SECRETARIA DE ESTADO DE CIÊNCIA E TECNOLOGIA UNIVERSIDADE DO ESTADO DE MATO GROSSO CAMPUS DE NOVA XAVANTINA PROGRAMA DE PÓS-GRADUAÇÃO EM ECOLOGIA E CONSERVAÇÃO

ÚTEIS, M<mark>AS PO</mark>DEM MELHORAR: AVANÇOS METODOLÓGICOS PARA A MODELAGEM DE NICHO E DISTRIBUIÇÃO DE ESPÉCIES

João Carlos Pires de Oliveira

Tese apresentada à Coordenação do Programa de Pós-graduação em Ecologia e Conservação da Universidade do Estado de Mato Grosso - Campus de Nova Xavantina, como parte dos requisitos para a obtenção do título de Doutor em Ecologia e Conservação.

Orientador: Dr. Pedro Vasconcellos Eisenlohr Coorientador: Dr. Matheus de Souza Lima Ribeiro

SECRETARIA DE ESTADO DE CIÊNCIA E TECNOLOGIA UNIVERSIDADE DO ESTADO DE MATO GROSSO CAMPUS DE NOVA XAVANTINA PROGRAMA DE PÓS-GRADUAÇÃO EM ECOLOGIA E CONSERVAÇÃO

ÚTEIS, MAS PODEM MELHORAR: AVANÇOS METODOLÓGICOS PARA A MODELAGEM DE NICHO E MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIES

João Carlos Pires de Oliveira

Tese apresentada à Coordenação do Programa de Pós-graduação em Ecologia e Conservação da Universidade do Estado de Mato Grosso - *Campus* de Nova Xavantina, como parte dos requisitos para a obtenção do título de Doutor em Ecologia e Conservação.

Orientador: Dr. Pedro Vasconcellos Eisenlohr Coorientador: Dr. Matheus de Souza Lima Ribeiro

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)

This study was partially funded by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazilian Ministry of Education)

Pires-Oliveira, João Carlos.

ÚTEIS, MAS PODEM MELHORAR: AVANÇOS METODOLÓGICOS PARA A MODELAGEM DE NICHO E DISTRIBUIÇÃO DE ESPÉCIES / João Carlos Pires de Oliveira. - Cáceres, 2023.

71f.: il.

Universidade do Estado de Mato Grosso, Programa de Pósgraduação em Ecologia e Conservação, Nova Xavantina.

Orientador: Pedro Vasconcellos Eisenlohr. Coorientador: Matheus de Souza Lima Ribeiro.

1. avaliação de modelos. 2. importância das variáveis. 3. sinal ecológico. 4. viés de amostragem. I. Eisenlohr, Pedro Vasconcellos. II. Ribeiro, Matheus de Souza Lima. III. Título.

UNEMAT / MTSCB CDU 502

TERMO DE APROVAÇÃO ÚTEIS, MAS PODEM MELHORAR: AVANÇOS METODOLÓGICOS PARA A MODELAGEM DE NICHO E MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIE

João Carlos Pires de Oliveira

Tese apresentada ao Programa de Pós-graduação em Ecologia e Conservação da Universidade do Estado de Mato Grosso - Campus de Nova Xavantina, como parte dos requisitos para a obtenção do título de Doutora em Ecologia e Conservação.

Danca avamina dana	Aprovada em 04 de agosto de 2023.
Banca examinadora:	
	Prof. Dr. Pedro Vasconcellos Eisenlhor
	(Orientador) Universidade do Estado de Mato Grosso
	Prof ^a . Dra. Levi Carina Terribile
	(Membro titular-Externo) Universidade Federal de Jataí
	Prof. Dr. Daniel de Paiva Silva
	(Membro titular-Externo) Instituto Federal Goiano
	Prof. Dr. Alessandro Ribeiro de Moraes
	(Membro titular) Instituto Federal Goiano
	Prof. Dra. Mônica Aparecida Cupertino Eisenlohr
	(Membro titular) Universidade Federal de Mato Grosso
	Prof ^a . Dra. Vanessa Leite Rezende
(Membro	suplente-Externo) Universidade do Estado de Minas Gerais
	Prof. Dr. Joaquim Manoel da Silva
(Membro suplente) Universidade do Estado de Mato Grosso

Dedico este trabalho

Ao garotinho inocente, que lá em 1999 disse que queria estudar os bichos, pois queria entender como a natureza funcionava, e que graças à sua escolha aqui estou.

"O 'estudo' é a única coisa que ninguém nunca vai tomar de você". (Meu Pai)

AGRADECIMENTOS

Agradeço primeiramente aos meus pais Silvio e Abadia por sempre terem me apoiado desde o início dessa jornada acadêmica. Sou grato pelas palavras de incentivo e acolhimento que sempre me direcionaram.

Sou grato pela participação e colaboração dos meus orientadores, Prof. Dr. Pedro Vasconcellos Eisenlohr e Prof. Dr. Matheus de Souza Lima Ribeiro. O Prof. Pedro, por sua inestimável colaboração na minha formação científica e cidadã desde meu mestrado e também durante o desenvolvimento deste doutorado; o Prof. Matheus, por todo o seu apoio, disponibilidade e suporte metodológico e conceitual direcionado para o desenvolvimento da presente tese.

Sou imensamente grato aos colegas que me acompanharam até aqui, sempre apresentando pontos de vista diferentes para que eu pudesse construir uma visão de mundo mais assertiva. Em especial, agradeço a Loana Araújo e Islandia Pereira pelas muitas horas de conversas extra-acadêmicas para "refrescar o cérebo", e Loana mais ainda por realizar correções para o melhoramento da presente tese. Também agradeço a Jôine Cariele pelas inúmeras horas de "fofoca" com sua visão diferenciada de mundo que sempre acrescentou demais à minha formação como pesquisador e ser humano.

Agradeço a todas as instituições públicas que me proporcionaram a oportunidade de cursar este doutorado. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa e também pela disponibilidade de acesso ao acervo literário disponibilizado por meio das suas plataformas. Agradeço à Universidade do Estado de Mato Grosso (UNEMAT), pela infraestrutura e suporte necessário para o desenvolvimento da presente tese. Por fim, mas não menos importante, agradeço ao Programa de Pós-Graduação em Ecologia e Conservação, por todo o suporte disponibilizado para que meu doutorado pudesse ser desenvolvido.

Sumário

Lista de figuras	11
Lista de tabelas	12
Abstract	14
1. Introdução Geral	15
ista de tabelas bstract 1. Introdução Geral 2. Referências Bibliográficas da Introdução Geral apítulo 1: PSEUDO-AUSÊNCIAS PODEM SER UTILIZADAS PARA AVALIA IODELOS DE NICHO ECOLÓGICO? Resumo Abstract 1. Introdução 2. Material e Métodos Área de Estudo Variáveis Ambientais Espécies Virtuais Modelagem de nicho e Distribuição de espécies Avaliação dos Modelos Análises das métricas de avaliação dos modelos. 3. Resultados 4. Discussão 5. Conclusão 6. Referências bibliográficas 7. Apêndice apítulo 2: VIÉS DE AMOSTRAGEM, PSEUDO-AUSÊNCIAS E CAPACIDAD REDITIVA: O QUE EMERGE DESSA COMBINAÇÃO? Resumo Abstract Introdução Material E Métodos Área de Estudo Variáveis Ambientais Espécies Virtuais e Amostragens das Ocorrências Modelagem e Projeções	18
Capítulo 1: PSEUDO-AUSÊNCIAS PODEM SER UTILIZADAS PARA A	VALIAR
,	
•	
, , ,	
,	
_	
-	
Abstract	52
Lista de figuras	
Abstract	
,	
Variáveis Ambientais	55
- · · · · · · · · · · · · · · · · · · ·	
Resultados	60

LISTA DE FIGURAS

Capítulo 1
Figura 1. Mapa da área de estudo
Figura 2. Gráfico com médias e desvios padrão das correlações das métricas de avaliação
Figura 3. Gráfico com valores da correlação de <i>Spearman</i> para cada prevalência37
Figura 4. Gráfico médias e desvios padrão das correlações para dois diferentes thresholds 38
Capítulo 2
Figura 1. Distribuições contínuas das cincos Espécies Virtuais (EV)
Figura 2. Representação esquemática das ocorrências enviesadas
Figura 3. Representação do efeito do viés ao comparar valores da métrica de sobreposição
do conjunto de dados sem viés de amostragem com dados enviesados para modelos
construídos com dados de ausências reais e pseudo-ausência
Figura 4. Comparação do tamanho do efeito, em relação à intensidade de viés para cada
número de registros de ocorrência, obtidos da comparação par a par62

LISTA DE TABELAS

\sim	7.1	
Cai	pítu	10 1
~ ••		

Tabela 1. Matriz de Confusão.	26
Tabela 2. Tabela com Métricas Avaliativas	32-33
Capítulo 2	
Tabela 1. Tabela comparativa da relação entre métricas de avaliação e mét	rica de
sobreposição para modelos construídos com dados de ausências reais	60

RESUMO

A Modelagem de Nicho Ecológico (ENM) é amplamente utilizada em várias áreas do conhecimento que objetivam entender as propriedades do nicho e da distribuição das espécies. O desenvolvimento de ferramentas e de conceitos para modelar nicho e distribuição de espécies ocorreu rapidamente ao longo de sua curta história, que teve início na década de 1990, sob o objetivo de incorporar mais significado ecológico e precisão aos mapas de predição dos modelos. Portanto, a busca por algoritmos mais precisos se tornou inevitável, bem como a busca por melhores formas de avaliar esses modelos. Entretanto, os modeladores não formam uma unidade para recomendar o arcabouço metodológico e teórico, e isso gera imprecisões na construção e interpretação dos resultados dos modelos. Nesse sentido, ainda precisamos avançar conceitualmente e metodologicamente para tornar os ENM mais confiáveis. Assim, buscamos responder à seguinte pergunta: podemos construir e interpretar ENM mais robustos, que proporcionem interpretações ecológicas válidas em suas múltiplas esferas de utilização? Acreditamos que sim e propomos novas metodologias para: 1) Avaliar e propor alternativas para proporcionar uma avaliação dos modelos mais acurados; e 2) Identificar o efeito do viés de amostragem sobre os desempenhos preditivos dos modelos e apresentar soluções.

Palavras-chave: avaliação de modelos, importância das variáveis, sinal ecológico, viés de amostragem.

ABSTRACT

Ecological Niche Modeling (ENM) is widely used in various fields of knowledge that aim to understand species' niche properties and distribution. The development of tools and concepts for modeling niche and species distribution has progressed rapidly over its short history, which began in the 1990s, with the goal of incorporating more ecological significance and accuracy into model prediction maps. Therefore, the quest for more accurate algorithms became inevitable, as well as the search for better ways to evaluate these models. However, modelers do not form a unified front in recommending the methodological and theoretical framework, which leads to inaccuracies in the construction and interpretation of model results. In this regard, we still need to advance conceptually and methodologically to make ENMs more reliable. Thus, we seek to answer the following question: can we build and interpret more robust ENMs that provide valid ecological interpretations in their various spheres of use? We believe we can and propose new methodologies to: 1) Evaluate and suggest alternatives to provide a more accurate model assessment; and 2) Identify the sampling bias effect on model predictive performance and present solutions.

Keywords: ecological signal, collection bias, model evaluation, variable importance.

1. Introdução Geral

A utilização da modelagem de nicho e distribuição de espécies [ecological niche models (ENM) e species distribution models (SDM) em inglês, respectivamente) é uma realidade irreversível. Atualmente, a aplicação destes métodos tem se estendido muito além do seu objetivo inicial de identificar onde uma espécie pode ser encontrada e o seu "perfil" ambiental (Busby, 1991). Seu universo de utilização encontra-se presente em estudos sobre identificação de populações desconhecidas (Siqueira et al., 2009), dinâmica das espécies no passado (Banks et al., 2008), projeção da distribuição das espécies para o futuro (Dormann, 2007), previsão do possível impacto das mudanças climáticas sobre a agricultura e biodiversidade nativa (Castro et al., 2020; Fraga et al., 2012), identificação de risco de invasões biológicas, disseminação de doenças e proliferação de pragas (Estay et al., 2014; Jiménez-Valverde et al., 2011; Peterson, 2014), estimativa de parâmetros das populações (Lima-Ribeiro and Felizola Diniz-Filho, 2013; Pellissier et al., 2012; Wisz et al., 2007), análise de interações biológicas entre espécies (Anderson et al., 2002; Dormann et al., 2018), diversificação e especiação (Silva et al., 2014); processos de dispersão (Génard & Lescourret, 2013), processos de extinções (Lima-Ribeiro et al., 2013; Nogués-Bravo et al., 2008), identificação de hotspots, refúgios e locais prioritários para a conservação (Carnaval et al., 2009; Nóbrega and De Marco, 2011; Terribile et al., 2012), teste de hipóteses relacionadas ao nicho das espécies, tais como similaridade e equivalências (Jakob et al., 2010; Martínez-Meyer and Peterson, 2006), estudos sobre reintrodução e translocação de espécies (Gomides et al., 2021; Richmond et al., 2010) e teste de hipóteses filogeográficas (Alvarado-Serrano and Knowles, 2014; Collevatti et al., 2015). Essas ferramentas operam por meio de modelagens numéricas, geralmente estatísticas, dos locais onde as espécies ocorrem – seja por simples presença e ausência ou pela abundância - em função das condições ambientais presentes nesses locais de ocorrência (Franklin, 2010).

Após a criação do BIOCLIM, proposto por Busby (1991), muitos outros algoritmos foram desenvolvidos, tais como algoritmos de análises multivariadas – análise de fator do nicho ecológico (ENFA) (Hirzel et al., 2002), algoritmos de regressão – modelos lineares generalizados (GLMs) (Guisan et al., 2002) e aprendizagem de máquinas – máxima entropia (MAXENT) (Phillips et al., 2004). Tanto as técnicas supracitadas, quanto as que

se sucederam, foram utilizadas com objetivos e modos operantes semelhantes, consistindo em calibrar o modelo com dados de presença e ausência ou pseudo-ausências (Engler et al., 2004). Desta forma, a busca por algoritmos mais acurados se tornou inevitável.

Assim como as técnicas de modelagem, as métricas de avaliação evoluíram ao longo do tempo, visando identificar aquelas que fossem capazes de identificar a capacidade de um modelo predizer corretamente os padrões de distribuição ou de abundância das espécies (Allouche et al., 2006; Fielding and Bell, 1997; Liu et al., 2013; Wunderlich et al., 2019). Para mensurar a capacidade dos modelos em discriminar presenças de ausências, eles são avaliados por meio de suas projeções no espaço geográfico utilizando a matriz de confusão (CM). Na CM são computados todos os erros e acertos dos modelos ao predizer as presenças e ausências (Fielding and Bell, 1997). A CM requer uma binarização das predições do modelo, que geralmente são contínuas e a escolha de um limiar de decisão. Assim, células da área de estudo com valores de adequabilidade acima desse limiar são consideradas presença da espécie e células com valores abaixo desse valor são consideradas ausência da espécie (Liu et al., 2005). As métricas que avaliam os modelos baseadas na CM almejam quantificar a proporção de erros e acertos dos modelos ao prever presenças e ausências/pseudo-ausências (Allouche et al., 2006; Fielding and Bell, 1997). Existem poucas métricas que quantificam a capacidade preditiva dos modelos que não são baseadas na CM (Boyce et al., 2002; Hirzel et al., 2006; Scherrer et al., 2020).

Após três décadas desde a contribuição de Busby et al. (1991), avanços conceituais e metodológicos foram alcançados no sentido de se melhorar a capacidade preditiva dos modelos e de proporcionar interpretações ecológicas úteis. Além disso, ocorreram grandes avanços buscando incorporar mais significado biológico aos mapas gerados pelas predições dos modelos e também na padronização das diretrizes utilizadas para modelar e interpretar os ENM(Araújo et al., 2019; Jarnevich et al., 2017; Sillero et al., 2021). Contudo, aspectos conceituais sobre os ENM e SDM ainda geram dúvidas e confusão, apesar de diversas tentativas para evitar ambiguidades (Peterson and Soberón, 2012; Pires-Oliveira et al., 2019; Warren, 2012). Além disso, dúvidas sobre quais atributos das espécies são de fato modelados ainda são comuns e podem levar a interpretações equivocadas e/ou distorcidas (Sillero, 2011).

