Oecologia Australis
22(2): 117–129, 2018
10.4257/oeco.2018.2202.02

SELEÇÃO DAS VARIÁVEIS PREDITORAS PARA MODELAGEM CORRELATIVA DE DISTRIBUIÇÃO DE ESPÉCIES NA AMÉRICA DO SUL

Débora Samira Gongora Negrão[1] *& Peter Löwenberg-Neto*[2]*

[1] Universidade de São Paulo, Instituto de Biociências, Pós-Graduação em Ecologia, Rua do Matão, 321, Travessa 14, CEP 05508-090, São Paulo, SP, Brasil.

[2] Universidade Federal da Integração Latino-Americana, Pós-Graduação em Biodiversidade Neotropical, Laboratório de Biogeografia e Macroecologia, Av. Tarquínio Joslin dos Santos, 1000, CEP 85870-901, Foz do Iguaçu, PR, Brasil.

E-mails: deborasamira@gmail.com; peter.lowenberg@unila.edu.br (*autor correspondente)

Resumo: Modelagem de distribuição geográfica é uma técnica utilizada para estimar a área de distribuição de espécies no espaço geográfico. A etapa de seleção das variáveis preditoras é fundamental para a construção do modelo conceitual no procedimento de modelagem correlativa. O presente estudo teve por objetivo descrever como os pesquisadores selecionaram as variáveis preditoras em seus estudos conduzidos na extensão geográfica da América do Sul. Foi realizado levantamento e revisão da literatura de estudos publicados no período de 2002 a 2014. Os artigos foram triados em duas categorias: seleção arbitrária e seleção baseada em critério. Os artigos que utilizaram critérios foram subcategorizados em quatro grupos: axiomático, biológico, estatístico ou metodológico. Do total de artigos analisados (n = 177), em 31% os autores não justificaram a seleção das variáveis. Em 24% dos artigos os autores combinaram critérios, o critério estatístico foi adotado de modo não combinado em 23% dos artigos e o biológico em 13% dos artigos. Quando os autores utilizaram algum critério, o conjunto apresentou menor número de variáveis. Os autores que adotaram o critério biológico apresentaram explicações imprecisas da relação entre as variáveis e os organismos. Pela primeira vez foi diagnosticado o procedimento de combinar critérios e este interpretado como solução para a) balancear o número de variáveis do conjunto e b) contornar o déficit de conhecimento das espécies. A maior frequência de utilização do critério estatístico, isolado e combinado, corroborou o esperado para a modelagem correlativa. Admite-se como prática adequada a seleção de variáveis utilizando princípios e análises estatísticas para a construção de conjuntos com seis a nove variáveis preditoras. A adoção plena do critério biológico necessita de estudos que descrevam com maior rigor teórico a relação entre as variáveis e as respostas ecofisiológicas dos organismos.

Palavras-chave: camada matricial; espaço geográfico; fenomenologia; modelo empírico; sobreajuste.

SELECTING PREDICTORS IN CORRELATIVE SPECIES DISTRIBUTION MODELING IN SOUTH AMERICA. Species distribution modelling is a technique that estimates species geographical ranges. Variable selection is a fundamental step in conceptual model formulation and in correlative modelling procedure. Herein we describe how authors selected variables in their modelling studies within the extent of the South American continent. We analysed 177 papers that were published between 2002 and 2014. For each paper, we searched for explicit information on authors' reasoning for selecting variables. We classified papers into two categories: arbitrary and criterion-based. Criterion-based papers were

subclassified into axiomatic, biological, methodological, or statistical categories. From all papers analysed (n = 177), authors did not justify or comment on their variable sets in 31% of the papers. For 24% of the papers authors combined criteria; statistical criterion alone was adopted in 23% of the papers; and biological criterion was adopted in 13% of the papers. We observed that when authors selected variables based on a criterion, it decreased the number of variables. Authors that adopted biological criterion showed unprecise explanations on the relation between variables and organisms. We highlighted the emergent procedure of combining criteria which was interpreted as a way a) to balance the number of variables and avoid overfitting and overestimated predictions and b) to overcome the deficit on biological knowledge about species. The statistical criterion was more frequently adopted, and this fact corroborates our expectation based on the empirical nature of the correlative approach. For a sounding use of the biological criterion it is necessary a more theoretical and rigorous description of the relationship between variables and ecophysiological responses of the organisms

Keywords: empirical model; geographic space; overfitting; phenomenology; raster data.

INTRODUÇÃO

A modelagem de distribuição geográfica de espécies é um conjunto de procedimentos que faz uso da informação dos organismos e do ambiente para estimar áreas onde as espécies têm maior ou menor potencial de ocorrer (Elith & Leathwick 2009). O produto da modelagem é especialmente importante no contexto da biogeografia da conservação (De Marco Jr. & Siqueira 2009), pois auxilia a preencher lacunas de conhecimento sobre a ocorrência geográfica das espécies que é fundamental nas estratégias de manejo e conservação da biodiversidade (Löwenberg-Neto & Loyola 2016). Ademais, a crescente oferta de variáveis ambientais que remontam ao passado (*e.g.*, 6.000, 21.000 anos atrás) e ao futuro (*e.g.*, 2050, 2070), permite executar transferências temporais e estimar respostas espaciais dos organismos frente às mudanças climáticas (Faleiro *et al.* 2013, Diniz-Filho *et al.* 2016).

Conforme o propósito do modelo, a modelagem da distribuição das espécies pode ser conduzida sob duas abordagens: mecanicista ou correlativa. Na abordagem mecanicista os requerimentos morfológicos, fisiológicos e comportamentais são obtidos experimentalmente (Kearney 2006) e, então, vinculados às variáveis ambientais para estimar a distribuição geográfica das espécies (Kearney & Porter 2004, 2009). Ela exige a formulação de modelos conceituais baseados em processos fisiológicos, que são mais realísticos e generalistas, e suas predições são avaliadas por seu rigor teórico das relações de causa e efeito (Guisan & Zimmermann 2000). Esta abordagem é promissora, no entanto, ainda permanece em estágios iniciais de desenvolvimento, principalmente porque exige conhecimento mais detalhado da relação entre a aptidão biológica (*fitness*) das espécies com o meio ambiente (Kearney 2006, Buckley *et al.* 2010).

