Como determinar a distribuição potencial de espécies sob uma abordagem conservacionista?

PAULO DE MARCO JÚNIOR[1] *
MARINEZ FERREIRA DE SIQUEIRA[2]

[1] Laboratório de Ecologia Teórica e Síntese, Departamento de Biologia Geral, ICB, Universidade Federal de Goiás, Goiânia, Brasil.
[2] CRIA, São Paulo, Brasil.
* e-mail: pdemarco@icb.ufg.br

RESUMO

Os modelos de distribuição potencial são ferramentas importantes para determinar a distribuição de espécies ameaçadas com fins conservacionistas e para avaliar abordagens teóricas sobre processos biogeográficos. Esses modelos estão baseados na distribuição dos pontos de ocorrência da espécie no sub-espaço de condições de seu nicho ecológico e produzem funções para predizer em que locais no espaço geográfico é provável sua ocorrência. Apresentamos aqui conceitos e técnicas necessários ao emprego adequado desses modelos através de uma revisão da literatura sobre as principais questões atuais nessa área de pesquisa, além de uma comparação entre algumas técnicas de uso corrente (GARP, Maxent, SVM e BIOCLIM) utilizando dados sobre *Caryocar brasiliense* (Pequi). Problemas como a qualidade, a quantidade, os vícios amostrais, a escolha do tipo de informação ambiental, a escolha da estratégia de modelagem e a escolha da técnica para avaliar os resultados do modelo são avaliados. Alguns métodos foram originalmente desenvolvidos para dados de presença/ausência, enquanto que para a maioria dos problemas reais, apenas dados de presença são disponíveis. Com um pequeno número de amostras, como o de estudos de espécies ameaçadas, modelos mais simples são desejáveis (e.g. Similaridade e BIOCLIM). A possibilidade de testes externos (independentes) através da partição dos dados originais é um avanço importante para avaliar a qualidade do modelo final, mas técnicas baseadas em procedimento "jackknife" são adequadas para espécies raras. A análise da sensibilidade e especificidade dos modelos, com técnicas derivadas do ROC, ainda permitem a identificação de um valor limite para a determinação da amplitude de distribuição das espécies. O uso de todos esses procedimentos deve ser considerado, não só para garantir a utilidade desses modelos em uma abordagem conservacionista, mas também para uma melhor comparação dos resultados garantindo a robustez das conclusões atingidas.

ABSTRACT

Potential distribution models are important tools to determine the distribution of threatened species for conservationist purposes and to evaluate theoretical approaches about biogeographic processes.

These models are based on the distribution of species occurrence points in the environmental conditions sub-space of their ecological niche and build functions to predict in which sites its occurrence are expected. We present here the concepts and techniques need to the adequate use of these models through a literature review about the main in this research area, besides a comparison among some of these methods (GARP, Maxent, SVM and BIOCLIM) using data from Caryocar brasiliense *(Pequi). Problems related to data quality, quantity, sampling biases, choice of environmental information, modelling strategy and the technique to evaluate model results are presented. Some methods are originally developed to presence/absence data, while to the majority of real problems, only presence data are available. Under the restriction of a small number of samples, as in threatened species study, simpler models are desirable (e.g. environmental similarity and BIOCLIM). The possibility to external tests (independent) through the partition of original data bases are an important advance to evaluate the quality of the final model, but methods based on the jackknife procedure are more adequate to rare species. The analysis of sensitivity and specificity of the models, derived from ROC methodology, allow the identification of a threshold limit to the determination of species range distribution. The use of all these procedures should be considered, not only to guarantee the usefulness of these models in a conservationist approach, but to better comparison of the results to guarantee the robustness of attained conclusions.*

Introdução

A modelagem de distribuição potencial (MDP) se tornou um procedimento comum para determinar a amplitude da distribuição geográfica das espécies. Uma lista de aplicações atuais para esses métodos não vai ser completa, principalmente porque seu uso está ainda em crescimento, com inovações que permitem novas abordagens. Apesar disso, a Tabela 1 apresenta um resumo de alguns dos usos mais importantes que podem ser encontrados na literatura. É possível perceber que o MDP é útil em uma variedade de áreas, mas que há um domínio nas atividades ligadas à biologia da conservação, que será o foco principal dessa contribuição.

Há duas razões principais do aumento do uso de MDP nos últimos anos. O primeiro é o aumento da disponibilidade de métodos estatísticos poderosos e técnicas computacionais que podem ser aplicados mesmo com apenas dados da presença da espécie, recolhidos de informações de museus/herbários e levantamentos de fauna e flora (Guisan & Thuiller, 2005; Guisan *et al*., 2006). A segunda razão é a disponibilidade de dados ambientais em diferentes níveis de resolução e para uma vasta área de território, que permite produzir predições para, virtualmente, qualquer área terrestre do globo.

Nesse trabalho apresentamos uma base teórica geral de como essas técnicas podem funcionar e os principais problemas que têm sido levantados em seu uso, com a esperança de que a sua difusão em problemas de conservação de espécies seja acompanhada pela avaliação criteriosa de seus resultados. Também apresentamos um exercício para a modelagem de Pequi (*Carycocar brasiliense* Cambess.) utilizando algumas das técnicas de uso corrente para exemplificar a execução desse tipo de abordagem.

O nicho ecológico como modelo teórico

Na maior parte das aplicações de MDP se considera que o nicho ecológico é o modelo básico que sustenta a possibilidade de produzir predições sobre a ocorrência de espécies (Peterson, 2001; Thuiller *et al*., 2005; Elith *et al*., 2006; Stockwell, 2006). O argumento é simples e bem fundamentado: o nicho ecológico é definido como o conjunto de condições e recursos nos quais os indivíduos de uma espécie são capazes de sobreviver, crescer e reproduzir. Logo, o conhecimento dessas condições e recursos deve servir para predizer os locais de ocorrência da espécie. Apesar disso, muita confusão sobre o uso desse conceito para a modelagem foi resultado de equívocos sobre conceitos acessórios como o de nicho realizado.

Em geral, os MDP podem ser considerados como o ajuste a uma função entre os pontos de ocorrência de uma espécie e um conjunto multivariado de dados ambientais (Phillips *et al*., 2006). Como em geral só estão disponíveis dados de presença (ver mais à frente sobre dados de ocorrência) essas funções devem representar as características ambientais nos pontos de ocorrência ou o "nicho" da espécie.

