LaMaB: Laboratório de Mastozoologia e Biogeografia
Departamento de Ciências Biológicas
Universidade Federal do Espírito Santo

TUTORIAL DE MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES: GUIA TEÓRICO

Jeronymo Dalapicolla[1]

[1]*E-mail: jdalapicolla@usp.br*

APRESENTAÇÃO

Esse tutorial foi feito em novembro de 2015, durante a disciplina *LCF5883 – Modelagem e Distribuição de Espécies para a Conservação da Biodiversidade* na Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ/USP) em Piracicaba-SP. Essa disciplina foi lecionada pelos professores: Dra. Katia M. P. M. de Barros Ferraz (ESALQ/USP), Dr. Milton Cezar Ribeiro (UNESP/Rio Claro) e a doutoranda Flávia Pinto (UNESP/Rio Claro). Além disso esse material conta com dicas de Alan Braz (Universidade Federal do Rio de Janeiro), Ana Carolina Loss (Universidade Federal do Espírito Santo), João Paulo Hoope (Universidade Federal do Espírito Santo) e algumas imagens de Flávia Pinto (UNESP/Rio Claro) e de artigos clássicos da área.

As técnicas de modelagem são novas e muito dinâmicas, assim as metodologias usadas aqui podem vir a ser consideradas antiquadas no futuro. O intuito desse material é ajudar as pessoas que estão começando nessa área. Há um tutorial com um guia prático para a construção de modelos de distribuição de espécies usando o MaxEnt e o ArcGIS. Esse tutorial será referido aqui como a parte 2 do tutorial. É aconselhado ler este tutorial teórico antes de iniciar o guia prático.

Esse material será atualizado periodicamente por mim, dentro do possível, acrescentando técnicas novas, corrigindo erros de digitação e problemas metodológicos que surgirem, por isso o *feedback* é tão importante. Qualquer dúvida, esclarecimento ou sugestão podem me escrever. A intenção no futuro é trocar o ArcGIS pelo QGIS e construir os modelos dentro do ambiente R. Espero que esse material seja útil a vocês como foi para mim.

Jeronymo Dalapicolla.

Dalapicolla, J. 2016. Tutorial de modelos de distribuição de espécie: guia teórico. Laboratório de Mastozoologia e Biogeografia, Universidade Federal do Espírito Santo, Vitória. Disponível em: http://blog.ufes.br/lamab/tutoriais

1) O QUE SÃO OS MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES - MDE (*SPECIES DISTRIBUTION MODELS - SDM*):

Os modelos são associações entre as variáveis ambientais e os registros de ocorrência de uma espécie-alvo para identificar as condições ambientais dentro das quais as populações dessa espécie podem ser mantidas indefinidamente. A ferramenta permite estimar a distribuição espacial do ambiente que é favorável a uma determinada espécie para uma determinada área de estudo. As associações podem ser realizadas por diferentes algoritmos (funções matemáticas implementadas em diferentes *softwares*). Os modelos gerados podem ser chamados de envelopes bioclimáticos ou envelopes climáticos quando incluir apenas variáveis climáticas.

Os modelos são utilizados mais frequentemente de quatro maneiras: (1) para calcular a adequabilidade relativa do habitat ocupado pelas espécies – distribuição atual; (2) para estimar a adequabilidade relativa do habitat em áreas geográficas que não são ocupadas pelas espécies – distribuição potencial/espécies invasoras; (3) para estimar mudanças na adequação de habitat ao longo do tempo, dado um cenário específico para a mudança ambiental – paleomodelagem e modelagem futura; e (4) como estimativa do nicho ecológico da espécie (Warren & Seifert 2011).

O tutorial dará mais foco para o uso de MDE na estudo da evolução e da biogeografia, pois é o foco principal do laboratório. Há três formas principais dos modelos se integrarem com a evolução/biogeografia e especialmente com a filogeografia segundo Alvarado-Serrano & Knowles (2014):

a) Avaliações visuais de concordância com a variação genética: usado para interpretar padrões de variação genética com base na concordância *post hoc* entre os padrões de divergência genética e projeções da distribuição das espécies. Visualmente são indicadas barreiras ou corredores entre populações a partir dos modelos (*e. g.*, Lamb et al. 2008).

b) Identificação de efeitos da paisagem: pode-se calcular o impacto da paisagem sobre a conectividade da população. Por exemplo, informações sobre a adequabilidade do habitat pode ser traduzidas em caminhos de migração prováveis entre as populações usando Análise de Corredores de Menor Custo (*e. g.*, Chan et al 2011; Dalapicolla 2014).

c) Áreas de estabilidade: As previsões para a distribuição das espécies em diferentes períodos de tempo pode ser usado para identificar as regiões de estabilidade ambiental em que uma espécie pode (a princípio) ter persistido, em contraste com áreas instáveis, ou seja,

áreas onde as mudanças climáticas teria feito a região inabitável durante determinados períodos (*e. g.*, Carnaval & Moritz 2008).

2) PREMISSAS DOS MDE:

a) Espécies em equilíbrio: A espécie deve estar em "equilíbrio" em relação às condições ambientais atuais. Uma espécie está em equilíbrio com o ambiente físico se ela ocupar todas as zonas ambientalmente adequadas e se ausentar de todas as áreas inadequadas. O grau de equilíbrio depende tanto interações bióticas (por exemplo, competição e parasitas) quanto a capacidade de dispersão (organismos com maior capacidade de dispersão são mais propícios a estarem mais perto do equilíbrio do que os organismos com menor capacidade de dispersão; Araújo & Pearson 2005). Ao utilizar o conceito de "equilíbrio", devemos lembrar que a distribuição das espécies muda ao longo do tempo, de modo que o termo não deve ser utilizado para implicar estase.

b) Registro de distribuição completo: Os registros de ocorrência utilizados nos MDE devem proporcionar uma amostra de todo espaço ambiental ocupada pela espécie. Nos casos em que há poucos registros ocorrência, devido a levantamento de dados incompleto, os registros disponíveis não fornecerão as informações suficientes sobre as condições ambientais ocupados pela espécie.

c) Conservadorismo de nicho ecológico ou climático (*Niche Conservatism*): é definido como a tendência de uma espécie reter ou permanecer com seu nicho ecológico sem modificações por uma escala de tempo evolutivo (Wiens & Graham 2005). Isto significa dizer que populações atuais têm o mesmo nicho ecológico que suas populações ancestrais. Essa ideia é uma condição básica para a paleomodelagem, modelagem futura e para a modelagem de espécies invasoras ou exóticas, onde a *transferibilidade* temporal e espacial do modelo só poderia acontecer se o nicho ecológico da espécie (requisitos para a sobrevivência) fosse conservado no tempo (projeção no passado/futuro) e no espaço (projeção na área invadida).

d) Conceito de nicho ecológico (*Ecological Niche*): Para utilizar os MDE e a premissa do conservadorismo de nicho ecológico é necessário seguir uma definição de nicho, da mesma forma que se segue um conceito de espécie em sistemática. Há várias opções e a partir deles as interpretações dos modelos poderão mudar. Mais sobre os conceitos de nicho ecológico será discutido no próximo tópico.

e) Registros de distribuição amostrados aleatoriamente: Um pressuposto fundamental dos MDE é que toda a área de interesse tenha sido sistematicamente ou aleatoriamente amostrados (Phillips et al. 2009; Royle et al. 2012). Na prática, os modelos são quase invariavelmente construído com pontos de ocorrência que são espacialmente tendenciosos para as áreas mais facilmente acessadas ou melhor pesquisadas (Phillips et al. 2009; Kramer-Schadt et al. 2013). Isso gera erros de *autocorrelação espacial* que devem ser corrigidos. Mais sobre os erros comuns dos MDE será discutido em um dos tópicos abaixo.

f) Amostras de treino e teste independentes: as amostras usadas para a fase de treino e de teste do modelo devem ser independentes e não correlacionados, seja espacialmente ou em termos de esforço de coleta. Mais sobre as etapas dos MDE será discutida em um dos tópicos abaixo.