Na abordagem correlativa as condições ambientais das espécies são estimadas pela intersecção espacial entre os pontos de ocorrência e os valores das variáveis ambientais (Elith & Leathwick 2009). Esta é a abordagem mais utilizada, pois não necessita do conhecimento prévio do nicho fundamental das espécies (Kearney 2006) e há grande oferta de dados primários de ocorrência de espécies em bancos de dados (*e.g.*, SpeciesLink, GBIF), o que facilita a execução do procedimento de modelagem (Pearson 2010). Na abordagem correlativa o modelo conceitual é empírico e baseado no fenômeno observado na natureza. O modelo empírico não tem o compromisso de descrever a relação de causa-efeito e nem de explorar os mecanismos ecofisiológicos que fundamentam a ocorrência geográfica como na abordagem mecanicista e sim de apresentar estatisticamente uma aproximação da realidade (Guisan & Zimmermann 2000).

A modelagem correlativa segue o seguinte procedimento (Elith & Leathwick 2009, Franklin 2009, Peterson *et al.* 2011): preparação dos dados, modelagem do nicho, projeção do modelo e avaliação, e transferabilidade do modelo, caso haja transferência espacial (*e.g.*, espécies invaso-

ras) ou temporal (*e.g.*, mudança climática). A etapa de preparação é quando os dados de presença das espécies, que correspondem aos pontos de ocorrência geográfica, são obtidos e organizados e as variáveis, em formato matricial (*e.g.*, raster), são selecionadas para compor o conjunto de variáveis preditoras do modelo. Os pontos de ocorrência geográfica da espécie são interpolados com os pixels das variáveis matriciais e, para cada ponto, é obtido um vetor de valores. A precisão nos dados de entrada e do conjunto das variáveis preditoras são fundamentais para o bom desempenho dos modelos, pois eles são os dados primários que influenciam no nível de ajuste entre o observado e o estimado (Austin 2002, Araújo & Guisan 2006, Bradie & Leung 2017).

As variáveis representam a variação gradual de um parâmetro associada à disposição espacial em uma dada área de estudo. Elas podem ser classificadas em três modos:

1) Variáveis cenopoéticas e bionômicas (Hutchinson 1980) – As variáveis cenopoéticas consistem naquelas variáveis abióticas que compõem o ambiente e estão disponíveis de forma equivalente para todos os organismos; e variáveis bionômicas são as variáveis dinâmicamente ligadas à população que podem tanto ser consumidas quanto modificadas pela presença ou densidade dos indivíduos.
2) Variáveis indiretas, diretas ou recurso (Austin 2002) – Nesta classificação o foco é na relação da variável com efeitos fisiológicos no organismo, podendo ser indireta, quando o efeito é correlativo com outras variáveis (*e.g.*, elevação, latitude), direta, quando afeta primariamente o funcionamento do organismo (*e.g.*, temperatura, pH), e recurso, quando a variável é consumível pelo organismo (*e.g.*, água, luz, nutrientes).
3) Variáveis proximais e distais (Austin 2002) – Nesta classificação o foco é quanto à posição da variável preditora na cadeia do processo que relaciona a resposta do organismo e as variáveis ambientais. Uma variável é dita proximal quando a resposta do organismo é consequência direta e objetiva da variável (*e.g.*, concentração de fosfato solúvel na raiz da planta) e distal quando a resposta do organismo é consequência de uma cadeia de múltiplas ligações causais (*e.g.*, variáveis indiretas).

Ressalta-se aqui que as classificações apresentadas não são mutuamente excludentes.

Uma forma de selecionar as variáveis é verificar quais delas melhor representam ou descrevem a ocorrência dos pontos de entrada observados. Deste modo, é possível utilizar-se de princípios estatísticos e de análises estatísticas para montar o conjunto de preditores, por exemplo, variáveis que não sejam colineares, ranquear e selecionar as variáveis que melhor descrevem a variação do conjunto (*e.g.*, Petitpierre *et al.* 2017). Outra forma de selecionar o conjunto de preditores é elencar variáveis que façam sentido fisiológico ou ecológico para a espécie. Neste contexto, é comum observar o uso de variáveis de médias de temperatura e precipitação, para contemplar aspectos biológicos gerais, em conjunto com variáveis de amplitude, variação (*e.g.*, sazonalidade) e variáveis derivadas (*e.g.*, precipitação no trimestre mais quente) para contemplar atributos mais específicos e limitações fisiológicas (*e.g.*, Vasconcelos 2014). Entretanto, as informações precisas sobre os atributos ecológicos das espécies são escassas e não há conhecimento em que escala geográfica ocorre a interação (Austin & Van Niel 2011). A seleção das variáveis acaba ocorrendo pela percepção subjetiva do pesquisador sobre a relação entre os atributos gerais (*e.g.*, ectotérmicos ocorrem em áreas quentes) e a natureza da variável ambiental (*e.g.*, climática, topográfica) (Austin 2007). Ademais, a seleção de variáveis conduzida por especialistas nos grupos taxonômicos demonstrou ser ineficaz e não melhorou a capacidade preditiva nem a transferabilidade dos modelos quando comparada com a seleção por critérios estatísticos (Seoane *et al.* 2005).

Os modelos de distribuição de espécies são executados em um contexto multidimensional, ou seja, pode-se utilizar um grande conjunto de variáveis (Seoane & Bustamante 2001), o que é esperado quando não existe conhecimento dos fatores que afetam a distribuição das espécies. A recomendação é de não utilizar muitas variáveis, pois tende a gerar um sobreajuste na estimativa geográfica (Peterson *et al.* 2011), fazendo com que as distribuições das espécies sejam sub-representadas, particularmente quando o número de variáveis é maior ou igual ao número de observações (Guisan & Zimmermann 2000). Por

outro lado, um conjunto com poucas variáveis tende a aumentar o risco de não caracterizar adequadamente o habitat ou o nicho da espécie e estimar amplitude geográfica excessivamente grande da espécie (Beaumont *et al.* 2005, Peterson *et al.* 2011). Uma alternativa para usar menos variáveis mantendo a caracterização do nicho da espécie é usar autovetores como variáveis preditoras (Cruz-Cárdenas *et al.* 2014). Neste procedimento, o conjunto de variáveis tem suas dimensões reduzidas em componentes ortogonais (*e.g.*, Silva *et al.* 2014) que representam a maior porção de informação dos dados. A desvantagem em proceder desta forma é que a "supervariável" é apenas uma transformação dos dados do conjunto inicial e, portanto, a questão da escolha das variáveis inicialmente permanece em aberto.