Quanto aos aspectos metodológicos, a maioria das metodologias propostas foram desenvolvidas e testadas em estudos de caso em que não se tem controle de todos os

fatores que determinam as condicionantes do nicho e da distribuição geográfica das espécies, o que dificulta a reprodutibilidade e comparação entre estudos (Meynard et al., 2019). Além disso, a utilização de dados do "mundo real" traz problemas como viés de amostragem e também incertezas atreladas às coordenadas geográficas e delimitação e identificações taxonômicas, o que pode dificultar a avaliação dos modelos de forma adequada e levar a interpretações ecológicas equivocadas (Aguirre-Gutiérrez et al., 2013; Boria et al., 2014; Leitão et al., 2011). Também podemos encontrar problemas associados aos preditores, tais como autocorrelação espacial e colinearidade entre as variáveis (De Marco and Nóbrega, 2018; de Oliveira et al., 2014). Todos esses fatores adicionam incertezas ao resultado final da modelagem e podem levar o modelador a conclusões imprecisas, sendo raros os trabalhos que quantificaram e/ou consideraram essas incertezas em seus estudos (Diniz-Filho et al., 2009; Terribile et al., 2012; Thuiller et al., 2019).

A presente tese é norteada por uma pergunta geral: podemos construir e interpretar ENM mais robustos, que proporcionem interpretações ecológicas válidas em suas múltiplas esferas de utilização? Acreditamos que sim, pois mesmo com conceitos incompletos e metodologias imprecisas os ENM apresentaram resultados satisfatórios quando aplicados ao mundo real. Portanto, é desejável que tenhamos metodologias amplamente testadas em condições ideais e confrontadas com condições encontradas no mundo real (Meynard et al., 2019), desenvolvendo tecnologias para substituir aquelas que não apresentarem resultados satisfatórios (Bohl et al., 2019; Vollering et al., 2019).

Estruturamos a tese em dois capítulos e neles iremos: 1) testar a nossa capacidade atual de avaliar os ENM e propor uma alternativa viável para avaliar os modelos em um cenário hipotético simulado – com ausências reais e utilizando pseudo-ausências (Capítulo 1) e 2) avaliar o efeito da estruturação espacial dos pontos de ocorrência (viés de amostragem) sobre a capacidade dos modelos de prever corretamente a distribuição das espécies e avaliar a capacidade das métricas atualmente disponíveis de avaliar corretamente os ENM (Capítulo 2).

2. Referências Bibliográficas da Introdução Geral

Aguirre-Gutiérrez, J., Carvalheiro, L.G., Polce, C., van Loon, E.E., Raes, N., Reemer, M., Biesmeijer, J.C., 2013. Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria – Dutch Hoverflies as a Case Study. PLoS ONE 8, e63708. https://doi.org/10.1371/journal.pone.0063708

- Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x
- Alvarado-Serrano, D.F., Knowles, L.L., 2014. Ecological niche models in phylogeographic studies: applications, advances and precautions. Mol. Ecol. Resour. 14, 233–248. https://doi.org/10.1111/1755-0998.12184
- Anderson, R.P., Peterson, A.T., Gómez-Laverde, M., 2002. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos 98, 3–16. https://doi.org/10.1034/j.1600-0706.2002.t01-1-980116.x
- Araújo, M.B., Anderson, R.P., Márcia Barbosa, A., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., Zimmermann, N.E., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. Sci. Adv. 5, eaat4858. https://doi.org/10.1126/sciadv.aat4858
- Banks, W.E., d'Errico, F., Peterson, A.T., Kageyama, M., Colombeau, G., 2008. Reconstructing ecological niches and geographic distributions of caribou (Rangifer tarandus) and red deer (Cervus elaphus) during the Last Glacial Maximum. Quat. Sci. Rev., Ice Age Refugia and Quaternary Extinctions: An Issue of Quaternary Evolutionary Palaeoecology 27, 2568–2575. https://doi.org/10.1016/j.quascirev.2008.09.013
- Bohl, C.L., Kass, J.M., Anderson, R.P., 2019. A new null model approach to quantify performance and significance for ecological niche models of species distributions. J. Biogeogr. 46, 1101–1111. https://doi.org/10.1111/jbi.13573
- Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012
- Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecol. Model. 157, 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4
- Busby, J.R., 1991. BIOCLIM: A bioclimate analysis and prediction system. Nat. Conserv. Cost Eff. Biol. Surv. Data Anal. 6.
- Carnaval, A.C., Hickerson, M.J., Haddad, C.F.B., Rodrigues, M.T., Moritz, C., 2009. Stability Predicts Genetic Diversity in the Brazilian Atlantic Forest Hotspot. Science 323, 785–789. https://doi.org/10.1126/science.1166955
- Castro, M.B., Barbosa, A.C.M.C., Pompeu, P.V., Eisenlohr, P.V., de Assis Pereira, G., Apgaua, D.M.G., Pires-Oliveira, J.C., Barbosa, J.P.R.A.D., Fontes, M.A.L., dos Santos, R.M., Tng, D.Y.P., 2020. Will the emblematic southern conifer Araucaria angustifolia survive to climate change in Brazil? Biodivers. Conserv. 29, 591–607. https://doi.org/10.1007/s10531-019-01900-x
- Collevatti, R.G., Terribile, L.C., Diniz-Filho, J.A.F., Lima-Ribeiro, M.S., 2015.

 Multi-model inference in comparative phylogeography: an integrative approach based on multiple lines of evidence. Front. Genet. 6. https://doi.org/10.3389/fgene.2015.00031
- De Marco, P., Nóbrega, C.C., 2018. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. PLOS ONE 13, e0202403. https://doi.org/10.1371/journal.pone.0202403
- de Oliveira, G., Rangel, T.F., Lima-Ribeiro, M.S., Terribile, L.C., Diniz-Filho, J.A.F., 2014. Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: A new approach based on environmentally equidistant records. Ecography 37, 637–647.

- https://doi.org/10.1111/j.1600-0587.2013.00564.x
- Diniz-Filho, J.A.F., Mauricio Bini, L., Fernando Rangel, T., Loyola, R.D., Hof, C., Nogués-Bravo, D., Araújo, M.B., 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. Ecography 32, 897–906. https://doi.org/10.1111/j.1600-0587.2009.06196.x
- Dormann, C.F., 2007. Promising the future? Global change projections of species distributions. Basic Appl. Ecol. 8, 387–397. https://doi.org/10.1016/j.baae.2006.11.001
- Dormann, C.F., Bobrowski, M., Dehling, D.M., Harris, D.J., Hartig, F., Lischke, H., Moretti, M.D., Pagel, J., Pinkert, S., Schleuning, M., Schmidt, S.I., Sheppard, C.S., Steinbauer, M.J., Zeuss, D., Kraan, C., 2018. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. Glob. Ecol. Biogeogr. 27, 1004–1016. https://doi.org/10.1111/geb.12759
- Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274. https://doi.org/10.1111/j.0021-8901.2004.00881.x
- Estay, S.A., Labra, F.A., Sepulveda, R.D., Bacigalupe, L.D., 2014. Evaluating Habitat Suitability for the Establishment of Monochamus spp. through Climate-Based Niche Modeling. PLoS ONE 9, e102592. https://doi.org/10.1371/journal.pone.0102592
- Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49. https://doi.org/10.1017/S0376892997000088
- Fraga, H., Malheiro, A.C., Moutinho-Pereira, J., Santos, J.A., 2012. An overview of climate change impacts on European viticulture. Food Energy Secur. 1, 94–110. https://doi.org/10.1002/fes3.14
- Franklin, J., 2010. Mapping species distributions: spatial inference and prediction. Cambridge University Press.
- Génard, M., Lescourret, F., 2013. Combining Niche and Dispersal in a Simple Model (NDM) of Species Distribution. PLoS ONE 8, e79948. https://doi.org/10.1371/journal.pone.0079948
- Gomides, S.C., Machado, T.M., Evangelista-Vale, J.C., Martins-Oliveira, A.T., Pires-Oliveira, J.C., Muller, A., Barros da Rosa, L., Lucas Santos-Silva, D., Eisenlohr, P.V., 2021. Assessing species reintroduction sites based on future climate suitability for food resources. Conserv. Biol. cobi.13796. https://doi.org/10.1111/cobi.13796
- Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1
- Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-Niche Factor Analysis: How to Compute Habitat-Suitability Maps without Absence Data? Ecology 83, 2027. https://doi.org/10.2307/3071784
- Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. 199, 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017
- Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D., Willis, S.G., 2014. Improving species distribution models: the value of data on abundance. Methods Ecol. Evol. 5, 506–513.
- Jakob, S.S., Heibl, C., Rödder, D., Blattner, F.R., 2010. Population demography influences

- climatic niche evolution: evidence from diploid American *Hordeum* species (Poaceae). Mol. Ecol. 19, 1423–1438. https://doi.org/10.1111/j.1365-294X.2010.04582.x
- Jarnevich, C.S., Talbert, M., Morisette, J., Aldridge, C., Brown, C.S., Kumar, S., Manier, D., Talbert, C., Holcombe, T., 2017. Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: An example with background selection. Ecol. Model. 363, 48–56. https://doi.org/10.1016/j.ecolmodel.2017.08.017
- Jiménez-Valverde, A., Lobo, J.M., Hortal, J., 2008. Not as good as they seem: The importance of concepts in species distribution modelling. Divers. Distrib. 14, 885–890. https://doi.org/10.1111/j.1472-4642.2008.00496.x
- Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P., Lobo, J.M., 2011. Use of niche models in invasive species risk assessments. Biol. Invasions 13, 2785–2797. https://doi.org/10.1007/s10530-011-9963-4
- Leitão, P.J., Moreira, F., Osborne, P.E., 2011. Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. Int. J. Geogr. Inf. Sci. 25, 439–454. https://doi.org/10.1080/13658816.2010.531020
- Lima-Ribeiro, M.S., Felizola Diniz-Filho, J.A., 2013. American megafaunal extinctions and human arrival: Improved evaluation using a meta-analytical approach. Quat. Int. 299, 38–52. https://doi.org/10.1016/j.quaint.2013.03.007
- Lima-Ribeiro, M.S., Nogués-Bravo, D., Terribile, L.C., Batra, P., Diniz-Filho, J.A.F., 2013. Climate and humans set the place and time of Proboscidean extinction in late Quaternary of South America. Palaeogeogr. Palaeoclimatol. Palaeoecol. 392, 546–556. https://doi.org/10.1016/j.palaeo.2013.10.008
- Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393. https://doi.org/10.1111/j.0906-7590.2005.03957.x
- Liu, C., White, M., Newell, G., Griffioen, P., 2013. Species distribution modelling for conservation planning in Victoria, Australia. Ecol. Model. 249, 68–74. https://doi.org/10.1016/j.ecolmodel.2012.07.003
- Martínez-Meyer, E., Peterson, A.T., 2006. Conservatism of ecological niche characteristics in North American plant species over the Pleistocene-to-Recent transition. J. Biogeogr. 33, 1779–1789. https://doi.org/10.1111/j.1365-2699.2006.01482 33 10.x
- Meynard, C.N., Leroy, B., Kaplan, D.M., 2019. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? Ecography 42, 2021–2036. https://doi.org/10.1111/ecog.04385
- Nóbrega, C.C., De Marco, P., 2011. Unprotecting the rare species: a niche-based gap analysis for odonates in a core Cerrado area: Niche-based gap analysis for odonates. Divers. Distrib. 17, 491–505. https://doi.org/10.1111/j.1472-4642.2011.00749.x
- Nogués-Bravo, D., Rodríguez, J., Hortal, J., Batra, P., Araújo, M.B., 2008. Climate Change, Humans, and the Extinction of the Woolly Mammoth. PLoS Biol. 6, e79. https://doi.org/10.1371/journal.pbio.0060079
- Pellissier, L., Pradervand, J.-N., Pottier, J., Dubuis, A., Maiorano, L., Guisan, A., 2012. Climate-based empirical models show biased predictions of butterfly communities along environmental gradients. Ecography 35, 684–692. https://doi.org/10.1111/j.1600-0587.2011.07047.x
- Peterson, A.T., 2014. Mapping disease transmission risk: enriching models using biogeography and ecology. JHU Press.

- Peterson, A.T., Soberón, J., 2012. Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. Nat. Conserv. 10, 102–107. https://doi.org/10.4322/natcon.2012.019
- Phillips, S.J., Dudík, M., Schapire, R.E., 2004. A Maximum Entropy Approach to Species Distribution Modeling. Twenty-First Int. Conf. Mach. Learn. ICML 04 83-. https://doi.org/10.1145/1015330.1015412
- Pires-Oliveira, J.C., José-Silva, L., Rocha, D.S.B., Eisenlohr, P.V., 2019. Concepts and methods in environmental suitability modeling, an important tool for forest conservation, in: Eisenlohr, P V (Ed.), Forest Conservation: Methods, Management and Challenges. Nova Publishers Ltda, New York, pp. 269–290.
- Richmond, O.M.W., McEntee, J.P., Hijmans, R.J., Brashares, J.S., 2010. Is the Climate Right for Pleistocene Rewilding? Using Species Distribution Models to Extrapolate Climatic Suitability for Mammals across Continents. PLoS ONE 5, e12899. https://doi.org/10.1371/journal.pone.0012899
- Scherrer, D., Mod, H.K., Guisan, A., 2020. How to evaluate community predictions without thresholding? Methods Ecol. Evol. 11, 51–63. https://doi.org/10.1111/2041-210X.13312
- Sillero, N., 2011. What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecol. Model. 222, 1343–1346. https://doi.org/10.1016/j.ecolmodel.2011.01.018
- Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C.G., Sousa-Guedes, D., Martínez-Freiría, F., Real, R., Barbosa, A.M., 2021. Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. Ecol. Model. 456, 109671. https://doi.org/10.1016/j.ecolmodel.2021.109671
- Silva, D.P., Vilela, B., De Marco, 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, e113246. https://doi.org/10.1371/journal.pone.0113246
- Siqueira, M.F. de, Durigan, G., de Marco Júnior, P., Peterson, A.T., 2009. Something from nothing: Using landscape similarity and ecological niche modeling to find rare plant species. J. Nat. Conserv. 17, 25–32. https://doi.org/10.1016/j.jnc.2008.11.001
- Terribile, L.C. et al., 2012. Areas of climate stability of species ranges in the Brazilian Cerrado: disentangling uncertainties through time. Nat. Conserv. 10, 152–159. https://doi.org/10.4322/natcon.2012.025
- Thuiller, W., Guéguen, M., Renaud, J., Karger, D.N., Zimmermann, N.E., 2019. Uncertainty in ensembles of global biodiversity scenarios. Nat. Commun. 10, 1446. https://doi.org/10.1038/s41467-019-09519-w
- Vollering, J., Halvorsen, R., Auestad, I., Rydgren, K., 2019. Bunching up the background betters bias in species distribution models. Ecography 42, 1717–1727. https://doi.org/10.1111/ecog.04503
- Warren, D.L., 2012. In defense of "niche modeling." Trends Ecol. Evol. 27, 497–500. https://doi.org/10.1016/j.tree.2012.03.010
- Wisz, M.S., Walther, B.A., Rahbek, C., 2007. Using potential distributions to explore determinants of Western Palaearctic migratory songbird species richness in sub-Saharan Africa. J. Biogeogr. 34, 828–841. https://doi.org/10.1111/j.1365-2699.2006.01661.x
- Wunderlich, R.F., Lin, Y.-P., Anthony, J., Petway, J.R., 2019. Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models. Nat. Conserv. 35, 97–116. https://doi.org/10.3897/natureconservation.35.33918

CAPÍTULO 1: PSEUDO-AUSÊNCIAS DEVEM SER UTILIZADAS PARA AVALIAR MODELOS DE NICHO ECOLÓGICO?

