Curso

Modelagem de biodiversidade: distribuição geográfica potencial de espécies

Marinez Ferreira de Siqueira - Centro de Referência em Informação Ambiental – CRIA
Francisco Candido Cardoso Barreto - Doutorando em Entomologia - Universidade Federal de Viçosa - MG

Sumário

Conceito de nicho e seu uso em modelagem

A primeiro idéia de nicho (Grinnell 1917) foi definida como sendo simplesmente os locais (habitats) onde os requisitos para uma determinada espécie viver e se reproduzir estão presentes. (Elton 1927) adicionou à idéia de lugar o nivel trófico que a espécie ocupa no ambiente biológico. (Gause 1934) adicionou a intensidade da competição entre espécies. Finalmente o termo nicho ecológico (Hutchinson 1957) foi definido como sendo um espaço com um hipervolume n-dimensional onde cada dimensão representa o intervalo de condições ambientais ou de recursos necessários para a sobrevivência e reprodução da espécie, tais como: temperatura, umidade, salinidade, pH, recursos alimentares, locais para nidificação, intensidade luminosa, pressão predatória, densidade populacional, entre outras.

Dentro deste conceito temos o conceito de nicho fundamental da espécie, que inclui os intervalos das condições ambientais necessárias para a existência da espécie, sem considerar a influência de interações bióticas, tais como competição e predação. O nicho realizado descreve a parte do nicho fundamental no qual a espécie realmente ocorre, ou seja, é delimitado por fatores bióticos. Desse modo, a área definida pelo nicho fundamental é, via de regra, maior que o nicho realizado.

Existem na literatura vários trabalhos que enfatizam a importância em esclarecer o objeto da modelagem (Austin 2002a, Oksanen and Minchin 2002, Rushton et al. 2004, Soberon and Peterson 2005, Araújo and Guisan 2006, Austin et al. 2006, Phillips et al. 2006). Na prática, é importante saber que um modelo de nicho representa uma aproximação do nicho ecológico da espécie, nas dimensões das camadas ambientais utilizadas, ou seja, é utilizado um sub-espaço de condições do nicho ecológico na realização da modelagem. Neste tipo de modelagem não entram fatores históricos, barreiras geográficas, competição, predação etc. Ou seja, os resultados da modelagem correspondem a uma previsão, baseada em dados do nicho realizado, que se aproxima do nicho fundamental da espécie e a área projetada representa a distribuição potencial da espécie baseada nas camadas ambientais utilizadas na modelagem (Figura 1).

As aplicações de modelagem são inúmeras, essas técnicas têm aplicação direta em conservação e planejamento de reservas, manejo de espécies invasoras, epidemiologia, biogeografia, aspectos evolutivos das espécies, entre outras.

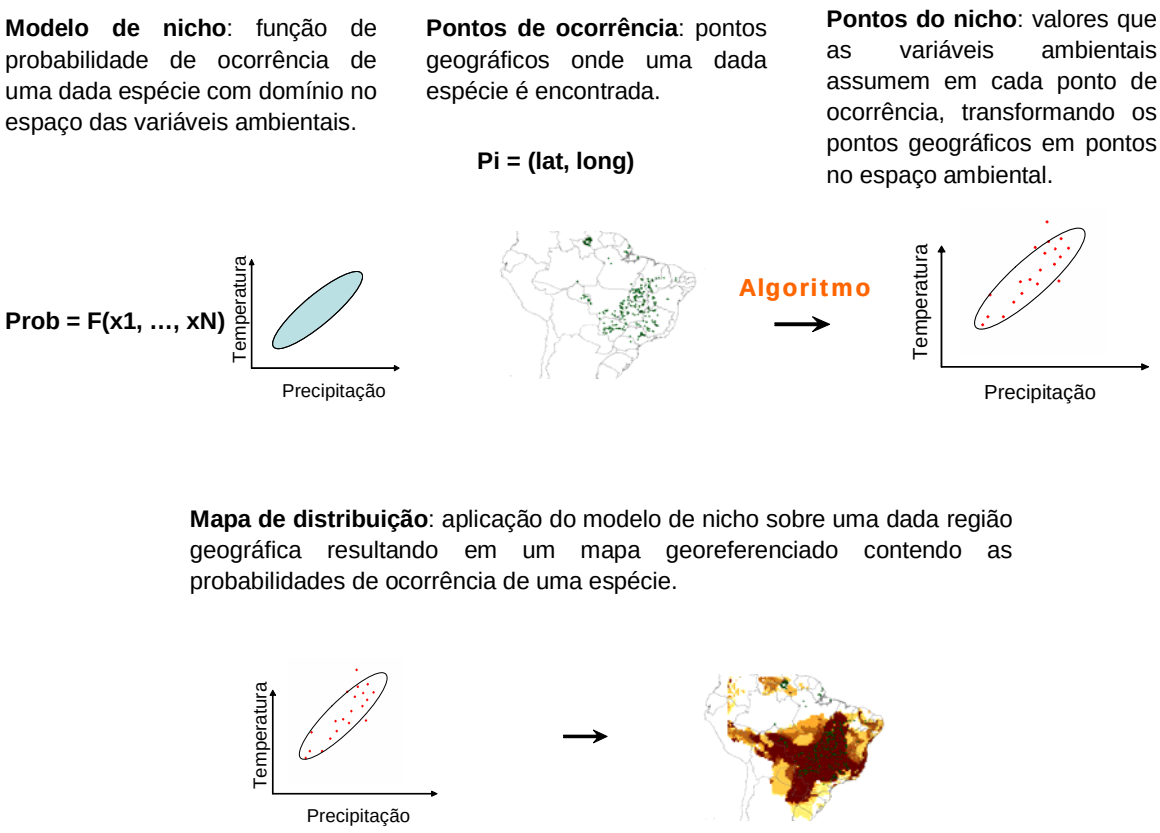

Figura 1: Modelagem de biodiversidade baseada em dados abióticos. Fonte: Mauro Muñoz – Inter-American Workshop on Environmental Data Access - http://www.cria.org.br/eventos/iaed/agenda

Software e ambientes para Modelagem

- Ambiente Openmodeller *(http://openmodeller.sourceforge.net/)* (Sutton et al. 2007). Este ambiente é independente de plataforma, gratuito, uso livre e de código aberto. Apresenta vários algoritmos disponíveis para modelagem de distribuição potencial de espécies. Fornece tratamento automatizado para dados em diferentes sistemas de coordenadas e de projeção, tanto para dados bióticos (registros de ocorrência de espécies) quanto abióticos (mapas temáticos) e também para dados abióticos em diferentes resoluções.
- Software Maxent (*http://www.cs.princeton.edu/~schapire/maxent/*). (Phillips et al. 2006). Plataforma windows, gratuito, uso livre, código fechado. Apenas 1 algoritmo de modelagem e não apresenta tratamento automatizado para dados em diferentes formatos.
- Software DesktopGarp *(http://www.lifemapper.org/desktopgarp/).* (Pereira 2002).

Plataforma windows, gratuito, uso livre, código fechado. Apenas 1 algoritmo de modelagem e não apresenta tratamento automatizado para dados em diferentes formatos. Não está mais sendo atualzido. Toda atualização do algoritmo GARP é feita na versão implementada no *openModeller*.

- Ambiente BIOMOD (Thuiller 2003). Plataforma windows, código fechado. Vários algoritmos de modelagem. Obs: não se encontra mais o site para download.
- Software FloraMap Plataforma windows, pago, código fechado. *(http://gisweb.ciat.cgiar.org/SIG/marksim_floramap.htm).* Apenas 1 algoritmo de modelagem e não apresenta tratamento automatizado para dados em diferentes formatos. Utiliza um conjunto próprio de dados climáticos.

Dados usados em modelagem

Para realizar modelagem de distribuição potencial de espécies são usadas as coordenadas dos registros de ocorrência das espécies e mapas temáticos que resumem a informação ambiental sobre a distribuição dessas espécies.

Dados bióticos: registros de ocorrência de espécies

Existem dois tipos de dados bióticos associados aos Modelos de Distribuição de Espécies (MDEs), registros de presença e os registros de ausência de uma determinada espécie. Esses dados são utilizados para alimentar, calibrar e avaliar os MDEs.

Os registros de presença são comumente gerados através de coletas de espécimes ou de observações de campo. Já os registros de ausência são extremamente raros, e foram um dos primeiros problemas enfrentados pela modelagem de distribuição de espécies (Rushton et al. 2004).

Quando existem dados de presença e dados de ausência, é possível utilizar métodos estatísticos para modelar a distribuição (Corsi et al. 2000, Guisan and Zimmermann 2000, Elith and Burgman 2002, Scott et al. 2002) e nesta categoria estão o GLM (Modelo Linear Generalizado) (Austin et al. 1994) e o GAM (Modelo Linear Aditivo) (Yee and Mitchell 1991).

Quando só existem dados de presença, outra categoria de modelos tem que ser utilizada. Nesta categoria está o DOMAIN (Carpenter et al. 1993), o BIOCLIM (Busby 1986, Nix 1986) e os algoritmos baseados em Distância Ambiental disponíveis no *openModeller*[1].

[1] http://openmodeller.sourceforge.net/

Ainda existe uma terceira categoria de algoritmos, que são os chamados híbridos ou intermediários. Esses algoritmos usam dados de presença e de pseudo-ausências para gerar os MDEs. Exemplo desse tipo de algoritmos são o GARP (Stockwell and Peters 1999) e o MAXENT (Phillips et al. 2004, Phillips et al. 2006). Esses dados de pseudo-ausências (também conhecidos por *background pixels*) são pontos escolhidos aleatoriamente na área de estudo (Ferrier and Watson 1996, Ferrier et al. 2002) e que são usados como ausências durante a modelagem.