O modo ou os critérios utilizados pelos autores para selecionar as variáveis ambientais e construir o conjunto de preditores foram pouco discutidas à luz do procedimento de modelagem. Sobre a qualidade dos dados de ocorrência das espécies há ampla discussão sobre o efeito do tamanho amostral dos pontos (Stockwell & Peterson 2002, Hernandez *et al.* 2006), utilização de dados de presença somente (Tsoar *et al.* 2007), autocorrelação espacial (Dormann *et al.* 2007), completude dos banco de dados (Phillips *et al.* 2009, Newbold 2010) e sobre os vieses de amostragem e modelagem com dados de presença e/ou ausência (Elith & Leathwick 2009, Lobo *et al.* 2010). Ademais, houve profícua discussão sobre o desempenho de diferentes algoritmos para modelagem (Elith & Graham 2009), métodos de consenso (Marmion *et al.* 2009), parâmetros e interpretações para avaliar a acurácia dos modelos (Liu *et al.* 2005, Lobo *et al.* 2010, Hijmans 2012). No que se refere às variáveis ambientais, a discussão recai sobre a influência da resolução das variáveis na estimativa das distribuições (Guisan *et al.* 2007, Austin & Van Niel 2011), na performance de diferentes conjuntos de variáveis preditoras na transferabilidade espacial (Seoane *et al.* 2005, Peterson & Nakazawa 2008, Petitpierre *et al.* 2017) e na importância das variáveis usadas, quantificadas em um contexto pós-algoritmo (Bradie & Leung 2017).

Dentre os temas relacionados às variáveis ambientais, permanece desconhecido qual é a lógica ou qual é o critério que os autores têm adotado para selecionar as variáveis na composição do conjunto de preditores para o procedimento de modelagem correlativa. O presente estudo teve por objetivo descrever como os autores de estudos que modelaram a distribuição geográfica de espécies selecionaram as variáveis. A descrição foi feita a partir da revisão sistemática da literatura para estudos que modelaram as distribuições dentro dos limites da América do Sul. Espera-se que os autores tenham adotado com maior frequência o critério estatístico, já que a abordagem correlativa está relacionada com o modelo conceitual empírico, no qual se busca a melhor aproximação estatística do fenômeno observado na natureza (Guisan & Zimmermann 2000).

MATERIAL E MÉTODOS

Extensão geográfica do estudo

A América do Sul foi adotada como extensão geográfica do estudo por dois motivos: 1) possibilidade de caracterizar o perfil dos pesquisadores que realizaram estudos na área; e 2) possibilidade de controlar por peculiaridades geográficas da área. A América do Sul compreende um continente de forma aproximadamente triangular com grande amplitude latitudinal e margeada ocidentalmente pela Cordilheira dos Andes. Por consequência, o continente alberga combinações climáticas típicas (Peel *et al.* 2007), como, por exemplo, formações de clima temperado e de clima árido em baixas latitudes (*e.g.*, Florestas de Magdalena e Caatinga), faixa desértica na porção central da costa do Pacífico e faixa úmida adjacente (*e.g.*, Deserto do Atacama e Floresta Valdiviana). Além das combinações climáticas típicas, a América do Sul apresenta uma estruturação espacial singular de constância relativa de habitats (Walter & Breckle 1985) e das condições contemporâneas análogas (Löwenberg-Neto 2018). Deste modo, acredita-se que o conjunto de características geográficas possa ter influenciado na escolha de variáveis, tanto para os pesquisadores que selecionaram variáveis com sentido biológico quanto para os que lançaram mão de análises estatísticas.

Obtenção da literatura e artigos analisados

Os artigos científicos foram listados a partir do

sítio eletrônico especializado *Thomson Reuters Web of Science database* (https://www.webofknowledge.com/) (*e.g.*, Alexandre *et al.* 2013) em janeiro de 2015. As buscas foram realizadas utilizando o campo "tópico" e combinando três palavras-chave: "*species distribution model** AND *South America*" e "*ecological niche model** AND *South America*". Existe um debate na literatura sobre o termo correto do procedimento de modelagem e duas terminologias são empregadas: modelagem de distribuição de espécies (*species distribution modelling* - SDM) (Elith & Leathwick 2009) e modelagem do nicho ecológico (*ecological niche modelling* - ENM) (Soberón & Peterson 2005). A diferença prática está relacionada com a porção do diagrama BAM que é estimada. Ou seja, os trabalhos que buscam obter a área de distribuição potencial ou a área abioticamente adequada seriam classificados como ENM, e aqueles que buscam obter a área de distribuição atual ocupada como SDM (Peterson 2006, Peterson & Soberón 2012). Ademais, há os estudos que buscam estimar o nicho fundamental das espécies por meio da abordagem mecanicista (*e.g.*, Kearney & Porter 2004). Foi defendido que as abordagens mecanicistas são as que mais se aproximam do nicho da espécie e que, portanto, as abordagens correlativas deveriam fazer o uso do termo "habitat" e não "nicho" na modelagem (Kearney 2006). Na busca por artigos, as palavras-chave foram utilizadas em conjunto, pois apesar de haver fortes argumentos para a distinção conceitual (Sillero 2011, Peterson & Soberón 2012) tal procedimento é ainda recente ao tomar como referência a abrangência temporal da presente análise (2002 até 2015) e, na prática, os autores têm recorrentemente utilizado os termos como sinônimos (*e.g.*, Fitzpatrick *et al.* 2007, Elith & Leathwick 2009, Raes 2012, Barbosa & Schneck 2015). No corpo textual foi utilizado o termo "modelagem de distribuição de espécies" (*sensu* Franklin 2009) ao invés de "modelagem de nicho ecológico" por ser mais inclusivo e, por consequência, retratar melhor a heterogeneidade de abordagens apresentada nos trabalhos. Dos artigos obtidos na busca foram excluídos os artigos que usaram a abordagem mecanicista (n = 3), artigos cuja extensão geográfica de estudo era fora dos limites da América do Sul e artigos de abordagem teórica sem ensaio com dados empíricos.