3) TEORIA E CONCEITOS DE NICHO ECOLÓGICO:

a) Grinnell (1917): A primeira ideia de nicho. Foi definida como sendo simplesmente os locais (habitats) onde os requisitos para uma determinada espécie viver e se reproduzir estão presentes;

b) Elton (1927): adicionou o nível trófico à ideia de nicho;

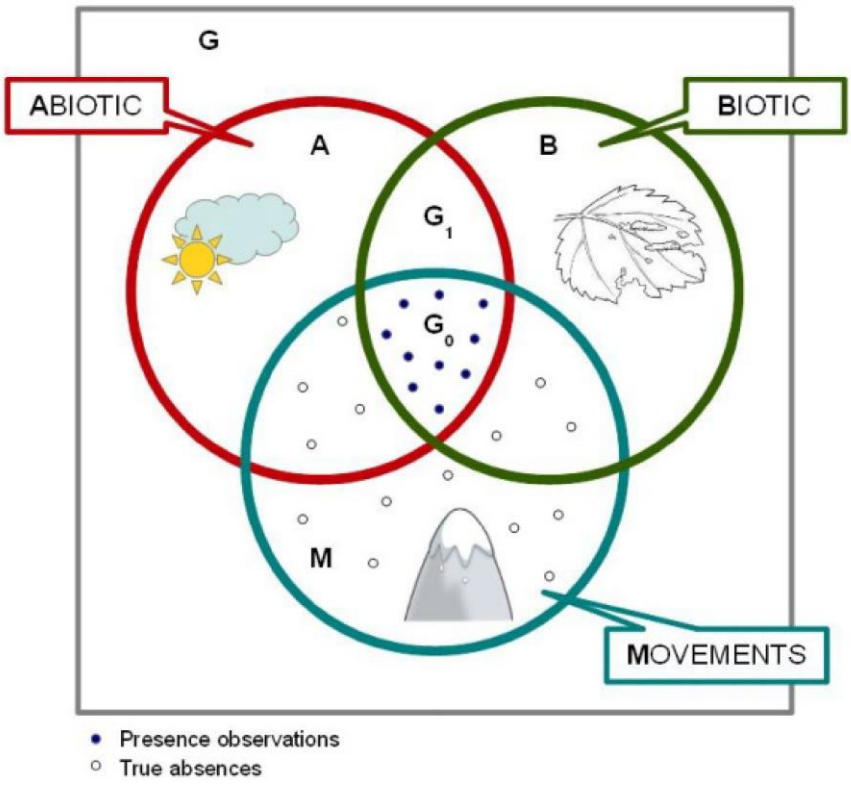
c) Gause (1934): adicionou a intensidade da competição entre espécies;

d) Hutchinson (1957): o termo nicho ecológico 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 outros. Dentro deste conceito existe 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.

De acordo com o conceito de nicho ecológico utilizado, a interpretação dos resultados da modelagem pode ser diferente. Por exemplo, muitos trabalhos levam em consideração a definição de nicho de Hutchinson (1957), onde foi criado o conceito de nicho ecológico

fundamental, ou seja, aquele que a espécie pode ocupar, e o de nicho ecológico realizado ou efetivo, que é aquele que a espécie realmente ocupa, devido às barreiras para dispersão ou interações ecológicas desfavoráveis (Araújo & Guisan 2006). Alguns autores citam que os modelos de nicho proporcionam uma aproximação ao nicho fundamental da espécie (Soberón & Peterson 2005) enquanto outros defendem que a modelagem seja a representação espacial do nicho efetivo (Guisan & Zimmermann 2000; Pearson & Dawson 2003). Baseados no conceito de nicho de Elton (1927), alguns autores defendem que, na ausência de dados sobre interações biológicas e com apenas dados climáticos, a representação da modelagem não seria de um nicho ecológico (Soberón 2007). Neste caso, as áreas indicadas na modelagem seriam de distribuição potencial ou preditiva (Jiménez-Valverde et al. 2008; Giannini et al. 2012), ou ainda de hábitat e locais satisfatórios para a ocorrência de espécie (Phillips 2008), baseado no conceito de nicho de Grinnell (1917).

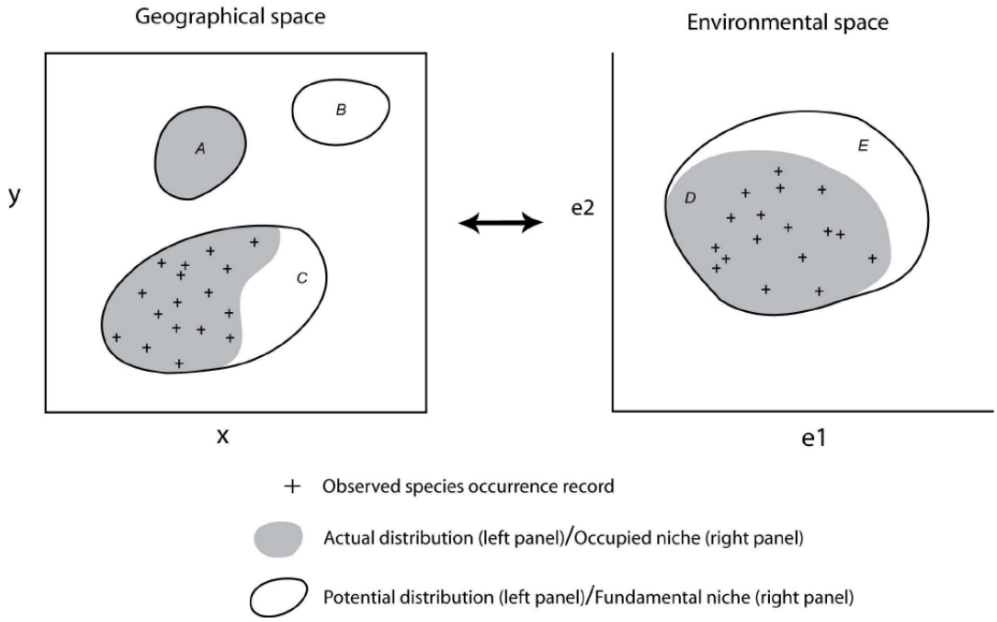
O mais importante é saber que um modelo representa uma aproximação do nicho ecológico da espécie apenas nas dimensões das camadas ambientais utilizadas, ou seja, é utilizado um subespaço do nicho ecológico na realização da modelagem. Neste tipo de modelagem não entram fatores históricos de ocupação, barreiras geográficas, interação entre espécies (competição, predação, doenças, mutualismo etc). Dessa forma, os resultados da modelagem correspondem a uma previsão, baseada em dados de parte do nicho realizado, que se aproxima do nicho fundamental da espécie, mas não o é. A área projetada representa a distribuição potencial da espécie baseada nas camadas ambientais utilizadas na modelagem. Atualmente apenas um grupo de pesquisa, que ainda é influente, defende que os modelos representam o nicho ecológico, todos os demais acreditam que os modelos representam a distribuição atual. Soberón e Nakamura (2009) acrescentaram a capacidade de dispersão na interpretação de modelos e criaram o diagrama BAM (Biotic, Abiotic and Moviments).