João Carlos Pires-Oliveira¹, Matheus S. Lima-Ribeiro², Pedro V. Eisenlohr^{1,3}

¹ Programa de Pós-Graduação em Ecologia e Conservação, Universidade do Estado de Mato Grosso, Brasil

² Universidade Federal de Jataí, Brasil

³ Universidade do Estado de Mato Grosso, *Campus* Alta Floresta, Brasil

Resumo

Os modelos de nicho e distribuição potencial de espécies (MNDPE) possuem ampla aplicação em variados campos da biologia, tais como predições de distribuições potenciais no espaço e no tempo, passado, presente e futuro, identificação de propriedades do nicho, estudos de espécies invasoras, previsão de doenças, identificação de locais com bom potencial para criação de unidades de conservação e estudos referentes aos efeitos das mudanças climáticas e em todos os casos nos quais os modelos precisam ser avaliados. A avaliação é feita, normalmente, computando acertos e erros dos modelos em relação às presenças e ausências das espécies modeladas por meio da matriz de confusão (CM). Porém, devido à escassez de dados de ausências reais, pontos aleatórios ou estruturados no espaço geográfico e ambiental (pseudo-ausências) são utilizados para calibrar e avaliar os MNDPE. A utilização de pseudo-ausências é amplamente criticada e apontada com grande fonte de incerteza na construção e avaliação dos MNDPE. Neste trabalho, utilizamos espécies virtuais (EVs) para avaliarmos se as métricas de avaliação, baseadas na CM, são afetadas negativamente pela utilização de pseudo-ausências e apresentamos alternativas robusta conceitual e metodologicamente para avaliar os modelos com dados apenas de presenças, quando dados de ausências reais não estão disponíveis. Também avaliamos se métricas de avaliação focadas em dados de presença apresentam menos capacidade avaliativa. Por fim, avaliamos se métricas de apenas presença são afetadas pelas pseudo-ausência quando estas foram utilizadas para calibrar modelos. Nossos resultados demonstraram que as pseudo-ausências reduzem a capacidade avaliativa das métricas focadas em presenças e ausências, com o dado de ausência sendo o componente que mais gera incerteza. As métricas de apenas presenças não são inferiores às métricas e apresentam resultados superiores em alguns cenários. Finalmente, constatamos que as métricas focadas em apenas presenças também apresentam perdas da sua capacidade avaliativa ao serem utilizadas para avaliar modelos calibrados e avaliados com pseudo-ausências e para essas métricas o componente mais relevante foi a prevalência das espécies. Nossos resultados reforçam os cuidados que devemos ter ao interpretar os resultados das avaliações dos MNDPE, principalmente quando usamos pseudo-ausências.

Palavras-chave: avaliação de modelos, capacidade preditiva, modelos presença e ausência, métricas de avaliativas.

ABSTRACT

Niche and Potential Species Distribution Models (NPSDM) have wide application in several fields of biology, such as predictions of potential distributions in space and time, past, present and future, identification of niche properties, studies of invasive species, disease forecasting, identification of priority areas for the creation of conservation units and studies regarding the effects of climate change and in all cases the models need to be evaluated. The evaluation is usually carried out by measuring the model's successes and errors, in relation to the presence and absence of the modeled species, using the confusion matrix (CM). However, when true absence data is not available, random or structured points in geographic space (pseudo-absences) are used to calibrate and evaluate the NPSDM. The use of pseudo-absences is widely criticized and identified as a great source of uncertainty in the construction and evaluation of the NPSDM. In this work, we use virtual species (VE) to evaluate whether the evaluation metrics, based on the CM, are negatively affected by the use of pseudo-absences and we present conceptually and methodologically robust alternatives to evaluate the models with only presences data, when data from true absences are not available. We also evaluated whether evaluation metrics focused on presence data have less evaluative capacity when compared with metrics focused on presences and absences. Finally, we evaluated whether presence-only metrics are affected by pseudo-absences when these were used to calibrate models presence-absence based. Our results demonstrated that pseudo-absences reduce the evaluative capacity of metrics focused on presences and absences, with the absence data being the component that most generates uncertainty. Presence-only metrics are not inferior to presence-absence metrics and show superior results in some scenarios. Finally, we found that the metrics focused on just presences also lost their evaluative capacity when used to evaluate calibrated models and evaluated with pseudo-absences, and for these metrics the most relevant component generator of uncertainty was species prevalence. Our results reinforce the care that we must take when interpreting the results of the NPSDM assessments, especially when pseudo-absences are used.

Keywords: evaluative metrics, model evaluation, predictive ability, presence and absence models.

1. Introdução

Os modelos de nicho ecológico e modelagem de distribuição potencial de espécies (ENM e SDM) são ferramentas valiosas na biologia e em áreas correlatas, sendo amplamente utilizados para prever distribuições no espaço e no tempo, abrangendo o passado, presente e futuro (Capinha and Pateiro-López, 2014; Sobral-Souza et al., 2015). Esses modelos contribuem para a identificação de propriedades do nicho, estudos de espécies invasoras, previsão de doenças, identificação de locais para unidades de conservação e pesquisas sobre mudanças climáticas (Martínez-Meyer and Peterson, 2006; Petitpierre et al., 2017; Peterson, 2014; Prieto-Torres et al., 2018; Castro et al., 2020). Independentemente do objetivo de uso, é crucial avaliar a capacidade preditiva dos modelos, considerando erros e acertos, o que representa um desafio na modelagem de nicho e distribuição de espécies (Araújo & Guisan, 2006; Elith and Leathwick, 2009). A avaliação geralmente é realizada por meio de métricas baseadas na matriz de confusão (CM) (Tabela 1) (Fielding and Bell, 1997), que indicam as taxas de acertos e erros ao prever a presença e ausência da espécie modelada.

Tabela 1: Matriz de confusão utilizada para computar os erros e acertos das predições dos modelos. As linhas são as predições dos modelos e as colunas são os dados reais utilizados para testar o desempenho dos modelos. A letra *a* informa que o modelo prediz a ocorrência da espécie onde há uma ocorrência real. A letra *b* corresponde a um erro (*comissão*), em que o modelo prediz a ocorrência, mas há uma ausência da espécie naquela localidade (falso positivo). A letra *c* corresponde a um erro (*omissão*) em que o modelo prediz a ausência da espécie e há um dado de ocorrência no local (falso negativo). A letra *d* mostra que o modelo prediz a ausência da espécie e há uma ausência da espécie neste local.

Predito	Observado	
	Dados de A	Avaliação
	Presente	Ausente
Presente	а	b
Ausente	c	d

A matriz de confusão (CM), em sua concepção original, requer a disponibilidade de dados de presença e ausência das espécies, além da conversão das projeções contínuas dos modelos em projeções binárias de presença/ausência (Fielding & Bell, 1997). Com exceção das métricas propostas por Boyce et al. (2002) e Scherrer et al. (2020), todas as outras métricas de avaliação dependem da utilização de pelo menos um componente da CM para serem calculadas (Allouche et al., 2006; Fielding & Bell, 1997; Wunderlich et al., 2019). Poucas métricas foram desenvolvidas para usar apenas dados de presença (Leroy et al., 2018; Liu et al., 2013; Márcia Barbosa et al., 2013; Pearson et al., 2007) e dados de ausência são raramente disponíveis e podem não ser confiáveis (Wisz & Guisan, 2009).

Obter ausências confiáveis requer conhecimento completo do nicho fundamental da espécie para identificar inequivocamente os locais onde ela está ausente e isso é não aplicável metodologicamente, e em geral, dados subamostrados são utilizados (Bini et al., 2006). Nesse contexto, existem dois cenários para a avaliação dos modelos: 1) o uso de pseudo-ausências, que são pontos (aleatórios/estruturados) amostrados na área de estudo para substituir as supostas ausências reais das espécies (Guisan et al., 2002); 2) o uso de métricas que não são focadas em dados de ausência (Leroy et al., 2018; Liu et al., 2013; Márcia Barbosa et al., 2013; Pearson et al., 2007).

O primeiro cenário, amplamente utilizado, pode levar a problemas na avaliação dos modelos, uma vez que todas as etapas na seleção e uso de pseudo-ausências afetam o desempenho dos modelos e, consequentemente, das métricas utilizadas para avaliá-los, podendo levar a interpretações equivocadas devido à superestimação da capacidade preditiva dos modelos (Iturbide et al., 2018; Leroy et al., 2018; Wisz & Guisan, 2009). Adicionalmente às questões metodológicas da utilização das pseudo-ausências, estão as questões conceituais, pois elas podem deturpar ou inflar valores das métricas de avaliação, levando a utilização de modelos avaliados como "bons modelos" que podem apenas acertaram dados artificiais de pseudo-ausências (Golicher et al., 2012).

Por outro lado, o segundo cenário é uma alternativa desejável, uma vez que identificar ausências reais para espécies não é uma tarefa simples. Em um cenário de "mundo real", é desejável desenvolver e utilizar métricas que não dependam de ausências, em contraste com aquelas que dependem dessas ausências/pseudo-ausências, pois dados de presenças já estão amplamente disponíveis (Leroy et al., 2018; Liu et al., 2013; Pearson et al., 2007). Esses autores apresentaram novas métricas de avaliação que, embora dependam da CM, não dependem ou não são focadas em dados de ausências para serem computadas. No entanto, Pearson et al. (2007) propuseram uma métrica que considera apenas a presença, restrita a situações em que há poucos pontos de ocorrência da espécie disponíveis. Já Leroy et al. (2018) e Liu et al. (2013) não compararam o desempenho das métricas que dependem de ausências reais, tampouco investigaram o desempenho das métricas que não dependem de pseudo-ausências ao avaliar modelos de presença-ausência que foram calibrados e avaliados utilizando pseudo-ausências.

Utilizamos a abordagem de Espécies Virtuais (EV) (Hirzel & Guisan, 2002; Miller, 2014) para investigar os efeitos do viés de coleta nos modelos e métricas de avaliação. Esta

abordagem permite que controlemos todas as condições que determinam a distribuição e é a forma mais adequada de testar os métodos da ENM (Leroy et al., 2018; Meynard et al., 2019). Assim, nossas perguntas foram: 1) O uso de pseudo-ausências afeta negativamente as métricas que requerem ausências na calibração e avaliação de modelos (presença-ausência)? 2) As métricas que dependem apenas das presenças são alternativas robustas para avaliar modelos com dados apenas de presenças (presença-apenas)? 3) Métricas de apenas presença têm sua capacidade avaliativa reduzida ao avaliar modelos que dependem de ausências reais calibrados com pseudo-ausências? As pseudo-ausências podem reduzir a capacidade das métricas de avaliar corretamente os modelos, inflando artificialmente seus valores e não refletindo a alta capacidade preditiva (Golicher et al., 2012; Leroy et al., 2018). Esperávamos encontrar modelos com projeções menos acuradas quando pseudo-ausências forem utilizadas (Warren et al., 2008). Além disso, esperávamos que as métricas baseadas apenas nas presenças não sejam inferiores às métricas que dependem de ausências quando dados de ausências não estão disponíveis (Lobo et al., 2010). Por fim, antecipamos que as métricas apenas presença também tenham sua capacidade avaliativa reduzida devido ao impacto negativo das pseudo-ausências no desempenho dos modelos e, consequentemente, na avaliação desses modelos (Chefaoui & Lobo, 2008; Golicher et al., 2012; Lobo & Tognelli, 2011). Ao responder a essas questões, evidenciamos as problemáticas da utilização de pseudo-ausências na avaliação de modelos de distribuição de espécies (ENM e SDM). Nosso objetivo é apresentar alternativas (métricas) conceitualmente e metodologicamente robustas para avaliar esses modelos, eliminando a dependência de ausências ou pseudo-ausências.

2. Material e Métodos

Área de Estudo

Utilizamos como área de estudo a região da Pan-Amazônia localizada no norte da América do Sul (Figura 1). Optamos por uma região com ocorrência natural para incorporarmos realismos às nossas análises. Embora não tenhamos trabalhado com espécies reais, a utilização de uma área de estudo real não interfere nos resultados do presente estudo e evita que tenhamos que simular todos os componentes de uma área de estudo, sem que tivéssemos qualquer ganho na robustez dos resultados.

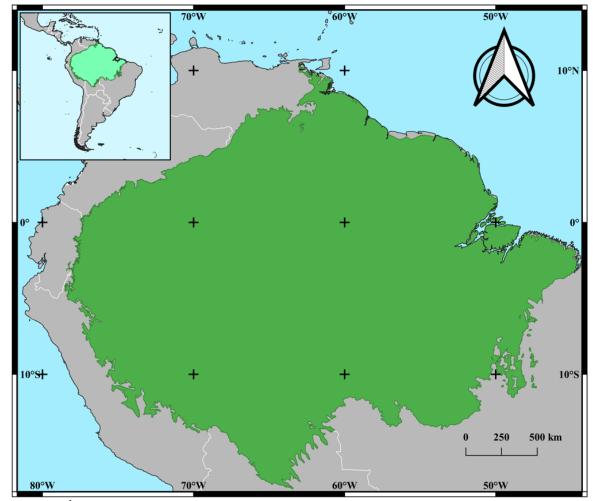


Figura 1: Área da região amazônica panamericana utilizada para a realização do presente estudo.

Variáveis Ambientais

Utilizamos as 19 variáveis bioclimáticas (BIO) da base de dados *WorldClim 2.0* (Fick & Hijmans, 2017) e 3 variáveis de topografia: elevação, obtida da base de dados *EarthEnv* (Amatulli et al., 2018); declividade e orientação de vertente, obtidas pela função *terrain* do pacote *raster* do R (Hijmans, 2018; R Core Team, 2018). Todas as variáveis foram recortadas para a região da Pan-Amazônia e aquelas que tinham resolução superior foram reamostradas para a resolução de 5 minutos de arco (~10 km). A reamostragem foi feita combinando múltiplas células menores para formar uma célula maior com ~ 10 km e tiramos a média dos valores das células menores para compor o valor da célula maior.

Espécies Virtuais

Geramos cinco espécies virtuais (VS) sobre as 22 variáveis ambientais de clima e topografía utilizando a função *generateSpFromPCA* disponível em ambiente R por meio do pacote *virtualspecies* (Leroy et al., 2016). Essa função permite criar o "nicho fundamental" de uma espécie baseado em uma Análise de Componentes Principais – PCA, de tal forma que é possível escolher os limites de tolerância dessa espécie no espaço ambiental e, então, projetar seu nicho no espaço geográfico para gerar a distribuição da espécie na área de estudo. Inicialmente é criado um gradiente de adequabilidade para a VS (contínuo) que varia de 0 a 1. Os mapas contínuos foram convertidos em mapas binários (0 ou 1), em que valores 0 indicam ausência da espécie e valores 1 indicam presença da espécie. A binarização foi feita utilizando a abordagem *threshold* por meio da função *convertToPA* do pacote *virtualspecies*.

Convertemos os mapas contínuos em binários utilizando a proporção de células da área de estudo preditas como presenças (prevalência). Esse critério permite que escolhamos a proporção de células ocupadas pela VS (prevalência) e então a função convertToPA escolhe um threshold atenda à essa delimitação da prevalência escolhida. Nesse estudo criamos distribuições com quatro amplitudes para nossas VS, sendo selecionadas 10%, 15%, 20% e 25% de prevalência, pois dessa forma teremos distribuições variando de restrito (10% de prevalência) até uma espécie ampla (25% de prevalência). Após termos mapas binários com todas as prevalências citadas anteriormente, utilizamos a função sampleOccurrences, também do pacote virtualspecies, para amostrar aleatoriamente pontos de "presença" e "ausência" sobre cada um dos cinco mapas binários. Nós amostramos 50, 75 e 100 registros de ocorrências e ausências para cada prevalência, pois o número reduzido ou elevado de registros pode reduzir a efetividade das métricas (Jiménez-Valverde, 2020; Liu et al., 2019).

Os procedimentos de binarização e amostragem dos registros de presença e ausência foram repetidos 10 vezes para minimizar efeitos de estruturação no espaço geográfico proveniente da aleatorização. Utilizamos esses pontos de presenças e ausências nos procedimentos de modelagem descritos a seguir.

Modelagem de nicho e Distribuição de espécies

Obtivemos dois conjuntos de dados que foram utilizados para treinar os modelos, um composto por presenças e ausências da VS (dados PEA) e outro composto por presenças da VS e utilizando pseudo-ausências para substituir as ausências (dados PPA). Amostramos as pseudo-ausências em número igual ao das presenças utilizando amostragem aleatória na área de estudo, evitando células que já tinham um ponto de presença. Amostramos 10,000 pontos de *background* também pela estratégia aleatória. A amostragem do *background* difere da amostragem de pseudo-ausências, pois pontos de *background* podem cair em células que contêm presenças, mas com pseudo-ausências isso não ocorre..

Na etapa de modelagem utilizamos 15 algoritmos, que incluem aqueles que necessitam de dados de presença e ausência, presença e *background* e apenas presença (Tabela 1 no Apêndice). Todos os algoritmos foram mantidos na sua configuração padrão (*default*). Para cada algoritmo, executamos cem repetições do procedimento de modelagem para cada número de presença, para cada uma das dez estratégias de amostragem, totalizando 15,000 modelos para cada VS. Para cada repetição de modelagem os dados de presença e ausência foram divididos em dados de teste e treino, conforme descrito na seção abaixo. Com os modelos calibrados, nós os projetamos no espaço geográfico para obtermos mapas contínuos de adequabilidade com a provável distribuição das espécies preditas pelos modelos.

Avaliação dos Modelos

Após obtermos os mapas com as predições contínuas, avaliamos os modelos com nove métricas disponíveis e apresentamos uma nova métrica de avaliação que não depende de ausências reais da espécie modelada para ser computada. Essa nova métrica denomina-se *Pearson's D* e é uma variação da métrica proposta originalmente por Pearson et al. (2007). Em sua proposta original, a métrica *Pearson's D* é restrita a aplicações com poucos registros de ocorrência (< 25), pois é utilizada na abordagem leave-one-out em que um ponto de ocorrência é excluído durante a construção do modelo e a medida de precisão é computada se o modelo acerta ou erra esse único ponto excluído.

$$D = \sum X_i (1 - PR_i)$$
 Equação 1.