Acesso aos dados bióticos (paginas acessadas em 08/03/2007)

Para dados de presença de espécies, existem grandes coleções científicas nos quais esses dados estão armazenados, e parte desses dados, atualmente é disponível via Internet.

- GBIF (Global Biodiversity Information Facility) Portal Home: http://www.gbif.org/. Já existe uma forma automatizada de busca neste portal na interface desktop do *openModeller*.
- MOBOT – Missouri Botanical Garden, dados de plantas disponíveis em (http://mobot.mobot.org/W3T/Search/vast.html).
- NYBG – New York Botânical Garden, dados de plantas disponíveis em (http://sciweb.nybg.org/science2/vii2.asp).
- SpeciesLink – dados de plantas, animais e microorganismos disponíveis em (http://splink.cria.org.br/tools?criaLANG=pt). Já existe uma forma automatizada de busca neste portal na interface desktop do *openModeller*.
- CONABIO – dados de plantas e de animais disponíveis em (http://www.conabio.gob.mx/).

Possíveis problemas associados aos dados bióticos

Existem vários problemas associados aos dados de presença que precisam ser levados em conta na hora de sua utilização em modelagem, são eles:

- Erros na identificação da espécie.
- Erros ou imprecisões associado ao georeferenciamento dos dados.
- Tendências na coleta de dados (por exemplo, por facilidade de acesso, dados coletados ao longo de rodovias ou de rios, etc).
- Falta de correspondência temporal entre as coletas, muitas vezes representando uma discrepância de várias décadas entre os registros de espécimes.

Esses problemas, em sua maioria, estão associados a falta uma estratégia de coleta pré-definida, o que leva muitas vezes a problemas de amostragem (Stockwell and Peterson 2002), com os dados apresentando tendências ou vícios de coleta (Austin 2002a, Reddy and Dávalos 2003) ou ainda imprecisão na localização do ponto (Chapman et al. 2005). No caso

de museus e herbários, a maior parte dos dados é referenciado pela localidade ou até mesmo pelo município em que ocorrem. Apenas recentemente o uso de GPS em coletas de campo passou a ser empregado. A discrepância temporal entre os dados de coleções é outro fator bastante comum. Esses problemas se potencializam principalmente quando se trabalha com dados oriundos de diferentes fontes. De qualquer maneira, todos esses problemas devem ser levados em conta na hora de realizar e/ou interpretar a modelagem.

Portanto, espécies fáceis de serem identificadas, sem problemas taxonômicos/nomenclaturais; dados georeferenciados com GPS, coletas espacialmente bem distribuídas e dados coletados em escala temporal próxima constituem-se no melhor conjunto de dados para modelagem.

É claro que isso nem sempre é possível, e em se tratando de regiões tropicais e megadiversas, esses problemas são situações comuns e correspondem, muitas vezes, a maior parte dos dados disponíveis para modelagem. Por outro lado, é justamente nessas regiões, que são, via de regra, mal coletadas, que as técnicas de modelagem tornam-se mais valiosas para a conservação (Graham et al. 2004).

O uso de gazeteers para georeferenciar dados de coleções científicas, principalmente através da localidade de ocorrência da espécie, aumentou muito a quantidade de dados utilizáveis em modelagem. Ferramentas de limpeza de dados também têm ajudado bastante na escolha do melhor conjunto de dados para modelagem. A rede specieslink[2] disponibiliza uma série de ferramentas nessa linha para usuários em geral. Para as coleções que compõem a da rede, são fornecidas ferramentas ainda mais completas para facilitar e agilizar o trabalho do curador na limpeza de dados, tornando os dados das coleções bem mais precisos.

Dados abióticos: mapas temáticos

Nos últimos anos houve um grande aumento na geração e disponibilização de dados via sensoriamento remoto. Paralelo a esse aumento de dados, o crescimento no uso de Sistemas de Informação Geográfica (SIGs) impulsionaram bastante o uso de MDEs. O uso de dados de sensoriamento remoto também tornou possível o estudo de áreas de difícil acesso (Rushton et al. 2004, Guisan and Thuiller 2005), sendo uma fonte alternativa, e bastante útil, de informação para países com problemas de coleta e de grandes dimensões como é o caso do Brasil.

---

[2] http://splink.cria.org.br/

Os dados abióticos tradicionalmente usados em modelagem são as camadas climáticas, topográficas e mais recentemente, uso e ocupação de solo (Guisan et al. 1999, Hirzel et al. 2002, Zaniewski et al. 2002).

Mais recentemente dados oriundos de satélites ambientais deram um novo impulso à geração de MDEs (Stoms and Hargrove 2000, Paruelo et al. 2001, Parra et al. 2004, Roura-Pascual et al. 2005, Stockwell 2006). São eles os dados oriundos dos satélites NOOA e Modis que geraram dados de índices de vegetação do tipo NDVI (*Normalized Difference Vegetation Index*)(Verhoef et al. 1996, Strahler et al. 1999, Egbert et al. 2000, UMD 2001, DEH 2004), EVI (vegetation index isolines), assim como dados de umidade de solo.

Acesso a dados abióticos (paginas acessadas em 08/03/2007)

- IPCC – Intergovernmental Panel on Climate Change, dados de clima (presente e projeção do passado e do futuro disponíveis em (http://www.ipcc-data.org/), resolução de 50Km.
- Hidro1K - Elevation Derivative Database na resolução de 1Km (para o mundo todo), fonte Earth Resources Observation and Science (EROS), disponível em (http://eros.usgs.gov/)
- Worldclim - dados climáticos mensais (para o mundo todo) de temperatura (máximas, médias e mínimas), precipitação e altitude em 4 diferentes resoluções (10', 5', 2.5' e 30") disponíveis em (http://www.worldclim.org/) (Hijmans et al. 2004, Hijmans 2005).
- USGS – US Geological Surveys, dados de imagens de satélite, Advanced Very High Resolution Radiometer (AVHRR) disponíveis em (http://edc.usgs.gov/products/satellite/avhrr.html).
- KSID – Kansas Satellite Image Database, dados disponíveis em (http://www.kars.ku.edu/products/ksid/modis-products.shtml).
- CPTEC/INPE – dados climáticos e dados do satélite ambiental modis disponíveis em (http://www.cptec.inpe.br/clima/) e em (http://satelite.cptec.inpe.br/htmldocs/modis/modis_dsa.htm).

Possíveis problemas associados aos dados abióticos

Vários erros, ou falta de informação suficiente, podem estar associados aos dados abióticos, alguns deles são passíveis de serem levantados, outros não, de qualquer forma é bom prestarmos atenção na fonte dos dados, se existe metadados associados aos dados e se a metodologia utilizada para gerar a informação está disponível. Os principais problemas associados às camadas temáticas são:

- Erros na manipulação dos dados. Esse tipo de erro não é muito visível para o usuário comum, mas é uma boa prática utilizar SIGs para visualizar os mapas e verificar as tabelas de conteúdo dos dados.

- Problemas com os modelos de clima, utilizados na geração das variáveis climáticas. Principalmente quando vamos trabalhar com projeções climáticas do futuro ou do passado. Aqui é fundamental que exista informação disponível sobre a metodologia utilizada, de preferência publicada, para gerar as projeções climáticas.
- Dados gerados através de interpolação de informação com dados de baixa resolução. Idem ao anterior. É importante que o usuário saiba qual a resolução original dos dados para saber qual a resolução dos mapas gerados pela modelagem.
- Problemas de escala (dados com baixa resolução) para a necessidade da questão a ser analisada pelos MDEs. Esse problema foi bastante sério no passado, hoje já existem dados em resolução de muito mais detalhe disponíveis para download.
- Falta de dados suficientes para descrever todos os parâmetros associados ao nicho fundamental da espécie. A recente geração e inclusão de dados ambientais oriundos de sensoriamento remoto em modelagem têm ajudado bastante nessa questão.
- Problemas associados ao acesso, para o usuário comum, a dados já tratados e em formato compatível para uso em MDE. Isso ainda é difícil, dados prontos para serem usados em modelagem não é muito fácil de achar, principalmente quando falamos de dados do meio físico gerados por sensoriamento remoto, por isso é importante a formação de parcerias com as instituições geradoras de dados.
- Tamanho dos arquivos: dados ambientais podem ser bastante grandes e pesados para serem armazenados pelo usuário. Cada vez mais os dados ambientais estão sendo gerados em resoluções de maior detalhe, o que os torna bastante pesados e, muitas vezes, impraticável para o usuário comum mantê-los em seu computador pessoal. O CRIA vem trabalhando no projeto openModeller para que a modelagem possa ser feita via Internet, assim os dados de grande porte ficariam armazenadas em servidores com maior capacidade de disco.

Algoritmos disponíveis para modelagem

Conforme descrito acima, existem pelo menos três categorias de algoritmos atualmente utilizados em modelagem, destes, apenas alguns serão abordados no curso, ou por serem os mais amplamente usados na literatura e/ou por estarem implementados no *openModeller*, são eles:

Bioclim: prevê as condições habitáveis de uma dada espécie baseado em um envelope bioclimático, que consiste em uma região (reto-linear) representando a amplitude (ou uma porcentagem dela) de cada dimensão ambiental, baseado no cálculo da média e do desvio padrão dos valores ambientais presentes nos registros de ocorrência da espécie. Para cada variável ambiental o algoritmo encontra a média e o desvio padrão (assumindo que os

valores tenham distribuição normal) associado aos pontos de ocorrência. Cada variável tem seu próprio envelope representado pelo intervalo [m - c*s, m + c*s], onde 'm' é a média; 'c' é o parâmetro de corte fornecido; e 's' é o desvio padrão da média. Além deste envelope, cada variável ambiental tem um valor adicional referente aos limites máximo e mínimo relacionado aos pontos de ocorrência. Sendo assim, cada pixel (célula) pode ser classificadao como:

- Habitável (suitable): se todos os valores ambientais estiverem dentro do envelope calculado;
- Tolerável (marginal): se um ou mais valores ambientais estiverem fora do envelope, mas dentro dos valores limites máximos e mínimos das variáveis ambientais.
- Inabitável (unsuitable): se um ou mais valores associados estiverem fora dos valores limites máximos e mínimos das variáveis ambientais.