Exemplo acima é de um Diagrama BAM. G: região geográfica em estudo; A e B: zonas onde as variáveis climáticas e bióticas, respectivamente, são adequadas para a espécie; M: área que a espécie tem sido capaz de alcançar em um determinado período de tempo; G0: área real da distribuição das espécies (atual); G1: região que tem tanto condições bióticas e abióticas adequadas para a espécie, que podem ser potencialmente invadidas se as condições M mudar (distribuição potencial). Os pontos brancos (TRUE ABSENCES) são pontos onde a espécie pode dispersar, mas não consegue persistir por muito tempo por falta de condições bióticas, abióticas ou ambas. Os modelos de distribuição que tem apenas pontos de presença (biótico) e variáveis ambientais climáticas (abióticos) seriam representações mais parecidas da distribuição potencial junto da distribuição geográfica atual (G0 + G1). Os modelos que nas variáveis ambientais acrescentarem dados de paisagem, conectividade entre ambientes, que representariam o M, estariam modelando algo próximo da distribuição atual da espécie (G0).

Outro conceito que surge com a teoria de nicho de Hutchinson é de *Espaço Geográfico* e *Espaço Ambiental*. O espaço geográfico refere-se à localização espacial da espécie, geralmente referenciadas usando coordenadas x e y. O espaço ambiental refere-se ao nicho

dimensional de Hutchinson, ilustrado na figura abaixo em apenas duas dimensões por uma questão didática (definida por dois fatores ambientais, e1 e e2).



As cruzes representam os registros observados de ocorrência da espécie-alvo. O sombreamento cinza no espaço geográfico representa a distribuição real da espécie (ou seja, áreas ocupadas pela espécie). Observe que algumas áreas de distribuição real podem ser desconhecidas (por exemplo, a área A está ocupada mas a espécie não foi detectada lá). A área cinzenta no espaço ambiental representa a parte do nicho ocupado pela espécie: o nicho realizado. Novamente, observe que os registros de ocorrência observados podem não identificar a extensão do nicho ocupado (por exemplo, a área sombreada ao redor da letra D não inclui locais conhecidos). A linha contínua no espaço ambiental retrata o nicho fundamental, que representa toda a gama de condições abióticas dentro das quais a espécie é viável. Dentro do espaço geográfico, as linhas contínuas representam áreas com condições abióticas adequadas para a ocorrência da espécie. Essa é a distribuição potencial da espécie. Algumas regiões da distribuição potencial pode não ser habitada pelas espécies devido a interações bióticas ou limitações de dispersão. Por exemplo, a área B é ambientalmente adequada para a espécie, mas não faz parte da distribuição, talvez porque a espécie tenha sido incapaz de se dispersar e alcançar esta área. Similarmente, a área não sombreada em torno da letra C não é habitada, talvez devido à competição com outra espécie. Assim, a área não sombreada em torno da letra E identifica as partes do nicho que estão desocupados, por exemplo devido a interações bióticas ou restrições geográficas à dispersão das espécies.

4) TIPOS DE MODELOS:

a) Modelagem correlativa (*Correlative modelling*): Modelagem clássica que usa dois tipos de dados para modelar os requisitos ambientais de uma espécie e estimar sua distribuição geográfica potencial: as localidades (registros de ocorrência) de presença da espécie (geralmente não usa informações sobre localidades onde a espécie está ausente) e as variáveis ambientais para a região de estudo, especialmente as climáticas. É criada uma correlação entre os dois conjuntos de variáveis para estimar a distribuição;

b) Modelagem mecanicista (*Mechanistic modelling*): abordagens que usam propriedades biofísicas dos organismos (principalmente fisiologia) para ligar diretamente traços funcionais com as condições ambientais para determinar as áreas onde possam existir tais espécies (Alvarado-Serrano & Knowles 2014). Em suma, a partir de requerimentos fisiológicos (quantidade de água, capacidade de digerir certas plantas) são criado modelos sobre onde a espécie pode ocorrer.

5) TIPOS DE DADOS UTILIZADOS NA MODELAGEM CORRELATIVA:

a) Dados bióticos: São os registros de ocorrência da espécie. Existem dois tipos de dados bióticos associados aos Modelos de Distribuição de Espécies (MDE), registros de presença e os registros de ausência de uma determinada espécie. Os registros de presença são comumente gerados através de coletas de espécimes ou de observações em 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). Os dados de ausência nem sempre refletem uma ausência real ou inadequação do ambiente para ocorrência da espécie, podendo simplesmente indicar que a espécie não chegou a um determinado local ou que há falta de inventários no local. Portanto, esses dados devem ser usados com extrema cautela na geração de MDEs (Peterson et al. 2008). Normalmente os MDE são construídos apenas com dados de presença, o que pode gerar outros problemas durante a modelagem. Alternativamente, os "dados de ausência" podem ser gerados sob a forma de pseudoausências (alguns autores consideram os pontos de *background* como sinônimos de pseudoausências) e usados para a modelagem. As pseudoausências são áreas sem informações definidas sobre a ocorrência da espécie, mas que supostamente as espécies estão ausentes. Esses pontos têm como objetivo fornecer contrastes estatísticos com a área de presença para as análises (Soberón & Peterson 2005).

b) Dados abióticos: São as variáveis ambientais. Normalmente são imagens (também chamadas de camadas ou *layers*) de uma região. Essas imagens são divididas em pequenas quadrículas chamadas células (*cells*) ou pixels e cada pixel possui um valor numérico, correspondente ao valor da variável ambiental representada. Existem dois tipos de variáveis/camadas ambientais, as contínuas e as categóricas. Os dados ambientais contínuos (ou numéricos) são os mais usados em MDE, como os dados climáticos (temperatura e precipitação) e os topográficos (elevação e inclinação do terreno). Neles o valor do pixel representa o valor da variável naquela área. Nos dados categóricos os valores do pixels representam uma categoria em que a área foi classificada. As variáveis mais usadas para a modelagem são as do *WorldClim*, que são 19 variáveis oriundas de 2 informações apenas, precipitação e temperatura. Por causa disso há muita correlação entre as 19 variáveis (uma variável é influenciada por outra), e essa é uma das principais críticas ao uso delas. Contudo, as variáveis *WorldClim* têm sido amplamente utilizados para a geração de MDE, incluindo exemplos bem sucedidos com pequenos mamíferos não-voadores na região Neotropical (Anderson & Gonzalez 2011; Anderson & Raza 2010; Shcheglovitova & Anderson 2013). É necessário tomar cuidado com o uso de variáveis categóricas, elas reduzem o número de graus de liberdade do modelo e podem aumentar o erro. Se for usá-las é importante ter muitos registros de ocorrência e usar variáveis com poucas categorias.

As variáveis ambientais podem ser geradas de duas formas: pela *interpolação* ou por *sensoriamento remoto*. A interpolação é um método que permite construir um novo conjunto de dados a partir de dados pontuais previamente conhecidos. No caso, os dados ambientais são recolhidos em estações pontuais ao redor do globo ou de uma área específica. A partir daí são construídas funções que aproximadamente se "encaixem" nestes dados pontuais, conferindo-lhes, então, a continuidade desejada para criar resoluções mais fina (maior detalhamento). Todos os dados climáticos do *WorldClim* são interpolados. Os dados de sensoriamento remoto são derivados de imagens de satélite e representariam dados mais precisos sobre a superfície do planeta (mas há o problema de nuvens, poluição e outros fenômenos que podem alterar as leituras dos satélites). É recomendado, se possível, utilizar variáveis ambientais dos dois tipos para a construção dos modelos.

6) O PROGRAMA MAXENT:

MaxEnt utiliza o princípio da máxima entropia em dados de presença para estimar um conjunto de funções que se relacionam com variáveis ambientais do habitat a fim de se

aproximar da distribuição geográfica potencial das espécies (Phillips et al. 2006). O 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 da espécie-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.