TABELA 1 – Alguns exemplos de aplicações dos modelos de distribuição potencial de espécies (MDP) retirados da literatura recente.

ÁREA	MÉTODO*	EXEMPLOS
Predição de distribuição de espécies raras ou ameaçadas de extinção	Maxent, GARP	Engler *et al.*, 2004; Guisan *et al.*, 2006; Pearson *et al.* 2007
Guiar levantamentos para detectar espécies novas ou raras e novos padrões de distribuição	GARP, Distância Euclidiana	Raxworthy *et al.*, 2003
Escolha de espécies para recuperação de áreas degradadas	GARP	Siqueira, 2005
Escolha de áreas prioritárias para conservação	GARP, ENFA, GLM	Araujo *et al.*, 2004; Martinez *et al.*, 2006
Determinação de áreas com maior risco de invasão por espécies exóticas	GARP, BIOCLIM, Maxent	Broennimann *et al.*, 2007; Herborg *et al.*, 2007; Loo *et al.*, 2007; Peterson, 2003; Rouget *et al.*, 2001; Sutherst & Maywald, 2005
Análise do efeito das mudanças climáticas globais sobre a biodiversidade	GARP, BIOCLIM, GLM	Heikkinen *et al.*, 2006; Hijmans & Graham, 2006; Parra-Olea *et al.*, 2005; Roura-Pascual *et al.*, 2004; Thuiller *et al.*, 2005
Predição de áreas ideais para plantio	GARP, Maxent, BIOCLIM, DOMAIN	Villordon *et al.*, 2006

A abordagem mais coerente sobre a relação entre MDP e nicho foi apresentada por Soberón (2007) que separa esse conceito da mesma forma que Hutchinson (1957; 1981) o fez em seu trabalho clássico: o sub-espaço de condições (ou *cenopoético*) e o sub-espaço de recursos. Os dados ambientais disponíveis devem apenas representar o sub-espaço de condições e não o nicho completo da espécie. Além disso, é bastante provável que os pontos de ocorrência tomados representem áreas em que as condições são favoráveis, mas podem existir outras áreas com condições semelhantes, mas que a presença da espécie é impedida por interações interespecíficas. Hutchinson identificou esse conceito como nicho *pós-interativo* ou *realizado* e é ele a base correta para a modelagem com os dados disponíveis. Soberón (2007) reforça esse argumento distinguindo entre nicho Grinnelliano (que apenas leva em consideração as condições do ambiente) e nicho Eltoniano (que leva em conta as interações entre espécies). Incluir interações ecológicas dentro dos MDP é área de intensa pesquisa que tem sido desenvolvida principalmente na predição de ocorrência de uma espécie em relação à outra fortemente relacionada (e.g. herbívoros e suas plantas hospedeiras).

A necessidade de embasar as estratégias de MDP na teoria do nicho facilita a interpretação e discussão dos resultados desses modelos dentro de um contexto conservacionista. Um exemplo importante é o argumento recentemente apresentado de que espécies exóticas podem ter uma mudança de nicho nas novas áreas invadidas fora de sua distribuição original (Broennimann *et al*. 2007). Esse resultado pode representar que nossa habilidade de predizer a invasão inicial de uma espécie pode ainda ser válida, mas que esses modelos dificilmente servirão para predizer a expansão subsequente de uma espécie em uma nova área. Os modelos gerados para o nicho original seriam uma sub-estimativa da área real de ocupação. Por outro lado, esse resultado pode ainda representar o efeito da liberação competitiva ou outros mecanismos semelhantes agindo nas novas populações e, talvez, estarem mais próximo do nicho real da espécie, sem as limitações impostas pelas interações em seu habitat original.

"In data we trust"

A frase que encabeça essa seção evidencia a necessária preocupação com os dados quando se trabalha com modelagem e a MDP é particularmente sensível à qualidade e ao tipo de dado disponível. A Tabela 2 resume alguns dos problemas mais comuns encontrados nessa área.

TABELA 2 – Principais problemas encontrados nos dados utilizados em modelos de distribuição potencial de espécie (MDP).

PROBLEMA	EXPLICAÇÃO	RISCO
Precisão nos dados de ocorrência	Muitas informações da literatura apresentam apenas o município coletado e o georeferenciamento é feito pela sede do município.	Em municípios muitos extensos (e.g. na Amazônia) esses erros podem representar uma enorme diferença de características ambientais.
Vício dos dados de ocorrência	Os coletores tendem a se distribuir ao redor de grandes cidades ou estradas.	Os vícios devem gerar um modelo mais restrito que a distribuição real da espécie.
Erros de identificação	Dados de museus e herbários podem conter erros de identificação.	Descrição incorreta da relação com os fatores ambientais.
Resolução dos dados ambientais	Dados ambientais em uma resolução muito pequena podem gerar um alisamento da variação ambiental real.	Descrição pobre ou incorreta da relação com os fatores ambientais.
Dados ambientais não relacionados à espécie	As espécies podem ser limitadas em sua distribuição por variáveis não disponíveis para modelagem.	Descrição pobre da relação com os fatores ambientais.

A qualidade e a quantidade dos dados de distribuição afeta fortemente os resultados do MDP (Suarez-Seoane *et al*., 2002; Luoto *et al*., 2005), assim como a resolução e escolha das variáveis ambientais (Robertson *et al*., 2003; Elith *et al*., 2006; Austin, 2007). Todos os estudos demonstram um aumento da acurácia dos modelos com o aumento do número de pontos de ocorrência disponível (Stockwell & Peterson, 2002; Hernandez *et al*., 2006; Pearson *et al*., 2007). Essas escolhas possivelmente afetam mais o resultado dos modelos do que o efeito da seleção de um tipo de abordagem de modelagem. A geografia da espécie (amplitude de variação, padrão de autocorrelação espacial) pode também afetar a eficiência do MDP e se espera que espécies de distribuição mais ampla sejam mais suscetíveis a vícios nos dados ambientais gerados por amostragem deficiente (Luoto *et al*., 2005; Segurado *et al*., 2006).