Onde X_i corresponte ao sucesso absoluto (acerto do ponto deixado de fora da modelagem) ou fracasso absoluto (erro do ponto deixado de fora da modelagem). PRi corresponde ao número de células da área de estudo preditas como presença da espécie

dividido pelo número total de células da área de estudo (prevalência). Desta forma, a métrica *Pearson's D* computa o sucesso ou fracasso do modelo em questão.

Porém, baseados na CM, podemos computar a taxa de acerto das presenças, conhecida como sensibilidade, ou *true positive rate* - TPR, e substituir o componente Xi por essa taxa de acerto nas presenças. Desta forma, poderemos aplicar a métrica Pearson's D a cenários onde temos mais de 25 registros de ocorrência das espécies. A métrica *Pearson's D* avalia o modelo focando apenas nas presenças da espécie modelada e penaliza o modelo que prediz uma área de ocorrência excessiva (é trivial que o modelo acerte muitas presenças prevendo uma elevada prevalência). Todas as equações e breves descrições das métricas utilizadas aqui estão na Tabela 2.

Tabela 2: Métricas de avaliação analisadas (esquerda externa), bem como seus nomes, fórmulas das métricas (esquerda centro), suas descrições (direita centro) e o tipo da métrica (direita externa). OP são métricas que

dependem apenas de dados de presença e PA são métricas que dependem de presenças e ausências.

Métrica	Equação	Descrição	Tipo
Sensibilidade - <i>TPR</i>	$\frac{a}{a+c}$	<i>a</i> corresponde aos acertos nas presenças e <i>c</i> corresponde aos acertos nas ausências.	OP
Área sob a curva <i>lift</i> (AU <i>l</i>)	$1/n_i n_r \sum_{i=1}^{n_r} \sum_{j=1}^{n_1} \phi(X_i, X_j)$	$\phi(X, Y)$ é igual a 1 se Y>X, 1/2 se Y=X, e 0 caso contrário; X_i e Y_j são o valor previsto para o site de ausência i e site de presença j .	ОР
Pearson's D	TPR(1 - PR)	TPR é a sensibilidade e PR corresponde a proporção de células da área de estudo preditas como presença (Prevalência).	OP
Máxima distância da curva <i>lift</i> (MaxVD <i>l</i>)	max(TPR-PR)	TPR é a sensibilidade e PR corresponde a proporção de células da área de estudo preditas como presença (Prevalência).	OP
F-measure baseado em dados somente	(2*a)/(c+a+co*b) $co = pre/(aus*prev)$	Onde <i>a</i> corresponde aos acertos nas presenças, <i>b</i> são os erros nas presenças e <i>c</i> corresponde aos acertos	ОР

presença		nas ausências. Onde <i>pre</i> é o número de presenças, <i>aus</i> o número de ausências e <i>prev</i> é a prevalência predita.	
Especificidade - TNR	$\frac{b}{b+d}$	b corresponde aos erros nas presenças e d corresponde aos erros nas ausências da matriz de confusão.	PA
True skill statistics (TSS)	(TPR + TNR) - I	TPR corresponde à sensibilidade e TNR corresponde à especificidade.	PA
Symmetric extremal dependence index (SEDI)	$\frac{log(TNR) - log(TPR) - log(1-TNR) + log(1-TPR)}{log(TNR) + log(TPR) + log(1-TNR) + log(1-TPR)}$	TPR é a sensibilidade e TNR a especificidade.	PA
Área sob a curva ROC (AUC)	$1/n_{i} n_{0} \sum_{i=1}^{n_{0}} \sum_{j=1}^{n_{1}} \phi(X_{i}, X_{j})$	$\phi(X, Y)$ é igual a 1 se Y>X, 1/2 se Y=X, e 0 caso contrário; X_i e Y_j são o valor previsto para o site de ausência i e site de presença j .	PA
F-measure	(2*a)/(c+a+b)	a corresponde aos acertos nas presenças, b são os erros nas presenças e c corresponde aos acertos nas ausências.	PA

Os dados de ocorrência e ausência/pseudo-ausências das espécies foram divididos em dois conjuntos: um para calibrar o modelo (conjunto de treino) e outro para avaliar o modelo (conjunto de testes). Para isso adotamos a estratégia *bootstrap* de particionamento dos dados, que foi repetida 100 (cem) vezes, em que 70% dos dados foram utilizados para treinar os modelos e 30% foram utilizados para testar os modelos (Fielding & Bell, 1997).

Dentre as métricas utilizadas apenas duas não dependem da escolha de um limiar (threshold) para serem computadas (AUC e AUl) (Liu et al., 2013, 2005). As demais necessitam da escolha de um threshold para construir a CM (Tabela 1) e consequentemente as taxas de acerto e erros do modelo. Utilizamos o threshold que maximiza a soma dos acertos das presenças e ausências dos modelos (Maximum Sensibility plus Specificity – MaxSpecSens) (Liu et al., 2005). Apresentamos um threshold que maximiza os acertos das presenças (TPR), mas que controle a sobrepredição denominado Threshold Maximum D -

TMD que visa maximizar a métrica que apresentamos como uma alternativa para avaliar os modelos utilizando apenas dados de presenças (*Pearson's D*).

Análises das métricas de avaliação dos modelos

A fim de avaliar se cada métrica de acurácia dos modelos está relacionada à capacidade preditiva, realizamos uma correlação entre os valores da métrica de avaliação e a métrica de sobreposição I (índice I). A métrica de sobreposição I quantifica o quanto dois mapas são semelhantes. Essa métrica varia de 0 a 1, em que o valor 0 representa dissimilaridade completa entre dois mapas e o valor 1 representa similaridade completa (Broennimann et al., 2012; Warren et al., 2008). Portanto, é desejado que uma boa métrica de avaliação espacial dos modelos tenha elevados valores de correlação com a métrica de sobreposição. Essa correlação é útil para determinar se altos valores nas métricas de avaliação confirmam modelos que preveem com precisão a área de distribuição da espécie em questão (alta sobreposição). Devido à falta de normalidade dos dados em todas as métricas e cenários, utilizamos o coeficiente de correlação de *Spearman rho* (ρ). A correlação foi calculada para cada algoritmo de modelagem, considerando todas as combinações possíveis de tipo de dados de ausências, prevalência e número de pontos.

Utilizamos a Análise de Variância Fatorial de Medidas Repetidas presente no pacote 'ez' do Ambiente R sob arcabouço do pacote 'RVAideMemoire' para testar nossas perguntas, modificando apenas os fatores da variável preditora (Hervé, 2022; Lawrence, 2016; R Core Team, 2018). Primeiramente, embora estivéssemos interessados primordialmente em avaliar o efeito da utilização de pseudo-ausências, incluímos em nossa análise os fatores secundários (número de pontos de ocorrência e prevalências). Realizamos 9999 aleatorizações para computar o valor da probabilidade baseado no método de Monte Carlo. Fizemos a comparação par a par para todas as métricas que apresentaram significância para a interação do tipo de dado de ausência com a prevalência, bem como para as métricas que obtiveram significância apenas para o tipo dado isoladamente. Testamos se o resultado da avaliação obtida com dados de ausências reais era superior àquela utilizando pseudo-ausências. Para isso, mantivemos fixos os fatores tipos de dados de ausências, testando a capacidade avaliativa de dados PEA > PPA e fizemos as comparações par a par entre cada prevalência. Este procedimento foi realizado separadamente para os diferentes números de registros de ocorrências. Utilizamos a função

"perm.t.test" do pacote 'RVAideMemoire' do R para fazer todas as comparações par a par deste estudo e modificamos apenas as variáveis preditoras e respostas (Hervé, 2022; Lawrence, 2016; R Core Team, 2018).

Para testar se as métricas de apenas presença são alternativas viáveis, classificamos as métricas em dois grupos, em que as métricas PA correspondem às métricas que são computadas com foco em presenças e ausências reais ou pseudo-ausências e as métricas PO correspondem às métricas que são focadas na avaliação com dados apenas de presenças. Aqui os fatores da Anova foram o tipo de métrica (PA ou PO), a prevalência da espécie e também o número de pontos de ocorrências. A significância estatística foi obtida utilizando 9999 permutações pelo método de Monte Carlo. Realizamos esse procedimento separadamente para os *thresholds MaxSpecSens* e *TMD*.

Realizamos uma ANOVA para verificar o desempenho das métricas apenas presença ao avaliar modelos construídos utilizando pseudo-ausências. Os fatores utilizados para realização dessa análise foram o tipo de dado de ausência utilizado para calibrar os modelos, a prevalência da espécie e o número de pontos de ocorrência. Realizamos as comparações par a par entre o tipo de dado de ausência testando se valores de correlação obtidos por meio de dados PEA eram maiores do que para dados PPA mantendo fixos os valores de prevalência e número de pontos de ocorrência. Realizamos a comparação par a par para verificar entre quais categorias havia diferença. A significância estatística foi obtida por meio de 9999 permutações para a Anova e para a comparação par a par.

Todas as análises de manipulação das variáveis, criação da VS, amostragem dos pontos de ocorrência e ausência, modelagem e avaliação foram conduzidas em ambiente R versão 3.6.3 (R Core Team, 2018). As versões dos pacotes utilizados em todas as etapas deste trabalho estão disponíveis na Tabela 2 do Apêndice.

3. Resultados

Quanto às métricas que dependem de dados de presença e ausência (primeira pergunta), nossos resultados demonstram que há performance superior para as métricas avaliativas dos ENM e SDM quando dados de ausências reais são utilizados para todas as métricas que dependem desse tipo de dado (p-valor < 0.05: Tabela Suplementar 3). As tendências de distribuição dos dados também demonstram isso (Figura 2 e Figuras Suplementares 1 e 2).

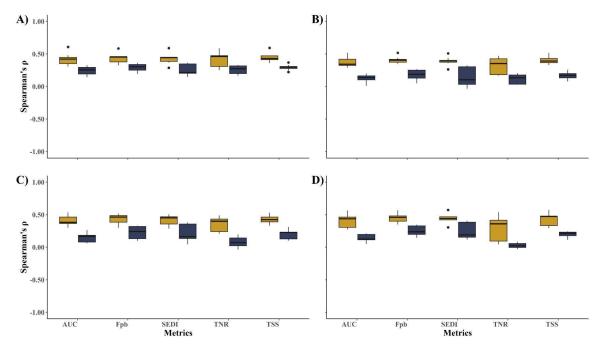


Figura 2: Valores das correlações de Spearman (eixo Y) das métricas de avaliação dos modelos (eixo X) com a métrica que mede a capacidade preditiva real dos modelos (Índice *I*) para modelos gerados com o conjunto de dados de 50 pontos de ocorrência. As figuras obtidas para modelos construídos com 75 e 100 registros de ocorrências estão disponíveis nas Figuras Suplementares 1 e 2. As letras **A** (superior esquerdo), **B** (superior direito), **C** (inferior esquerdo) e **D** (inferior direito) são valores obtidos de espécies com prevalência de 10%, 15%, 20% e 25%, respectivamente. As cores amarelo âmbar e azul marinho correspondem aos valores de correlação da métrica de avaliação proveniente de modelos construídos com dados de ausências reais (dados PEA) e modelos construídos com dados de pseudo-ausências (dados PPA), respectivamente.

A ANOVA revelou que as métricas AUC, Fpb, SEDI e TSS apresentaram diferenças para a interação com a prevalência e para o tipo de dado de ausência isoladamente (Tabelas Suplementares 3, 4, 5 e 7). A métrica TNR apresentou significância apenas para o tipo de dado de ausência e não para a interação (*p-valor* > 0.05 e *p-valor* > 0.05, respectivamente: Tabela Suplementar 6). A comparação para a par demonstrou que dados PEA proporcionam melhor relação entre a capacidade preditiva dos modelos e as métricas de avaliação para todas as prevalências ao utilizarmos 50 registros de ocorrências.

As métricas que dependem de limiar (Fcp, SEDI e TSS) mostraram-se dependentes do tipo de dado de ausência (Figura Suplementar 3 do Apêndice), enquanto ao AUC, que não depende de limiar, mostrou-se igualmente dependente de todos os fatores de modelagem e o componente que mais afetou a relação entre a AUC e a capacidade preditiva, isoladamente, foi o tipo de dado de ausência (Figura Suplementar 3 do Apêndice).

Não encontramos diferenças significativas entre as métricas que dependem de dados de ausências reais e aquelas que não possuem essa dependência (p-valor > 0.05). Há

uma tendência de o grupo de métricas que não depende de ausências reais serem mais eficientes que aquelas que dependem das ausências, especialmente quando analisadas espécies com prevalências menores (10% e 15%), independentemente do número de registros utilizado (Figura 3).

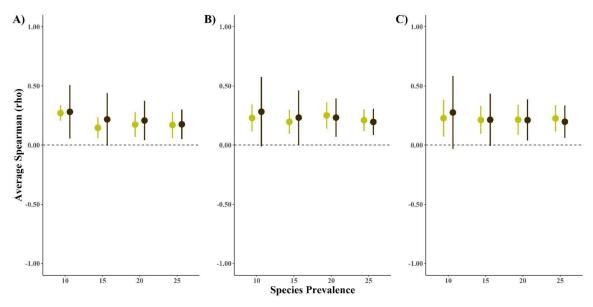


Figura 3: Valores da correlação de *Spearman* (eixo Y) para cada prevalência (eixo X). Pontos correspondem às médias dos dados e as barras são os desvios padrão das médias. Amarelos são os valores obtidos para as métricas que são baseadas em ausências reais e marrons correspondem aos valores obtidos para as métricas que são baseadas em dados de presenças apenas. Em **A**, **B** e **C** temos dados dos modelos construídos com 50, 75 e 100 registros de ocorrência, respectivamente.

Destacamos a métrica *Pearson's D*, que apresenta resultados médios superiores em todos os cenários de números de registros de ocorrências e utilização de *threshold* e *MaxVDl* em alguns cenários (Figura 4). *Pearson's D* e *MaxVDl* apresentam redução na sua capacidade avaliativa ao aumentar a prevalência da espécie (Figura 4).

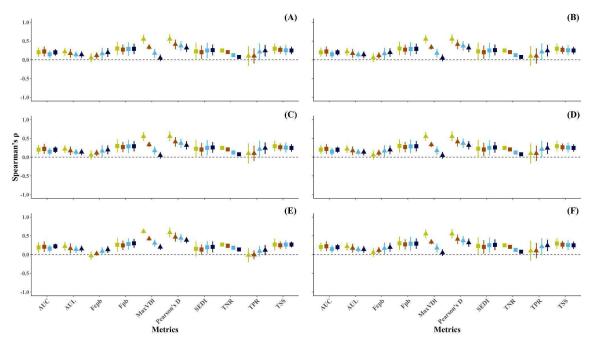


Figura 4: Média das correlações (eixo Y) das métricas de avaliação (eixo x) com a métrica de sobreposição (*I*) para quatro prevalências (10% = amarelo, 15% = marrom, 20% = azul claro e 25% = azul escuro). Quadrados e triângulos correspondem às métricas presença-ausência (PA) e presença-apenas (PO), respectivamente. As barras correspondem aos respectivos desvios padrão. Na coluna da esquerda (**A**, **C** e **E**) temos o *threshold* que maximiza os valores de sensibilidade (TPR) mais especificidade (TNR) denominado *MaxSpecSens* e na coluna da direita (**B**, **D** e **F**) temos o *threshold* que maximiza os valores de sensibilidade (TPR) na menor área predita denominado *Threshold Maximum D - TMD*. Nas linhas temos, de cima para baixo, os valores obtidos com 50, 75 e 100 registros de ocorrências, respectivamente.

Com exceção da métrica TPR, todas as outras métricas baseadas em apenas presenças são afetadas pelo tipo de dado de ausência utilizado para calibrar os modelos (*p-valor* < 0.05) (Figura 5 e Figura Suplementar 5 e 6). Dados de ausências reais produziram modelos que foram melhores avaliados em todos os cenários de prevalência e número de registros de ocorrência para *Pearson's D, MaxVDl* e Fcp (*p-valor* < 0.05). A AU*l* também foi afetada pelo tipo de dado de ausência utilizado para calibrar os modelos, mas o padrão não foi totalmente consistente.