Um modelo é então gerado com as probabilidades de 1.0, 0.5 e 0.0 (habitável, tolerável e inabitável, respectivamente).

Distância ambiental: normaliza os valores das variáveis ambientais e calcula a distância entre as condições ambientais para cada ponto de ocorrência e seleciona a menor distância (distância mínima). Ou calcula o ponto médio no espaço ambiental considerando todos os pontos de presença fornecidos e calcula a distância entre o ponto médio e cada ponto no espaço ambiental (média da distância). Em ambos os casos, se o valor de distância calculado estiver entre o valor 0 e o valor do parâmetro fornecido (valor máximo de distância que será utilizado), então a probabilidade de ocorrência estará entre [0,1], no caso da média da distância, com decaimento linear. Se o valor for superior ao valor do parâmetro, então a probabilidade será zero.

Environmental Distance: algoritmo genérico baseado em dissimilaridade ambiental com quatro diferentes métricas para o cálculo de distância: Euclidiana, Mahalanobis, Chebyshev e Manhattan.

Algoritmo genético (GARP): A idéia é que uma população de soluções candidatas para a resolução de um problema evolua e seus indivíduos sejam melhorados através da aplicação de operadores eurísticos inspirados na variação genética e na seleção natural. O GARP opera sobre um conjunto de regras, realizando uma “seleção natural”, excluindo regras menos eficientes e criando novos conjuntos de regras a partir das regras “sobreviventes” (Stockwell and Peters 1999). As regras são baseadas nos valores das camadas ambientais presentes nos registros de ocorrência das espécies. Durante a execução do algoritmo as regras são modificadas aleatoriamente através de operadores heurísticos de recombinação e mutação. As novas regras geradas a partir da recombinação e da mutação são diferentes das

regras originais e por isso apresentam diferentes valores de adaptação. Estes valores de adaptação resultantes podem ser tanto melhores quanto piores que os valores originais, baseado nos pontos de teste. Quando um número pré-determinado de iterações é atingido, o algoritmo é encerrado. É criado como resultado um conjunto de regras a partir dos indivíduos sobreviventes. Este conjunto de regras representa o modelo de nicho da espécie. Este modelo é aplicado de volta ao espaço geográfico, indicando as regiões onde a espécie está provavelmente presente ou ausente. Diferente da concepção original dos algoritmos genéticos em que a solução para o problema é representada apenas pelo indivíduo mais apto da população, o GARP considera toda a população (conjunto de regras sobreviventes) como solução para o problema de modelagem. Para garantir uma maior eficácia na geração dos modelos (principalmente para adequar o uso do algoritmo à questão a ser modelada) alguns parâmetros devem ser utilizados/modificados pelo usuário antes de rodar o modelo. Alguns exemplos disso serão mostrados na parte prática do uso dos algoritmos.

Máxima Entropia (MAXENT): Baseia-se no princípio da máxima entropia, que diz que a melhor aproximação para uma distribuição de probabilidades desconhecida é aquela que satisfaça qualquer restrição à distribuição. Entropia baseia-se na quantidade de escolhas envolvendo a seleção de um evento. Trata-se de um método pra realizar previsões ou inferências a partir de informações incompletas. É aplicado em áreas como astronomia, reconstrução de imagens e processamento de sinais. A aplicação de máxima entropia na geração de MDE é estimar a probabilidade de ocorrência da espécie encontrando a distribuição de probabilidade da máxima entropia (que é a distribuição mais próxima da distribuição uniforme), submetidas a um conjunto de restrições que representam a informação incompleta sobre a distribuição alvo. A informação disponível sobre a distribuição da espécie constitui um conjunto de valores tomados como **verdades** (oriundos dos dados de presença) e suas restrições são os valores esperados de cada valor que devem corresponder às médias para o conjunto de dados tomados da distribuição alvo. Os valores reais correspondem aos valores dos pixels da área de estudo na qual a espécie está presente, ou seja, aos valores das camadas ambientais utilizadas nesses pixels. É importante notar que para este algoritmo, os registros de ocorrência das espécies fornecem os dados que serão tratados sempre como verdades pelo algoritmo. No caso de MDE, isso nem sempre é verdade, dada a quantidade de ruidos que temos nos bancos de dados de distribuição de espécies. Portanto, cuidado extra deve ser tomado ao se utilizar este algoritmo para modelagem de distribuição de espécies baseados em dados não muito confiáveis ou que não representem bem a distribuição da espécie em questão.

Suport Vector Machine (SVM): este algoritmo (Máquina de Vetores de Suporte)

caracteriza-se por ser um conjunto de métodos de aprendizagem supervisionado relacionados que pertecem a família dos classificadores lineares generalizados. As SVMs foram introduzidas recentemente como uma técnica para resolver problemas de reconhecimento de padrões. Esta estratégia de aprendizagem, introduzida por (Vapnik 1995) é um método muito poderoso que em poucos anos desde sua introdução tem superado a maioria dos sistemas em uma ampla variedade de aplicações (Cristianini and Shawe-Taylor 2000). De acordo com a teoria de SVMs, enquanto técnicas tradicionais para reconhecimento de padrões são baseadas na minimização do *risco empírico,* isto é, tenta otimizar o desempenho sobre o conjunto de treinamento, as SVMs minimizam o *risco estrutural,* isto é, a probabilidade de classicar de forma errada padrões ainda não vistos pela distribuição de probabilidade dos dados. O objetivo de classição Vetor Suporte é elaborar uma forma computacionalmente eficiente de aprender `bons' hiperplanos de separação em um espaço de características de alta dimensão, onde `bons' hiperplanos entendemos ser aqueles que otimizam os limites de generalização e por `computacionalmente eficiente' significa algoritmos capazes de tratar amostras de tamanho da ordem de 100.000 instâncias. A teoria da generalização dá uma orientação clara sobre como controlar a capacidade, e logo como prevenir modelos ruins, controlando as medidas das margens dos hiperplanos, enquanto a teoria da otimização fornece as técnicas matemáticas necessárias para encontrar hiperplanos otimizando essas medidas. Existem diferentes limites de generalização, o que motiva o desenvolvimento de algoritmos diferentes: alguém pode por exemplo otimizar a margem máxima, ou a distribuição da margem ou ainda o número de vetores suporte etc (Cristianini and Shawe-Taylor 2000). Uma propriedade especial das SVMs é que eles simultaneamente minimizam erros de classificação empírica e maximizam a margem geométrica. Os mapas gerados pelas SVMs geram vetores com alta dimensão espacial onde é construído um hiperplano de separação máxima. Dois hiperplanos paralelos são construídos em cada lado do hiperplano que separa os dados. Os hiperplanos de separação maximizam a distância que separa os dados do hiperplano central. Os modelos gerados pela SVM só dependem de um subconjunto de dados de treino, que são aqueles que estão dentro da margem formada pelos dois hiperplanos paralelos, os demais dados serão descartados da solução do problema. Sendo assim, o algoritmo utilizará apenas os dados mais informativos (entre todos os dados fornecidos no treinamento) para as SVMs, para gerar o MDE. Esta característica torna esta técnica especialmente interessante para utilização em situações onde a confiabilidade dos dados de entrada (registros de ocorrência da espécie e/ou variáveis ambientais) é duvidosa ou incompleta, o que é especialmente comum em se tratando de levantamento de biodiversidade em regiões tropicais. É claro que, para qualquer técnica de modelagem, quanto menos ruído nos dados, melhor será o resultado. Mas é

importante sabermos que ruído é sempre uma constante nesse tipo de dado, então, é importante escolher a técnica que melhor se adeque ao conjunto de dados disponível.

Uso dos algoritmos GARP, Maxent e SVM

A parte prática deste curso será focada no uso dos algoritmos mais usados na literatura recente sobre modelagem, são eles o GARP e o MAXENT e mais recentemente, o SVM. O algoritmo genético GARP apresenta duas versões disponíveis para download, uma presente no ambiente de modelagem *openModeller* e outra disponível em versão *desktop.* O algoritmo para modelagem MAXENT, por enquanto só apresenta uma versão *desktop,* está prevista a implementação deste algoritmo no ambiente *openModeller* O SVM apresenta uma versão implementada no *openmodeller*.

O uso do GARP no ambiente de modelagem *openModeller* traz uma série de vantagens para o usuário. Este ambiente de modelagem trata uma série de problemas envolvendo a compatibilidade de dados ambientais de diferentes fontes e de diferentes resoluções, que, de outra forma, teria que ser previamente resolvida pelo usuário. Para uso tanto da versão *desktop* do GARP quanto a versão do MAXENT é necessário um grande trabalho prévio com os dados ambientais para que o modelo possa ser rodado. Ambos precisam dos dados ambientais em perfeita sintonia quanto ao tamanho, formato dos dados, sistema de coordenadas e de projeção, datum, tamanho de célula e coordenada de origem. Para compatibilizar tudo isso, é necessário que o usuário tenha um SIG e saiba como usá-lo.