O programa é popular porque é fácil de usar e produz resultados robustos com dados de distribuição esparsos, amostrados de forma irregular e com pequenos erros na sua localização (Elith et al. 2006). Tais limitações são comuns em dados de distribuição de espécies raras ou ameaçadas, em dados provenientes de áreas de difícil acesso e em dados de museus (Graham et al. 2008; Wisz et al. 2008). MaxEnt tem a vantagem de usar apenas dados de presença, portanto, não exigindo dados das ausências confirmadas de áreas específicas. Em geral, estas características levaram MaxEnt a ser considerado como um dos melhores programas para MDE. Em resumo: (1) ele é de uso comum, passível de comparação com outros trabalhos; e (2) teve bom desempenho para amostras de pequenas dimensões (Wisz et al. 2008) e também para amostras maiores e (3) é possível ajustar as definições que afetam a complexidade dos modelos (Elith et al. 2010; Shcheglovitova & Anderson 2013). Para saber mais como o programa funciona, as fórmulas matemáticas e outros princípios leia Merow et al. (2013).

7) ETAPAS PARA A CONTRUÇÃO DOS MODELOS:

Normalmente a construção dos MDE envolvem três etapas: **Pré-Análise**, **Modelagem** e **Pós-Análise**.

I) PRÉ-ANÁLISE: Compreende a seleção, organização e limpeza (extração de ruídos) dos pontos de ocorrência, além da seleção dos dados ambientais para a modelagem. Pode ser realizada manualmente ou por meio de *software*. As técnicas e o passo-a-passo para realizar essa etapa se encontram na parte 2 desse tutorial. Agora é importante conhecer os termos e os problemas mais comuns na realização dessa etapa:

a) Incerteza dos dados bióticos: Como os dados de presença de espécies são os únicos dados bióticos usados pelo MaxEnt, eles serão tratados sempre como exatos pelo algoritmo. No caso de MDE, isso nem sempre é verdade, dada a quantidade de ruídos que temos nos bancos de dados de distribuição de espécies. Portanto, cuidado extra deve ser tomado com a qualidade dos pontos ao se utilizar este algoritmo para modelagem. Deve-se levar isso em consideração ao analisar os resultados, principalmente quando os pontos não são muito confiáveis ou caso eles não representem bem a distribuição da espécie em questão. **Solução:** Remoção de dados imprecisos. Há ferramentas para remoção de dados redundantes (ou seja, que estão na mesma localidade ou próximo e que caem dentro do mesmo pixel de dados ambientais) que podem enviesar os modelos. Sempre é necessário conferir os pontos, as fontes, as localidades e usar pontos que não sejam sede de município. Dicas de como fazer isso estão na parte 2 do tutorial.

b) Viés Amostral ou Autocorrelação Espacial (*Sample bias*): Os pontos de ocorrência são muitos mais frequentes em áreas de fácil acesso ou com maior número de levantamento de fauna/flora. Isso significa que alguns lugares são mais propensos a serem vistoriados do que outros; tal viés é autocorrelacionado espacialmente (Reddy e Dávalos 2003). Esse viés, conhecido como *Viés de Seleção da Amostra*, *Viés de Levantamento* ou *Efeito Museu*, pode afetar seriamente a qualidade do modelo (Phillips et al. 2009). A premissa para a modelagem é que os pontos de ocorrência sejam coletados aleatoriamente por toda a extensão da distribuição da espécie e normalmente isso não é possível. Fazer uma modelagem com aglomerados de pontos (*Clusters*) em áreas superamostradas produz substancialmente valores mais altos de AUC. Além disso, outros trabalhos indicaram que os modelos feitos com viés amostral exibiram fortes sinais de sobreajuste (*overfitting*) (Veloz 2009). **Solução:** Aplicação de filtros espaciais e escolher os pontos de *background*. A escolha de um melhor *background* será tratado com detalhes adiante. Os filtros terão a função de eliminar pontos próximos e que tenham pouca variação nos valores das camadas ambientais, assim não há tanta perda de informação, nem há tendência de se dar mais peso a certos valores ambientais que se repetem. Há várias formas de se criar um filtro, as mais

recomentadas são a criação de uma malha de linhas, de extensão variável que formará um *grid*, e selecionar um ponto de ocorrência dentro de cada *grid*. Outra forma é a criação de *buffers* circulares em cada ponto de tamanho variável (alguns autores usam o tamanho do *home-range* das espécies-alvo) e posterior eliminação dos pontos que se sobrepõem no mesmo *buffer*. Existem outras formas que serão discutidas na parte 2 desse tutorial, mas não há consenso de qual dessas técnicas de filtro seja a melhor. O importante é justificar, de preferência com características ecológicas da espécie-alvo, do porquê usar um tipo de filtro e não outro.

c) Escala e Resolução: A escala refere-se ao tamanho da área da área do modelo (lembre-se da escala de um mapa), enquanto resolução é o tamanho da unidade de amostragem em que os dados são registrados, em outras palavras, qual o tamanho do pixel que seu modelo será construído (Austin 2007). MDEs construídos em grandes extensões espaciais muitas vezes dependem de dados de sensoriamento remoto de resolução grosseira ou com variáveis ambientais interpolados, criando vícios inerentes e problemas de amostragem. A escolha da escala também pode determinar se a distribuição da espécie inteira será incluída no modelo (Merow et al. 2014). Na parte 2 do tutorial voltamos a falar do assunto para a escolher a melhor resolução para as camadas ambientais.

d) Área de estudo/fundo (*Background*): é a área de estudo que se utiliza na modelagem. O tamanho dessa área (escala) é importante em algoritmos como o MaxEnt porque uma amostra aleatória de pontos é tomada a partir de toda a região de estudo (amostra *background*) para funcionar como pontos de pseudoausência e para a validação do modelo (Anderson & Raza 2010). Essa amostra do *background* será usada para caracterizar as condições ambientais disponíveis na região de estudo e para comparar com as condições ambientais onde a espécie-alvo está presente. Se a área de *background* for muito grande em relação à distribuição de pontos de presença (por exemplo, o *background* da América do Sul para uma espécie que ocorre só na Mata Atlântica), o MaxEnt usará pontos por exemplo do Andes, da Caatinga, e da Patagônia, para caracterizar a área ao redor da espécie e para fazer as comparações, aumentando o viés e o sobreajuste (*overfitting*). No default do programa o MaxEnt usa um prior, Q (X), que assume que a espécie é igualmente susceptível de estar em qualquer lugar na paisagem. Isto assume que cada pixel tem a mesma probabilidade de ser selecionado como ponto de *background*. Modificar a amostra de *background*, portanto, equivale a modificar as expectativas prévias para a distribuição das espécies. Os trabalhos publicados sugerem que a área de estudo não deve incluir áreas

onde a espécie está ausente devido a limitações de dispersão ou interações bióticas (especialmente competição). Isso ocorre porque pontos de *background* retirados de ambientes inadequados em tais regiões fornecerão um sinal falso negativo que interfere na modelagem. **Solução:** Escolher um *background* ecológica e geograficamente coerente com a espécie-alvo. Se é usado as configurações *default* do MaxEnt para o *background* (um prior uniforme para a espaço geográfico), a área de estudo deve conter apenas os locais onde a espécie tem a probabilidade de alcançar (dispersar). Os usuários podem especificar exatamente o número e a extensão espacial (*Bias Grid* ou *Bias File*) a partir do qual os pontos de *background* serão escolhidos para a criação dos modelos. Essas técnicas são mostradas na parte 2 do tutorial. A seleção da área de estudo deverá ser estritamente relevante para a ecologia das espécies e para o objetivo do estudo. Um *background* bem selecionado deve refletir o espaço geográfico acessível às espécies ao longo de um determinado período de tempo (Fourcade et al. 2014). Por exemplo, se alguém está interessado em determinar a melhor localização para uma Reserva Biológica para uma planta em extinção, o *background* deve ser delineado apenas com a distribuição geográfica da espécie conhecida Se alguém estivesse interessado em quais regiões com clima mediterrâneo no mundo essa planta poderia invadir na ausência de limitação de dispersão (hipoteticamente), o *background* deveria ser delineado contendo todas as áreas de clima mediterrâneo do mundo. Outro exemplo, se uma espécie só ocorre na Mata Atlântica, mas quero modelá-la para toda a América do Sul, tenho que reduzir o *background* para a área de ocorrência (só Mata Atlântica) e depois projetá-la para toda América do Sul, como no estudo de espécies invasoras, mudanças climáticas e paleomodelagem, levando em conta o princípio da transferibilidade. Normalmente o que alguns trabalhos fazem é usar apenas o *background* da América do Sul, o que superestima os valores de AUC e aumenta o sobreajuste dos modelos.