Categorização dos artigos

A leitura do corpo textual dos artigos foi realizada em duas etapas, tendo por tarefa inicial obter informações sobre a lógica ou procedimento adotado pelos autores para a construção do conjunto de variáveis preditoras. Nesta etapa os artigos foram triados em duas categorias: seleção arbitrária das variáveis, quando os autores não informaram no texto o critério de escolha das variáveis ou quando eles não apresentavam nenhuma justificativa ou comentário sobre o conjunto de preditores; e seleção de variáveis baseada em critério, quando os autores explicaram a escolha das variáveis. Na segunda etapa os artigos cujos autores selecionaram as variáveis baseando-se em critérios foram triados em quatro subcategorias: critério axiomático, quando os autores justificaram que as variáveis foram utilizadas em estudos anteriores ou similares; critério biológico, quando os autores justificavam uma relação causal entre as variáveis ambientais e aspectos fisiológicos e/ou ecológicos da espécie; critério estatístico, quando os autores utilizaram princípios ou análises estatísticas para a escolha das variáveis; e critério metodológico, quando os autores justificaram a escolha das variáveis por condicionantes ou limitações no procedimento de modelagem. A quantificação das frequências foi realizada tomando-se como unidade amostral o artigo. No entanto, em artigos cujos autores tenham executado mais de um modelo com diferentes critérios e/ou subcritérios, a unidade contada foi a de modelos. Portanto é possível que um artigo tenha sido quantificado em mais de um critério ou subcritério.

Descrição e análises estatísticas

Foram confeccionados histogramas para descrever a frequência de quantidade de variáveis e a frequência de critérios adotados nos estudos. O valor da mediana do número de variáveis selecionadas foi usado como parâmetro para dividir os dados em dois grupos para verificar se havia associação entre usar mais ou menos variáveis e o tipo de seleção das variáveis (arbitrária *vs.* criteriosa). Neste caso foi executado

um teste de qui-quadrado de Pearson com correção de continuidade de Yates. Uma análise de correlação de Pearson foi executada para verificar se havia relação entre o número de variáveis selecionadas e o número de pontos de ocorrência das espécies. Ademais, a razão entre o número de pontos e o número de variáveis foi calculada como parâmetro para descrever os dados analisados. Uma baixa razão entre pontos e variáveis poderia indicar que os autores delinearam seus modelos com alto grau de sobreajuste (Peterson *et al.* 2011).

RESULTADOS

A listagem de artigos que retornou do sítio eletronico especializado foi utilizada para a obtenção de 273 artigos científicos que, após triagem de escopo, resultou em 177 artigos analisados (Material Suplementar 1) publicados entre 2002 e 2014 (Material Suplementar 2). Em 11 artigos (6%) os autores não mencionaram as variáveis utilizadas no procedimento de modelagem.

Perfil dos artigos e dos conjuntos de variáveis preditoras

Os artigos analisados tiveram como objetivos mais frequentes a estimativa da distribuição geográfica da espécie para descrever padrões espaciais e estimar a resposta espacial das espécies frente a mudanças no clima. Dentre outros objetivos, os artigos buscaram estimar a distribuição geográfica per se das espécies, utilizar as técnicas correlativas para estimar o nicho ecológico e identificar áreas de possível invasão de espécies exóticas. A lista completa de temas e objetivos pode ser encontrada no Material Suplementar 3.

O conjunto de variáveis preditoras contou em média com 10,84 (± 6,28 D.P.) variáveis e que variaram entre duas e 35 variáveis (mediana = 9). O histograma bimodal apresentou um pico em sete variáveis e um segundo pico em 19 variáveis (Figura 1), que é o número de variáveis do conjunto bioclimático de abrangência global mais popular nos estudos de modelagem (Worldclim; Hijmans *et al.* 2005). Alguns estudos que utilizaram o conjunto bioclimático frequentemente adicionaram a variável "altitude", o que

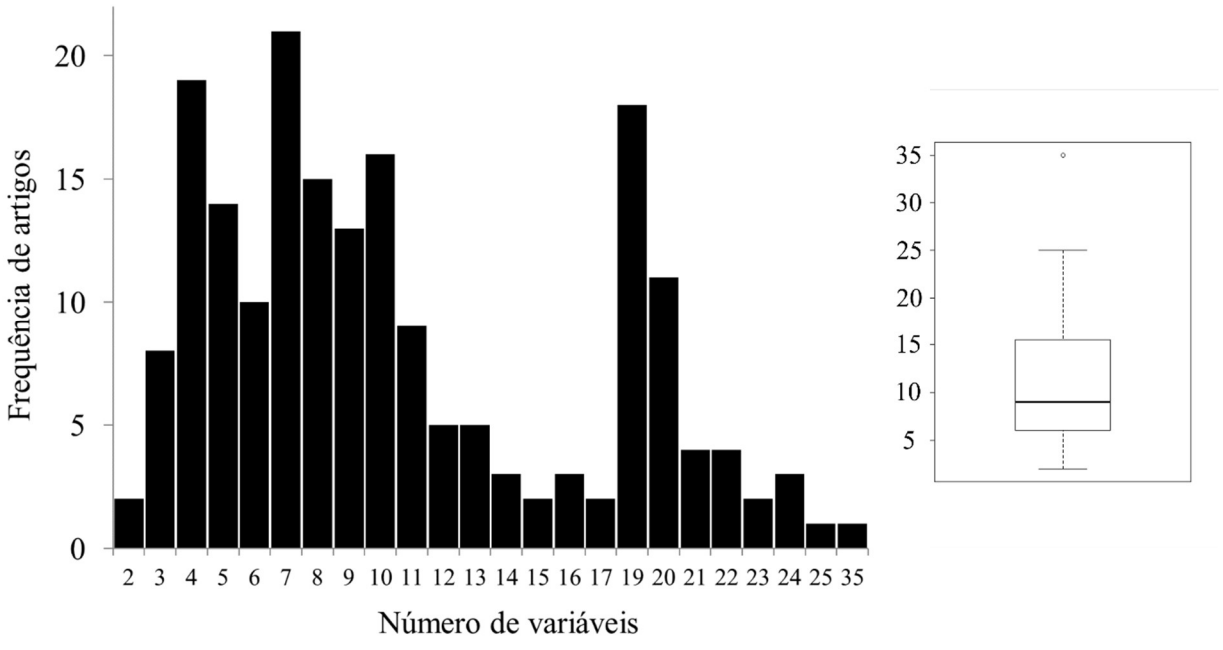

Figura 1. Frequência de artigos e número de variáveis utilizadas pelos autores em seus procedimentos de modelagem. À direita, a distribuição dos dados é mostrada em um box plot: mediana = 9, primeiro quartil = 6, terceiro quartil = 15, mínimo = 1, máximo = 35, outlier = 35.

Figure 1. *Frequency of papers and number of variables that authors used in their modelling procedure. On the right, data distribution is shown in a box plot: median = 9, first quartile = 6, third quartile = 15, min = 1, max = 25, and outlier = 35.*

explica a alta frequência de utilização de 20 variáveis nos estudos.