Em uma comparação com nove técnicas muito usadas em MDP, Pearson *et al*. (2006) mostraram que o tipo de modelo classificado em relação à entrada de dados de ocorrência (só presença *vs.* presença/ausência) e os pressupostos na hora de extrapolar a distribuição foram os critérios mais importantes para explicar as diferenças nas predições dos modelos. Em especial, os modelos que usam só dados de presença tenderam a apresentar com maior frequência perda de área total de distribuição predita para 2030.

Um dos problemas importantes associados à relação entre qualidade de dados ambientais e modelagem está em que, normalmente, os pesquisadores estão produzindo predição não para uma espécie isolada, mas para um conjunto de espécies, selecionados de acordo com os critérios da pesquisa particular que está sendo desenvolvida. Alguns estudos mostram que a inclusão de muitas variáveis em modelos do tipo BIOCLIM leva sistematicamente a uma diminuição do tamanho da área de distribuição predita (Beaumont *et al*., 2005). Em uma situação como essa, e baseado na lógica de nicho ecológico que deveria embasar a MDP, a sugestão lógica seria escolher um conjunto de variáveis que deveria afetar diretamente a espécie sob estudo. Esse conjunto mínimo deveria ser escolhido baseado nos conhecimentos sobre fisiologia e ecologia geral da espécie que está sendo avaliada. No entanto, ao modelar um conjunto grande de espécies, as particularidades de cada uma serão necessariamente esquecidas, mesmo que seja devido à necessidade de manter um certo nível de comparabilidade no estudo. A generalidade da proposta provavelmente gera uma escolha de variáveis ambientais para a modelagem que, ao representar uma espécie "média", falha em descrever o nicho da maioria das espécies.

Em problemas de conservação de espécies espera-se que o número de pontos de ocorrência disponíveis deva ser considerado o principal limitante para MDP. As técnicas disponíveis têm sido substancialmente melhoradas para tratar esse problema, principalmente no que se refere aos métodos para avaliar os modelos gerados. Um argumento estatisticamente simples é que quanto menos dados estiverem disponíveis, menos parâmetros podem ser ajustados nos modelos. A consequência disso é que modelos mais simples (como os métodos de distâncias e envelopes bioclimáticos) devem ser considerados mais adequados. Mesmo assim, Pearson *et al*. (2007) demonstrou uma alta eficiência preditiva do Maxent com números de pontos de ocorrência entre 5 e 15, o que é compatível com muitos problemas de predição de espécies raras ou ameaçadas de extinção atuais.

As diferentes abordagens de modelagem

Há uma variedade de formas de modelagem aplicadas ao problema de predizer a distribuição de uma espécie. Uma primeira classificação apropriada seria distinguir modelos que foram originalmente delineados para dados de presença/ausência daqueles que foram construídos apenas para dados de presença. A maior parte dos modelos baseados em presença/ausência são derivados de técnicas estatísticas clássicas e bem conhecidas. Bons exemplos desse tipo de abordagem é o uso da regressão logística (Pearce & Ferrier, 2000; Stephenson *et al*., 2006), modelos lineares gerais ("General Linear Models" – GLM) (Guisan *et al*., 2002; Thuiller, 2003; Brotons *et al*., 2004) e de sua extensão mais complexa os modelos aditivos generalizados (Generalized additive models – GAM) (Guisan *et al*. 2002; Lehmann *et al*. 2002; Leathwick *et al*. 2006). Excetuando GAM, esses modelos se baseiam na existência de uma função simples para a relação entre a presença/ausência da espécie e um conjunto de variáveis ambientais. Essas estratégias podem produzir modelos realistas e simples para essa função de alta interpretabilidade na compreensão de processos naturais. Um bom exemplo seria o uso de regressão logística com dados ambientais e incluindo termos quadráticos (que geram uma resposta semelhante à curva de Gauss para a probabilidade de presença dependendo dos parâmetros ajustados).

Os modelos GAM, por outro lado, mantém a estrutura estatística dos modelos generalizados, mas inclui uma modelagem baseada em funções *spline* de ordens maiores. Essa estratégia gera modelos que perdem muito em interpretabilidade, mas garante maior ajuste aos dados.

Todos esses modelos, originalmente baseados em presença e ausência real, podem ser utilizados com dados reais de presença e dados de ausência simulados, ou como são usualmente referidos na literatura, pseudo-ausência. O uso de pseudo-ausência necessariamente inclui uma taxa de erro no modelo, diretamente relacionado com o tamanho da área no espaço "ecológico" definido pelas variáveis ambientais estudadas nos quais a espécie ocorre, mas que não apareceu nos dados de ocorrência. O uso de pseudo-ausências nos modelos acima poderá incluir essas áreas como "falsos zeros". Novamente, a intensidade dos vícios de amostragem nos dados de ocorrência limita diretamente o sucesso desse tipo de abordagem (Jimenez-Valverde & Lobo, 2006).

Dos modelos que foram inicialmente concebidos para dados apenas de presença, a melhor forma de classificá-los é em relação ao grau de complexidade nos processos que envolvem. Os métodos de distâncias (ou modelos de similaridade ambiental) são as representações mais simples da lógica de nicho ecológico, por estarem baseados na existência de um ponto de ótimo ecológico para cada espécie definido pelo centróide dos pontos de ocorrência no espaço ecológico. A distância entre esse ótimo estimado e os valores observados para cada célula da grade ambiental para a área geográfica estudada é inversamente relacionada à adequabilidade do ambiente naquele local. A distância euclidiana gera um envelope circular ao redor do ótimo no espaço ecológico e a distância de Mahalanobis um envelope elipsoidal (Farber & Kadmon, 2003). A distância de Mahalanobis inclui uma maior complexidade porque leva em conta a matriz de covariância entre as variáveis ambientais nos pontos de ocorrência. Isso permite interpretar o modelo como uma expressão das restrições ambientais que a espécie sofre incluindo as correlações entre variáveis, mas exige que o número de pontos seja maior que o número de variáveis ambientais (o que pode ser um problema para espécies raras).