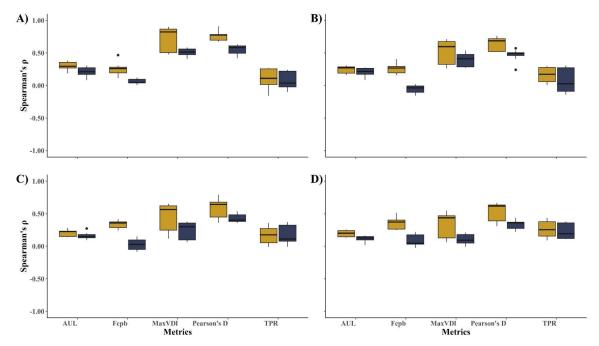


Figura 5: Valores das correlações de Spearman (eixo Y) das métricas de avaliação dos modelos (eixo X) com a métrica que mede a capacidade preditiva real dos modelos (Índice *I*) para modelos gerados com o conjunto de dados de 50 pontos de ocorrência. As figuras obtidas para modelos construídos com 75 e 100 registros de ocorrências estão disponíveis nas Figuras Suplementares 5 e 6. As letras **A** (superior esquerdo), **B** (superior direito), **C** (inferior esquerdo) e **D** (inferior direito) são valores obtidos de espécies com prevalência de 10%, 15%, 20% e 25%, respectivamente. As cores amarelo âmbar e azul-marinho correspondem aos valores de correlação da métrica de avaliação proveniente de modelos construídos com dados de ausências reais (dados PEA) e modelos construídos com dados de pseudo-ausências (dados PPA), respectivamente.

À exceção da TPR, a prevalência foi o componente com maior efeito (Figura Suplementar 5 do Apêndice). A métrica Fcpb foi a única que apresentou significância para a interação entre a prevalência e tipo de dados de ausências; as demais demonstraram significância apenas para o tipo de dados de ausência individualmente.

4. Discussão

As pseudo-ausências afetam negativamente a capacidade das métricas avaliativas dos modelos. Nossos resultados demonstram também que as métricas avaliativas que requerem apenas dados de presenças são superiores às métricas que requerem dados de presença e ausência. Finalmente, demonstramos que a utilização de pseudo-ausências afeta também a capacidade avaliativa das métricas que dependem apenas de presenças, quando pseudo-ausências são empregadas para calibrar os modelos.

A utilização de pseudo-ausências foi apontada como uma estratégia prejudicial na história de utilização dos ENM e SDM, tanto na avaliação, quanto na calibragem (Golicher et al., 2012; Leroy et al., 2018; Lobo & Tognelli, 2011). Nossos resultados demonstraram

que as pseudo-ausências interagem com outros componentes, tais como prevalência e número de pontos de ocorrência, adicionando incertezas à avaliação dos modelos e reduzindo a confiabilidade sobre os resultados, e demonstra que a avaliação dos modelos ainda é um desafio a ser superado (Araújo & Guisan, 2006; Elith & Leathwick, 2009; Liang et al., 2018).

A utilização de pseudo-ausências afeta a capacidade das métricas de avaliação de avaliar corretamente os modelos que dependem de ausências. Isso é particularmente preocupante, uma vez que as métricas de avaliação são atualmente a principal ferramenta para medir a capacidade preditiva dos modelos. Ao utilizar métricas que se concentram em ausências, sem ter dados de ausências reais, corremos o risco de considerar como "bons modelos" aqueles que na verdade não possuem uma capacidade preditiva significativa (Barve et al., 2011; Golicher et al., 2012; Leroy et al., 2018).

O componente "dado de ausência" se mostrou como o fator individual mais importante para quatro das cinco métricas de presença e ausência avaliadas (AUC, Fpb, SEDI e TSS), o que demonstra o potencial impacto negativo que essa abordagem pode ter (Hazen et al., 2021; Stokland et al., 2011; Wisz & Guisan, 2009). Nossos resultados reforçam os argumentos para que pseudo-ausências não sejam utilizados na avaliação dos modelos (Chefaoui & Lobo, 2008; Leroy et al., 2018; Wisz & Guisan, 2009).

As métricas de apenas presença não serem inferiores, quando comparadas às métricas presença-ausência, era algo esperado por nós e isso foi confirmado por nossos resultados. Isso reforça a recomendação sobre a importância de se utilizar dados de ausências reais e que na impossibilidade de utilizar tais dados a utilização de pseudo-ausências na avaliação dos modelos deve ser evitada (Liu et al., 2013; Barbosa et al., 2013). Cuidados com a prevalência devem ser tomados, pois essa propriedade das espécies afeta fortemente o desempenho das métricas de apenas presença, bem como já havia sido notificado para métricas de presença e ausência (Figura 3) (Barve et al., 2011; Iturbide et al., 2018; Jiménez-Valverde, 2021; Leroy et al., 2018). As métricas baseadas em presenças, *MaxVDl* e *Pearson's D*, são superiores, em média, às demais métricas (Figura 4), o que às tornam alternativas conceitualmente e metodologicamente mais robustas (Liu et al., 2013; Pearson et al., 2007).

Ao analisarmos nossa terceira pergunta, observamos que as pseudo-ausências também afetam negativamente a relação entre as métricas focadas nos dados de presença quando elas são utilizadas para avaliar modelos de presença e ausência (Figura 5). Isso acontece porque, embora as métricas em si não sejam severamente afetadas pelas

pseudo-ausências, o desempenho dos modelos é prejudicado. Portanto, a utilização de modelos de presença e ausência se torna problemática quando ausências reais não estão disponíveis, independentemente da métrica de avaliação utilizada. Recomendamos, respectivamente, a utilização de ausências reais e que os modelos sejam avaliados apenas com base em presenças quando ausências reais não estão disponíveis (Leroy et al., 2018; Liu et al., 2013; Pearson et al., 2007).

A prevalência da espécie é apontada como um dos principais fatores que afeta a acurácia dos ENMs e SDMs. O fato de ser ela o componente que mais gerou incerteza na relação da capacidade avaliativa da maioria das métricas focada em apenas presenças (AU*l*, *MaxVDl*, Fcpb e *Pearson's D*) (Figura Suplementar 4) é algo que merece atenção, pois pode afetar a confiabilidade sobre a avaliação dos modelos (Jiménez-Valverde, 2021; Lawson et al., 2014; Vollering et al., 2019). A métrica TPR, curiosamente, foi majoritariamente afetada pelo número de registros de ocorrência. Tal resultado é especialmente problemático, pois os registros dos organismos no mundo real são heterogêneos em quantidade e qualidade e isso pode deturpar as avaliações dos modelos focada nessa métrica (Liu et al., 2019; Ochoa-Ochoa et al., 2016; Watling et al., 2015).

5. Conclusão

A utilização de pseudo-ausências na calibração dos modelos que dependem de ausências é problemática para a capacidade preditiva e avaliativa dos ENMs e SDMs. Assim, não recomendamos a sua utilização para avaliar os modelos que dependem da utilização de ausências reais. Embora as métricas de presença também sejam afetadas negativamente pela utilização de pseudo-ausências, elas dispõem de uma abordagem mais coerente, conceitual e metodologicamente, e são recomendadas quando ausências reais confiáveis não estão disponíveis. Dentre as métricas que não são focadas em ausências reais, destacamos a métrica *Pearson's D* que possui bom desempenho na maioria das condições. Alertamos que todas as métricas de avaliação devem ser usadas com parcimônia, respeitando critérios de boas práticas na construção dos modelos, bem como na interpretação e notificação sobre o desempenho dos modelos, pois mesmo a melhor métrica não será a melhor em todos os cenários. Finalmente, encorajamos que mais estudos sejam realizados sobre os efeitos potenciais de cada fonte de incerteza dos modelos, para que possamos entender melhor como cada um deles afeta nossa capacidade de

6. Referências bibliográficas

- Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x
- Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., Jetz, W., 2018. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific Data 5, 180040. https://doi.org/10.1038/sdata.2018.40
- Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33, 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x
- Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J., Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological Modelling 222, 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011
- Bini, L.M., Diniz-Filho, J.A.F., Rangel, T.F.L.V.B., Bastos, R.P., Pinto, M.P., 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. Diversity https://en.org/10.1111/j.1366-9516.2006.00286.x
- Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecological Modelling 157, 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4
- Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.J., Randin, C., Zimmermann, N.E., Graham, C.H., Guisan, A., 2012. Measuring ecological niche overlap from occurrence and spatial environmental data. Glob. Ecol. Biogeogr. 21, 481–497. https://doi.org/10.1111/j.1466-8238.2011.00698.x
- Capinha, C., Pateiro-López, B., 2014. Predicting species distributions in new areas or time periods with alpha-shapes. Ecological Informatics 24, 231–237. https://doi.org/10.1016/j.ecoinf.2014.06.001
- Castro, M.B., Barbosa, A.C.M.C., Pompeu, P.V., Eisenlohr, P.V., de Assis Pereira, G., Apgaua, D.M.G., Pires-Oliveira, J.C., Barbosa, J.P.R.A.D., Fontes, M.A.L., dos Santos, R.M., Tng, D.Y.P., 2020. Will the emblematic southern conifer Araucaria angustifolia survive to climate change in Brazil? Biodiversity and Conservation 29, 591–607. https://doi.org/10.1007/s10531-019-01900-x
- Chefaoui, R.M., Lobo, J.M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling 210, 478–486. https://doi.org/10.1016/j.ecolmodel.2007.08.010
- 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, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159
- Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37,

- 4302–4315. https://doi.org/10.1002/joc.5086
- Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38–49. https://doi.org/10.1017/S0376892997000088
- Golicher, D., Ford, A., Cayuela, L., Newton, A., 2012. Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve. International Journal of Geographical Information Science 26, 2049–2063. https://doi.org/10.1080/13658816.2012.719626
- Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1
- Hazen, E.L., Abrahms, B., Brodie, S., Carroll, G., Welch, H., Bograd, S.J., 2021. Where did they not go? Considerations for generating pseudo-absences for telemetry-based habitat models. Mov Ecol 9, 5. https://doi.org/10.1186/s40462-021-00240-2
- Hervé, M., 2022. RVAideMemoire: Testing and Plotting Procedures for Biostatistics.
- Hijmans, R.J., 2018. raster: Geographic Data Analysis and Modeling.
- Hirzel, A.H., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling 157, 331–341. https://doi.org/10.1016/S0304-3800(02)00203-X
- Iturbide, M., Bedia, J., Gutiérrez, J.M., 2018. Background sampling and transferability of species distribution model ensembles under climate change. Global and Planetary Change 166, 19–29. https://doi.org/10.1016/j.gloplacha.2018.03.008
- Jiménez-Valverde, A., 2021. Prevalence affects the evaluation of discrimination capacity in presence-absence species distribution models. Biodiversity and Conservation 30, 1331–1340. https://doi.org/10.1007/s10531-021-02144-4
- Jiménez-Valverde, A., 2020. Sample size for the evaluation of presence-absence models. Ecological Indicators 114, 106289. https://doi.org/10.1016/j.ecolind.2020.106289
- Lawrence, M.A., 2016. ez: Easy Analysis and Visualization of Factorial Experiments.
- Lawson, C.R., Hodgson, J.A., Wilson, R.J., Richards, S.A., 2014. Prevalence, thresholds and the performance of presence-absence models. Methods in Ecology and Evolution 5, 54–64. https://doi.org/10.1111/2041-210X.12123
- Leroy, B., Delsol, R., Hugueny, B., Meynard, C.N., Barhoumi, C., Barbet-Massin, M., Bellard, C., 2018. Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. Journal of Biogeography 45, 1994–2002. https://doi.org/10.1111/jbi.13402
- Leroy, B., Meynard, C.N., Bellard, C., Courchamp, F., 2016. virtualspecies, an R package to generate virtual species distributions. Ecography 39, 599–607. https://doi.org/10.1111/ecog.01388
- Liang, W., Papeş, M., Tran, L., Grant, J., Washington-Allen, R., Stewart, S., Wiggins, G., 2018. The effect of pseudo-absence selection method on transferability of species distribution models in the context of non-adaptive niche shift. Ecological Modelling 388, 1–9. https://doi.org/10.1016/j.ecolmodel.2018.09.018
- Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of ocurrence in the prediction of species distributions. Ecography 28, 385–393. https://doi.org/10.1111/j.0906-7590.2005.03957.x
- Liu, C., Newell, G., White, M., 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or

- background sites. Ecography 42, 535-548. https://doi.org/10.1111/ecog.03188
- Liu, C., White, M., Newell, G., Griffioen, P., 2013. Species distribution modelling for conservation planning in Victoria, Australia. Ecological Modelling 249, 68–74. https://doi.org/10.1016/j.ecolmodel.2012.07.003
- Lobo, J.M., Jiménez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33, 103–114. https://doi.org/10.1111/j.1600-0587.2009.06039.x
- Lobo, J.M., Tognelli, M.F., 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. Journal for Nature Conservation 19, 1–7. https://doi.org/10.1016/j.jnc.2010.03.002
- Márcia Barbosa, A., Real, R., Muñoz, A.-R., Brown, J.A., 2013. New measures for assessing model equilibrium and prediction mismatch in species distribution models. Diversity Distrib. 19, 1333–1338. https://doi.org/10.1111/ddi.12100
- Martínez-Meyer, E., Peterson, A.T., 2006. Conservatism of ecological niche characteristics in North American plant species over the Pleistocene-to-Recent transition. Journal of Biogeography 33, 1779–1789. https://doi.org/10.1111/j.1365-2699.2006.01482_33_10.x
- Meynard, C.N., Leroy, B., Kaplan, D.M., 2019. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? Ecography 42, 2021–2036. https://doi.org/10.1111/ecog.04385
- Miller, J.A., 2014. Virtual species distribution models: Using simulated data to evaluate aspects of model performance. Progress in Physical Geography: Earth and Environment 38, 117–128. https://doi.org/10.1177/0309133314521448
- Ochoa-Ochoa, L.M., Flores-Villela, O.A., Bezaury-Creel, J.E., 2016. Using one vs. many, sensitivity and uncertainty analyses of species distribution models with focus on conservation area networks. Ecological Modelling 320, 372–382. https://doi.org/10.1016/j.ecolmodel.2015.10.031
- Pearson, R.G., Raxworthy, C.J., Nakamura, M., Townsend Peterson, A., 2007. Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. Journal of Biogeography 34, 102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x
- Peterson, A.T., 2014. Mapping disease transmission risk: enriching models using biogeography and ecology. JHU Press.
- 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, 275–287. https://doi.org/10.1111/geb.12530
- Prieto-Torres, D.A., Nori, J., Rojas-Soto, O.R., 2018. Identifying priority conservation areas for birds associated to endangered Neotropical dry forests. Biological Conservation 228, 205–214. https://doi.org/10.1016/j.biocon.2018.10.025
- R Core Team, 2018. R: A language and environment for statistical computing.
- Scherrer, D., Mod, H.K., Guisan, A., 2020. How to evaluate community predictions without thresholding? Methods Ecol Evol 11, 51–63. https://doi.org/10.1111/2041-210X.13312
- Sobral-Souza, T., Lima-Ribeiro, M.S., Solferini, V.N., 2015. Biogeography of Neotropical Rainforests: past connections between Amazon and Atlantic Forest detected by ecological niche modeling. Evolutionary Ecology 29, 643–655. https://doi.org/10.1007/s10682-015-9780-9
- Stokland, J.N., Halvorsen, R., Støa, B., 2011. Species distribution modelling-Effect of

- design and sample size of pseudo-absence observations. Ecological Modelling 222, 1800–1809. https://doi.org/10.1016/j.ecolmodel.2011.02.025
- Vollering, J., Halvorsen, R., Auestad, I., Rydgren, K., 2019. Bunching up the background betters bias in species distribution models. Ecography 42, 1717–1727. https://doi.org/10.1111/ecog.04503
- Warren, D.L., Glor, R.E., Turelli, M., 2008. Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. Evolution 62, 2868–2883. https://doi.org/10.1111/j.1558-5646.2008.00482.x
- Watling, J.I., Brandt, L.A., Bucklin, D.N., Fujisaki, I., Mazzotti, F.J., Romañach, S.S., Speroterra, C., 2015. Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. Ecological Modelling 309–310, 48–59. https://doi.org/10.1016/j.ecolmodel.2015.03.017
- Wisz, M.S., Guisan, A., 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecology 9, 1–13. https://doi.org/10.1186/1472-6785-9-8
- Wunderlich, R.F., Lin, Y.-P., Anthony, J., Petway, J.R., 2019. Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models. Nature Conservation 35, 97–116. https://doi.org/10.3897/natureconservation.35.33918

7. APÊNDICE

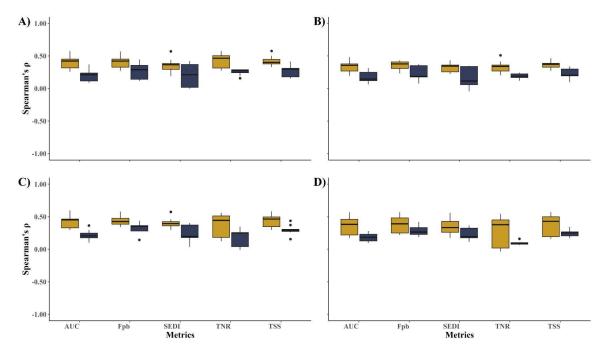


Figura Suplementar 1: Valores das correlações de Spearman (eixo Y) das métricas de avaliação dos modelos (eixo X) com a métrica que mede a capacidade preditiva real dos modelos (Índice *I*) para modelos gerados com o conjunto de dados de 75 pontos de ocorrência. As letras **A** (superior esquerdo), **B** (superior direito), **C** (inferior esquerdo) e **D** (inferior direito) são valores obtidos de espécies com prevalência de 10%, 15%, 20% e 25%, respectivamente. As cores amarelo âmbar e azul marinho correspondem aos valores de correlação da métrica de avaliação proveniente de modelos construídos com dados de ausências reais (dados PEA) e modelos construídos com dados de pseudo-ausências (dados PPA), respectivamente.