Por outro lado, a versão do GARP no openModeller ainda apresenta problemas com o tempo de processamento, sendo bem mais lento que a versão do desktop GARP e também bem mais lento que o software MAXENT. Já a SVM implementada no *openModeller* apresenta um tempo de processamento bastante rápido.

Validação dos modelos gerados

Para avaliar a qualidade do modelo gerado é preciso gerar um conjunto independente de dados. Existem pelo menos duas formas de se fazer isso:

- Coletar novos dados (trabalho de campo ou levantamento de literatura)
- Dividir seu conjunto de dados em duas partes antes de modelar.

Independente da estratégia escolhida, você terá dois conjuntos de dados independentes, um que irá gerar o modelo (conjunto de treino) e outro que irá testá-lo (conjunto de teste).

A forma mais comum de se dividir o conjunto de dados é através da aplicação de números aleatórios à tabela de registros de ocorrências, 50% dos dados para gerar o modelo e 50% para testar (Fielding and Bell 1997, Hirzel and Guisan 2002).

Atualmente tem-se usado técnicas de bootstrapping (Efron 1979) para avaliar a precisão dos modelos de distribuição usando curvas ROC. Esta técnica envolve a partição dos dados aleatoriamente em vários conjuntos de treino e teste. Neste curso, iremos utilizar a técnica de bootstrapping implementada no DIVA-GIS, que consiste em 5 divisões do conjunto de dados de entrada, sendo 75% dos dados usados para treinar o modelo e 25% para testar. Como não se tem dados de ausência, a técnica utilizada é gerar aleatoriamente, na área de estudo, pontos de pseudo-ausência. No caso da implementação do DIVA-GIS, são gerados 5 pontos de ausência, para cada ponto de presença utilizado.

Procedimento para quantificar o erro gerado

- Gerar um modelo com o conjunto de treino
- Quantificar os componentes de erro através de uma matriz de confusão (Figura 2)

Matriz de Confusão

	Presença real	Ausência real
Presença prevista	*a*	*b*
Ausência prevista	*c*	*d*

a = verdadeiro positivo
b = falso positivo
c = falso negativo
d = verdadeiro negativo

a e d são acertos
b e c são erros

taxa de falso positivo = erro de sobreprevisão = b/(b+d)
taxa de falso negativo = erro de omissão = c/(a+c)
sensitivity = a/(a+c)
specificity = d/(b+d)

Figura 2: Matriz de confusão, onde a e d são previsões corretas; b é erro de sobreprevisão (falsos positivos) e c é erro de omissão (falsos negativos).

Erro de omissão:

No geral, erros de omissão podem ser considerados erros verdadeiros (erro grave) (Figura3). Contudo, sob algumas circunstâncias, um registro de presença pode não ser muito confiável, devido à pelo menos três situações:

- A identificação da espécie (taxa), para alguns pontos, pode estar errada.
- O georeferenciamento de alguns pontos pode estar errado.
- A localização de um indivíduo pode estar fora do seu habitat usual (indivíduos em trânsito ou

introduzidos).

Obs: em todas as três situações, estes pontos podem representar um "outlyer" para o algoritmo. Ou seja, pontos com informação ambiental muito fora do padrão gerado pelos demais pontos do conjunto de treino. Nessas circunstâncias, um erro de omissão não seria um erro, e sim uma forma do algoritmo dar menos "importância" para pontos "ruins". Uma conseqüência direta disso é, caso se tenha dúvida sobre a qualidade dos registros de ocorrência da espécie, é que não é recomendado usar taxa de omissão de 0% ao se rodar um modelo. É preferível deixar uma margem de segurança (ex: até 10%) para que o algoritmo possa trabalhar melhor essa questão. Porém, na literatura mais atual há uma predominância de trabalhos utilizando 5% de taxa de omissão (Elith et al. 2006, Phillips et al. 2006)

Erro de sobreprevisão:

A sobreprevisão pode ou não ser um erro (Figura 3). A previsão de um modelo em uma área na qual não se tem registro de ocorrência da espécie pode ser causada por diferentes situações:

- A área é habitável para a espécie, mas não se tem um esforço amostral suficiente na região para afirmar se a espécie ocorre ou não, são as lacunas de conhecimento.
- A área é habitável para a espécie, mas fatores históricos ou ecológicos (barreiras geográficas, capacidade de dispersão) ou bióticos (competição, predação) impediram a espécie de chegar ou de se estabelecer na região.
- A área é inabitável mesmo, o que seria o verdadeiro erro de sobreprevisão.

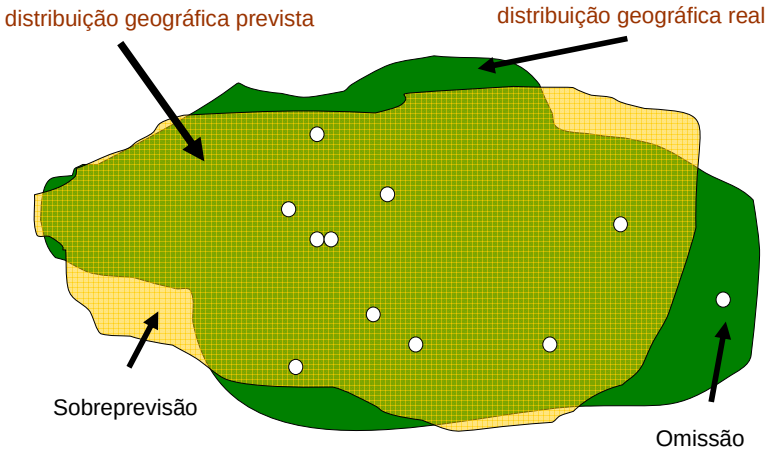

Figura 3: representação dos erros de omissão e sobreprevisão (comissão) do modelo.

Seleção dos melhores modelos

Algoritmos estocásticos, sujeitos a variações internas ao acaso na geração dos modelos, como o Garp, produzem resultados diferentes para o mesmo conjunto de dados de entrada. Se produzirmos vários modelos, podemos calcular o erro (omissão e sobreprevisão) e visualizar os valores gerados em um gráfico (Figura 4).

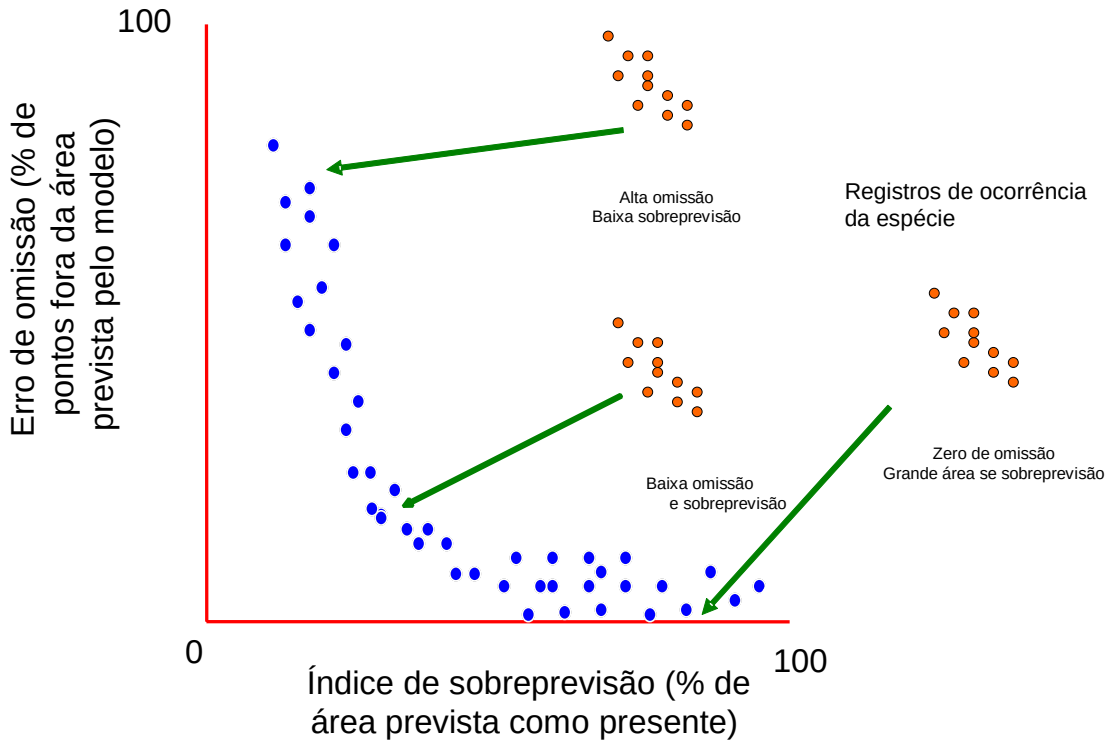

Figura 4: Curva típica para espécies com um número de ocorrência padrão. Fonte: Peterson and Martinez – Global Biodiversity Information Facility (GBIF) - http://www.gbif.org/prog/ocb/modeling_workshop/mexico_city/presentations

Os modelos com altas taxas de omissão são definitivamente ruins e devem ser desprezados a partir de um determinado nível de corte (ex: 10%). Os demais modelos, os que sobram, apresentam taxas altas e baixas de sobreprevisão. Os modelos com valores altos de sobreprevisão são os muito amplos (modelos muito inclusívos), os de valores baixos são muito restritos. Então, neste caso é recomendado tomar o valor da mediana dos valores de

sobreprevisão e estabelecer um determinado limite de corte para selecionar os melhores modelos (ex: 5 de cada lado da mediana, ou 50% de cada lado da mediana) (Figura 5). Estes valores podem ser aumentados ou diminuídos dependendo da pergunta inicial, se o interesse é por modelos amplos ou restritos.

