# AMOSTRAR MUITAS PARCELAS PEQUENAS OU POUCAS PARCELAS GRANDES PARA ASSOCIAR UMA COMUNIDADE BIOLÓGICA A VARIÁVEIS AMBIENTAIS? SIMULAÇÕES COM UMA COMUNIDADE DE ÁRVORES DO CERRADO

**F.S.C. Takahashi**

Universidade de Brasília, Instituto de Ciências Biológicas, Departamento de Ecologia-C.P. 04457 CEP 70919 - 970 - Brasília, DF - Brasil-fredtaka@gmail.com

## INTRODUÇÃO

Determinar os fatores que influenciam a distribuição das espécies em uma comunidade biológica é um dos principais temas em estudos ecológicos. A avaliação de dados obtidos neste contexto apresenta grande complexidade, considerando a natureza multidimensional dos dados que sintetizam a ocorrência ou abundância de múltiplas espécies em função de gradientes ambientais distintos. Diversas técnicas estatísticas multivariadas foram apresentadas como possíveis formas de analisar tais dados, com conjuntos diferentes de premissas e objetivos analíticos (Magnusson & Mourão, 2005).

Uma das análises mais utilizadas por ecólogos para estudar esta associação é a Análise de Correspondência Canônica (CCA) (Birks *et al.,* 1998). A descrição detalhada da técnica está disponível em Ter Braak (1986) e em Legendre e Legendre (1998), porém algumas de suas propriedades merecem destaque aqui. Um dos principais atrativos desta técnica é que a matriz de variáveis ambientais interfere no cálculo dos vetores de ordenação, de modo que a relação dos vetores de ordenação e a combinação linear das variáveis ambientais seja maximizada (Legendre e Legendre, 1998). Esta técnica de ordenação com análise de "gradientes diretos" permite avaliar hipóteses ecológicas formuladas *a priori* ao utilizar toda a variação na comunidade relacionada às variáveis ambientais (Legendre e Legendre, 1998).

Contudo a aplicação da CCA apresenta uma série de dificuldades. Decisões que influenciam marcantemente na qualidade dos resultados obtidos devem ser tomadas não só na análise dos dados, mas também para delinear sua coleta. Um ponto fundamental é determinar qual o tamanho da amostragem que deverá ser conduzida de forma a representar bem o sistema de estudo. Um conceito prevalente entre pesquisadores é que técnicas estatísticas complexas, como a CCA, necessitam de um número relativamente elevado de amostras para obtenção de análises válidas. Porém tal requisito, de grandeza subjetiva, é dificilmente alcançado em trabalhos em campo, sendo então o tamanho da amostragem escolhido com base em fatores logísticos e não técnicos (ex. maior amostragem que o tempo disponível permitir). Paralelamente à escolha do número de amostras de um experimento, é necessário estabelecer o tamanho destas unidades amostrais. Tal ponto apresenta dificuldades adicionais de decisão, devido à ausência de estudos teóricos ou empíricos enfocando a CCA que possam guiar a decisão. Neste cenário de recursos limitados (ie. tempo, material, equipe), fica evidente a dificuldade de alocar o esforço amostral entre número de amostras e tamanho desta. Com isto a decisão de se é melhor amostrar muitas unidades amostrais pequenas ou, em outro extremo, poucas unidades amostrais grandes é tomada de maneira arbitrária.

## OBJETIVOS

O objetivo deste trabalho foi testar diferentes delineamentos amostrais que podem ser empregados em estudos da associação da comunidade biológica com caracteres ambientais de modo a avaliar a consistência dos resultados obtidos. Enfoquei neste estudo uma comunidade simulada equivalente a de árvores do cerrado, de modo a definir neste sistema qual a melhor estratégia de amostragem: poucas parcelas grandes ou várias pequenas.