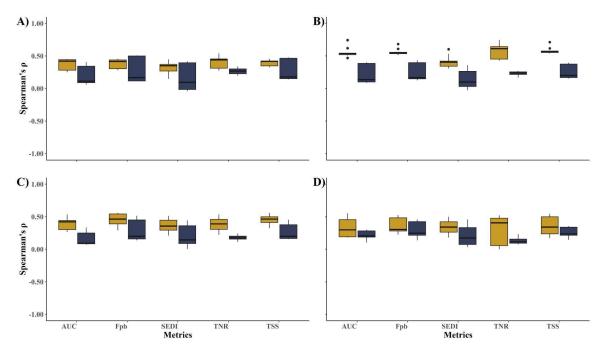


Figura Suplementar 2: Valores das correlações de Spearman (eixo Y) das métricas de avaliação dos modelos (eixo X) com a métrica que mede a capacidade preditiva real dos modelos (Índice *I*) para modelos gerados com o conjunto de dados de 100 pontos de ocorrência. As letras **A** (superior esquerdo), **B** (superior direito), **C** (inferior esquerdo) e **D** (inferior direito) são valores obtidos de espécies com prevalência de 10%, 15%, 20% e 25%, respectivamente. As cores amarelo âmbar e azul marinho correspondem aos valores de correlação da métrica de avaliação proveniente de modelos construídos com dados de ausências reais (dados PEA) e modelos construídos com dados de pseudo-ausências (dados PPA), respectivamente.

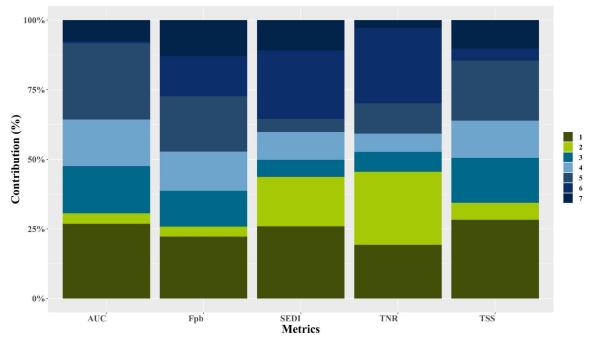


Figura Suplementar 3: Porcentagem de contribuição (eixo Y) de cada componente (Data = 1, NP = 2, NP*Data = 3, NP*Prev = 4, NP*Prev*Data = 5, Prev = 6, Prev*Data = 7) de cada métrica de avaliação (eixo X) para métricas de presença e ausência.

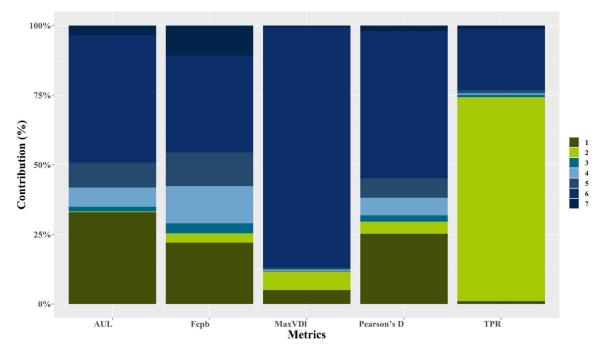


Figura Suplementar 4: Porcentagem de contribuição (eixo Y) de cada componente (Data = 1, NP = 2, NP*Data = 3, NP*Prev = 4, NP*Prev*Data = 5, Prev = 6, Prev*Data = 7) de cada métrica de avaliação (eixo X) para métricas de apenas presença.

Tabela Suplementar 1: Algoritmos utilizados neste trabalho com suas abreviações à esquerda e nomes

completos à direita.

Siglas	Nomenclatura		
glm	Modelo Linear Generalizado		
gam	Modelo Aditivo Generalizado		
mars	Regressão Adaptativa Multivariada Splines		
fda	Análise Discriminante Flexível		
cta	Análise de Árvore de Classificação		
gbm	Modelo de Impulso Generalizado		
nnet	Redes Neurais Artificiais		
rf	Floresta Aleatória		
bayes	Naive Bayes		

Tabela Suplementar 2: Pacotes R utilizados para realização da modelagem (à esquerda), versões dos pacotes

(ao centro) e algoritmo utilizado do referido pacote (a direita).

Pacote	Versão	Algoritmos Utilizados
stats	3.6.3	GLM
nnet	7.3-15	NNET
earth	5.3.0	MARS
randomForest	4.6-14	Random Forest
gam	1.16.1	GAM
mda	0.5-2	FDA
rpart	4.1-15	CTA
gbm	2.1.8	GBM
e1071	1.7-6	Bayes

Tabela Suplementar 3: Tabela da ANOVA fatorial, para a métrica AUC, realizada para avaliar o potencial efeito múltiplos componentes, e das suas interações, sobre a capacidade preditiva dos modelos de nicho

ecológico (ENM) e modelos de distribuição potencial de espécies (SDM). **Effect** corresponde ao componente avalia, **DFn** são os graus de liberdade no numerador, **DFd** são os graus de liberdade no denominador, **F** é o valor da estatística e **perm p** corresponde ao valor-*p* obtido por 9999 permutações.

Effect	DFn	DFd	F	perm_p
(Intercept)	1	15	238.37	1.00
NP	3	45	21.50	0.00
Prev	3	45	1.35	0.27
Data	1	15	52.66	0.00
NP:Prev	9	135	1.31	0.24
NP:Data	3	45	3.00	0.04
Prev:Data	3	45	15.43	0.00
NP:Prev:Data	9	135	1.76	0.08

(**Data** = Tipo de dado de ausência; **Prev** = Prevalência real da Espécie Virtual – VS; **NP** = Número de registro utilizados).

Tabela Suplementar 4: Tabela da ANOVA fatorial, para a métrica **Fcb**, realizada para avaliar o potencial efeito múltiplos componentes, e das suas interações, sobre a capacidade preditiva dos modelos de nicho ecológico (ENM) e modelos de distribuição potencial de espécies (SDM). **Effect** corresponde ao componente avalia, **DFn** são os graus de liberdade no numerador, **DFd** são os graus de liberdade no denominador, **F** é o valor da estatística e **perm p** corresponde ao valor-*p* obtido por 9999 permutações.

Effect	DFn	DFd	F	perm_p
(Intercept)	1	15	245.15	1.00
NP	3	45	34.19	0.00
Prev	3	45	5.05	0.00
Data	1	15	34.91	0.00
NP:Prev	9	135	4.41	0.00
NP:Data	3	45	0.52	0.66
Prev:Data	3	45	32.80	0.00
NP:Prev:Data	9	135	2.34	0.02

(**Data** = Tipo de dado de ausência; **Prev** = Prevalência real da Espécie Virtual – VS; **NP** = Número de registro utilizados).

Tabela Suplementar 5: Tabela da ANOVA fatorial, para a métrica **SEDI**, realizada para avaliar o potencial efeito múltiplos componentes, e das suas interações, sobre a capacidade preditiva dos modelos de nicho ecológico (ENM) e modelos de distribuição potencial de espécies (SDM). **Effect** corresponde ao componente avalia, **DFn** são os graus de liberdade no numerador, **DFd** são os graus de liberdade no denominador, **F** é o valor da estatística e **perm_p** corresponde ao valor-*p* obtido por 9999 permutações.

Effect	DFn	DFd	F	perm_p
(Intercept)	1	15	56.25	1.00
NP	3	45	31.58	0.00
Prev	3	45	9.16	0.00
Data	1	15	116.21	0.00
NP:Prev	9	135	20.34	0.00
NP:Data	3	45	8.55	0.00
Prev:Data	3	45	14.85	0.00
NP·Prev·Data	9	135	4 99	0.00

(**Data** = Tipo de dado de ausência; **Prev** = Prevalência real da Espécie Virtual – VS; **NP** = Número de registro utilizados).

Tabela Suplementar 6: Tabela da ANOVA fatorial, para a métrica TNR, realizada para avaliar o potencial

efeito múltiplos componentes, e das suas interações, sobre a capacidade preditiva dos modelos de nicho ecológico (ENM) e modelos de distribuição potencial de espécies (SDM). **Effect** corresponde ao componente avalia, **DFn** são os graus de liberdade no numerador, **DFd** são os graus de liberdade no denominador, **F** é o valor da estatística e **perm p** corresponde ao valor-*p* obtido por 9999 permutações.

Effect	DFn	DFd	F	perm_p
(Intercept)	1	15	85.40	1.00
NP	3	45	57.71	0.00
Prev	3	45	63.30	0.00
Data	1	15	166.47	0.00
NP:Prev	9	135	2.53	0.01
NP:Data	3	45	0.69	0.57
Prev:Data	3	45	78.46	0.00
NP:Prev:Data	9	135	4.10	0.00

(**Data** = Tipo de dado de ausência; **Prev** = Prevalência real da Espécie Virtual – VS; **NP** = Número de registro utilizados).

Tabela Suplementar 7: Tabela da ANOVA fatorial, para a métrica **TSS**, realizada para avaliar o potencial efeito múltiplos componentes, e das suas interações, sobre a capacidade preditiva dos modelos de nicho ecológico (ENM) e modelos de distribuição potencial de espécies (SDM). **Effect** corresponde ao componente avalia, **DFn** são os graus de liberdade no numerador, **DFd** são os graus de liberdade no denominador, **F** é o valor da estatística e **perm p** corresponde ao valor-*p* obtido por 9999 permutações.

Effect	DFn	DFd	\mathbf{F}	perm_p
(Intercept)	1	15	320.92	1.00
NP	3	45	41.90	0.00
Prev	3	45	3.05	0.04
Data	1	15	69.15	0.00
NP:Prev	9	135	3.40	0.00
NP:Data	3	45	0.21	0.89
Prev:Data	3	45	56.19	0.00
NP:Prev:Data	9	135	3.97	0.00

(**Data** = Tipo de dado de ausência; **Prev** = Prevalência real da Espécie Virtual – VS; **NP** = Número de registro utilizados).

50

Capítulo 2: VIÉS DE AMOSTRAGEM, PSEUDO-AUSÊNCIAS E CAPACIDADE PREDITIVA: O QUE EMERGE DESSA COMBINAÇÃO?

João Carlos Pires-Oliveira¹, Hugo Bampi², Matheus S. Lima-Ribeiro³, Pedro V. Eisenlohr⁴

Programa de Pós-Graduação em Ecologia e Conservação, Universidade do Estado de Mato Grosso, Brasil
 Programa de Pós-Graduação Ecologia e Evolução, Universidade Federal de Goiás, Brasil
 Universidade Federal de Jataí, Brasil
 Universidade do Estado de Mato Grosso, Campus Alta Floresta, Brasil

Resumo

A Modelagem de Nicho Ecológico (ENM) e Modelagem Distribuição Potencial de Espécies (SDM) tornou-se uma ferramenta poderosa nos variados campos da biologia, biogeografía, paleoecologia e conservação da biodiversidade. O desenvolvimento e aprimoramento das bases de dados online contribuíram muito para sua ascensão. Os ENM e SDM são avaliados por métricas que contabilizam erros e acertos dos modelos ao predizer locais com presenças e ausências das espécies. Entretanto, dados de ocorrência de bases de dados online podem conter viés de amostragem, com locais de acesso mais fácil sendo mais amostrados. O viés de amostragem é apontado como problemático para o desempenho dos modelos. Métodos foram desenvolvidos para reduzir os efeitos dos viés de amostragem sem que seus efeitos sobre a capacidade preditiva dos modelos fossem explorados, tampouco seus efeitos sobre as métricas avaliativas. Nós avaliamos, com espécies virtuais (VS), os efeitos do viés de amostragem sobre a relação das métricas avaliativas com a capacidade preditiva dos modelos. Nossos resultados demonstraram que o viés de amostragem reduz a capacidade preditiva dos modelos. Não encontramos um padrão consistente do efeito do viés de amostragem sobre a relação entre as métricas de avaliação e a capacidade preditiva dos modelos. O tamanho do efeito variou entre intensidades de viés. A utilização de pseudo-ausências potencializa o efeito do viés e a filtragem geográfica, por exemplo, pode amenizar tais efeitos. Concluímos que o viés de amostragem é problemático para a capacidade preditiva dos ENM e SDM, principalmente quando utilizamos pseudo-ausências.

Palavras-chave: avaliação de modelos, métricas de avaliação, modelos de nicho ecológico, modelos de distribuição potencial.

ABSTRACT

Ecological Niche Modeling (ENM) and Species Distribution Modeling (SDM) have become powerful tools in various fields of biology for studying the past, present, and future geographic distribution of different taxa. This has been simplified with the widespread availability of online databases. In ENM and SDM, models are typically evaluated using metrics that are computed by binarizing continuous output maps of the models and based on known occurrence/absence records of the species. The confusion matrix (CM) is used to account for the model's accuracy and errors in predicting the presence and absence of the species in different locations. However, occurrence data from online databases may have problems associated with sampling bias, where locations with easier access are more frequently sampled. Methods have been developed to reduce the effects of sampling bias, but the effects of bias on the predictive capacity of the models have not been explored, especially on the effects of bias on evaluative metrics. This evaluation can only be done if we know all the suitable environmental conditions for the species, and this is not possible if we use biased data. We used Virtual Species (VS), for which we know all the suitable conditions, to evaluate the effects of sampling bias on the relationship between evaluative metrics and the predictive capacity of the models. Our results demonstrate that sampling bias is detrimental to the predictive capacity of the models. We did not find a consistent pattern of significance of the sampling bias effect on the relationship between evaluative metrics and overlap metric. The observed effect size varied considerably between bias intensity. Finally, we found that the use of pseudo-absences enhances the effect of sampling bias, which is generally mitigated by reducing bias through geographic filtering, for example. We conclude that sampling bias is a problem for the predictive capacity of niche and species distribution models, especially when real absence data are not available.

Keywords: evaluation metrics, ecological niche models, model evaluation, potential distribution models.

Introdução

A Modelagem de Nicho Ecológico e Modelagem Distribuição Potencial de Espécies (ENM e SDM) tornou-se uma ferramenta poderosa nos variados campos da biologia, em estudos para compreendermos a distribuição geográfica passada, presente e futura dos mais variados táxons (Gomes et al., 2020; Nogués-Bravo et al., 2008; Siqueira et al., 2009). Correlacionar dados de ocorrência de uma espécie com variáveis das localidades onde essas espécies ocorrem tornam relativamente simples a aplicação da ENM e SDM (Elith & Franklin, 2013). Entretanto, os dados de ocorrência disponíveis nas diversas bases de dados *online* podem conter vieses de identificação e geolocalização, cujos quais podem causar prejuízos na avaliação e na capacidade preditiva dos modelos (Beck et al., 2014; Halvorsen et al., 2016; Hortal et al., 2008; Leitão et al., 2011).

A utilização de dados de ocorrência com viés de amostragem pode levar a estimativas imprecisas do nicho e da distribuição das espécies, comprometendo a confiabilidade dos resultados dos modelos na ENM e SDM (Dormann, 2007; Merckx et al., 2011; Segurado et al., 2006). Diversos métodos têm sido propostos para minimizar esses efeitos, como a filtragem ambiental e espacial dos dados (Aiello-Lammens et al., 2015; Boria et al., 2014; Varela et al., 2014; Zizka et al., 2019). No entanto, a produção de métodos para "correção do viés" tem levado geralmente à exclusão de dados de ocorrências (Aiello-Lammens et al., 2015; Nolan et al., 2022; Robertson et al., 2016; Zizka et al., 2019), mas questões sobre os possíveis efeitos do viés sobre as métricas de avaliação e a capacidade preditiva dos modelos não foram completamente elucidadas (Hortal et al., 2008). Note que dados de espécies reais foram utilizados para estudar os efeitos do viés de amostragem, mas essa abordagem é a forma mais precisa e adequada de avaliar questões metodológicas na ENM e SDM, uma vez que não temos conhecimento real de toda a distribuição de uma espécie (Colwell & Rangel, 2009; Costa et al., 2010; Meynard & Kaplan, 2013).