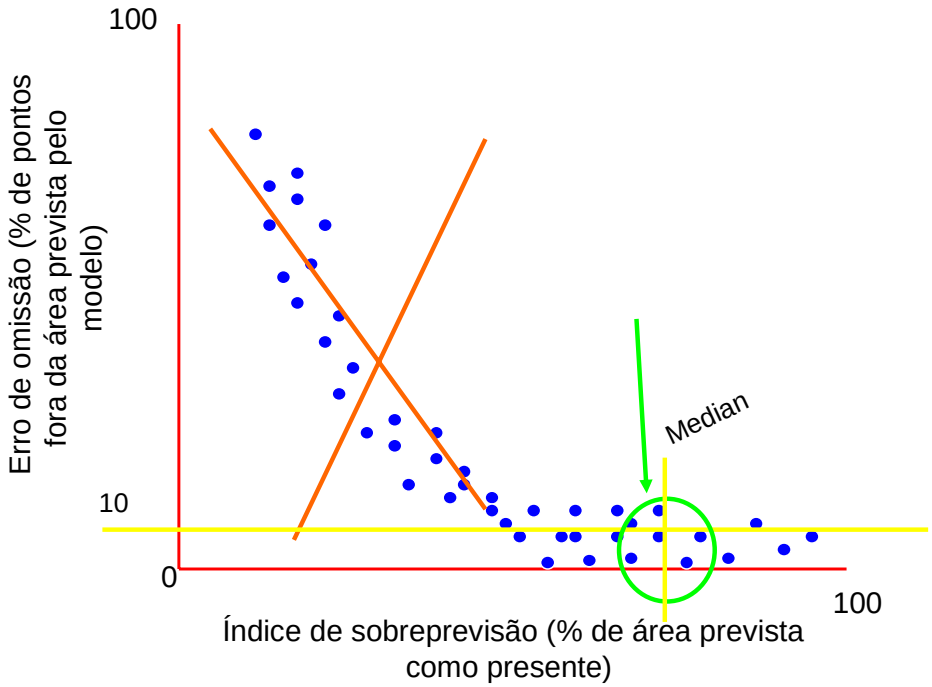

Figura 5: Escolha dos melhores modelos a partir de níveis de corte de omissão e sobreprevisão. Fonte: Peterson and Martinez – Global Biodiversity Information Facility (GBIF) - http://www.gbif.org/prog/ocb/modeling_workshop/mexico_city/presentations

Para avaliar a qualidade do modelo gerado

A validação de um modelo é também conhecida como teste externo e consiste da confrontação de seus resultados contra dados reais (ou semi-reais como veremos adiante) da distribuição da espécie. Isso distingue esse processo do chamado teste interno, também conhecido como verificação, no jargão da modelagem matemática. Discussões sobre a teoria para validação de modelos pode ser encontrada em (Pearce and Ferrier 2000) e (Pearson et al. 2006a).

Essa é uma das etapas mais importantes do processo. Sem a validação, a interpretação de um modelo perde seu sentido visto que tudo o que está representado pode estar incorreto ou com graus inaceitáveis de imprecisão.

O método mais comum para avaliar a qualidade do modelo é a matriz de confusão, descrita anteriormente, mas na hora de avaliar (ou comparar) o resultado de uma modelagem, é necessário observar a necessidade de se aplicar limites de corte no resultado da modelagem, para então decidir fazer uma avaliação dependente ou independente do limite de corte.

Avaliação dependente do limite de corte

As medidas tradicionais de capacidade discriminatória de um modelo dependem da adoção de um limite de decisão arbitrário (threshold) o que acaba inserindo mais um fator de subjetividade na interpretação do modelo. Uma boa revisão sobre limites de decisão pode ser encontrada em (Liu et al. 2005).

A escolha desse limite de decisão é usualmente baseada no conhecimento da probabilidade – subjetiva – da ocorrência da espécie de interesse (por exemplo, dados sobre ela ser uma espécie rara) assim como do julgamento das conseqüências em tomar decisões erradas. Por exemplo, se uma espécie é ameaçada e o objetivo do modelo é identificar áreas potenciais para maximizar o sucesso de sua reintrodução na natureza, precisamos identificar áreas adequadas que minimizem a chance de erro. Isso requer um limite de decisão alto, que selecione apenas as áreas com máxima probabilidade de ocorrência.

Outro objetivo interessante seria a necessidade de identificar locais onde há o planejamento de intervenção antrópica (ex. instalação de uma indústria, dentro de um programa de análise de impactos ambientais) que possa afetar negativamente a espécie de interesse. Nesse caso, faz-se necessário identificar o máximo de áreas adequadas à permanência da espécie quanto for possível. Logo, o limite de decisão será bem menos restrito e a taxa de sobreprevisão deve ser mais alta.

Todo modelo apresenta seus erros e acertos, que são avaliados em conjunto para determinarmos a qualidade do mesmo. A figura 6 apresenta novamente a matriz de confusão, ou seja, o esquema que reúne as possíveis formas de acerto e erro em relação a o que o modelo previu e a distribuição "real" da espécie na natureza.

	REGISTRO DE PRESENÇA	REGISTRO DE AUSÊNCIA	
PRESENÇA PREDITA	A	B	A + B
AUSÊNCIA PREDITA	C	D	C + D
	A + C	B + D	A + B + C + D

Figura 6. Representação esquemática dos erros e acertos de um modelo, também chamada de matriz de confusão.

Com isso podemos definir os seguintes componentes:

$Sensibilidade = \frac{A}{(A+C)}$
$Especificidade = \frac{D}{(B+D)}$
$Sobreprevisão = \frac{B}{(B+D)}$
$omissão = \frac{C}{(A+C)}$
$Acurácia* = \frac{A+D}{A+B+C+D}$

* Relembrando: Acurácia é o quão próximo do valor real o modelo se aproxima. Essa medida é sempre balanceada contra a Precisão, que é quanto que os valores preditos variam entre si (vide figura 7 abaixo).

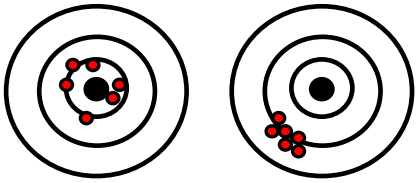

Figura 7. Exemplo esquemático da acurácia e da precisão de um modelo. O ponto central dos alvos é o valor real que desejamos atingir. Percebe-se que a precisão não possui relação com o valor real e que a acurácia de um modelo pode variar bastante. O modelo ideal seria aquele que é ao mesmo tempo acurado e preciso, esse é um campo ativo de pesquisas.

Esses componentes de erro e acerto do modelo relacionam-se com a definição de um limite de decisão para a interpretação do modelo da forma representada na figura 8.

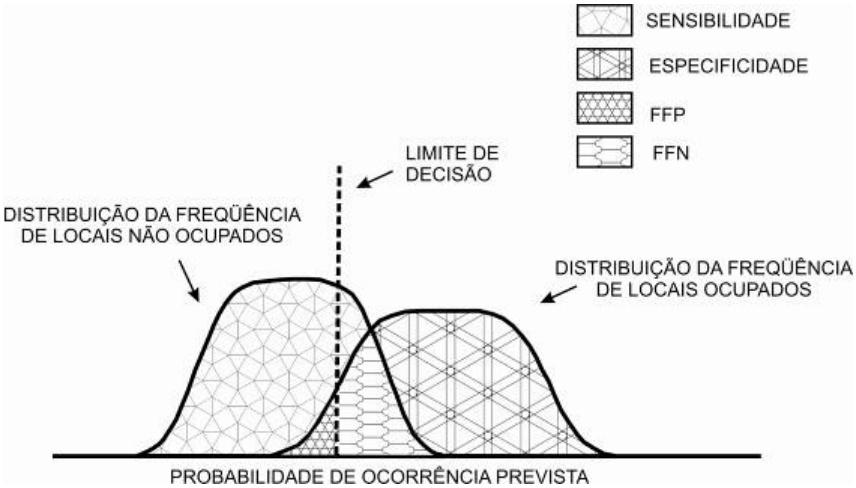

Figura 8. Relação entre sensibilidade, especificidade, taxa (ou fração) de falsos positivos (FFP), taxa de falsos negativos (FFN) e o limite de decisão (threshold) adotado.

No caso do GARP é necessário aplicar um limite de corte para avaliar o modelo gerado (geralmente é usado como limite de corte a área de 1 ou de 10 dos melhores modelos somados) (Anderson et al. 2003). Depois de estabelecido o limite de corte, testes estatísticos (X^2, proporção binomial entre outros) podem ser aplicados para avaliar a qualidade do modelo gerado. Estes testes avaliam a probabilidade do acerto dos pontos de teste (taxa de omissão) ser diferente do acaso em relação à área prevista do modelo gerado. Um modelo pode "acertar" todos os pontos de teste, mas em uma área tão grande que o resultado do teste pode não ser significativo. A recíproca é verdadeira, ou seja, o modelo pode não acertar todos os pontos, mas a área do modelo é tão pequena (em relação à área de estudo) que a chance de acertar aqueles pontos é significativamente diferente do acaso. Neste caso, um teste binomial uni-caudal pode ser usado para se determinar se um modelo prediz os pontos de teste significativamente melhor do que o acaso (Anderson et al. 2002a). Esta medida não pode ser usada diretamente para comparar modelos gerados por diferentes algoritmos porque diferentes algoritmos usam diferentes limites de corte que por sua vez geram diferentes áreas previstas como presença da espécie.