## MATERIAL E MÉTODOS

Como base para avaliar as diferentes formas de amostragem, criei um modelo representando a comunidade de árvores do cerrado e de fatores ambientais que influenciam esta comunidade. Inicialmente criei 10 matrizes (1000 linhas e 1000 colunas) cada uma representando uma variável ambiental hipotética. Estas foram elaboradas por meio de equações matemáticas escolhidas de modo que sejam representados gradientes ambientais com variações graduais ou bruscas formando diferentes padrões. Paralelamente, elaborei uma matriz de igual tamanho para representar a comunidade de

árvores. Nesta matriz, cada célula recebeu o valor de 0 a 54, representando desta forma se uma unidade de área estava vazia (valor 0) ou ocupada por alguma espécie de árvore (valores de 1 a 54, cada um representando uma espécie diferente). Para a alocação das espécies na matriz, utilizei uma função probabilística binomial em que a probabilidade de ocorrência de uma espécie em cada célula era proporcional ao valor de uma das variáveis ambientais nesta posição. Adicionalmente a esta presença de espécies condicionada a variáveis ambientais, aloquei uma menor quantidade de células para serem ocupadas por cada espécie independentemente de qualquer variável ambiental por meio de sorteio com probabilidade de ocorrência igual para todas as células. Para as células nas quais foram designadas mais de uma espécie, realizei sorteios para definir qual destas ocuparia cada célula. Defini as probabilidades de ocorrência associadas a gradientes ambientais e probabilidades de ocorrência independentes de posição de modo que a matriz resultante apresente abundâncias relativas de cada espécie aproximadamente iguais às observadas em amostragem fitossociológica numa área de cerrado *sensu stricto* de Brasília (Assunção & Felfili, 2004).

Para realizar as comparações das diferentes estratégias de amostragem, sorteei coordenadas da matriz, de forma a simular o sorteio de parcelas no ambiente natural. A partir destas coordenadas de referência, registrei os valores da matriz de espécies nas coordenadas sorteadas e dos valores adjacentes a este, conforme a opção de delineamento testada, simulando assim a amostragem de uma parcela no ambiente natural. Paralelamente, registrei os valores de cinco matrizes de variáveis ambientais nas coordenadas da parcela simulada, obtendo então os valores médios de cada variável ambiental em cada parcela. Repeti este procedimento de forma a simular a obtenção de amostras em um experimento real. Deste modo, pude simular diferentes estratégias de coleta de dados em campo, variando o tamanho da parcela e o número de parcelas.

A primeira estratégia testada (T1) consistia na obtenção de 125 parcelas contendo 4 células da matriz em cada parcela. A segunda estratégia (T2) consistia na obtenção de 56 parcelas contendo 9 células cada. A terceira abordagem testada (T3) consistia na obtenção de 32 parcelas com 16 células cada. Para avaliar qual das respostas obtidas melhor refletia a realidade do sistema simulado, realizei uma amostragem de referência (R), composta de 1000 parcelas com 49 células. Desta forma, com exceção de R, os demais esquemas de amostragem mantinham um total de células analisadas aproximadamente igual (T1: 500 células, T2: 504 células, T3: 512 células). Para avaliar a variabilidade dos resultados dentro de um mesmo esquema de amostragem, repeti cada um deles três vezes.