Os métodos de correção do viés baseados em filtragem geralmente diminuem o tamanho do conjunto de dados (exclusão de registros) e assumem que um conjunto de dados reduzido e com menos vieses melhora a capacidade preditiva dos modelos (Aiello-Lammens et al., 2015; Varela et al., 2014). A filtragem pode reduzir o efeito do viés de amostragem, mas não é uma boa alternativa quando dados de ocorrências são escassos e, além disso, não informam sobre os potenciais efeitos negativos da combinação entre viés e quantidade de registros disponíveis (Baker et al., 2022; Boria et al., 2014; Jiménez-Valverde, 2020; Nolan et al., 2022). O número de registros reduzido é uma fonte de incerteza reconhecida para a ENM

e SDM, podendo gerar estimativas imprecisas da distribuição e do nicho das espécies e é recomendado utilizar a completude dos pontos de ocorrência das espécies (Bean et al., 2012; Jiménez-Valverde, 2020; Moudrý & Šímová, 2012).

A disponibilidade de ausências reais das espécies é rara e pseudo-ausências foram apresentadas como um artificio metodológico para superar essa limitação e utilizar todos os elementos da matriz de confusão (CM) (Fielding & Bell, 1997, p. 7; Golicher et al., 2012; Guisan et al., 2002). Entretanto, a utilização de pseudo-ausências foi bastante criticada, pois são apontadas como fonte de grande incerteza prejudicial sobre a avaliação e sobre a capacidade preditiva dos modelos (Chefaoui & Lobo, 2008; Jarnevich et al., 2017; Leroy et al., 2018). Portanto, uma abordagem focada na não utilização de pseudo-ausências já foi defendida como estratégia para minimizar seus efeitos negativos sobre o desempenho dos modelos (Golicher et al., 2012; Liu et al., 2013; Pearson et al., 2007).

Uma abordagem com Espécies espécies Vvirtuais (EVs) é necessária para investigar os efeitos do viés de amostragem sobre a capacidade preditiva dos modelos (habilidade do modelo de prever corretamente a distribuição de uma espécie) e a possível interferência na relação das métricas avaliativas com a capacidade preditiva dos modelos (capacidade avaliativa) (Meynard et al., 2019; Meynard and Kaplan, 2013; Miller, 2014). Essa abordagem permite controlar outras fontes de incerteza, como as limitações no número de registros de ocorrência e as incertezas associadas ao tipo de dado de ausência utilizado (Golicher et al., 2012; Jiménez-Valverde, 2020; Leroy et al., 2018; Liu et al., 2019). Utilizamos EVs para responder a três questões: (1) se a intensidade do viés de amostragem afeta negativamente a capacidade preditiva dos modelos, (2) se a força do viés de amostragem interfere negativamente na relação das métricas de avaliação com a capacidade preditiva do modelo e (3) se a utilização de pseudo-ausências aumenta o efeito do viés de amostragem sobre a capacidade preditiva do modelo. O objetivo é entender como a projeção espacial dos modelos representa corretamente a real distribuição da espécie e como isso é refletido nas métricas de avaliação.

MATERIAL E MÉTODOS

Área de Estudo

A área de estudo considerada abrange todo o território continental brasileiro, que tem

uma extensão de 8.516.000 km2 e se estende em latitude de -73.99° a -28.84° e em longitude de -84.00° a 5.27°. Essa escolha foi feita para adicionar realismo aos modelos, já que as condições ambientais nessa região são heterogêneas, o que é comum em procedimentos de modelagem de espécies reais.

Variáveis Ambientais

Para este estudo, foram utilizadas 22 variáveis ambientais, das quais 19 eram relacionadas ao clima e três possuíam informações sobre relevo e topografia (ver Tabela Suplementar 3 do Apêndice). As variáveis climáticas foram obtidas a partir da base de dados *WorldClim* (Fick e Hijmans, 2017). Já as variáveis de relevo e topografia foram obtidas por meio da variável elevação da base de dados *SoilGrids* (Hengl et al., 2014) e das variáveis *aspect* e *slope* utilizando a função terrain do pacote 'raster' do R (Hijmans, 2018). Todas as variáveis foram recortadas para o território brasileiro e reamostradas para uma resolução de 5' de arco (cerca de 10 km).

Espécies Virtuais e Amostragens das Ocorrências

Utilizamos o *framework* do pacote 'virtualspecies' do R (Leroy et al., 2016) para criar cinco distribuições das EVs em diferentes regiões do Brasil (Figura 1). As distribuições foram geradas sobre os três primeiros eixos de uma análise de componentes principais (PCA) que utilizou as 22 variáveis ambientais mencionadas anteriormente. Para a criação das EVs, utilizamos a função "*generateSpFromPCA*" do pacote 'virtualspecies' e seguimos as recomendações para avaliação metodológica na ENM e SDM (Meynard et al., 2019). Como os preditores no processo de modelagem, retivemos os três eixos da PCA utilizados para gerar as EVs, para garantir que as variáveis que descrevem a distribuição da EV seriam as mesmas utilizadas pelos modelos.

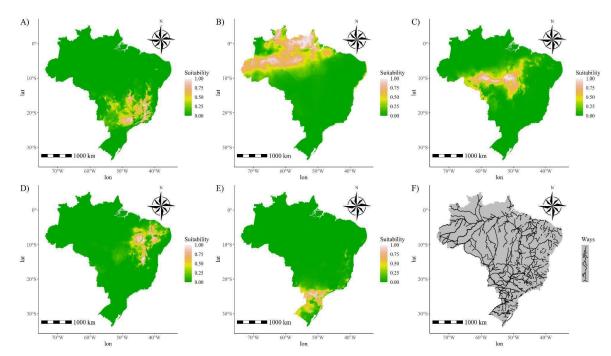


Figura 1: Distribuições contínuas das cincos Espécies Virtuais (EV) (A, B, C, D e E) e mapa do Brasil com suas vias - rodovias/ferrovias/hidrovias (linhas pretas) utilizadas para enviesar a amostragem de dados de ocorrência/ausência (F).

Após gerar cinco mapas de distribuição contínua de espécies virtuais (EV), os convertemos em mapas binários (0 ou 1), em que 1 indica a presença da EV e 0 indica a ausência. Em seguida, amostramos registros reais de ocorrência e ausência das EV para serem utilizados na etapa de modelagem, realizando amostras com quatro diferentes tamanhos (n = 25, n = 50, n = 100 e n = 200) (Liu et al., 2019; Proosdij et al., 2016). Para incorporar o efeito do viés de amostragem no procedimento de amostragem, utilizamos a malha viária (rodovias e hidrovias federais) disponível em https://ide.geobases.es.gov.br/ e https://www.gov.br/infraestrutura/pt-br/ (Figura 1-F). Criamos um buffer de 5 km de cada lado das vias e obtivemos as amostras enviesadas dentro desse buffer. Utilizamos o shape com as rodovias e hidrovias como "máscara" no processo de amostragem das ocorrências/ausências, aumentando a probabilidade de amostrar uma ocorrência/ausência nas proximidades das rodovias e hidrovias (Figura 2).

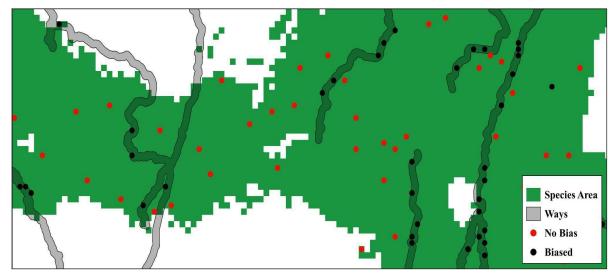


Figura 2: Representação esquemática das ocorrências enviesadas (pontos pretos), que têm maior probabilidade de serem amostradas próximas de rodovias (área sombreada), e das ocorrências não enviesadas (pontos vermelhos). Somente as ocorrências são mostradas, pois todos os pontos estão distribuídos sobre a área de distribuição da Espécie Virtual (área verde).

Utilizamos a função "sampleOccurrences" do pacote 'virtualspecies' para amostrar os registros de ocorrência/ausência, inserindo quatro níveis de viés no procedimento. Aumentamos a força do viés progressivamente, utilizando X0, X10, X100 e X1000, correspondendo a 0, 10, 100 e 1000 vezes mais chances de obter uma ocorrência/ausência próxima a um trecho da malha viária, respectivamente. Na etapa de modelagem seguinte, utilizamos apenas os registros de ocorrência das EV. As ausências reais das EV foram amostradas sob as mesmas intensidades de viés de amostragem das presenças.

Modelagem e Projeções

Na construção dos modelos, utilizamos 15 algoritmos amplamente utilizados (consultar Tabela Suplementar 2 do Apêndice). Mantivemos as configurações padrão dos pacotes nas versões informadas na Tabela Suplementar 2 do Apêndice para todos os algoritmos. Os algoritmos baseados em regressão (Modelos Lineares Generalizados - GLM, Modelo Aditivo Generalizado - GAM, Splines de Regressão Adaptativa Multivariada - MARS e Modelo de Impulso Generalizado - GBM) e aprendizado de máquina (Redes Neurais Artificiais - ANN, Random Forest - RF e *Naive Bayes*) requerem dados de presença e ausência das espécies modeladas. Por isso, elaboramos duas estratégias de modelagem. Na primeira abordagem, utilizamos dados reais de presença e ausência da EV modelada (dados PEA). Na segunda abordagem, calibramos e avaliamos os modelos utilizando dados de presença e pseudo-ausência (dados PPA) amostrados aleatoriamente em toda a área de estudo. O número

de pseudo-ausências amostradas foi sempre igual ao número de presenças utilizadas. Para os algoritmos que requerem background, como MaxEnt e SVM, amostramos 10.000 pontos aleatórios em toda a área de estudo. Os demais algoritmos foram *Bioclim*, *Domain* e Análise de Fator do Nicho Ecológico - ENFA.

Realizamos a modelagem para cada combinação de tipos de dados de ausência, números de registros das espécies e intensidade de viés, já que cada conjunto de dados (PEA e PPA) possui variados números de registros de presenças e ausências que foram amostrados com diferentes níveis de viés. Para isso, particionamos os registros de presença e ausência/pseudo-ausência em dois subconjuntos, sendo um com 70% dos dados de presença e ausências/pseudo-ausências e o segundo conjunto com os 30% remanescentes, procedimento amplamente utilizado na literatura (Elith e Franklin, 2013; Fielding e Bell, 1997; Guisan et al., 2017). Este particionamento foi repetido 100 vezes e, para cada partição dos dados, realizamos a modelagem para todos os algoritmos, utilizando os mesmos conjuntos de dados de treino para calibrar todos os modelos dos diferentes algoritmos e o mesmo conjunto de dados de teste para avaliar tais modelos.

Métricas de Avaliação e Sobreposição

Para avaliar o desempenho dos modelos, utilizamos dez métricas de avaliação descritas na Tabela Suplementar 1 do Apêndice. Essas métricas são baseadas na Matriz de Confusão (CM) e medem a proporção de acertos e erros dos modelos na predição dos dados de teste. Para avaliar a capacidade preditiva dos modelos, utilizamos uma métrica de sobreposição proposta por Warren et al. (2008), baseada no índice I. Essa métrica foi empregada para avaliar a habilidade dos modelos em prever corretamente a área de distribuição da espécie virtual, variando de 0 a 1 e medindo a similaridade entre dois mapas (Warren et al., 2008).

Análise de dados

Após treinarmos e avaliarmos os modelos, geramos valores das métricas de avaliação e da métrica de sobreposição para dados PEA e PPA, considerando diferentes intensidades de viés e números de registros de ocorrência/ausências. Em seguida, classificamos esses valores em quatro grupos de acordo com a intensidade do viés presente nos dados utilizados para gerar os modelos. Esses grupos consistem em um grupo controle (C = viés X0) e três

tratamentos (T1 = viés X10, T2 = viés X100, T3 = viés X1000).

Analisamos os efeitos do viés de amostragem sobre a capacidade preditiva dos modelos ao utilizar dados PEA e PPA por meio da métrica de sobreposição *I* (variável resposta), utilizando uma ANOVA fatorial de medidas repetidas. Nesta análise, os fatores foram o tipo de dado de ausência (data), o número de pontos de ocorrências (np) e o viés de amostragem (bias). Essa abordagem estatística foi adotada uma vez que os mesmos dados de ocorrência foram utilizados para gerar cada modelo, em cada repetição (*run*). Computamos o p-valor da ANOVA utilizando 9999 permutações adaptadas do pacote 'RVAideMemoire' do R

Realizamos a comparação par a par apenas para a interação entre os fatores data e bias, em consonância com nossa primeira pergunta, o que nos deu um total de 28 comparações. Aqui utilizamos testes *t* pareado. Obtivemos os p-valores dos testes *t* utilizando 9999 permutações utilizando a função "*perm.t.test*" do pacote 'RVAideMemoire' do R (Hervé, 2022; R Core Team, 2018). Aplicamos a correção de Benjamini & Yekutieli para teste de múltiplas hipóteses, por ser menos conservadora, porém mantendo elevadas taxas de controle do Erro Tipo I (Benjamini & Yekutieli, 2001).

Para avaliar o efeito do viés de amostragem sobre a relação das métricas de avaliação e a capacidade preditiva dos modelos utilizamos o coeficiente de correlação de *Spearman* (*rho*) entre as métricas de avaliação e a métrica de sobreposição como variável resposta para cada algoritmo de modelagem. Submetemos os valores de *rho* de cada métrica de avaliação a uma ANOVA fatorial de medidas repetidas, em que os fatores utilizados foram os mesmos descritos anteriormente. Utilizamos os testes *t* pareados para comparações par a par para identificar diferenças entre pares. Obtivemos o valor de *p* por meio de 999 permutações utilizando a função "*perm.t.test*" do pacote 'RVAideMemoire' do R (Hervé, 2022; R Core Team, 2018). O p-valor também foi corrigido utilizando o método de Benjamini & Yekutieli (Benjamini & Yekutieli, 2001).

Finalmente, avaliamos se o efeito do viés de amostragem seria maior ao utilizar pseudo-ausências comparando os grupos C, T1, T1 e T3 entre os tipos de dados PEA e PPA, utilizando teste t pareado baseado em 9999 permutações utilizando a função "perm.t.test" do pacote 'RVAideMemoire' do R (Hervé, 2022; R Core Team, 2018). A comparação foi feita para cada intensidade de viés e para cada número de pontos de ocorrências. Nesta etapa verificamos se a média dos valores da métrica de sobreposição I para dados PEA seria maior do que a média de dados PPA para cada intensidade de viés, bem como ao utilizar cada número de pontos. Utilizamos o valor de significância de 5% (α < 0.05) para todos os testes

realizados e aplicamos a correção de Benjamini & Yekutieli (Benjamini & Yekutieli, 2001).

RESULTADOS

Encontramos variação na capacidade preditiva dos modelos em todos os cenários avaliados, incluindo tipos de dados de ausência, número de pontos e intensidade do viés, bem como nas interações entre esses fatores (ANOVA Fatorial de Medidas Repetidas; *p-value* < 0.05). Todas as 28 comparações par a par para a interação entre dados de ausência e viés foram significativas (*p-value* < 0.05) (Tabela Suplementar 4). Na comparação entre o grupo controle (sem viés) e os outros tratamentos (enviesados com intensidade de 10, 100 e 1000 vezes), observamos que o tamanho do efeito do viés aumenta com a intensidade do viés, independentemente do tipo de dado de ausência, ou seja, ausências reais ou pseudo-ausências (Figura 3).

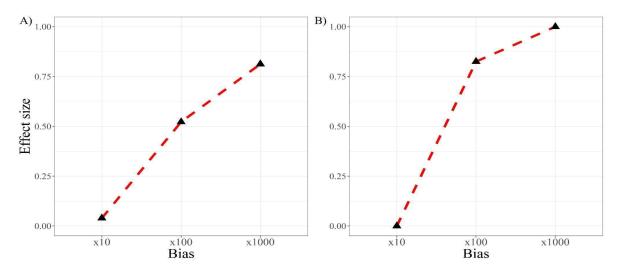


Figura 3: Representação do efeito do viés (eixo Y) ao comparar valores da métrica de sobreposição (I) do conjunto de dados sem viés de amostragem com dados enviesados em 10, 100 e 1000 vezes (eixo X) para modelos construídos com dados de ausências reais (A) e pseudo-ausências (B). O tamanho do efeito é expresso pela estatística T padronizada, variando de 0 a 1.