e) Transferibilidade (*Transferability*): refere-se à capacidade de um modelo ser transferido para um contexto diferente (por exemplo, para outro período de tempo após a mudança climática, ou para outra região em uma espécie invasora). Os modelos produzidos com uma região de estudo (*background*) muito grande ou com muita autocorrelação espacial são susceptíveis a ter baixa transmissibilidade. É um tema central para trabalhos com paleomodelagem, mudanças climáticas e espécies invasoras: controlar o sobreajuste e as autocorrelações ambientais e espaciais para ter modelos bons para as projeções.

d) Autocorrelação Ambiental: O projeto *WorldClim* usou dados de estações meteorológicas (geralmente de 1950 a 2000) interpolados usando uma técnica *splining* (placa fina de suavização) para produzir mapas de alta resolução (cerca de 1 km^2) de dados climáticos médios mensais. Estes dados mensais brutos foram processados em seguida para se obter 19 variáveis bioclimáticas que refletem diferentes aspectos da temperatura, da precipitação e da sazonalidade e que são importantes na determinação da distribuição das espécies (Hijmans et al., 2005; Anderson & Gonzalez 2011). Uma abordagem comum é usar todas as 19 variáveis do *WorldClim* (Hijmans et al., 2005), porque elas estão disponíveis em alta resolução e com escala global. No entanto, muitas vezes existem fortes correlações entre essas variáveis e criar modelos usando variáveis correlacionadas podem atrapalhar na transferibilidade, ou seja, o modelo pode se comportar de forma irregular quando eles são transferidos para um cenário em que as correlações são diferentes (passado, futuro ou outra área). **Solução:** Escolher as variáveis mais importantes para o conjunto de pontos. Ao invés de usar as 19 variáveis no modelo, deve se escolher as que estão menos correlacionadas entre sim (variáveis que não influenciam outra variável, que sejam independentes). Para isso alguns pesquisadores defendem a escolha de variáveis baseadas em relações conhecidas entre ambiente e fisiologia da espécie-alvo (por exemplo Kearney et al. 2008; Rodder et al. 2009). No entanto, como esses dados são frequentemente indisponível, métodos alternativos são muitas vezes necessários. A mais usada é a realização de uma Análise de Componentes Principais (PCA) para escolher as variáveis que mais contribuem para a distribuição da espécie e que são mais independentes (formam componentes diferentes). Remover as variáveis ambientais correlacionadas diminui a generalidade, criando modelos que tendem para erros de omissão em relação aos modelos construídos com um conjunto maior de variáveis (Warren et al. 2014). O próprio MaxEnt tem um procedimento de regularização (G-1 regularização) que equilibra o ajuste da complexidade do modelo com a correlação das variáveis ambientais, mas os estudos indicam que essa abordagem é insuficiente para evitar a colinearidade e sobreajustes causados por esse problema. Como realizar o PCA estão na parte 2 do tutorial.

II) MODELAGEM: Compreende a utilização de um algoritmo (no caso o MaxEnt) e os dados de *input* produzidos e trabalhados na etapa de pré-análise para a criação dos modelos. As técnicas e o passo-a-passo para realizar essa etapa se encontram na parte 2

desse tutorial. Agora é importante conhecer os termos e os problemas mais comuns na realização dessa etapa:

a) Treino (*Train*) e Teste (*Test*): para a criação de MDE são necessários dois conjuntos de dados bióticos (pontos de ocorrência) diferentes. Um conjunto será usado na parte de treino do modelo e outra no teste. Em resumo a etapa que o MaxEnt constrói os modelos é chamado de treino ou calibração (calibração/calibration é mais usado para designar os parâmetros do algoritmo para executar o modelo, qual botão apertar na interface do MaxEnt por exemplo, mas tem autores que usam calibração como sinônimo para treino), enquanto a etapa de teste ou avaliação acontece com o modelo já pronto. No teste o programa verifica qual a taxa de acerto do modelo gerado no treino, com outro conjunto de pontos de ocorrência. O método ideal para avaliar a precisão de um modelo é treinar/calibrar um modelo com um conjunto de dados e, em seguida, com testá-lo com um conjunto de dados independente. Existem pelo menos duas formas de se fazer isso: (a) coletar novos dados (trabalho de campo em áreas de alta adequabilidade ambiental indicadas pelo modelo) ou (b) dividir seu conjunto de dados em duas partes antes de modelar. Como a primeira é inviável financeira e metodologicamente, ainda mais para animais (o fato de não conseguir encontrar um espécie em uma área não significa que ela não está lá, pode ser problema na técnica de captura, isca, época do ano, fase da lua, etc.) a segunda opção é mais usada. Os MDE são criados usando os dados de treinamento/calibração e avaliados usando os dados de teste/avaliação. Os dados são geralmente separados em cada grupo aleatoriamente. Para a etapa de treino ou calibração se usa a maior parte dos pontos (70-80%) enquanto para o teste ou avaliação usa-se poucos pontos (30-20%). O MaxEnt faz essa divisão automaticamente. Antes de criar o modelo ele divide os registros de ocorrência em duas partes, uma vai para o treino e outra fica reservada para o teste. Os pontos utilizados em uma etapa nunca são utilizados na outra.

b) Replicações dos modelos: Como a modelagem trabalha com probabilidade, é necessária a criação de vários modelos para que depois o programa MaxEnt faça uma média de todas as replicações e gere um modelo final. O número de replicações varia, mas normalmente os trabalhos publicados trabalham com pelo menos 10 replicações. Há três algoritmos disponíveis para realizar as replicações: *Bootstrap*, *Crossvalidate* (correspondente ao *Jackknife*) e *Subsamples*. Essas técnicas são usadas para avaliar a precisão dos modelos de distribuição (Efron 1979). Basicamente, o que vai definir qual técnica aplicar é o número de dados disponíveis. *Subsamples* precisa de mais dados,

enquanto *Bootstrap* e *Crossvalidate* (*Jackknife*) podem ser feitas para poucos dados. **Bootstrap**: Esta técnica envolve a retirada aleatória de uma localidade no conjunto de treino e a duplicação de outra localidade do mesmo grupo para manter o mesmo número amostral. Ao fim da replicação, a localidade inicialmente retirada é reposta. Como há reposição, pode ser feita para amostras pequenas e médias (acima de 14 pontos de ocorrência). Contudo há presença de dados duplicados na construção de modelos. **Crossvalidate (Jackknife)**: os dados ocorrência são divididos aleatoriamente em uma série de grupos de tamanhos iguais chamadas "dobras" (*folds*) e os modelos são criados deixando de fora uma dobra de cada vez. As dobras deixadas de lado na análise são utilizados para o teste/avaliação do modelo. A avaliação cruzada tem uma grande vantagem sobre os outros: ele usa todos os dados para a fase de teste. O número de replicações será igual a n-1, onde n seria o número inicial de pontos de ocorrência. É mais indicado para modelos com poucos pontos, entre 8 e 13 pontos de ocorrência. **Subsamples**: os pontos de presença são divididos repetidamente em subconjuntos de treino e teste a cada replicação.