Analisei os dados obtidos do modelo de cerrado utilizando CCA. Para avaliar a qualidade do ajuste de cada estratégia de amostragem, calculei a proporção da inércia total que é associada aos dois primeiros eixos canônicos. Para calcular a influência que cada variável ambiental tinha na comunidade em cada estratégia de amostragem, calculei a correlação entre cada variável ambiental e o escore canônico de cada parcela. Estes valores, denominados *intraset correlations*, apresentam interpretação similar ao dos coeficientes canônicos, porém são mais confiáveis em situações em que há correlação entre as variáveis ambientais (Ter Braak, 1986). Avaliei também a posição relativa que cada espécie ocuparia ao longo de cada gradiente ambiental. Para isto, calculei o ranque das *weighted average* de cada espécie em função de cada variável ambiental. Em cada CCA, testei a significância dos *constraints* por meio de testes de permutação com modelo reduzido utilizando alfa igual a 0,001. Para sumarizar os resultados de uma mesma estratégia de amostragem, calculei o valor médio dos indicadores de qualidade de ajuste, influência de cada variável ambiental (ie. *intraset correlation*) e posição relativa das espécies nos gradientes ambientais (ie. ranque das espécies em função do seu *weighted average* em cada variável ambiental). Adicionalmente, para os últimos dois indicadores, calculei também o valor médio dos desvios padrão de todas as *intraset correlations* e ranques de espécies. Para avaliar os resultados das diferentes estratégias de amostragem (T1, T2 e T3), comparei os resultados destas com o obtido na amostragem de referência (R). Desta forma, calculei a correlação entre a *intraset correlation* de R com T1, T2 e T3 para os dois primeiros eixos da CCA. Paralelamente, calculei as diferenças dos ranques obtidos pelas 20 espécies mais abundantes ao longo de cada gradiente ambiental nas diferentes estratégias de amostragem, da seguinte forma:
Erro geral de ranqueamento = $\sum$ módulo[ranque da espécie(i) ao longo da variável(j) estimado em R - ranque da espécie(i) ao longo da variável(j) estimado em T]
sendo i de 1 a 20, representando as 20 espécies mais abundantes; j de 1 a 5 representando as 5 variáveis ambientais analisadas. Adicionalmente repeti esta avaliação desconsiderando as variáveis ambientais com baixo *intraset correlation*, uma vez que estas habitualmente não são utilizadas na interpretação dos resultados.
Em todas as etapas deste trabalho utilizei o software R versão 2.7.2 (R Foundation for Statistical Computing). Para as análises de correspondência canônica, utilizei o pacote "Vegan: Community Ecology Package" versão 1.15 - 0 (Oksanen *et. al*, 2008).

## RESULTADOS

Todos as CCA analisadas apresentaram associação significativa ($p < 0{,}001$) indicando que todas as estratégias de amostragem conseguiram captar alguma associação entre comunidade e ambiente físico. A qualidade dos ajustes foi superior em T3, assumindo valores próximos à R (média da qualidade do ajuste: R: 0,193; T1: 0,045; T2: 0,093; T3: 0,177). Isto indica que a estratégia T3 conseguiu explicar uma maior proporção da variação nos dados que os demais nos dois primeiros eixos, que são os que geralmente são utilizados na interpretação dos resultados.
A análise do peso das variáveis ambientais na comunidade avaliado pela *intraset correlation* indicou boa consistência dos resultados de T1 e T3. A correlação entre *intraset correlations* de cada estratégia com R apresentou os valores de 0,932; 0,698 e 0,824 para T1, T2 e T3 no eixo 1 e de 0,684; 0,634 e 0,891 para o eixo 2. Porém, a média dos desvios padrão dos valores de *intraset correlations* mais elevados no T1, indicam a menor confiabilidade dos resultados obtidos

nesta estratégia do que nas demais (T1: 0,326; T2: 0,248 e T3: 0,265). Os dois maiores *intraset correlations* associados aos dois primeiros eixos de R, T1 e T3 indicaram três variáveis ambientais como as mais influentes na comunidade, enquanto T2 conduziria a seleção de uma variável ambiental distinta (valores não apresentados).

A posição relativa das 20 espécies mais abundantes nos gradientes ambientais indicou que as estratégias T1 e T3 obtiveram resultados relativamente próximos enquanto T2 apresentou maior erro geral de ranqueamento (T1: 226; T2: 244 e T3: 221). O ranqueamento também pode ser avaliado considerando somente os gradientes ambientais que apresentaram maior relação com composição de espécies. Desta forma, considerando somente as duas variáveis ambientais mais importantes de cada eixo, os erros gerais de ranqueamento foram próximos entre os diferentes esquemas de amostragem (T1: 124; T2: 121; T3: 119). Porém, é importante destacar que T1 apresentou maior variação na estimativa de ranque do que as demais estratégias, o que resultaria numa menor confiabilidade de seus resultados (média dos desvios padrão dos ranques dentro de cada estratégia, utilizando todas as variáveis ambientais: T1: 4,972; T2: 4,810 e T3: 4,469; e desconsiderando variáveis ambientais com baixas *intraset correlations*: T1: 4,466; T2: 4,431 e T3: 4,201).