Avaliação independente de limite de corte

As curvas características de operação (ROC) são amplamente utilizadas na área de controle de qualidade, em processos industriais e na área de saúde, na padronização de doses-resposta, etc. O cálculo da área sob a curva (AUC) fornece uma medida única do desempenho do modelo, independente da escolha prévia de qualquer limite de decisão. Ela mede a capacidade discriminatória do modelo, nos permitindo interpretar seu resultado como a probabilidade de que ao sortearmos dois pontos, um sendo uma ocorrência verdadeira e o outro uma ausência verdadeira, o modelo consiga prever os dois

corretamente.

Uma das melhores fontes de explicação sobre as curvas características de operação pode ser encontrada no relatório de (Fawcett 2003).

Considere um problema de classificação (no jargão da modelagem, o modelo atua como um "classificador") onde cada exemplo ou é positivo ou negativo. Um classificador determina um valor real a cada exemplo para o qual um limite de decisão pode ser aplicado para predizer a que classe ele pertence; para facilitar nós adotamos os rótulos {Y, N} para as predições de classe. A sensibilidade de um classificador para um limite de decisão particular é a taxa de todos os exemplos positivos que são classificados Y, enquanto especificidade é a taxa de todos os exemplos de negativos que são classificados como N.

A Sensibilidade também é conhecida como a taxa de verdadeiros positivos, e representa ausência de erro de omissão. A quantidade 1-especificidade também é conhecida como a taxa de falso positivo, e representa o erro de sobreprevisão.

Esta análise caracteriza-se por avaliar a performance do modelo através de todos os possíveis limites de corte, gerando um único valor, que representa a área sob a curva (AUC), que pode então ser usado para comparações entre diferentes algoritmos.

A curva ROC é obtida plotando-se a sensitivity no eixo y e o valor 1-specificity no eixo x para todos os possíveis limites de corte. A área debaixo da curva (AUC) é normalmente determinada conectando os pontos com linhas diretas; isto é chamado o método de trapezóide (ao invés de métodos paramétricos que ajustam uma curva aos pontos). Esta área sob a curva (AUC) tem interpretação intuitiva, quanto mais próximo do valor um for a área, ou seja, quanto mais distante o resultado do modelo for da previsão aleatório, melhor o desempenho do modelo (Rushton et al. 2004, Phillips et al. 2006) (Figura 9).

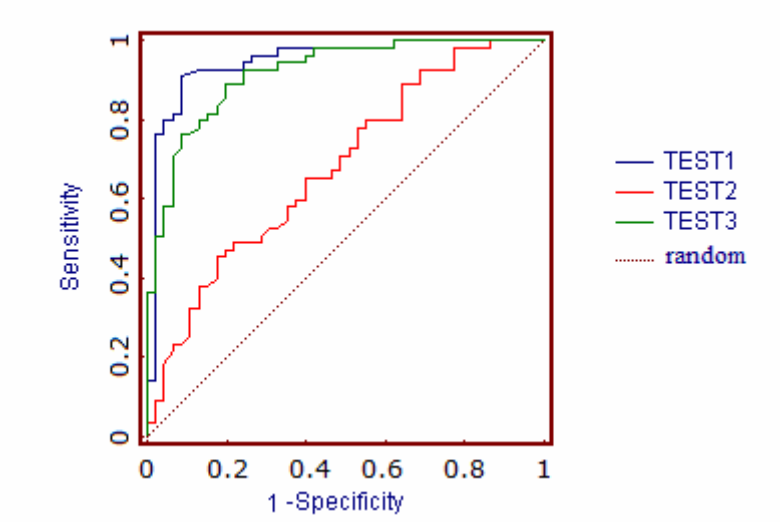

Figura 9: Exemplo de curva ROC gerada por dados de presença para três espécies. A interpretação é intuitiva, o teste um foi melhor que o teste três e que o teste dois, e o teste três foi melhor que o teste dois.

Quando possuímos apenas dados de presença, o uso do ROC parece inaplicável, pois sem dados de ausência, não há como calcular a especificidade. Podemos contornar esse problema considerando um problema distinto de classificação, ao invés de tentar avaliar o modelo sob sua capacidade de discernir presenças e ausências, medimos sua capacidade de distinguir presença contra o acaso.

Formalmente, para cada pixel x na área de estudo, é sorteado um ponto x-aleatório e para cada pixel x dentro da área de distribuição geográfica da espécie é definido um ponto x-presente.

O modelo de distribuição potencial de espécies fará suas predições para cada pixel correspondente a esses exemplos, mas sem ter informação sobre se este é uma presença ou um ponto escolhido ao acaso (aqui entra o termo pseudo-ausência).

Logo, obteremos previsões tanto para amostras positivas (x-presentes) quanto negativas (x-aleatório, pixels selecionados uniformemente ao acaso do "background" da área de modelagem). Juntas essas informações são suficientes para definir uma curva ROC.

O uso do ROC com dados de presença e ausência difere do cálculo feito apenas com dados de presença em relação ao valor máximo do AUC, que é nesse caso, menor do que 1 (Wiley et al., 2003). Se a distribuição da espécie cobre uma taxa *a* de pixels, então o máximo que a AUC pode ter é exatamente 1- *a*/2. Infelizmente, nós não sabemos o valor de *a*, logo não podemos dizer qual o valor ótimo que a AUC pode chegar sob essas circunstâncias (Wiley

et al. 2003).

Uma área igual a 1 representaria o "modelo perfeito"; uma área de 0,5 indica que o modelo seleciona ao acaso. De um ponto de vista prático, um teste de validação pode adotar os valores de AUC a seguir como indicadores da qualidade do modelo (Metz 1986):

0,90 - 1,0 = Excelente

0,80 - 0,90 = Bom

0,70 - 0,80 = Médio

0,60 - 0,70 = Ruim

0,50 - 0,60 = Muito ruim

Lembrando sempre que a validação de modelos ainda é uma área de pesquisa ativa e que novas técnicas podem surgir tão rápido quanto outras podem se tornar obsoletas. É necessário se manter continuamente informado sobre o assunto, acompanhando o ritmo das publicações mais recentes.

Existe uma grande variedade de programas gratuitos (ou não) disponíveis na internet que calculam as curvas ROC, a área sob a curva e testam se a previsão do modelo é ou não melhor do que a escolha de pontos ao acaso (utilizando principalmente o teste U de Mann-Whitney) e ajudam a selecionar qual modelo é melhor comparado a outro qualquer. Como exemplo temos o software AccuROC (disponível em http://www.accumetric.com/accurocw.htm") e o pacote ROCR, discutido abaixo.

Exemplo da elaboração da curca característica de operação e do cálculo da AUC

O formato da tabela de dados para a construção da curva característica de operação e calculo da AUC variam de acordo com cada um dos programas disponíveis para isso e também de acordo com o formato do arquivo de saída produzido por cada modelo.

Esse exemplo foi construído com os valores preditos para cada pixel na área de estudo e um indicador de presença ou ausência real de cada um desses pixels. De posse dessas informações utilizamos um pacote gratuito chamado ROCR (Sing et al. 2005) (disponível em http://rocr.bioinf.mpi-sb.mpg.de/ ou em um espelho CRAN, por exemplo, www.insecta.ufv.br/CRAN) que deve ser utilizado junto à plataforma R. Os dados estão resumidos na tabela 1 e a ROC na figura 10.

Tabela 1. Exemplo de dados para construção de uma curva ROC utilizando o pacote ROCR desenvolvido por Sing et al (2005) para uso na plataforma R. A primeira coluna contém os

valores preditos pra cada pixel e a segunda coluna contém um indicador 1 ou 0 para presença e ausência verdadeiras, respectivamente.

Predição	Estado
0.612548	1
0.364271	1
0.432136	0
0.140291	0
0.384896	0
0.244415	1
0.970641	1
0.890173	1
0.781781	1
0.868752	0
0.716681	1
0.360169	0
0.547983	1
0.38524	0
0.423739	0
0.1017	0
0.628096	1
0.74477	1
0.657733	1
0.49012	0
0.07237	0
0.172742	0

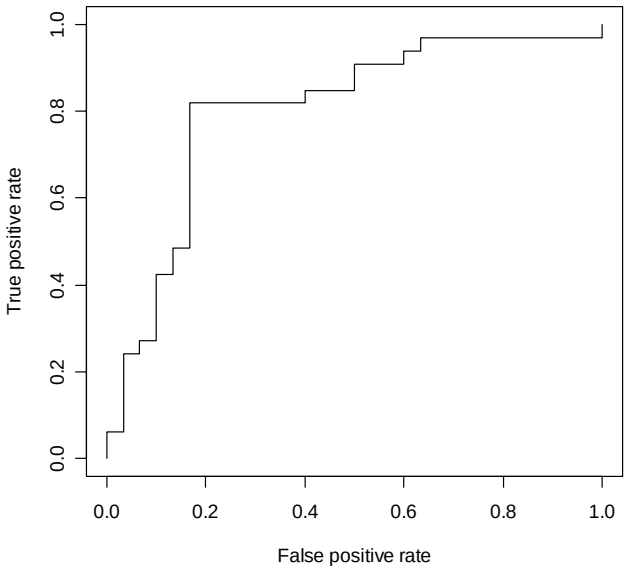

Figura 10. Gráfico ROC construído com os dados da tabela 1 utilizando o pacote ROCR na plataforma R.

A área sob o gráfico foi calculada como 0,803 - ou seja, esse foi um bom modelo.