c) Formato de Saída (*Output*): É o tipo de arquivo que o MaxEnt vai gerar no fim da modelagem. **Formato logístico** (*logistic format*): é um dos formatos de saída do modelo que MaxEnt produz. Nesse formato cada pixel do mapa tem valores que variam de 0 a 1 e pode ser interpretada como a probabilidade de presença de condições ambientais adequadas para as espécies-alvo. (Veloz 2009). **Formato acumulativo** *(Cumulative format)***:** Nesse formato cada pixel do mapa tem valores que variam indefinidamente, os valores de probabilidade de cada replicação são somados e não transformada em escala logaritmo. Fica mais difícil a interpretação dos dados nesse formato.

d) Complexidade dos Modelos (Sobreajuste ou *Overfitting*): ocorre quando o modelo se ajusta demasiadamente bem aos dados de treino e, por conseguinte, não consegue prever com precisão os dados de teste. Ocorre quando os modelos são construídos com muitos parâmetros, o que acaba diminuindo os graus de liberdade e afetando o poder de previsão. De igual modo, modelos *underfitted* (aqueles que não incluem complexidade suficiente) não fornecem discriminação adequada e, portanto, preveem mal a distribuição. Os estudos sugerem que o *underfitting* é menos frequente do que o *overfitting*, pelo menos em técnicas como MaxEnt (Elith et al. 2006; Anderson & Gonzalez, 2011; Warren & Seifert, 2011). A ambos modelos, *overfitted* e *underfitted*, falta generalidade, o que dificulta os estudos que envolvem a transferência de modelo para outro período de tempo ou região ou que tenham como objetivo comparar nicho de espécies (Radosavljevic et al. 2014). Na

modelagem com apenas dados de presença, o viés de amostragem geográfica e pequenas amostras costumam ter maior sobreajuste. Por esta razão, MaxEnt usa um processo chamado L1 regularização para restringir distribuições dos modelos, para situar-se dentro de um certo intervalo em torno da média empírica, em vez de combinar exatamente com ela (Warren & Seifert 2011). Mas em alguns casos essa regularização automática não é suficiente para evitar o *overfitting*, necessitando aumentar a regularização manualmente. **Solução:** Aumentar/alterar os índices de regularização/complexidade dos modelos (*Turning*). Enquanto reduzir o número de variáveis autocorrelacionadas dos modelos leva a modelos mais restritos e a erros de omissão, o aumento a complexidade/regularização da modelagem leva a modelos amplos, com tendência a aumentar os erros de sobreprevisão (Warren et al. 2014). O aumento em demasia da complexidade do modelo tem um efeito negativo também, ele aumenta a amplitude do modelo, elevando os valores dos índices usados para a validação. Isso sugere que modelos menos complexos podem fornecer estimativas menos extremas dos efeitos das projeções climáticas futuras ou do passado sobre as espécies-alvo (Warren et al. 2014). Os resultados sugerem que as abordagens de complexidade/regularização na modelagem devem ser escolhidas com o objetivo do trabalho em mente. Modelos que favorecem os erros de sobreprevisão (ou seja, incluindo o viés de amostragem e aumentando regularização) pode ser uma boa saída quando o objetivo é identificar amplamente áreas onde uma espécie possa ocorrer. Os modelos inclinados para erros de omissão (ou seja, redução das variáveis e da autocorrelação, reduzindo a regularização) podem ser ideais quando o objetivo é achar áreas com altíssima probabilidade de ocorrência (Warren et al. 2014). Sugestões de valores para regularização e onde modificar esses valores no MaxEnt estão na parte 2 do tutorial.

III) PÓS-ANÁLISE: Compreende a avaliação do modelo gerado na etapa de modelagem. As técnicas e o passo-a-passo para realizar essa etapa se encontram na parte 2 desse tutorial. Agora é importante conhecer os termos e os problemas mais comuns na realização dessa etapa:

a) Validação do Modelo (*Validation*): A validação é etapa de avaliação, são os testes estatísticos para analisar se os modelos propostos são bons. O MaxEnt tem um forma própria de avaliar os modelos e dá o resultado como uma tabela no *output*, mas alguns revisores acreditam que outras formas de validação são necessárias, independentemente do MaxEnt. Há dois tipos de validação, a dependente de limite de corte (*threshold*) que são

índices baseados na matriz de confusão e a avaliação independente de limite de corte (*threshold*) que são áreas sob as curvas ROC (AUC – *area under curve*). O MaxEnt gera gráficos de AUC na sua avaliação e também gera algumas estatísticas com o *threshold*, mas não todas. Para mais detalhes veja abaixo sobre limites de corte, matriz de confusão e AUC.

b) Limites de corte (*thresholds*): o modelo gerado pelo MaxEnt é um mapa feito de pixels. Cada pixel terá um número que corresponderá à probabilidade da sua espécie-alvo ocorrer na área do pixel. O número normalmente varia de 0 a 1 quando é escolhido a construção do mapa logístico. Nos modelos há áreas de alta probabilidade 0.75-0.90 e também áreas de baixas probabilidades 0.10- 0.20. Para a etapa de validação por meio da matriz de confusão é necessário transformar essas probabilidades em presença ou ausência, em um mapa binário (o valor de cada pixel será transformado em 0 ou 1, indicando ausência ou presença da espécie-alvo respectivamente). O limite de corte é o valor da probabilidade que você adotará para transformar o modelo em mapa binário, isto é, abaixo desse valor de corte a espécie com certeza não ocorrerá, e acima desse valor com certeza ela ocorrerá. É uma medida subjetiva e por isso é muito criticada. Valores alto de limite de corte são mais conservadores, geram resultados mais restritos e valores baixos de limite de corte são mais abrangentes e pouco específicos. Uma boa revisão sobre limites de corte pode ser encontrada em (Liu et al. 2005). A escolha do valor está relacionada com o propósito da modelagem. 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, ou seja, modelos menos inclusivos. Isso requer um limite de corte alto, que selecione apenas as áreas com altos valores de adequabilidade ambiental para a ocorrência da espécie. Já quando o interesse é avaliar o potencial invasivo de espécies exóticas, ou aumentar o conhecimento do nicho de uma espécie pouco coletada, podemos adotar um limite de corte mais baixo para aumentar a área de interesse em relação à adequabilidade ambiental de ocorrência das espécies. Os limites de corte mais utilizados na literatura são: o limite de corte mínimo (*Minimum training presence logistic threshold*), o limite de corte de 10% (*10 percentile training presence logistic threshold*) e o limite de corte máximo (*Maximum test sensitivity plus specificity logistic threshold*), mas cada um tem vantagens e desvantagens de uso. A escolha deve sempre estar alinhada com o que se pretende fazer com o modelo gerado, ou seja, depende da pergunta. Os valores mais recomendados que mostram modelos mais robustos são os limites máximo e médio.

c) Matriz de Confusão: Todo modelo apresenta erros e acertos que são avaliados em conjunto para determinarmos a qualidade do mesmo. A matriz de confusão é um esquema que reúne as possíveis formas de acerto e erro em relação ao que o modelo previu e a distribuição "real" da espécie na natureza, é uma forma de avaliar se o modelo é bom. São utilizados dados de presença e ausência para a construção da matriz. Quando não se tem dados de ausência, pode-se gerar pontos aleatoriamente na área de estudo (*background*) ou a partir de pontos de pseudoausência. Para criar a matriz de confusão é necessário definir um limite de corte (*threshold*) para o modelo produzido no MaxEnt. Por isso é dito que a matriz de confusão é uma forma de validação de modelos dependente do limite de corte (*threshold-dependent*).