Todas as estratégias de amostragem analisadas forneceram dados que permitiram interpretações realísticas por meio de CCA. Entretanto, a qualidade dos resultados obtidos em cada estratégia de amostragem foi bastante diferenciada. A estratégia de amostragem baseada em um número relativamente pequeno de parcelas grandes (T3) apresentou simultaneamente resultados menos variados e mais realistas, sendo por isto a melhor das opções analisadas. Vale notar inclusive, que esta estratégia de amostragem é a com maior facilidade de implementação em campo, uma vez que envolve menor coleta de dados ambientais e menor deslocamento entre sítios.

Os resultados intermediários de T2 foram compatíveis com o esperado, considerando que esta estratégia tinha balanceamento intermediário do número de amostras e tamanho das parcelas. Entretanto, o pior ajuste dos dados coletados na estratégia T1 merece atenção especial, uma vez que tal delineamento empregava o maior número de unidades amostrais independentes, o que geralmente é desejável para análises estatísticas. Dentre as possíveis causas está a grande variação na composição de espécies que ocorre ao amostrar parcelas muito pequenas. Desta forma, a ausência de espécies devido ao acaso em parcelas com condições ambientais favoráveis a sua ocorrência gera ruído nos dados. Conforme demonstrado por McCune (1997), a CCA é fortemente prejudicada por este tipo de variação. Outro fator que pode ter prejudicado o ajuste dos dados de T1 e favorecido T3 é o maior número de espécies detectadas em T1. Isto porque a presença de espécies raras tende também a gerar ruído nestas análises (Hirst & Jackson, 2007). A retirada de espécies pouco freqüentes em T1 potencialmente resultaria no melhor ajuste destes dados, porém, em muitos casos a informação referente às espécies pouco freqüentes é importante dentro dos objetivos do estudo.

A concepção de que em análises multivariadas complexas é necessária uma quantidade elevada de amostras independentes não se mostrou universalmente válida. Este estudo reforça a demonstração de McCune (1997) a respeito do efeito do ruído nos dados de forma que a avaliação do grau de ruído nos dados seja um fator importante no planejamento de experimentos. Desta forma, a realização de testes equivalentes aos apresentados no presente trabalho, empregando conjuntos de dados simulados, pode ser uma ferramenta importante para subsidiar a escolha do delineamento amostral. A construção de representações da comunidade biológica de interesse e de gradientes ambientais hipotéticos poderiam ser etapas valiosas no planejamento de um estudo, nas quais o pesquisador poderia representar o conhecimento atual do sistema e suas hipóteses. Com isto, seria possível testar a amostragem mais eficiente e também obter uma previsão de resultados possíveis.

## CONCLUSÃO

A amostragem de um número relativamente pequeno de parcelas grandes (32 parcelas com 16 células cada) proporcionou resultados mais realistas e menos variáveis neste ambiente simulado de cerrado. Porém, é importante destacar que sistemas biológicos com propriedades distintas do utilizado neste trabalho, podem ter requisitos de amostragem diferenciados. Desta forma é prudente a repetição do protocolo demonstrado aqui para a adequação ao sistema de interesse.

## REFERÊNCIAS


Assunção, S.L. & Felfili, J.M. 2004. Fitossociologia de um fragmento de cerrado *sensu stricto* na APA do Paranoá, DF, Brasil. Acta Bot. Bras. 18(4):903 - 909

Birks, H. J. B., S. M. Peglar & H. A. Austin. 1998. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986 - 1996. Disponível online: http://www.bio.umontreal.ca/Casgrain/cca_bib/index.html

Hirst, C.N. & Jackson, D.A. 2007. Reconstructing community relationships: the impact of sampling error, ordination approach, and gradient length. Diversity Distrib. 13:361 - 371

Legendre, P. & Legendre, L. 1998. Numerical ecology. 2a. ed. Amsterdam, Elsevier Science. 853 p.

Magnusson, W.E. & Mourão, G. 2005. Estatística sem Matemática, 2a. ed. Londrina, Editora Planta. 138 p.

McCune, B. 1997. Influence of noisy environmental data on canonical correspondence analysis.
Ecology 78: 2617 - 2623

Oksanen J., Kindt R., Legendre P., O'Hara B., Simpson G.L., Solymos P., Stevens M.H.H. & Wagner H. 2008. Vegan: Community Ecology Package. R package version 1.15 - 0. Disponível online: http://cran.r - project.org/

Ter Braak, C.J.F. 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167 - 1179