O próximo conjunto de métodos são os envelopes bioclimáticos sob as técnicas BIOCLIM e DOMAIN (Hirzel & Arlettaz, 2003; Beaumont *et al*., 2005; Luoto *et al*., 2005; Heikkinen *et al*., 2006). Nesses casos os envelopes gerados são retilineares baseados em determinar para cada variável um limite superior e inferior para a ocorrência da espécie (ver critérios para os limites mais à frente) e produzir uma predição final que assume que não existe correlação entre as variáveis nos pontos de ocorrência.

Uma extensão lógica dos dois conjuntos de métodos apresentados seria o uso de técnicas de análise multivariada para a predição da distribuição das espécies. Métodos baseados na análise de componentes principais como o ENFA (Hettrich & Rosenzweig, 2003; Brotons *et al*., 2004; Hargrove & Hoffman, 2004; Martinez *et al*., 2006) têm a vantagem de produzir envelopes mais interpretáveis e de representarem uma forma automática de estabelecer que variáveis são mais importantes na determinação da distribuição. Tanto na utilização de técnicas multivariadas quanto no uso da distância Euclidiana a escala de medida das variáveis vai afetar fortemente os resultados: as variáveis que variarem mais serão necessariamente aquelas que dominarão as análises. Assim, se incluirmos altitude (variando de 200 a 1.000m) e temperatura (variando de 15 a 25º), a altitude dominará totalmente os modelos. Para evitar isso, devem ser utilizadas técnicas já há muito estabelecidas nos estudos com análises multivariadas (Noy-Meir *et al*., 1975; Stoddard, 1979) através da padronização

das variáveis (subtrair a média e dividir pelo desvio padrão), que faz com que cada variável entre em "pé de igualdade" no modelo.

O Maxent (Maximum Entropy) inicia a lista dos modelos mais complexos: essa é uma técnica de apredizagem-automática (*machine-learning*) que estima a distribuição de probabilidades mais próxima à distribuição uniforme sob a restrição de que os valores esperados para cada variável ambiental estejam de acordo com os valores empíricos observados nos pontos de ocorrência. Phillips *et al*. (2006) lista onze vantagens dessa técnica e as mais importantes são: i) ela necessita apenas de dados de presença; ii) a variável gerada é continua dentro do intervalo 0 a 100 indicando adequabilidade relativa; iii) ela tem uma definição matemática concisa e é facilmente interpretável dentro dos conceitos clássicos de análise de probabilidades.

O processo de modelagem no Maxent envolve alguns critérios de otimização, podendo gerar um sobre-ajuste (*overfiting*) quando o número de dados é menor que o número de parâmetros ajustados. Uma constante β é usada como parâmetro de regularização e pode depender da variabilidade observada (Dudik *et al*., 2004), mas há ainda alguma controvérsia sobre como escolher parâmetros apropriados em um conjunto de muitas espécies.

Algumas técnicas multivariadas exploratórias, originalmente desenvolvidas para *data-mining* têm se tornado popular na MDP. Dentre essas se destaca o uso de regressão multivariadas por *splines* (Leathwick *et al*., 2005, 2006; Elith & Leathwick, 2007) que apresenta algumas características semelhantes a GAM e as Árvores de regressão ou classificação (*Classification and regression trees* – CART) (Thuiller, 2003; Gavin & Hu, 2005). Nesses casos, a lógica de nicho é praticamente abandonada em favor da busca do melhor modelo que se ajuste ao conjunto de dados, e de certa forma, sacrifica a interpretabilidade ecológica em favor da qualidade do ajuste.

Por fim, as redes neurais (Thuiller, 2003; Joy & Death, 2004), algoritmos genéticos gerais (Pearson *et al*., 2006; Termansen *et al*., 2006) e o GARP ("genetic algorithm for rule-set production") compartilham muito da estrutura teórica comum aos métodos de aprendizagem-automática, mas o GARP é sem dúvida o mais utilizado desses modelos (Peterjohn, 2001; Anderson *et al*., 2002; Anderson, 2003; Ganeshaiah *et al*., 2003; Peterson & Kluza, 2003; Elith *et al*., 2006; Stockman *et al*., 2006; Villordon *et al*., 2006; Pearson *et al*., 2007, apenas para citar alguns dos mais importantes). O GARP representa uma técnica híbrida que inclui técnicas estatísticas (regressão logística) e envelopes bioclimáticos dentro de uma estratégia de aprendizado-automático. O GARP não é uma técnica de modelagem para dados de presença porque o ajuste é feito através da geração de um conjunto de pseudo-ausências, mas apresenta técnicas mais sofisticadas para tratar esse problema.

O algoritmo GARP define o modelo de nicho ecológico das espécies através de um conjunto de regras que é considerada como um "indivíduo", e o conjunto de regras são considerados uma "população", segundo a terminologia definida para os algoritmos genéticos. Internamente, as regras são codificadas através das faixas de valores ou coeficientes relativos às variáveis ambientais e também ao valor da previsão da regra. Os coeficientes das variáveis ambientais correspondem aos "genes" que compõem os "cromossomos". A previsão das regras também é codificada como um gene, podendo sofrer alterações durante a execução do algoritmo. A qualidade de cada regra presente no modelo é avaliada por uma função de adaptação, que é calculada através da significância estatística obtida pela aplicação da regra ao conjunto de pontos de treinamento fornecidos ao algoritmo. Durante a execução do algoritmo as regras são modificadas aleatoriamente por operadores heurísticos de recombinação e mutação. Esses operadores criam novas regras, que quando aplicadas aos pontos de treinamento, obtém um valor diferente na função de adaptação, devido à mudança realizada em um de seus genes. Após a criação de novos cromossomos e inclusão destes na população existente, é executada uma operação de seleção natural. Nesta operação aqueles cromossomos que têm valor da função de adaptação abaixo de um certo limiar pré-definido são eliminados da população. Quando um número predeterminado de iterações é atingido, o algoritmo é encerrado e o resultado é apresentado como um conjunto de regras a partir dos indivíduos sobreviventes. Este modelo é aplicado de volta ao espaço geográfico, indicando as regiões onde a espécie está provavelmente presente ou ausente Pereira & Siqueira (no prelo).