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)

Como calcular o valores de a, b, c, d da matriz de confusão será mostrado na parte 2 do tutorial. Na matriz de confusão pode-se calcular os seguintes índices que são usados na validação: sensibilidade (*sensitivity*) (se o modelo detecta bem a presença da espécie-alvo), especificidade (*specificity*) (se o modelo detecta bem a ausência da espécie-alvo), acurácia (taxa de acertos quando comparadas com todas as tentativas), sobreprevisão (falso positivo ou erro-tipo I) e omissão (falso negativo ou erro-tipo II). Desses índices, os dois últimos são os principais e precisam de maiores explicações e o MaxEnt gera a taxa de erro de omissão para cada limite de corte.

d) Erro de omissão (*omission error*): É quando o modelo não previu (omitiu) a existência de uma localidade em que a espécie-alvo verdadeiramente ocorre, que está naqueles 30% de localidades separadas para a avaliação/teste do modelo (Fig. 1). No geral,

erros de omissão podem ser considerados erros verdadeiros (erro grave, tipo II) representado pela letra c na matriz de confusão. Contudo, sob algumas circunstâncias, um registro de presença pode não ser muito confiável, devido a pelo menos três situações: (a) a identificação da espécie pode estar errada; (b) o georreferenciamento de alguns pontos pode estar errado; (d) a localização de um indivíduo pode estar fora do seu habitat usual (indivíduos em trânsito ou introduzidos). Em todas as três situações, estes pontos podem representar um *outlier* para o algoritmo, ou seja, pontos com informação ambiental muito fora do padrão. Nessas circunstâncias, um erro de omissão não seria um erro, e sim uma forma do algoritmo dar menos "importância" para pontos "ruins", fora do padrão. Uma consequência direta disso, 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 há uma predominância de trabalhos utilizando 5% de taxa de omissão (Elith et al. 2006, Phillips et al. 2006). Os erros de omissão atualmente (2010-2015) são aceito em até 15-20%, se os demais índices estatísticos também forem bons.

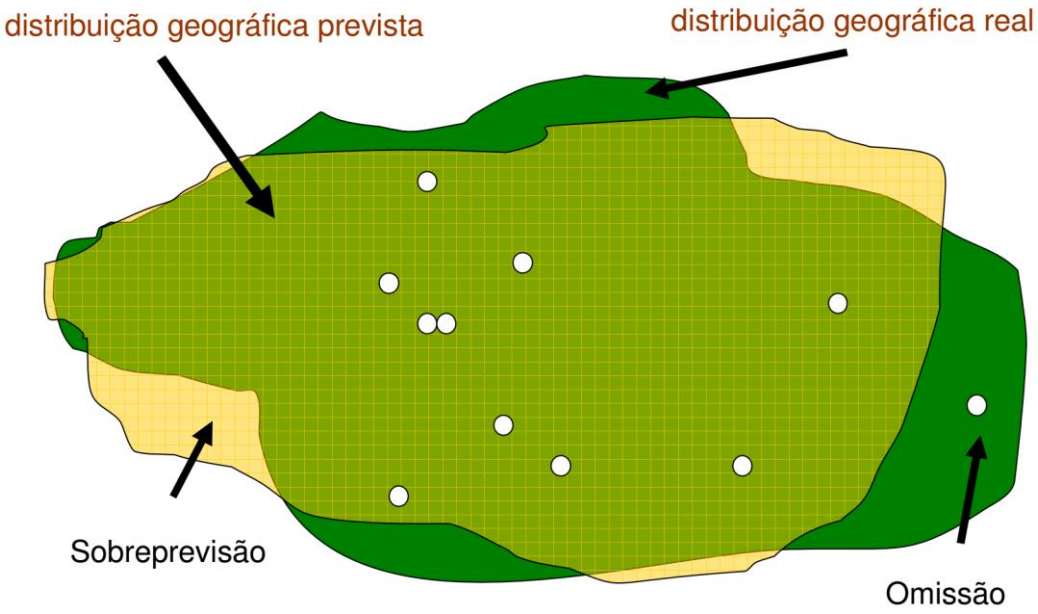


Figura 1: representação dos erros de omissão e sobreprevisão (comissão) do modelo.

e) Erro de sobreprevisão (*commission error*): é quando o modelo prevê uma região/localidade adequada para a espécie, mas não há registro dela no seu conjunto de

pontos para a validação/teste (Fig. 1). Os limites de corte baixos aumentam esse tipo de erro. A sobreprevisão pode ou não ser um erro. A previsão de um modelo em uma área na qual não se tem registro de ocorrência da espécie, representada pela letra b na matriz de confusão, pode ser causada por diferentes situações: (a) 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; (b) 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; (c) a área é inabitável mesmo, o que seria o verdadeiro erro de sobreprevisão.

f) Área sob a curva (AUC): A AUC é calculada com dados de teste da validação do MaxEnt e representa um índice independente de limite de corte que avalia a capacidade discriminatória de um modelo (Phillips et al. 2006). Alguns estudos têm questionado o uso da AUC em modelos construídos com algoritmos de presença-*background* (como o MaxEnt) (Lobo et al. 2008; Warren & Seifert 2011), mas ele fornece informações válidas e úteis em algumas circunstâncias (veja esclarecimentos em Peterson et al. (2011)). Por exemplo, a AUC é relevante e apropriado para comparações entre as configurações dos programas, usando uma espécie única e uma única região (Shcheglovitova & Anderson 2013). O uso da AUC é questionado por algumas razões: (1) ele ignora os valores de probabilidade prevista e sempre avalia com um bom ajuste também os modelos ruins; (2) pesa os erros de omissão e de sobreprevisão de forma igual; (3) ele não dá informações sobre os erros do modelo; e, mais importante ainda, (4) a extensão total da área de estudo dos modelos influenciam fortemente a taxa de ausências e causam valores maiores de AUC em áreas de estudo maiores (Lobo et al. 2008). Apesar dos problemas, a AUC ainda é utilizada principalmente na comparação entre algoritmos. Os valores da AUC variam entre 0 e 1, com a máxima precisão conseguida com valores de 1, e precisão não melhor do que o acaso com valores abaixo de 0.7. Quando utilizado com dados somente de presença ou presença-background, os valores máximos, teoricamente, deveriam ser <1 (Phillips et al. 2006).

O cálculo da área sob a curva (AUC) fornece uma medida única do desempenho do modelo, independente da escolha prévia de qualquer limite de corte, pois a curva é construída a partir de vários limites de corte (cada limite de corte é responsável por um ponto da curva). Este valor (AUC) mede a capacidade discriminatória do modelo, nos permitindo interpretar seu resultado como a probabilidade de que ao sortearmos dois pontos, um do conjunto de presença e outro do conjunto de ausência, o modelo consiga

prever os dois corretamente. Uma das melhores fontes de explicação sobre as curvas ROC pode ser encontrada no relatório de Fawcett (2003).