O número de pontos utilizados para a modelagem variou de dois a 5.233 pontos (mediana = 56, média = 176 ± 432 D.P.) e não apresentou correlação com o número de variáveis (r = -0,088; gl = 378; p = 0.08). A razão entre o número de pontos e o número de variáveis foi de, em média 19,01 (± 76,68 D.P., mediana = 5,42, mín = 0,14, máx = 1.046) pontos por variável.

As variáveis ambientais mais frequentemente utilizadas foram as do grupo das climáticas, sendo temperatura média anual, precipitação anual, precipitação sazonal e temperatura sazonal as mais empregadas. O segundo grupo mais utilizado foi de variáveis topográficas, seguido pelas variáveis de cobertura natural do solo e variáveis hidrográficas. Para uma lista completa de grupos e variáveis ver Material Suplementar 4.

Critérios usados para a seleção das variáveis

Do total de artigos analisados (n = 177), em 56 artigos (31%) os autores escolheram as variáveis sem justificativa explícita ou sem comentário sobre a seleção (Figura 2). Foi verificado que autores que selecionaram arbitrariamente as variáveis utilizaram maior número de variáveis do que autores que assumiram critérios (Figura 3). Em 121 artigos (68%) os autores utilizaram critérios para a seleção de variáveis e este uso foi relativamente mais frequente a partir do ano de 2009. O critério estatístico foi o mais empregado e, dentre os métodos estatísticos, os autores usaram com mais frequência a correlação de Pearson em um contexto pré-algoritmo e análise de Jackknife em um contexto pós-algoritmo (Tabela 1). O segundo critério mais utilizado foi o biológico, sendo que os autores dos estudos alegaram que as variáveis selecionadas eram "biologicamente importantes", seguido dos subcritérios de tolerância fisiológica e de interações ecológicas (Tabela 1).

Interessante foi observar a ocorrência de artigos cujos autores utilizaram critérios combinados. Neste caso, o procedimento de escolha das variáveis era submetido a dois ou mais critérios sequenciais. Dos 44 artigos de critérios combinados, 12 (27%) utilizaram o critério estatístico seguido do critério biológico, 11 (25%) utilizaram o critério biológico seguido do critério estatístico e oito (18%) utilizaram os critérios biológicos e estatísticos bidirecionalmente (Figura 2).

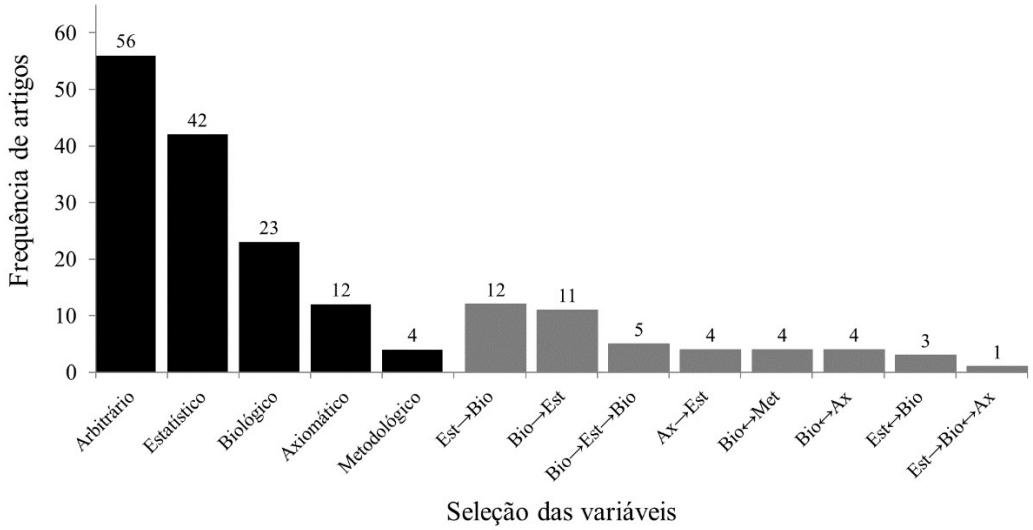

Figura 2. Número de publicações e os critérios empregados na seleção de variáveis ambientais. A seta unidirecional (→) indica que houve a adoção de mais de um critério consecutivamente. A seta bidirecional (↔) indica que ocorreu a adoção de dois critérios concomitantes ou sem uma ordem explícita entre eles. Est = Estatístico, Bio = Biológico, Ax = Axiomático e Met = Metodológico.

Figure 2. *Number of papers and criteria adopted by the authors to select environmental variables. Unidirectional arrow* (→) *indicates that authors adopted more than one criterion consecutively. Bidirectional arrow* (↔) *indicates that authors adopted two criteria in a non-ordered procedure. Est = Statistical, Bio = Biological, Ax = Axiomatic, and Met = Methodological.*

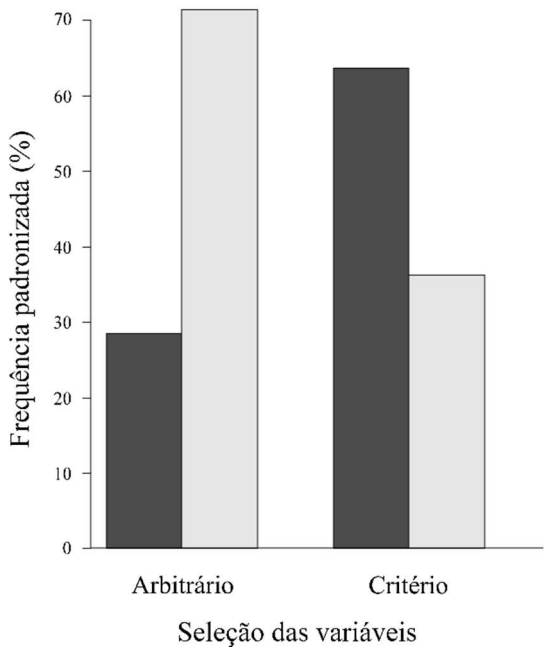

Figura 3. Frequência padronizada de artigos cujos autores utilizaram variáveis de modo arbitrário e adotando critério. Barras de cada classe foram separadas tomando-se como parâmetro a mediana (Figura 1): barras escuras < 9 variáveis e barras claras > 9 variáveis. O teste estatístico mostrou uma associação significativa (χ^2 = 18,2461; gl = 1; p = 0,00001) entre o uso de mais de nove variáveis e a seleção arbitrária e o uso de menos de nove variáveis e a adoção de critério.

