \qquad \text{(A2)}$$

[37] OTTO, S. P. & DAY, T. *A biologist's guide to mathematical modeling... Op. cit.*

[38] Disponível no sítio *http://maxima.sourceforge.net/*

[39] *http://wxmaxima.sourceforge.net/wiki/index.php/Main_Page*

[40] *http://www.gnu.org/licenses/licenses.html#GPL*

[41] Para este apêndice contei com o auxílio do matemático André Belloto.

A distribuição Poisson é discreta, o que parece mais razoável para amostras que possam ter números pequenos de indivíduos coletados. O valor esperado de abundância é igual ao seu único parâmetro, θ. Se conhecemos a fração de indivíduos amostrados na comunidade (*a*), bem como a abundância na comunidade de certa espécie, o valor esperado na amostra será o produto *a* x *n*. Substituindo o parâmetro da Poisson por este produto, temos a distribuição de probabilidades da abundância de uma espécie na amostra, *dado um certo valor de abundância n* na comunidade:

$$g(x \mid n) = \frac{(an)^x e^{-an}}{x!} \quad \text{(A3)}$$

A distribuição de probabilidades de cada abundância na amostra, *h(x)*, é a integral do produto das equações A1 e A3 (ver dedução da equação 3 no texto principal):

$$h(x) = \int_0^\infty f(n)\, g(x \mid n)\, dn = \int_0^\infty \lambda e^{-n\lambda} \frac{(an)^x e^{-an}}{x!} dn =$$

$$\int_0^\infty \frac{a^x \lambda}{x!} n^x e^{-n(a+\lambda)} dn \qquad \text{(A4)}$$

Para resolver esta integral, vamos simplificar os termos constantes, fazendo $b = a^x \lambda / x!$ e $c = a + \lambda$:

$$h(x) = \int_0^\infty b n^x e^{-nc}\, dn \quad \text{(A5)}$$

Com a substituição $y = nc$ temos:

$$h(x) = \int_0^\infty b\left(yc^{-1}\right)^x e^{-y} c^{-1}\, dy = \int_0^\infty b y^x c^{-(x+1)} e^{-y}\, dy \quad \text{(A6)}$$

As constantes podem ser retiradas da integral:

$$h(x) = bc^{-(x+1)} \int_0^\infty y^x e^{-y}\, dy \quad \text{(A7)}$$

A integral $\int_0^\infty y^x e^{-y} dy$ corresponde à função Gama de $x+1$, que é igual a fatorial de x quando este é inteiro[42], como no caso de abundâncias. Com esta substituição e as expressões originais das constantes *b* e *c* temos, finalmente:

[42] A função Gama é muito usada em estatística, e uma boa apresentação está em OTTO, S. P. & DAY, T. *Op. cit.* Para muitas outras integrais menos conhecidas, mas definidas, há catálogos como GRADSHTEYN I. S. & RYZHIK I. M. *Table of Integrals, Series and Products.* 6th ed. London: Academic Press, 2000.

$$h(x) = \frac{a^x \lambda}{x!}(\lambda + a)^{-(x+1)} x! = \frac{a^x \lambda}{(a+\lambda)^{x+1}} \quad \text{(A8)}$$

A equação A8 é a distribuição Poisson-exponencial, que atribui uma probabilidade a cada valor de abundância de uma espécie na amostra, incluindo a abundância zero, que é:

$$h(0) = \frac{a^0 \lambda}{(a+\lambda)^{0+1}} = \frac{\lambda}{a+\lambda} \quad \text{(A9)}$$

Em situações reais não temos acesso a tais observações de abundância zero, pois não sabemos quantas espécies da comunidade não foram incluídas na amostra. Precisamos então de um modelo que nos retorne as probabilidades em relação à soma das probabilidades dos valores observáveis, que é *1-h(0)*. Portanto, basta dividirmos o resultado da função original por este termo:

$$P(x) = \frac{a^x \lambda (a + \lambda)^{-(x+1)}}{1 - \frac{\lambda}{a + \lambda}} = \frac{\lambda a^{x-1}}{(\lambda + a)^x} \quad \text{(A10)}$$

E assim chegamos à função Poisson-exponencial truncada (equação 4), usada no exemplo do texto principal.