Outro algoritmo que começa a ser usado em MDP é o SVM (*Suport Vector Machine* – Máquina de Vetores de Suporte), que se caracteriza por ser um conjunto de métodos de aprendizagem supervisionado relacionados que pertecem à 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 & 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 dessa classificação é elaborar uma forma computacionalmente eficiente de aprender "bons" hiperplanos de separação em um espaço de características de alta dimensão. Por "bons" hiperplanos entendemos aqueles que otimizam os limites de generalização e por "computacionalmente eficiente" 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. Uma propriedade especial das SVMs é que eles simultaneamente minimizam erros de classificação empírica e maximizam a margem geométrica. Os modelos gerados pela SVM só dependem de um subconjunto de dados de treino, utilizando apenas os dados mais informativos para gerar o MDP. 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 esse tipo de ruído é sempre uma constante nesse tipo de dado, então, é importante escolher a técnica que seja mais adequada ao conjunto de dados disponível.

Critérios de escolha de limites e métodos de avaliação

Esse é, sem dúvida, o tópico de mais pesquisa atual e, portanto, o mais controverso. Em todos os métodos apresentados é necessário um critério para estabelecer o limite para a distribuição da espécie. Se existem dados de presença/ausência é possível determinar a melhor escolha como uma combinação das informações da omissão do modelo e de sua sobre-previsão: o melhor limite é aquele que minimiza a omissão e sobre-previsão. Evidentemente, essas duas propriedades estão ligadas e quanto maior a omissão menor a sobre-previsão, e vice-versa.

No entanto, essa estratégia é limitada quando apenas dados de presença estão disponíveis. A solução mais simples é a implementada nos métodos de envelopes bioclimáticos (BIOCLIM). A escolha de limites baseados em estabelecer uma taxa de omissão fixa tem como consequência lógica um maior controle da sobre-previsão. Como não é possível nenhuma inferência acerca da sobre-previsão com dados apenas de presença, o controle da omissão é a estratégia adequada ao problema. Esse controle pode ser feito utilizando uma estimativa baseada em intervalos de confiança ou através de uma escolha apropriada de percentis cobertos pelo modelo como implementado no DIVA-GIS (Hijmans *et al*., 2002; Ganeshaiah *et al*., 2003). Essa estratégia pode ser utilizada em quase todas as técnicas, sendo especialmente adequada para os métodos de distâncias.

No entanto, a escolha de limites baseados na omissão pode ser pouco efetiva se estamos tratando de espécies ameaçadas com poucos registros de ocorrência. O uso dos dados mais extremos (como a maior distância) pode ser a estratégia mais adequada tanto devido às limitações estatísticas quanto pela proposta mais "conservadora", apropriada à tomada de decisão sobre a conservação da espécie.

Uma técnica híbrida surge de produzir estimativas de omissão e sobre-previsão a partir de pseudo-ausências como desenvolvido no Maxent (Phillips *et al*., 2006). Nesse caso, foi desenvolvida uma abordagem baseada na técnica ROC (*Receiver Operating Characteristics*) no qual a sensibilidade do modelo é definida pela proporção de presenças verdadeiras do total de presenças preditas e a especificidade pela proporção de ausências verdadeiras em relação às ausências preditas. Uma curva ROC é produzida plotando a sensibilidade contra o complemento da especificidade (1-especificidade) para diferentes valores de limites da variável Maxent. A área abaixo dessa curva é conhecida como AUC e serve como uma medida de avaliação modelo independente do limite escolhido (Manel *et al*., 2001; Liu *et al*., 2005). O valor de AUC igual a 0.5 significa que o modelo não tem uma eficácia melhor do que uma seleção aleatória. O procedimento ROC permite a escolha de um limite ótimo pela leitura do limite que maximiza a soma da especificidade e sensibilidade (Manel *et al*., 2001) e foi considerado um dos cinco melhores métodos na determinação desses limites em MDP (Liu *et al*., 2005). A única questão é que as estimativas de especificidade são produzidas com a adição de 10000 pseudo-ausências e, portanto, elas incluem um erro sistemático no modelo.

Pearson *et al*. (2007) critica os métodos derivados do AUC e outras estratégias a partir da base teórica dos modelos e dos limites que os dados de presença determinam na interpretação dos resultados. Os autores argumentam que apenas a omissão é informativa nesse tipo de modelo e que falsos-positivos não devem ser considerados na sua avaliação de modelos de distribuição *potencial* que são construídos apenas para revelar áreas que podem ser ocupadas. Dentro da proposta teórica que gera esses modelos, os falsos-positivos devem ser resultado de fatores não incluídos como contingência histórica, limitação da dispersão e interações ecológicas (Anderson, 2003; Soberón, 2007). Avançando nesse argumento, Pearson *et al*. (2007) desenvolve uma técnica semelhante ao jackknife para avaliação dos modelos que é especialmente útil para análise de espécies raras ou com poucos pontos de ocorrência. Nesse método, a modelagem é repetida cada vez excluindo um dos *n* pontos de ocorrência, gerando *n-1* modelos independentes. O desempenho preditivo dos modelos pode ser então avaliado pela capacidade de prever a observação excluída em cada modelo. Essa técnica pode ser implementada em qualquer sistema de modelagem apresentada nesse trabalho.

No caso de espécies raras, o critério de escolha do limite para predição de sua ocorrência também tem um efeito muito grande nos resultados da modelagem. Sob o ponto de vista conservacionista, o peso da sobre-previsão da distribuição de uma espécie ameaçada, levando a diminuir sua estimativa de risco, é maior que o de omitir uma potencial presença sob os critério da IUCN (IUCN, 2004; Akcakaya *et al*., 2006). Nesses termos, o uso de critérios fixos como o do menor valor de adequabilidade de habitat no qual a espécie ocorreu, pode ser especialmente útil na modelagem de espécies ameaçadas de extinção. O trabalho de Pearson *et al*. (2007) suporta esse argumento, tendo encontrado que a melhor performance dos modelos foi obtida utilizando esse critério, quando avaliado pela técnica jackknife.