Figure 3. *Standardized frequency of papers whose authors selected variables arbitrarily and selected them adopting a criterion. Bars for each class were split based on median (Figure 1): black bars < 9 variables and gray bars > 9 variables. Statistical test showed a significant (χ^2 = 18.2461, df = 1, p = 0.00001) association between the use of more than nine variables and arbitrary selection; and between the use of less than nine variables and criterion-based selection.*

DISCUSSÃO

A abordagem correlativa apresenta como característica o modelo conceitual empírico, modelo no qual é priorizado aspectos como realidade e precisão (Guisan & Zimmermann 2000). A relação entre a ocorrência dos organismos e as variáveis preditoras é construída fenomenologicamente por correlação espacial entre os pontos e as camadas e, por isso, a natureza estatística da abordagem. Era esperado então que os autores tivessem selecionado as variáveis preditoras utilizando-se de algum critério, sobretudo o critério estatístico. Os resultados mostraram que um terço dos artigos analisados tiveram as variáveis selecionadas arbitrariamente. Isso contradiz a expectativa inicial do uso de critérios, no entanto, de algum modo a prática de seleção arbitrária corrobora o fato de que os estudos correlativos têm se prestado mais a gerar uma aproximação da realidade (*e.g.*, modelagem de habitat) do que a buscar uma relação mecanística e causal das variáveis com a fisiologia dos organismos.

A partir do ano de 2009, os estudos apresentaram maior frequência na utilização de critérios. É possível que um amadurecimento no modo mais crítico de seleção das variáveis tenha se consolidado como prática na literatura. A descoberta de que a utilização de poucas variáveis poderia gerar estimativas com áreas superestimadas e extrapoladas (Beaumont *et al.* 2005) e que a utilização de muitas variáveis poderia gerar resultados subestimados e cenários complexos de difícil interpretação (Peterson *et al.* 2011), podem ter levado os autores a desenvolver uma posição mais crítica ao selecionar o conjunto de preditores. Independente se esta foi ou não a causa, os resultados mostraram que com a adoção de critérios pelos autores os conjuntos de preditores apresentaram menor número de variáveis.

O critério estatístico foi o mais adotado pelos autores e teve por finalidade evitar a redundância das variáveis (Dormann *et al.* 2012). Desta forma, seja utilizando análise de correlação ou análise de componentes principais, o conjunto foi construído de modo a favorecer as variáveis que melhor representassem a disposição dos pontos de ocorrência no espaço geográfico. Alternativamente a utilização de autovetores como supervariáveis fortalece a busca pela melhor representação dos dados empíricos.

O critério biológico foi o segundo mais utilizado na seleção de variáveis. Apesar disso, a análise dos artigos revelou que há pouca precisão e há dificuldade de implementação deste critério. Em um contexto correlativo, o obstáculo está no estabelecimento da relação entre um atributo biológico e uma variável ambiental (Kerney 2006, Feng & Papes 2017). No caso de limites fisiológicos, como intolerância à seca ou frio, fica até mais fácil o estabelecimento de relação com variáveis proximais, pois os valores de limites fisiológicos são mais precisos. No entanto, ao buscar

Tabela 1. Sumário dos critérios e subcritérios apresentados pelos autores dos artigos ao selecionar variáveis ambientais para a modelagem de distribuição geográfica de espécies. N = número de artigos.

Table 1. *Summary of criteria and sub-criteria adopted by the authors for selecting environmental variables in the modelling procedure. N = number of papers.*

MODO DE SELEÇÃO	**N**
Arbitrário	56
Baseado em critério	121
Estatístico	
Quantificar redundância ou colinearidade	64
Correlação de Pearson	38
Análise de componentes principais (PCA)	10
Correlação (não especificou)	5
Correlação de Spearman	5
Análise de fatores com rotação *varimax*	2
Fator de inflação de variância (VIF)	2
Autocorrelação espacial	1
Média aritmética sem pesagem (UPMGA)	1
Ranquear variáveis	26
Jackknife	19
Análise de componentes principais (PCA)	6
Regressão logística	1
Random forest	1
Árvore de regressão	1
Reduzir dimensionalidade dos dados e criar novas variáveis	10
Análise de componentes principais (PCA)	10
Comparar modelos	4
Critério de informação de Akaike (AIC)	4
Verificar homogeneidade das variáveis	3
Desvio padrão	1
Qui-quadrado (χ^2)	1
Teste t	1
Biológico	
Variáveis que afetam ou limitam a amplitude geográfica da espécie	24
Variáveis de tendência sazonal ou anual	11
Biótopo e variáveis ambientais (não-climáticas)	9
Escala geográfica de análise	5
Variáveis que são biologicamente importantes, sem razão explícita.	22
Tolerância fisiológica	19
Fatores limitantes, valores extremos	10
Calcificação de conchas	5
Substitutos ou variáveis indiretas de energia e água	4
Interação ecológica	19
Crescimento e reprodução do organismo	7
Indicador de recurso alimentar	5
Variáveis antrópicas	3
Risco de doenças	2
Competição por alimento	1
Distribuição geográfica do hospedeiro	1
Produtividade	1
Axiomático	**24**
Baseado em variáveis utilizadas em estudos anteriores e estudos similares.	12
Variáveis frequentemente usadas em estudos de modelagem, padrão de riqueza de espécie e mudanças climáticas.	12
Metodológico	**11**
Construir modelo conservativo e realístico	5
*Evitar sobreajuste (*overfitting*)*	3
Variáveis de modelos globais de circulação atmosfera-oceano	3

relacionar temperatura com ectotermia, por exemplo, a relação é menos precisa e mais complexa, pois ectotermia pode apresentar relação distal e indireta, e o grau de precisão pode depender do organismo e da região. A falta de conhecimento sobre os aspectos biológicos dos organismos, associado ao desconhecimento do modelo de relação causal, tende a comprometer a interpretação das estimativas espaciais. Na falta de informações e relações causais mais claras, os autores tendem a assumir uma relação de fundo e geral, tal como que os organismos ocuparão áreas onde há suprimento da demanda energética e hídrica (Pearson & Dawnson 2003).

Possivelmente como alternativa para contornar a falta de conhecimento biológico e de precisão da relação entre a variável e o atributo biológico, os autores pragmaticamente lançaram mão de selecionar as variáveis baseando-se em critérios combinados. Neste caso, utilizou-se um dos critérios para compor o conjunto inicial de variáveis e realizou-se uma progressiva priorização das variáveis utilizando outros critérios, por exemplo, ponderando quanto à contribuição estatística e significado biológico da variável para o organismo estudado. O conjunto tende a ter número reduzido de variáveis, apesar de que a imprecisão da relação causal dos atributos biológicos e as variáveis ambientais pode permanecer.