Outros tipos de avaliação

- Avaliação do especialista na espécie modelada: este tipo de avaliação, que podemos chamar de avaliação biológica, é feita através de interpretação visual (Phillips et al. 2006) pelo especialista na espécie modelada. Esta avaliação é muito importante pois é bastante comum termos um resultado de modelo que estatisticamente é muito bom, mas que biologicamente não é. Existe na literatura algumas sugestões de como melhorar o resultado da modelagem retirando-se do modelo as áreas na qual a espécie não ocorre por fatores históricos (barreiras geográficas) (Peterson et al. 1999, Anderson et al. 2003). Pode-se ainda retirar áreas não habitadas por interações bióticas (competição com espécies similares) (Anderson et al. 2002b). O modelo ainda pode ser ajustado utilizando-se dados de uso da terra para excluir áreas deflorestadas ou alteradas, áreas urbanas, ou ainda áreas de cultivo (Anderson and Martinez-Meyer 2004).
- Aplicar o modelo de volta no campo: este é o teste mais robusto que se pode fazer com os resultados de modelagem. Avaliar a capacidade do modelo acertar a distribuição da espécie em áreas de alta probabilidade de ocorrência (Feria and Peterson 2002). A recíproca é verdadeira, ou seja, avaliar a capacidade do modelo em acertar a ausência da espécie em áreas não previstas pelo modelo.

Aplicações de modelagem em biodiversidade

- Utilização de modelos de distribuição potencial em análises biogeográficas (Siqueira and Durigan 2007).
- Conservação de espécies raras ou ameaçadas (Araújo and Williams 2000, Engler et al. 2004).
- Re-introdução de espécies (Hirzel et al. 2002)
- Perda de biodiversidade (Polasky and Solow 2001)
- Risco ambiental e/ou impactos de mudanças climáticas (Huntley et al. 1995, Magana et al. 1997, Sala et al. 2000, Peterson et al. 2002a, Oberhauser and Peterson 2003, Siqueira and Peterson 2003, Thomas et al. 2004, Pearson et al. 2006b).
- Avaliar o potencial invasivo de espécie exóticas (Peterson et al. 2003a, Peterson et al. 2003b).
- Estudar possíveis rotas de disseminação de doenças infecciosas (Petersen and Roehrig 2001, Peterson et al. 2002b)
- Auxiliar na determinação de áreas prioritárias para conservação (Bojorquez-Tapia et al. 1995, Egbert et al. 1998, Chen and Peterson 2002, Ortega-Huerta and Peterson 2004).

Leitura adicional recomendada

<u>Problemas metodológicos em modelagem:</u>

- Dificuldade de se obter dados de ausência de espécies para calibrar e validar modelos (Anderson 2003, Engler et al. 2004, Rushton et al. 2004, Phillips et al. 2006).
- Problemas na coleta de dados de ocorrência de espécies (Austin 2002b, Stockwell and Peterson 2002, Reddy and Dávalos 2003, Engler et al. 2004, Chapman et al. 2005, Barry and Elith 2006).
- Conseqüências de erros no posicionamento (georeferenciamento) dos pontos de coleta utilizados em modelagem (Iwashita 2007).
- Baixo número de amostragem (Hirzel et al. 2001, Hirzel and Guisan 2002, Stockwell and Peterson 2002, Austin et al. 2006).
- Problemas de escala (Collingham 2000, Engler et al. 2004, Chapman et al. 2005, Guisan and Thuiller 2005, Guisan et al. 2007).
- Problemas de auto-correlação dos dados (Isaaks and Srivastava 1989, Austin 2002a, Segurado et al. 2006)
- Quantidade e qualidade dos dados ambientais (Guisan et al. 2002, Chapman et al. 2005). Métodos para eliminar variáveis (Netter et al. 1996, Guisan and Thuiller 2005).

<u>Dados ambientais (mapas temáticos) utilizados em modelagem:</u>

- Influência dos fatores ambientais na distribuição de espécies (Bazzaz 1998).
- Uso de dados ambientais em modelagem (Bazzaz 1998). Uso de dados de sensoriamento remoto (Rushton et al. 2004, Guisan and Thuiller 2005). Uso de dados ambientais tradicionais

(Guisan et al. 1999, Hirzel et al. 2002, Zaniewski et al. 2002).

Análise dos resultados de modelagem

- Comparação entre resultados de modelagem de diferentes algoritmos (Manel et al. 2001, Hirzel et al. 2002, Anderson et al. 2003, Segurado and Araújo 2004, Guisan and Thuiller 2005, Austin et al. 2006, Elith et al. 2006, Phillips et al. 2006).
- Métodos para comparar modelos (Guisan et al. 2002, Rushton et al. 2004). Regressão linear múltipla (Netter et al. 1996), Matriz de Confusão, Análises ROC e índice Kappa (Fielding and Bell 1997).

Referências citadas no texto

Anderson, R. P. 2003. Real vs artefactual absences in species distributions: tests for Oryzomys albigularis (Rodentia: Muridae) in Venezuala. **30**:591-605.

Anderson, R. P., M. Laverde, and A. T. Peterson. 2002a. Geographical distributions of spiny pocket mice in South America: Insights from predictive models. Global Ecology and Biogeography **11**:131-141.

Anderson, R. P., M. Laverde, and A. T. Peterson. 2002b. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos **93**:3-16.

Anderson, R. P., D. Lew, and A. T. Peterson. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling **162**:211-232.

Anderson, R. P., and E. Martinez-Meyer. 2004. Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (Heteromys) of Ecuador. Biological Conservation **116**:167-179.

Araújo, M. B., and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogegraphy **33**:1677-1688.

Araújo, M. B., and P. H. Williams. 2000. Selecting areas for species persistence using occurrence data. Biological Conservation **96**:331-345.

Austin, M. 2002a. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling **157**:101-118.

Austin, M. P. 2002b. Case studies of the use of environmental gradients in vegetation and fauna modelling: theory and practice in Australia and New Zealand. Pages 73-82 *in* J. M. Scott, P. J. Heglund, F. Samson, J. Haufler, M. Morrison, M. Raphael, and B. Wall, editors. Predicting Species Occurences: Issues of Accuracy and Scale. Island Press, Covelo, CA.

Austin, M. P., L. Belbin, J. A. Meyers, M. D. Doherty, and M. Luoto. 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. Ecological Modelling **199**:197-216.

Austin, M. P., A. O. Nicholls, M. D. Doherty, and J. A. Meyers. 1994. Determining species response functions to an environmental gradient by means of a beta-function. Journal of Vegetation Science **5**:215-228.

Barry, S., and J. Elith. 2006. Error and uncertainty in habitat models. Journal of Applied Ecology **43**:413-423.

Bazzaz, F. A. 1998. Plant in changing environments: Linking physiological, population, and comunity ecology. Cambridge University Press, Cambridge, UK.

Bojorquez-Tapia, L. A., I. Azuara, E. Ezcurra, and O. A. Flores V. 1995. Identifying conservation priorities in Mexico through geographic information systems and modeling. Ecological

Applications **5**:215-231.
Busby, J. R. 1986. A biogeographical analysis of Nothofagus cunninghamii (Hook.) Oerst. in southeastern Australia. Australian Journal of Ecology **11**:1-7.
Carpenter, G., A. N. Gillison, and J. Winter. 1993. DOMAIN: A flexible modeling procedure for mapping potential distributions of animals and plants. Biodiversity and Conservation **2**:667-680.
Chapman, A. D., M. E. S. Munoz, and I. Koch. 2005. Environmental information: placing biodiversity phenomena in an ecological and environmental context. Biodiversity Informatics **2**:24-41.
Chen, G., and A. T. Peterson. 2002. Prioritization of areas in China for biodiversity conservation based on the distribution of endangered bird species. Bird Conservation International **12**:197-209.
Collingham, Y. 2000. Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. **37**:13-27.
Corsi, F., J. de Leeuw, and A. Skidmore. 2000. Modeling species distribution with GIS. Pages 389-434 *in* L. Boitani and T. Fuller, editors. Research Techniques in Animal Ecology. Controversies and consequences. Columbia University Press, New York.
Cristianini, N., and J. Shawe-Taylor. 2000. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.
DEH. 2004. Normalized Difference Vegetation Index. *in*. Department of Environmental and Heritage. http://www.deh.gov.au/erin/ndvi/ndvi.html.
Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife". The Annals of Statistics **7**:1-26.
Egbert, S. L., M. A. Ortega-Huerta, E. Mart¡nez-Meyer, K. P. Price, and A. T. Peterson. 2000. Time-series analysis of high-temporal resolution AVHRR NDVI imagery of Mexico. Pages 1978-1980 *in*.
Egbert, S. L., A. T. Peterson, V. Sanchez-Cordero, and K. P. Price. 1998. Modeling conservation priorities in Veracruz, Mexico. Pages 141-150 *in* S. Morain, editor. GIS in natural resource management: Balancing the technical-political equation. High Mountain Press, Santa Fe, New Mexico.
Elith, J., and M. Burgman. 2002. Predictions and their validation: Rare plants in the Central Highlands, Victoria. *in* J. M. Scott, P. J. Heglund, and M. L. Morrison, editors. Predicting Species Occurrences: Issues of Scale and Accuracy. Island Press, Washington, D.C.
Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. S. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**:129-151.
Elton, C. S. 1927. Animal Ecology. Sidgwich and Jackson, London.
Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology **41**:263-274.
Fawcett, T. 2003. ROC graphs: notes and practical considerations for data mining researchers. Palo Alto, CA: HP Laboratories.
Feria, T. P., and A. T. Peterson. 2002. Prediction of bird community composition based on point-occurrence data and inferential algorithms: a valuable tool in biodiversity assessments. Diversity and Distributions **8**:49-56.
Ferrier, S., and G. Watson. 1996. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. *in*. Canberra, Australia: NSW National Parks and Wildlife Service.

Ferrier, S., G. Watson, J. Pearce, and M. Drielsma. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. 1. Species-level modeling. Biological Conservation **11**:2275-2307.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation **24**:38-49.

Gause, G. F. 1934. The struggle for existence. Williams and Wilkins.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution **19**:497-503.