Um exemplo: modelando a distribuição de uma espécie típica do cerrado e de importância econômica

O problema

Atualmente existem vários algoritmos que podem ser aplicados em MDP. A comparação de modelos oriundos de diferentes algoritmos de modelagem pode ser um problema. Qual o melhor modelo? Que algoritmo melhor se aplica a uma determinada situação em modelagem? Para facilitar a escolha dos melhores modelos é interessante que existam *softwares* que realizem um processo de experimentação, ou seja, que realizem experimentos com os mesmos dados de entrada, utilizando diferentes algoritmos, em um ambiente controlado. Uma apresentação sobre este tipo de *software* pode ser encontrada em Sutton *et al*. (2007) e uma discussão sobre o processo de experimentação em MDP pode ser encontrada em Santana *et al*. (no prelo). A título de ilustração, foram gerados modelos com os mesmos dados de entrada, mesmos pontos de ocorrência da espécie e mesmas variáveis ambientais, para quatro diferentes algoritmos de modelagem.

Dados de ocorrência e variáveis ambientais

Neste exemplo foram utilizados 50 registros de ocorrência de *Caryocar brasiliense* Cambess. (Caryocaraceae) dentro do Estado de São Paulo utilizados em Siqueira & Durigan (2007). Esta é uma espécie típica do Cerrado brasileiro, conhecida popularmente como pequi e que é intensamente utilizada para alimentação.

Foram utilizados dados climáticos (temperatura anual média e precipitação anual) oriundos do Worldclim <http://www.worldclim.org/> (Hijmans *et al*. 2005) e topográficos (elevação, aspecto e inclinação do terreno), oriundos do US Geological Surveys <http://edc.usgs.gov> ambos com a mesma resolução, aproximadamente 1km.

Padrões de distribuição

Os resultados dos modelos são apresentados na figura 1. Os resultados mostram padrões que são semelhantes a muitas comparações recentes feitas entre modelos (Elith *et al*., 2006; e.g. Pearson *et al*., 2006). O Maxent tende a ser muito "limitado" aos dados produzindo um modelo de menor amplitude de distribuição. No entanto, é comum que seja um dos modelos que apresente maior valor de AUC, junto com o GARP (Elith *et al*., 2006).

Mesmo assim, há uma grande similaridade geral na distribuição gerada, como pode ser observado pelo limite sul da distribuição quando se compara, por exemplo, o GARP e BIOCLIM. Esses resultados também sugerem que um caminho interessante para a análise desse tipo de modelo, uma abordagem de "integração" dos resultados dessas diferentes abordagens, como o proposto por Araújo & New (2006). Um exemplo deste tipo de integração, chamado modelo de consenso entre vários algoritmos, já e um procedimento atuomatizado no ambiente computacional openModeller (versão 1.0.6) <http://openmodeller.sourceforge.net/>.

Conclusões

A teoria do nicho provê mais do que uma metáfora para a MDP, fornecendo uma base teórica que deve ser considerada de forma mais formal na interpretação dos resultados desse tipo de análise. A interpretação das mudanças de padrão de distribuição de espécies invasoras e falsos-positivos na ocorrência de espécies precisam ser avaliadas dentro da lógica da teoria do nicho e o reconhecimento das limitações de modelagem baseada apenas no sub-espaço de condições.

A escolha das técnicas de modelagem para MDP é um passo importante e existe uma variedade de técnicas que podem ser classificadas em relação à sua complexidade. O principal critério de escolha, no entanto, deve ser a qualidade e quantidade de dados de ocorrência da espécie a ser modelada: quanto menos dados mais simples deve ser o modelo utilizado.

A avaliação dos modelos é feita usualmente por técnicas baseadas no procedimento ROC, mas que vêm sendo criticadas. A utilização de técnicas mais simples e adequadas à limitação do uso apenas de dados de presença na MDP é desejável, principalmente as técnicas de avaliação por jackknife para espécies raras.

No geral, os diferentes modelos gerados como exemplo assemelham-se quanto à área prevista de distribuição potencial da espécie, mas os valores de AUC variaram bastante. No caso analisado, o melhor modelo foi o gerado pelo Maxent, apresentando o maior valor de AUC entre todos os algoritmos utilizados.