Este foi o primeiro trabalho que se tem conhecimento onde houve investigação e descrição da lógica e dos critérios de como os autores selecionaram o conjunto de variáveis empregado nos estudos com modelagem correlativa de distribuição geográfica de espécies. Em resumo, foi observado um número significativo de estudos nos quais os autores não apontaram os motivos ou não justificaram explicitamente a seleção do conjunto de variáveis ambientais no procedimento de modelagem. Isso demonstra que tanto os autores quanto os revisores dos artigos não sentiram necessidade de justificar o conjunto de variáveis no contexto correlativo. Diagnosticou-se que há uma recorrente dificuldade prática em se estabelecer a relação causal entre atributos biológicos dos organismos (*e.g.*, fisiologia) e as variáveis ambientais (*e.g.*, diretas, indiretas) na construção dos modelos correlativos. Quando o autor adotou um critério, o conjunto apresentou menor número de variáveis. Os autores adotaram com frequência o uso de critérios combinados e esta prática foi interpretada como uma solução para balancear o número de variáveis do conjunto, evitando sobreajuste e extrapolação, e para incluir especificidades biológicas ao poder das análises estatística na composição do conjunto de variáveis.

CONCLUSÃO

Para estudos correlativos o critério estatístico deve estar presente, seja na seleção pré ou pós-algorítmica, com a finalidade de balancear o modelo para que o conjunto de preditores não tenha tantas variáveis que gere sobreajuste nem tão poucas variáveis que gere áreas extrapoladas. Conjuntos com seis a nove variáveis se mostraram adequados para os procedimentos de modelagem correlativa. A dificuldade em empregar o critério biológico com rigor teórico representa uma força motivadora pela busca do entendimento das relações de causa e efeito entre as variáveis e os organismos e deve incentivar os estudos experimentais em modelagem baseada em processos nos próximos anos.

AGRADECIMENTOS

Os autores são gratos a M. V. Garey e L. R. R. Faria pela leitura crítica e contribuições nas versões do manuscrito. Ao PROBIC/UNILA pela concessão de bolsa de iniciação científica a D. S. G. N. À PRPPG e ao ILACVN por apoiar continuamente o programa de pesquisa em Biogeografia e Macroecologia.

REFERÊNCIAS

Alexandre, B. R., Lorini, M. L., & Grelle, C. E. V. 2013. Modelagem preditiva de distribuição de espécies ameaçadas de extinção: um panorama das pesquisas. Oecologia Australis, 17(4), 483–508. DOI: 10.4257/oeco.2013.1704.04

Araújo, M. B., & Guisan, A. 2006. Five (or so) challenges for species distribution modelling.

Journal of Biogeography, 33(10), 1677–1688. DOI: 10.1111/j.1365-2699.2006.01584.x

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling, 157(2-3), 101–118. DOI: 10.1016/S0304-3800(02)00205-3

Austin, M. P. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling, 200(1–2), 1–19. DOI: 10.1016/j.ecolmodel.2006.07.005

Austin, M. P., & Van Niel, K. P. 2011. Improving species distribution models for climate change studies: Variable selection and scale. Journal of Biogeography, 38(1), 1–8. DOI: 10.1111/j.1365-2699.2010.02416.x

Barbosa, F. G., & Schneck, F. 2015. Characteristics of the top-cited papers in species distribution predictive models. Ecological Modelling, 313, 77–83. DOI: 10.1016/j.ecolmodel.2015.06.014

Bradie, J., & Leung, B. 2017. A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. Journal of Biogeography, 44(6), 1344–1361. DOI: 10.1111/jbi.12894

Beaumont, L. J., Hughes, L., & Poulsen, M. 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. DOI: 10.1016/j.ecolmodel.2005.01.030

Buckley, L. B., Urban, M. C., Angilleta, M. J., Croizer, A. L. G., Rissler, L. J., & Sears, M. W. 2010. Can mechanism inform species' distribution models? Ecology Letters, 13(8), 1041–1054. DOI: 10.1111/j.1461-0248.2010.01479.x

Cruz-Cárdenas, G., López-Mata, L., Villaseñor, J. L., & Ortiz, E. 2014. Potential species distribution modeling and the use of principal component analysis as predictor variables. Revista Mexicana de Biodiversidad, 85(1), 189–199. DOI: 10.7550/rmb.36723

De Marco Jr., P., & Siqueira, M. F. 2009. Como determinar a distribuição potencial de espécies sob uma abordagem conservacionista? Megadiversidade, 5(1–2), 65–76.

Diniz-Filho, J. A. F., Araújo, M. B., & Terribile, L. C. 2016. Macroecologia e mudanças climáticas - avanços recentes e novas abordagens. In: C. J. B. Carvalho & E. A. B. Almeida (Eds.), Biogeografia da América do Sul: análise de tempo, espaço e forma. pp. 157–168. São Paulo: Editora Roca.

Dormann, C. F., McPerson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., & Wilson, R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30(5), 609–628. DOI: 10.1111/j.2007.0906-7590.05171.x

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., & Singer, A. 2012. Correlation and process in species distribution models: Bridging a dichotomy. Journal of Biogeography, 39(12), 2119–2131. DOI: 10.1111/j.1365-2699.2011.02659.x

Elith, J., & Graham, C. H. 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography, 32(1), 66–77. DOI: 10.1111/j.1600-0587.2008.05505.x

Elith, J., & Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40(1), 677–697. DOI: 10.1146/annurev.ecolsys.110308.120159

Faleiro, F. V., Machado, R. B., & Loyola, R. D. 2013. Defining spatial conservation priorities in the face of land-use and climate change. Biological Conservation, 158, 248–257. DOI: 10.1016/j.biocon.2012.09.020

Feng, X., & Papes, M. 2017. Physiological limits in an ecological niche modeling framework: A case study of water temperature and salinity constraints of freshwater bivalves invasive in USA. Ecological Modelling, 346, 48–57. DOI: 10.1016/j.ecolmodel.2016.11.008

Franklin, J. 2009. Mapping species distributions: spatial inference and prediction. New York: Cambridge University Press: p. 340.