A curva ROC é obtida plotando-se a sensibilidade no eixo y e o valor 1-especificidade no eixo x para todos os possíveis limites de corte. A sensibilidade também é conhecida como a taxa de verdadeiros positivos, e representa ausência de erro de omissão. Já a especificidade, ou a variante 1-especificidade, também é conhecida como a taxa de falso positivo, e representa o erro de sobreprevisão. A área abaixo da curva (AUC) é normalmente determinada conectando os pontos com linhas diretas e o valor da área é calculado pelo método de trapezoide. 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. 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 (aceitável para a publicação acima de 0.75);

0,60 - 0,70 = Ruim;

0,50 - 0,60 = Muito ruim;

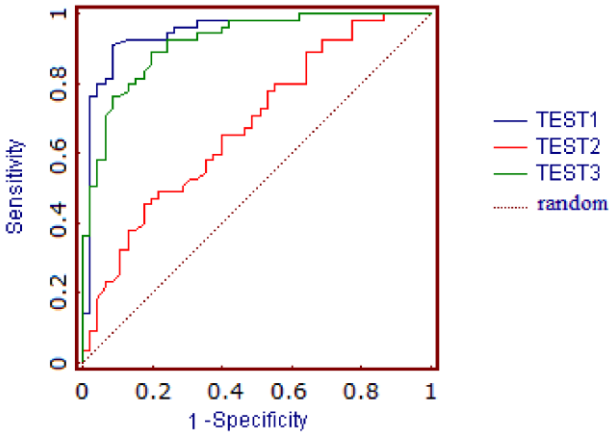


Figura 2: Exemplo de curva ROC gerada por dados de presença para três espécies. A interpretação é intuitiva, o Teste 1 foi melhor que o Teste 3 e que o Teste 2, e o Teste 3 foi melhor que o Teste 2.

8) REFERÊNCIAS:

Dalapicolla, J. 2016. Tutorial de modelos de distribuição de espécie: guia teórico. Laboratório de Mastozoologia e Biogeografia, Universidade Federal do Espírito Santo, Vitória. Disponível em: http://blog.ufes.br/lamab/tutoriais

Alvarado-Serrano, D. F., & Knowles, L. L. (2014). Ecological niche models in phylogeographic studies: applications, advances and precautions. Molecular Ecology Resources, 14(2), 233-248.

Anderson, R. P., & Gonzalez, I. (2011). Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. Ecological Modelling, 222(15), 2796-2811.

Anderson, R. P., & Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. Journal of Biogeography, 37(7), 1378-1393.

Araújo, M. B., & Pearson, R. G. (2005). Equilibrium of species' distributions with climate. Ecography, 28(5), 693-695.

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

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.

Austin, Mike. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecological modelling, v. 200, n. 1, p. 1-19.

Busby, J. R. 1986. A biogeographical analysis of Nothofagus cunninghamii (Hook.) Oerst. In southeastern Australia. Australian Journal of Ecology 11:1-7.

Carnaval, A. C., & Moritz, C. (2008). Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. Journal of Biogeography, 35(7), 1187-1201.

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.

Chan, L. M., Brown, J. L., & Yoder, A. D. (2011). Integrating statistical genetic and geospatial methods brings new power to phylogeography. Molecular phylogenetics and evolution, 59(2), 523-537.

Dalapicolla. J. Papel da hidrografia e do clima na estrutura genética do roedor semiaquático Nectomys squamipes. Dissertação de Mestrado. Universidade Federal do Espírito Santo. 2014.

de Souza Muñoz, M. E., De Giovanni, R., de Siqueira, M. F., Sutton, T., Brewer, P., Pereira, R. S., ... & Canhos, V. P. (2011). openModeller: a generic approach to species' potential distribution modelling. GeoInformatica, 15(1), 111-135.

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. ScachettiPereira, 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.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. Methods in ecology and evolution, 1(4), 330-342.

Elton, C. S. 1927. Animal Ecology. Sidgwich and Jackson, London.

Fawcett, T. 2003. ROC graphs: notes and practical considerations for data mining researchers. Palo Alto, CA: HP Laboratories.

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.

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias.

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

Giannini TC, Siqueira MF, Acosta AL, Barreto FCC, Saraiva AM, Alves-dos-Santos I (2012) Desafios atuais da modelagem preditiva de distribuição de espécies. Rodriguésia-Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, 63.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., & Loiselle, B. A. (2008). The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45(1), 239-247.

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

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

Guo QH, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of sudden oak death in California. Ecological Modelling, 182, 75–90.

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

Hirzel AH, Hausser J, Chessel D, Perrin N (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology, 83, 2027–2036.

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

Jiménez-Valverde A, Lobo JM, Hortal J (2008) Not as good as they seem: the importance of concepts in species distribution modeling. Diversity and Distributions, 14, 885-890.

Kearney, M., Phillips, B.L., Tracy, C.R., Christian, K.A., Betts, G. & Porter, W.P. (2008) Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. Ecography, 31, 423–434.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., & Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and Distributions, 19(11), 1366-1379.

Lamb, J. M., Ralph, T. M., Goodman, S. M., Bogdanowicz, W., Fahr, J., Gajewska, M., ... & Taylor, P. J. (2008). Phylogeography and predicted distribution of African-Arabian and Malagasy populations of giant mastiff bats, Otomops spp.(Chiroptera: Molossidae). Acta Chiropterologica, 10(1), 21-40.

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. Ecography, 28(3), 385-393.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), 145-151.

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography, 36(10), 1058- 1069.

Merow, Cory et al. What do we gain from simplicity versus complexity in species distribution models? Ecography, v. 37, n. 12, p. 1267-1281, 2014.

Metz, C. E. 1986. ROC methodology in radiologic imaging. Investigative Radiolgy 21:720-733. Pearson RG, Dawson TE (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography, 12, 361–371.

Peterson, A. T., M. Papes, and J. Soberón. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological Modelling 213:63-72.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B., 2011. Ecological niches and geographic distributions. Monographs in Population Biology, Princeton University Press, Princeton, NJ.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31(2), 161-175.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications, 19(1), 181-197.

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

Phillips, SJ (2008) Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). Ecography, 31, 272-278.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. Journal of biogeography, 41(4), 629-643.

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography, 30(11), 1719-1727.

Rodder, D., Schmidtlein, S., Veith, M. & L otters, S. (2009) Alien invasive slider turtle in unpredicted habitat: a matter of niche shift or of predictors studied? PLoS ONE, 4, e7843.

Royle, J.A., Chandler, R.B., Yackulic, C. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods in Ecology and Evolution, 3, 545–554.

Rushton, S. P., S. J. Ormerod, and G. Kerby. 2004. New paradigms for modelling species distributions? Journal of Applied Ecology 41:193-200.

Shcheglovitova, M., & Anderson, R. P. (2013). Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. Ecological Modelling, 269, 9-17.

Soberón, J; Nakamura, M. Niches and distributional areas: concepts, methods, and assumptions. Proceedings of the National Academy of Sciences, v. 106, n. Supplement 2, p. 19644- 19650, 2009.

Soberón J (2007) Grinnellian and Eltonian niches and geographic distributions of species. Ecology letters, 10, 1115-1123.

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., and D. Peters. 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. International Journal of Geographic Information Systems 13:143-158.

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography, 36(12), 2290-2299.

Walker PA, Cocks KD (1991) HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. Global Ecology and Biogeography Letters, 1, 108–118.

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecological Applications, 21(2), 335-342.

Warren, D. L., Wright, A. N., Seifert, S. N., & Shaffer, H. B. (2014). Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. Diversity and distributions, 20(3), 334-343.

Wiens JJ, Graham CH (2005) Niche conservatism: integrating evolution, ecology, and conservation biology. Annual Review of Ecology, Evolution, and Systematics, 36, 519-539.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., & Guisan, A. (2008). Effects of sample size on the performance of species distribution models. Diversity and Distributions, 14(5), 763-773.

Yee, T. W., and N. D. Mitchell. 1991. Generalized additive models in plant ecology. Journal of Vegetation Science 2:587-602.