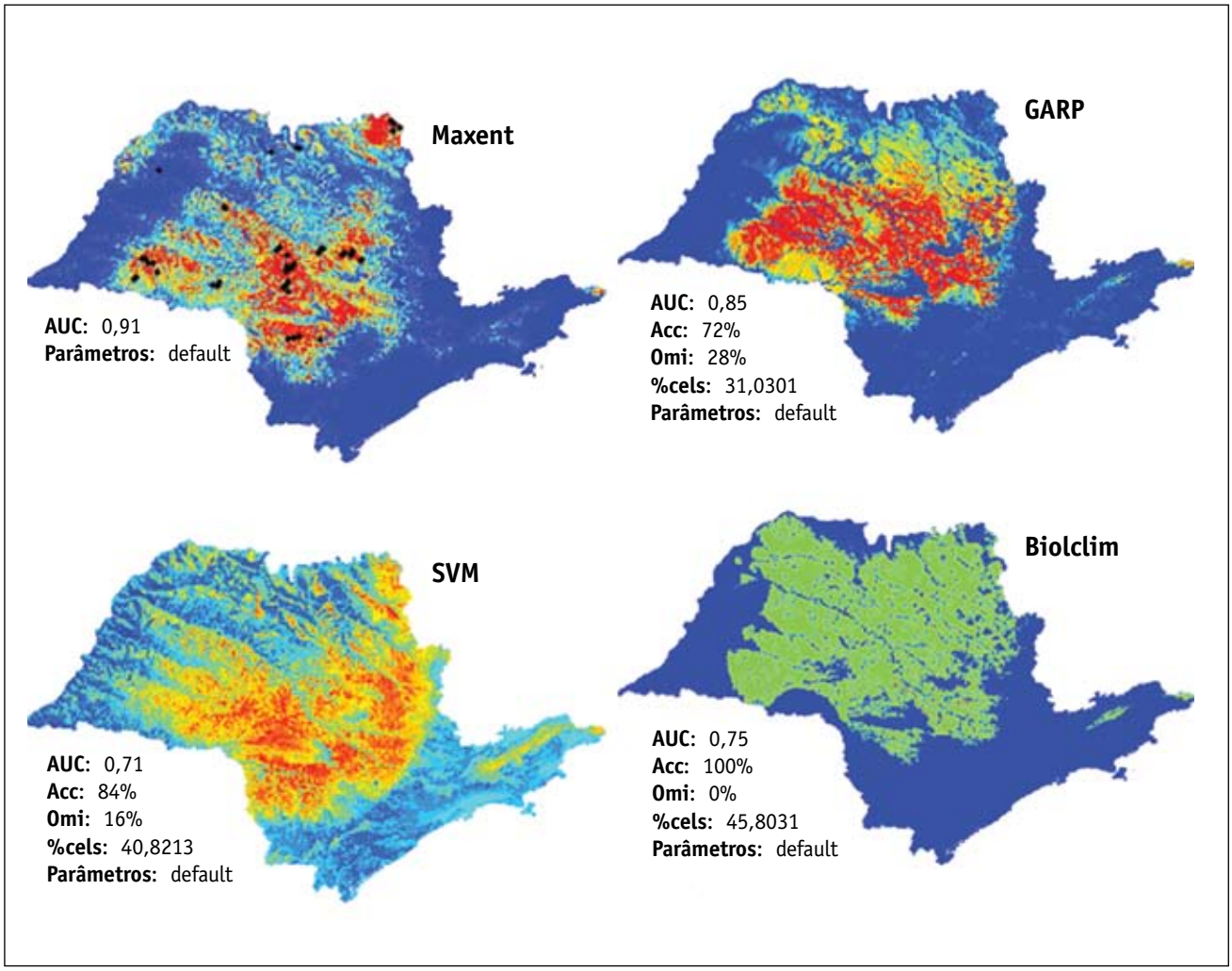

FIGURA 1 – Resultado de modelagens utilizando o mesmo conjunto de dados de entrada, modificando-se apenas os algorítmos de modelagem para gerar os MDP. AUC ("Area under curve"); Acc (acurácia); Omi (taxa de omissão) e %cels (porcentagem de células preditas como presente). Todos os modelos foram gerados com os parâmetros padrões (default) dos diferentes algorítmos. Os pontos pretos presentes no modelo gerado pelo Maxent representam os registros de ocorrência da espécie e que foram utilizados por todos os demais algorítmos.

Agradecimentos

Esse trabalho foi financiado pelo CNPq (bolsa de produtividade e financiamento direto a PDMJr), Projeto BioImpacto BBVA Espanha e Projeto openModeller – CRI – FAPESP (para MFS).

Referências bibliográficas

Akcakaya, H.R., S.H.M. Butchart, G.M. Mace, S.N. Stuart & C. Hilton-Taylor. 2006. Use and misuse of the IUCN Red List Criteria in projecting climate change impacts on biodiversity. Global Change Biology 12: 2037-2043.

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

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

Araujo, M.B., M. Cabeza, W. Thuiller, L. Hannah & P.H. Williams. 2004. Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. Global Change Biology 10: 1618-1626.

Araujo, M.B. & M. New. 2006. Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22: 42-47.

Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological Modelling 200: 1-19.

Beaumont, L.J., L. Hughes & M. Poulsen. 2005. Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. Ecological Modelling 186: 250-269.

Broennimann, O., U.A. Treier, H. Muller-Scharer, W. Thuiller, A.T. Peterson & A. Guisan. 2007. Evidence of climatic niche shift during biological invasion. Ecology Letters 10: 701-709.

Brotons, L., W. Thuiller, M.B. Araujo & A.H. Hirzel. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27: 437-448.

Cristianini, N. & J. Shawe-Taylor. 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, London.

Dudik, M., S.J. Phillips & R.E. Schapire. 2004. Performance guarantees for regularized maximum entropy density estimation. Proceedings of the 17th Annual Conference on Computational Learning Theory 655-662.

Elith, J. & J. Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions 13: 265-275.

Elith, J., C.H. Graham, R.P. Anderson, M. Dudik, 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. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberon, S. Williams, M.S. Wisz & N.E. Zimmermann. 2006. Novel methods improve prediction of species distributions from occurrence data. Ecography 29: 129-151.

Engler, R., A. Guisan & L. Rechsteiner, L. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41: 263-274.

Farber, O. & R. Kadmon. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. Ecological Modelling 160: 115-130.

Ganeshaiah, K.N., N. Barve, N. Nath, K. Chandrashekara, M. Swamy & R.U. Shaanker. 2003. Predicting the potential geographical distribution of the sugarcane woolly aphid using GARP and DIVA-GIS. Current Science 85: 1526-1528.

Gavin, D.G. & F.S. Hu. 2005. Bioclimatic modelling using Gaussian mixture distributions and multiscale segmentation. Global Ecology and Biogeography 14: 491-501.

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

Guisan, A., O. Broennimann, R. Engler, M. Vust, N.G. Yoccoz, A. Lehmann & N. E. Zimmermann. 2006. Using niche-based models to improve the sampling of rare species. Conservation Biology 20: 501-511.

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

Hargrove, W.W. & F.M. Hoffman. 2004. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. Environmental Management 34: S39-S60.

Heikkinen, R.K., M. Luoto, M.B. Araujo, R. Virkkala, W. Thuiller & M.T. Sykes. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography 30: 751-777.

Herborg, L.M., C.L. Jerde, D.M. Lodge, G.M. Ruiz & H.J. MacIsaac. 2007. Predicting invasion risk using measures of introduction effort and environmental niche models. Ecological Applications 17: 663-674.

Hernandez, P.A., C.H. Graham, L.L. Master & D.L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29: 773-785.

Hettrich, A. & S. Rosenzweig. 2003. Multivariate statistics as a tool for model-based prediction of floodplain vegetation and fauna. Ecological Modelling 169: 73-87.

Hijmans, R.J. & C.H. Graham. 2006. The ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biology 12: 2272-2281.

Hijmans, R.J., L. Guarino & E. Rojas. 2002. DIVA-GIS, version 2. A geographic information system for the analysis of biodiversity data. Manual. International Potato Center, Lima, Peru.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones & A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Hirzel, A.H. & R. Arlettaz. 2003. Modeling habitat suitability for complex species distributions by environmental-distance geometric mean. Environmental Management 32: 614-623.

Hutchinson, G.E. 1957. Concluding remarks. Cold Spring Harbor Symposium of Quantitative Biology 22: 415-427.

Hutchinson, G.E. 1981. Introducción a la Ecología de Poblaciones. Blume Ecologia, Barcelona.