Fitzpatrick, M. C., Weltzin, J. F., Sanders, N. J., & Dunn, R. R. 2007. The biogeography of prediction error: Why does the introduced range of the fire ant over-predict its native

range? Global Ecology and Biogeography, 16(1), 24–33. DOI: 10.1111/j.1466-8238.2006.00258.x

Guisan, A., & Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135(2–3), 147–186. DOI: 10.1016/S0304-3800(00)00354-9

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., & NCEAS Species Distribution Modelling Group. 2007. Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions, 13(3), 332–340. DOI: 10.1111/j.1472-4642.2007.00342.x

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography, 29(5), 773–785. DOI: 10.1111/j.0906-7590.2006.04700.x

Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology, 93(3), 679–688. DOI: 10.1890/11-0826.1

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25(15), 1965–1978. DOI: 10.1002/joc.1276

Hutchinson, G. E. 1980. An introduction to population ecology. New Haven and London: Yale University Press: p. 260.

Kearney, M. 2006. Habitat, environment and niche: what are we modelling? Oikos, 115(1), 186–191. DOI: 10.1111/j.2006.0030-1299.14908.x

Kearney, M., & Porter, W. P. 2004. Mapping the fundamental niche: physiology, climate, and the distribution of a nocturnal lizard. Ecology, 85(11), 3119–3131. DOI: 10.1890/03-0820

Kearney, M., & Porter, W. P. 2009. Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. Ecology Letters, 12(4), 334–350. DOI: 10.1111/j.1461-0248.2008.01277.x

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography, 33(1), 103–114. DOI: 10.1111/j.1600-0587.2009.06039.x

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. DOI: 10.1111/j.0906-7590.2005.03957.x

Löwenberg-Neto, P. 2018. A metric to quantify analogous conditions and rank environmental variables. Biodiversity Informatics, 13(1), 11–26. DOI: 10.17161/bi.v13i0.6744

Löwenberg-Neto, P., & Loyola, R. D. 2016. Biogeo grafia da Conservação. In: C. J. B. Carvalho & E. A. B. Almeida (Eds.), Biogeografia da América do Sul: análise de tempo, espaço e forma. pp. 169–178. São Paulo: Editora Roca.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. 2009. Evalua tion of consensus methods in predictive species distribution modelling. Diversity and Distributions, 15(1), 59–69. DOI: 10.1111/j.1472-4642.2008.00491.x

Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography, 34(1), 3–22. DOI: 10.1177/0309133309355630

Pearson, R. G. 2010. Species' distribution modeling for conservation educators and practitioners. Lessons in Conservation, 3, 54–89.

Pearson, R. G. & Dawson, T. P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology & Biogeography, 12(5), 361–371. DOI: 10.1046/j.1466-822X.2003.00042.x

Peel, M. C., Finlayson, B. L., & McMahon, T. A. 2007. Updated world map of the Köppen-Geiger climate classification. Hydrology and Earth Systems Sciences, 11(5), 1633–1644. DOI: 10.5194/hess-11-1633-2007

Peterson, A. T. 2006. Uses and requirements of ecological niche models and related distributional models. Biodiversity Informatics, 3, 59–72. DOI: 10.17161/bi.v3i0.29

Peterson, A. T., & Nakazawa, Y. 2008. Environmental data sets matter in ecological niche modelling: an example with Solenopsis invicta and Solenopsis richteri. Global ecology and Biogeography, 17(1), 135–144. DOI: 10.1111/j.1466-8238.2007.00347.x

Peterson, A. T., & Soberón, J. 2012. Species distribution modeling and ecological niche nodeling: getting the concepts right. Natureza & Conservação, 10(2), 102–107. DOI: 10.4322/n

atcon.2012.019

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer. E., Nakamura, M., & Araújo, M. A. 2011. Ecological niches and geographic distributions. Princeton: Princeton University Press. 328p.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., & Guisan, A. 2017. Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. Global Ecology and Biogeography, 26(3), 275–287. DOI: 10.1111/geb.12530

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. DOI: 10.1890/07-2153.1

Raes, N. 2012. Partial versus full species distribution models. Natureza & Conservação, 10(2), 127–138. DOI: 10.4322/natcon.2012.020

Seoane, J., & Bustamante, J. 2001. Modelos predictivos de la distribución de especies: una revisión de sus limitaciones. Ecología, 15, 99–21.

Seoane, J., Bustamante, J., & Díaz-Delgado, R. 2005. Effect of expert opinion on the predictive ability of environmental models of bird distribution. Conservation Biology, 19, 512–522. DOI: 10.1111/j.1523-1739.2005.00364.x

Sillero, N. 2011. What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecological Modelling, 222(8), 1343–1346. DOI: 10.1016/j.ecolmodel.2011.01.018

Silva, D. P., Vilela, B., De Marco Jr., P., & Nemésio, A. 2014. Using ecological niche models and niche analyses to understand speciation patterns: the case of sister neotropical orchid bees. PloS One, 9(11), e113246. DOI: 10.1371/journal.pone.0113246

Soberón, J., & Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics, 2, 1–10. DOI: 10.17161/bi.v2i0.4

Stockwell, D. R. B., & Peterson, A. T. 2002 Effects of sample size on accuracy of species distribution models. Ecological Modelling, 148(1), 1–13. DOI: 10.1016/S0304-3800(01)00388-X

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. 2007. A comparative evaluation of presence-only methods for modelling species distribution. Diversity and Distributions, 13(4), 397–405. DOI: 10.1111/j.1472-4642.2007.00346.x

Vasconcelos, T. S. 2014. Tracking climatically suitable areas for an endemic Cerrado snake under climate change. Natureza & Conservação, 12(1), 47–52. DOI: 10.4322/natcon.2014.009

Walter, H., & Breckle, S. W. 1985. Ecological Systems of the Geobiosphere: 1 ecological principles in global perspective. Berlin Heidelberg: Springer: p. 244.

Material Suplementar 1. Lista das publicações analisadas no estudo.

Supplementary Material 1. *Summary of analyzed publications.*

Material Suplementar 2. Número de publicações analisadas e ano de publicação. A busca por artigos foi realizada em janeiro de 2015, ano que não consta no gráfico.

Supplementary material 2. *Number of papers per year of publication. Papers were searched in January 2015, which was not shown in the graph.*

Material Suplementar 3. Lista de temas e objetivos apresentados nos estudos analisados ordenados pelo número de artigos.

Supplementary Material 3. *List of subjects and objectives presented in the analyzed papers. They were ordered by the number of papers.*

Material Suplementar 4. Lista de grupos e variáveis ambientais selecionadas nos estudos analisados. Elas foram ordenadas pelo número de artigos.

Supplementary Material 4. *Summary of groups and environmental variables selected in the analyzed studies. They were ordered by number of papers.*

Submetido: 06 /06/2017
Aceito: 11/04/2018
Editor Associado: Marcelo M. Weber