Grinnell, J. 1917. Field tests of theories concerning distributional control. American Naturalist **51**:115-128.

Guisan, A., T. C. Edwards Jr, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling **157**:89-100.

Guisan, A., C. H. Graham, J. Elith, and F. Huettmann. 2007. Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions **13**:332-340.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters **8**:993-1009.

Guisan, A., S. B. Weiss, and A. D. Weiss. 1999. GLM versus CCA spatial modelling of plant species distribution. **143**:107-122.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135**:147-186.

Hijmans, R. J. 2005. Very high resolution interpolated climate surfaces for global land areas. **25**:1965-1978.

Hijmans, R. J., S. Cameron, and J. Parra. 2004. WorldClim, a new high-resolution global climate database. *in* Inter-American Workshop on Environmental Data Access.

Hirzel, A. H., and A. Guisan. 2002. Wich is the optimal sampling strategy for habitat suitability modelling? Ecological Modelling **157**:331-341.

Hirzel, A. H., J. Hausser, D. Chessel, and N. Perrin. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology **83**:2027-2036.

Hirzel, A. H., V. Helfer, and F. Metral. 2001. Assessing habitat-suitability models with a virtual species. Ecological Modelling **145**:111-121.

Huntley, B., P. M. Berry, W. Cramer, and A. P. McDonald. 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. JOURNAL OF BIOGEOGRAPHY **22**:967-1001.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology **22**:415-427.

Isaaks, E. H., and R. M. Srivastava. 1989. Applied Geostatistics, 2 edition. Oxford University Press, Oxford.

Iwashita, F. 2007. Sensibilidade de modelos de distribuição de espécies a erros de posicionamento de dados de coleta. master degree. Instituto de Pesquisas Espaciais - INPE, São José dos Campos, SP.

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurence in the prediction of species distributions. Ecography **28**:385-393.

Magana, V., C. Conde, O. Sanchez, and C. Gay. 1997. Assessment of current and future regional climate scenarios for Mexico. Climate research **9**:107-114.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. **38**:921-931.

Metz, C. E. 1986. ROC methodology in radiologic imaging. Investigative Radiolgy **21**:720-733.

Netter, J., M. N. Kutner, C. J. Nachtssheim, and W. Wasserman. 1996. Applied linear statistical models, 4 edition. WCB/McGraw-Hill, Boston.

Nix, H. A. 1986. A biogeographic analysis of Australian elapid snakes. Pages 4-15 *in* R. Longmore, editor. Atlas of Australian Elapid Snakes. Australian Government Publishing Service, Canberra.

Oberhauser, K., and A. T. Peterson. 2003. Modelling current and future potencial wintering distributions of eastern North American monarch butterflies. PNAS (Proceedings of the National Academy of Sciences of the United States of America) **100**:14063-14068.

Oksanen, J., and P. R. Minchin. 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecological Modelling **157**:119-129.

Ortega-Huerta, M. A., and A. T. Peterson. 2004. Modelling spatial patterns of biodiversity for conservation prioritization in North-eastern Mexico. Diversity and Distributions **10**:39-54.

Parra, J. L., C. C. Graham, and J. F. Freile. 2004. Evaluating alternative data sets for ecological niche models of birds in the Andes. Ecography **27**:350-360.

Paruelo, J. M., E. G. Jobb gy, and O. E. Sala. 2001. Current distribution of ecosystem functional types in temperate South America. Ecosystems **4**:683-698.

Pearce, J., and S. Ferrier. 2000. Evaluating the predictive perfomance of habitat models developed using logistic regression. Ecological Modelling **133**:225-245.

Pearson, G. R., C. J. Raxworthy, M. Nakamura, and A. T. Peterson. 2006a. Predicting species distributios from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogegraphy **34**:102-117.

Pearson, R. G., W. Thuiller, M. B. Araújo, E. Martinez, L. Brotons, C. McClean, L. Miles, P. Segurado, T. Dawson, and D. Lees. 2006b. Model-based uncertainty in species' range prediction. JOURNAL OF BIOGEOGRAPHY **33**:1704-1711.

Pereira, R. S. 2002. Desktop Garp. *in*. University of Kansas Biodiversity Research Center, Lawrence, Kansas.

Petersen, L. R., and J. T. Roehrig. 2001. West Nile virus: A reemerging global pathogen. Emerging Infectious Diseases **7**:611-614.

Peterson, A. T., M. A. Ortega-Huerta, J. Bartley, V. Sanchez-Cordero, J. Soberć¢n, R. H. Buddemeier, and D. R. B. Stockwell. 2002a. Future projections for Mexican faunas under global climate change scenarios. Nature **416**:626-629.

Peterson, A. T., M. Papes, and D. A. Kluza. 2003a. Predicting the potential invasive distributions of four alien plant species in North America. Weed Science **51**:863-868.

Peterson, A. T., V. Sanchez-Cordero, C. B. Beard, and J. M. Ramsey. 2002b. Ecologic niche modeling and potential reservoirs for Chagas disease, Mexico. Emerging Infectious Diseases **8**:662-667.

Peterson, A. T., R. Scachetti-Pereira, and D. A. Kluza. 2003b. Assessment of Invasive Invasive Potential of Homalodisca coagulata in Western North America and South America. Biota Neotropica **3**.

Peterson, A. T., J. Soberón, and V. Sanchez-Cordero. 1999. Conservatism of ecological niches in evolutionary time. Science **285**:1265-1267.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling **190**:231-259.

Phillips, S. J., M. Dudļk, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. Pages 655-662 *in* Proceedings of the 21st International Conference on Machine Learning

21st International Conference on Machine Learning. ACM Press, New York.

Polasky, S., and A. R. Solow. 2001. The value of information in reserve site selection. Biodiversity and Conservation **10**:1051-1058.

Reddy, S., and L. M. Dávalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. **30**:1719-1727.

Roura-Pascual, N., A. SUAREZ, C. Gć¢mez, P. Pons, Y. Touyama, A. L. Wild, and A. T. Peterson. 2005. Geographic potential of Argentine ants (Linepithema humile Mayr) in the face of

global climate change. Proceedings of the Royal Society of London B **271**:2527-2535.
Rushton, S. P., S. J. Ormerod, and G. Kerby. 2004. New paradigms for modelling species distributions? Journal of Applied Ecology **41**:193-200.
Sala, O. E., F. S. Chapin-III, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, A. Kinzig, R. Leemans, D. M. Lodge, H. A. Mooney, M. n. Oesterheld, N. L. Poff, M. T. Sykes, B. H. Walker, M. Walker, and D. H. Wall. 2000. Global biodiversity scenarios for the year 2100. Science **287**:1770-1774.
Scott, J. M., P. J. Heglund, F. Samson, J. Haufler, M. Morrison, M. Raphael, and B. Wall. 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Pages 868 *in*. Island Press, Covelo, CA.
Segurado, P., and M. B. Araújo. 2004. An evaluation of methods for modelling species distributions. JOURNAL OF BIOGEOGRAPHY **31**:1555-1568.
Segurado, P., M. B. Araújo, and W. E. Kunin. 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology **43**:433-444.
Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. Bioinformatics **21**:3940-3941.
Siqueira, M. F., and G. Durigan. 2007. Modelagem da distribuição geográfica de espécies lenhosas de cerrado no Estado de São Paulo. Revista Brasileira de Botânica **30**:239-249.
Siqueira, M. F. d., and A. T. Peterson. 2003. Consequences of Global Climate Change for Geographic Distributions of Cerrado Tree Species. Biota Neotropica **3**.
Soberon, J. M., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics **2**:1-10.
Stockwell, D. R. B. 2006. Improving ecological niche models by data mining large environmental datasets for surrogate models. Ecological Modelling.
Stockwell, D. R. B., and D. Peters. 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. International Journal of Geographic Information Systems **13**:143-158.
Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling **148**:1-13.
Stoms, D. M., and W. W. Hargrove. 2000. Potential NDVI as a baseline for monitoring ecosystem functioning. International Journal of Remote Sensing **21**:401-407.
Strahler, A. H., W. Lucht, C. B. Shaaf, T. Tsang, F. Gao, X. Li, J. P. Muller, P. Lewis, and M. J. Barnsley. 1999. MODIS BRDF/Albedo Product: Algorithm Theoretical Basis Document (Version 5). *in*.
Sutton, T., R. Giovanii, and M. F. Siqueira. 2007. Introducing openModeller. OSGeo Journal **1**:1-6.
Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. N. Erasmus, M. F. d. Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A. S. v. Jaarsveld, G. F. Midgley, L. Miles, M. A. Ortega-Huerta, A. T. Peterson, O. L. Phillips, and S. E. Williams. 2004. Extinction risk from climate change. Nature **427**:145-148.
Thuiller, W. 2003. BIOMOD - optimizing prediction of species distributions and projecting potential future shifts under global change. Global Change Biology **9**:1353-1362.
UMD. 2001. AVHRR NDVI Data Set. *in*. University of Maryland, http://glcf.umiacs.umd.edu/index.shtml, College Park, Maryland.
Vapnik, V. 1995. The Nature of Statistical Learning Theory. SpringerVerlag.
Verhoef, W., M. Menenti, and S. Azzali. 1996. A colour composite of NOAA-AVHRR-NDVI based on time series analysis (1981-1992). International Journal of Remote Sensing **17**:231-235.
Wiley, E. O., K. M. McNyset, A. T. Peterson, and C. R. Robins. 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography **16**:120-127.
Yee, T. W., and N. D. Mitchell. 1991. Generalized additive models in plant ecology. Journal of

Vegetation Science **2**:587-602.
Zaniewski, A. E., A. Lehmann, and J. M. Overton. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling **157**:261-280.