IUCN. 2004 IUCN red list of threatened species. http://www.redlist. org/ 2005.

Jimenez-Valverde, A. & J.M. Lobo. 2006. The ghost of unbalanced species distribution data in geographical model predictions. Diversity and Distributions 12: 521-524.

Joy, M.K. & R.G. Death. 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology 49: 1036-1052.

Leathwick, J.R., D. Rowe, J. Richardson, J. Elith & T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biology 50: 2034-2052.

Leathwick, J.R., J. Elith & T. Hastie. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199: 188-196.

Lehmann, A., J.M. Overton & J.R. Leathwick. 2002. GRASP: generalized regression analysis and spatial prediction. Ecological Modelling 157: 189-207.

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

Loo, S.E., R. Mac Nally & P.S. Lake. 2007. Forecasting New Zealand mudsnail invasion range: Model comparisons using native and invaded ranges. Ecological Applications 17: 181-189.

Luoto, M., J. Poyry, R.K. Heikkinen & K. Saarinen. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. Global Ecology and Biogeography 14: 575-584.

Manel, S., H.C. Williams & S.J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38: 921-931.

Martinez, I., F. Carreno, A. Escudero & A. Rubio. 2006. Are threatened lichen species well-protected in Spain? Effectiveness of a protected areas network. Biological Conservation 133: 500-511.

Noy-Meir, I., D. Wlaker & W.T. Williams. 1975. Data transformations in ecological ordination. II. On the meaning of data standardization. Journal of Ecology 63: 779-800.

Parra-Olea, G., E. Martinez-Meyer & G.F.P. de Leon. 2005. Forecasting climate change effects on salamander distribution in the highlands of central Mexico. Biotropica 37: 202-208.

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

Pearson, R.G., W. Thuiller, M.B. Araujo, E. Martinez-Meyer, L. Brotons, C. McClean, L. Miles, P. Segurado, T.P. Dawson & D. C. Lees. 2006. Model based uncertainty in species range prediction. Journal of Biogeography 33: 1704-1708.

Pearson, R.G., C.J. Raxworthy, M. Nakamura & A.T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34: 102-117.

Pereira, R.S. & M.F. Siqueira, M.F. no prelo. Algoritmos Genéticos. Megadiversidade.

Peterjohn, B.G. 2001. Some considerations on the use of ecological models to predict species' geographic distributions. Condor 103: 661-663.

Peterson, A.T. 2001. Predicting species' geographic distributions based on ecological niche modeling. Condor 103: 599-605.

Peterson, A.T. 2003. Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology 78: 419-433.

Peterson, A.T. & D.A. Kluza. 2003. New distributional modelling approaches for gap analysis. Animal Conservation 6: 47-54.

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

Raxworthy, C.J., E. Martinez-Meyer, N. Horning, R.A. Nussbaum, G.E. Schneider, M. Ortega-Huerta & A.T. Peterson. 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426: 837-841.

Robertson, M.P., C.I. Peter, M.H. Villet & B.S. Ripley. 2003. Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques. Ecological Modelling 164: 153-167.

Rouget, M., D.M. Richardson, S.J. Milton & D. Polakow. 2001. Predicting invasion dynamics of four alien *Pinus* species in a highly fragmented semi-arid shrubland in South Africa. Plant Ecology 152: 79-92.

Santana, F.S., M.F. Siqueira, A.M. Saraiva & P.L.P. Correa. no prelo. A reference business process for ecological niche modelling. Ecological Informatics.

Segurado, P., M.B. Araujo & W.E. Kunin. 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology 43: 433-444.

Siqueira, M.F. 2005. Uso de modelagem de nicho fundamental na avaliação do padrão de distribuição geográfica de espécies vegetais. Tese de Doutorado. Universidade de São Paulo, Escola de Engenharia de São Carlos. 107pp.

Siqueira, M.F. & G. Durigan. 2007. Modelagem da distribuição geográfica de espécies lenhosas de cerrado no Estado de São Paulo. Revista Brasileira de Botânica 30: 249.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. Ecology Letters 10: 1115-1123.

Stephenson, C.M., M.L. MacKenzie, C. Edwards & J.M.J. Travis. 2006. Modelling establishment probabilities of an exotic plant, *Rhododendron ponticum*, invading a heterogeneous, woodland landscape using logistic regression with spatial autocorrelation. Ecological Modelling 193: 747-758.

Stockman, A.K., D.A. Beamer & J.E. Bond. 2006. An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. Diversity and Distributions 12: 81-89.

Stockwell, D.R.B. 2006. Improving ecological niche models by data mining large environmental datasets for surrogate models. Ecological Modelling 192: 188-196.

Stockwell, D.R.B. & A.T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148: 1-13.

Stoddard, A.M. 1979. Standardization of measures prior to cluster analysis. Biometrics 35: 765-773.

Suarez-Seoane, S., P.E. Osborne & J.C. Alonso. 2002. Large-scale habitat selection by agricultural steppe birds in Spain: identifying species-habitat responses using generalized additive models. Journal of Applied Ecology 39: 755-771.

Sutherst, R.W. & G. Maywald, G. 2005. A climate model of the red imported fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae): Implications for invasion of new regions, particularly Oceania. Environmental Entomology 34: 317-335.

Sutton, T., R. Giovanii & M.F. Siqueira. 2007. Introducing openModeller. OSGeo Journal 1: 1-6.

Termansen, M., C.J. McClean & C.D. Preston. 2006. The use of genetic algorithms and Bayesian classification to model species distributions. Ecological Modelling 192: 410-424.

Thuiller, W. 2003. BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9: 1353-1362.

Thuiller, W., S. Lavorel & M.B. Araujo. 2005. Niche properties and geographical extent as predictors of species sensitivity to climate change. Global Ecology and Biogeography 14: 347-357.

Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer Verlag

Villordon, A., W. Njuguna, S. Gichuki, P. Ndolo, H. Kulembeka, S.C. Jeremiah, D. LaBonte, B. Yada, P. Tukamuhabwa & R.O.M. Mwanga. 2006. Using GIS-based tools and distribution modeling to determine sweetpotato germplasm exploration and documentation priorities in sub-Saharan Africa. Hortscience 41: 1377-1381.