Após a análise de nossos resultados, constatamos a significância de todos os fatores e interações avaliados na relação entre o viés de amostragem e as métricas de avaliação para a capacidade preditiva do modelo (ANOVA, p-value < 0.05). Optamos por focar no cenário de dados PEA, considerado ideal para avaliar metodologias em ENM e SDM. Entretanto, não encontramos um padrão consistente de efeito do viés de amostragem na relação entre as métricas de avaliação e a métrica de sobreposição. Apenas quatro das dez métricas de avaliação analisadas apresentaram significância em algumas comparações par a par (ver

Tabela 1). Todas as métricas de avaliação significativas dependem da utilização de ausências/pseudo-ausências, exceto TPR (Sensibilidade), que mede a proporção de presenças corretamente identificadas pelo modelo. Das métricas que requerem ausências, apenas SEDI e Sørensen não apresentaram significância estatística.

Tabela 1: Tabela comparativa da relação entre métricas de avaliação e métrica de sobreposição para modelos construídos com dados de ausências reais (PEA), por meio de testes t pareados baseados em 9999 permutações. As células preenchidas marcam a significância (p-value < 0.05) da comparação entre o grupo controle de viés X0 com os demais grupos de viés (X10, X100 e X1000). Para mais informações sobre cada métrica, consulte a Tabela 1 no Apêndice.

Métrica		Viés		
	X10	X100	X1000	
AUC				
$\mathrm{AU}l$				
Jaccard				
MaxVD <i>l</i>				
Pearson_D				
SEDI				
Sørensen				
TNR				
TPR	'			
TSS			•	

AUC: Área Sob a Curva ROC; AUl: Área Sob a Curva *lift*; Jaccard: Índice de Similaridade de Jaccard; MaxVDl: Máxima Distância da Curva *lift*; Pearson_D: Adaptado de Índice de Pearson et al. (2007); SEDI: Symmetric Extremal Dependence Index; Índice de Similaridade de Sørensen; TNR: Taxa de Verdadeiros Negativos (Especificidade); TPR: Taxa de Verdadeiros Positivos (Sensibilidade); TSS: Estatísticas de Habilidade Verdadeira

Após avaliarmos se o viés de amostragem é potencializado com o uso de dados de pseudo-ausência, todas as comparações foram estatisticamente significativas. O tamanho do efeito variou consideravelmente de acordo com a intensidade do viés e o número de pontos de ocorrência, sendo que observamos uma diminuição do tamanho do efeito à medida que o número de pontos de ocorrência aumenta (Figura 4). Além disso, notamos que o tamanho do efeito foi geralmente maior ao utilizar dados PPA em comparação com dados PEA, como evidenciado na Figura 3.

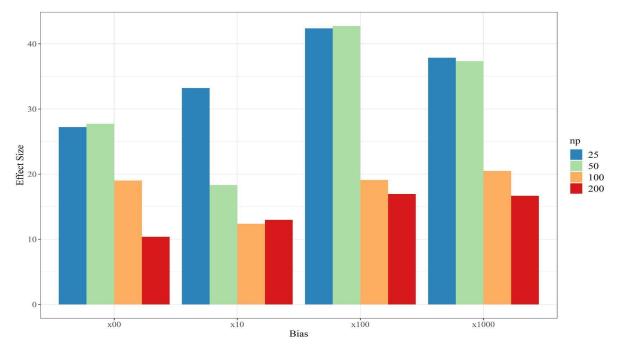


Figura 4: Comparação do tamanho do efeito (eixo y) em relação à intensidade de viés (eixo x) para cada número de registros de ocorrência (np), obtidos da comparação par a par por meio de testes t. Tamanho do efeito correspondente ao valor da estatística T.

DISCUSSÃO

Nossos resultados indicam que o viés de amostragem impacta negativamente a capacidade preditiva dos modelos no espaço geográfico, sendo que o tamanho do efeito é influenciado pelo tipo de dado de ausência utilizado e pelo número de registros de ocorrência. Notamos também que o viés prejudica as métricas de avaliação dos modelos que dependem de dados de ausências para serem computadas e que são baseadas em classificação, embora tenhamos observado uma heterogeneidade nos resultados, o que não confirmou totalmente a nossa hipótese, já que esperávamos um prejuízo para todas as métricas. Além disso, a ausência de efeito significativo do viés de amostragem para métricas que utilizam apenas dados de presenças as coloca como boas alternativas para avaliar modelos construídos com dados enviesados, mesmo quando presenças e ausências estão disponíveis.

Nossos resultados do primeiro cenário, que investigou a capacidade preditiva dos modelos no espaço geográfico, corroboram com os de Baker et al. (2022), que avaliaram o viés de amostragem no espaço ambiental (espaço do nicho) e identificaram efeitos significativos sobre a capacidade preditiva dos modelos. Esses resultados indicam que os efeitos do viés podem prejudicar além da avaliação e comprometer a estimativa do nicho,

uma aplicação frequente da ENM e SDM (Adhikari et al., 2019; Costa et al., 2010; Warren, 2012). Observamos que a intensidade do viés está diretamente relacionada com o tamanho do efeito, e técnicas de filtragem para reduzir a intensidade do viés, como as mencionadas por Aiello-Lammens et al. (2015), Vollering et al. (2019), e Zizka et al. (2019), podem reduzir o viés e melhorar a capacidade preditiva dos modelos. Embora Boria et al. (2014) tenham sugerido que o efeito do viés pode aumentar com o número de pontos, nosso estudo não confirmou essa suposição, visto que observamos uma redução no tamanho do efeito à medida que o número de registros de ocorrência aumenta.

Ao avaliarmos os efeitos do viés de amostragem sobre as métricas de avaliação, constatamos que há grande variabilidade nos resultados, sendo que algumas métricas são mais afetadas pelo viés do que outras, dependendo do tipo de dado de ausência e do cenário considerado. No entanto, notamos um padrão recorrente: O viés é mais problemático em condições de alta intensidade, o que corrobora com os achados relacionados à capacidade preditiva dos modelos e reforça a importância do uso de filtros para reduzir a intensidade do viés (Aiello-Lammens et al., 2015; Nolan et al., 2022; Varela et al., 2014). Ademais, enfatizamos que métricas de avaliação que não dependem de dados de ausência são menos suscetíveis aos efeitos do viés de amostragem, o que as torna alternativas conceitual e metodologicamente interessantes, como já foi defendido para a ENM e SDM (Leroy et al., 2018; Pearson et al., 2007).

Nossa análise dos efeitos da utilização de pseudo-ausências para calibrar e avaliar modelos revelou críticas persistentes à prática, destacando seus potenciais efeitos metodológicos no desempenho dos modelos (Chefaoui & Lobo, 2008; Golicher et al., 2012; Lobo & Tognelli, 2011; Vollering et al., 2019). Descobrimos que o uso de pseudo-ausências em cenários em que os dados contêm apenas informações de presença é mais prejudicial do que em cenários com ausências enviesadas, o que é particularmente preocupante, dado que essa é a realidade mais comum em uma variedade de contextos da ENM e SDM. (Vollering et al., 2019).

Considerando as limitações do uso de dados de pseudo-ausências e a importância de dados de ausências reais na construção de modelos de nicho e distribuição potencial de espécies, concluímos que o viés de amostragem pode afetar a capacidade preditiva desses modelos. Notavelmente, modelos construídos com dados de ausências reais enviesados ainda apresentam resultados superiores aos construídos com pseudo-ausências.

Demonstramos que o viés de amostragem não segue um padrão consistente na relação entre as métricas de avaliação e a métrica de sobreposição, mesmo considerando métricas que

dependem apenas de dados de presença, como AUl, MaxVDl e Pearson D. A avaliação dos modelos baseada apenas na Sensibilidade (TPR) não é suficiente para garantir a consistência avaliativa dos ENM e SDM. Por outro lado, as três métricas que dependem apenas de dados de presença, juntamente com SEDI e Sørensen, que dependem de ausências, mostraram-se boas alternativas conceituais e metodológicas, já que não foram afetadas pelo viés de amostragem em nenhum dos cenários avaliados.

Nossos resultados mostraram que a utilização de pseudo-ausências potencializa o efeito do viés de amostragem, e que esse efeito pode ser reduzido por meio de filtragem geográfica. Com um maior número de dados filtrados, é possível ter um conjunto de dados mais consistente, o que resulta em melhores estimativas de distribuição potencial e nicho fundamental. É importante destacar que o desafio consiste em ter um número elevado de pontos após realizar os procedimentos de filtragem. Isso pode ser superado com esforços de amostragem direcionados a localidades onde a espécie estudada ainda não foi amostrada, mas que há possibilidade de sua presença (Siqueira et al., 2009).

REFERÊNCIAS BIBLIOGRÁFICAS

- Adhikari, D., Tiwary, R., Singh, P.P., Upadhaya, K., Singh, B., Haridasan, K.E., Bhatt, B.B., Chettri, A., Barik, S.K., 2019. Ecological niche modeling as a cumulative environmental impact assessment tool for biodiversity assessment and conservation planning: A case study of critically endangered plant Lagerstroemia minuticarpa in the Indian Eastern Himalaya. J. Environ. Manage. 243, 299–307. https://doi.org/10.1016/j.jenvman.2019.05.036
- Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B., Anderson, R.P., 2015. spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography 38, 541–545. https://doi.org/10.1111/ecog.01132
- Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x
- Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., Jetz, W., 2018. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci. Data 5, 180040. https://doi.org/10.1038/sdata.2018.40
- Baker, D.J., Maclean, I.M.D., Goodall, M., Gaston, K.J., 2022. Correlations between spatial sampling biases and environmental niches affect species distribution models. Glob. Ecol. Biogeogr. 31, 1038–1050. https://doi.org/10.1111/geb.13491
- Bean, W.T., Stafford, R., Brashares, J.S., 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35, 250–258. https://doi.org/10.1111/j.1600-0587.2011.06545.x
- Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform. 19, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002
- Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing

- under dependency. Ann. Stat. 29. https://doi.org/10.1214/aos/1013699998
- Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012
- Chefaoui, R.M., Lobo, J.M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecol. Model. 210, 478–486. https://doi.org/10.1016/j.ecolmodel.2007.08.010
- Colwell, R.K., Rangel, T.F., 2009. Hutchinson's duality: The once and future niche. Proc. Natl. Acad. Sci. 106, 19651–19658. https://doi.org/10.1073/PNAS.0901650106
- Costa, G.C., Nogueira, C., Machado, R.B., Colli, G.R., 2010. Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. Biodivers. Conserv. 19, 883–899. https://doi.org/10.1007/s10531-009-9746-8
- Dormann, C.F., 2007. Promising the future? Global change projections of species distributions. Basic Appl. Ecol. 8, 387–397. https://doi.org/10.1016/j.baae.2006.11.001
- Elith, J., Franklin, J., 2013. Species Distribution Modeling. Encycl. Biodivers. Second Ed. 692–705. https://doi.org/10.1016/B978-0-12-384719-5.00318-X
- Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49. https://doi.org/10.1017/S0376892997000088
- Golicher, D., Ford, A., Cayuela, L., Newton, A., 2012. Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve. Int. J. Geogr. Inf. Sci. 26, 2049–2063. https://doi.org/10.1080/13658816.2012.719626
- Gomes, V.H.F., Mayle, F.E., Gosling, W.D., Vieira, I.C.G., Salomão, R.P., Steege, H., 2020. Modelling the distribution of Amazonian tree species in response to long-term climate change during the Mid-Late Holocene. J. Biogeogr. 47, 1530–1540. https://doi.org/10.1111/jbi.13833
- Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1
- Halvorsen, R., Mazzoni, S., Dirksen, J.W., Næsset, E., Gobakken, T., Ohlson, M., 2016. How important are choice of model selection method and spatial autocorrelation of presence data for distribution modelling by MaxEnt? Ecol. Model. 328, 108–118. https://doi.org/10.1016/j.ecolmodel.2016.02.021
- Hervé, M., 2022. RVAideMemoire: Testing and Plotting Procedures for Biostatistics.
- Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117, 847–858. https://doi.org/10.1111/j.0030-1299.2008.16434.x
- Jarnevich, C.S., Talbert, M., Morisette, J., Aldridge, C., Brown, C.S., Kumar, S., Manier, D., Talbert, C., Holcombe, T., 2017. Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: An example with background selection. Ecol. Model. 363, 48–56. https://doi.org/10.1016/j.ecolmodel.2017.08.017
- Jiménez-Valverde, A., 2020. Sample size for the evaluation of presence-absence models. Ecol. Indic. 114, 106289. https://doi.org/10.1016/j.ecolind.2020.106289
- Leitão, P.J., Moreira, F., Osborne, P.E., 2011. Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. Int. J. Geogr. Inf. Sci. 25, 439–454. https://doi.org/10.1080/13658816.2010.531020

- Leroy, B., Delsol, R., Hugueny, B., Meynard, C.N., Barhoumi, C., Barbet-Massin, M., Bellard, C., 2018. Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. J. Biogeogr. 45, 1994–2002. https://doi.org/10.1111/jbi.13402
- Liu, C., Newell, G., White, M., 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. Ecography 42, 535–548. https://doi.org/10.1111/ecog.03188
- Liu, C., White, M., Newell, G., Griffioen, P., 2013. Species distribution modelling for conservation planning in Victoria, Australia. Ecol. Model. 249, 68–74. https://doi.org/10.1016/j.ecolmodel.2012.07.003
- Lobo, J.M., Tognelli, M.F., 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. J. Nat. Conserv. 19, 1–7. https://doi.org/10.1016/j.jnc.2010.03.002
- Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecol. Model. 222, 588–597. https://doi.org/10.1016/j.ecolmodel.2010.11.016
- Meynard, C.N., Kaplan, D.M., 2013. Using virtual species to study species distributions and model performance. J. Biogeogr. 40, 1–8. https://doi.org/10.1111/jbi.12006
- Meynard, C.N., Leroy, B., Kaplan, D.M., 2019. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? Ecography 42, 2021–2036. https://doi.org/10.1111/ecog.04385
- Miller, J.A., 2014. Virtual species distribution models: Using simulated data to evaluate aspects of model performance. Prog. Phys. Geogr. Earth Environ. 38, 117–128. https://doi.org/10.1177/0309133314521448
- Moudrý, V., Šímová, P., 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. Int. J. Geogr. Inf. Sci. 26, 2083–2095. https://doi.org/10.1080/13658816.2012.721553
- Nogués-Bravo, D., Rodríguez, J., Hortal, J., Batra, P., Araújo, M.B., 2008. Climate Change, Humans, and the Extinction of the Woolly Mammoth. PLoS Biol. 6, e79. https://doi.org/10.1371/journal.pbio.0060079
- Nolan, V., Gilbert, F., Reader, T., 2022. Solving sampling bias problems in presence–absence or presence-only species data using zero-inflated models. J. Biogeogr. 49, 215–232. https://doi.org/10.1111/jbi.14268
- Pearson, R.G., Raxworthy, C.J., Nakamura, M., Townsend Peterson, A., 2007. Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. J. Biogeogr. 34, 102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x
- Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J., Raes, N., 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography (Cop.). 39, 542–552. https://doi.org/10.1111/ecog.01509
- R Core Team, 2018. R: A language and environment for statistical computing.
- Robertson, M.P., Visser, V., Hui, C., 2016. Biogeo: An R package for assessing and improving data quality of occurrence record datasets. Ecography 39, 394–401. https://doi.org/10.1111/ecog.02118
- Segurado, P., Araujo, M.B., Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. J. Appl. Ecol. 43, 433–444. https://doi.org/10.1111/j.1365-2664.2006.01162.x
- Siqueira, M.F. de, Durigan, G., de Marco Júnior, P., Peterson, A.T., 2009. Something from

- nothing: Using landscape similarity and ecological niche modeling to find rare plant species. J. Nat. Conserv. 17, 25–32. https://doi.org/10.1016/j.jnc.2008.11.001
- Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37, 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x
- Vollering, J., Halvorsen, R., Auestad, I., Rydgren, K., 2019. Bunching up the background betters bias in species distribution models. Ecography 42, 1717–1727. https://doi.org/10.1111/ecog.04503
- Warren, D.L., 2012. In defense of "niche modeling." Trends Ecol. Evol. 27, 497–500. https://doi.org/10.1016/j.tree.2012.03.010
- Wunderlich, R.F., Lin, Y.-P., Anthony, J., Petway, J.R., 2019. Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models. Nat. Conserv. 35, 97–116. https://doi.org/10.3897/natureconservation.35.33918
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., Antonelli, A., 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 744–751. https://doi.org/10.1111/2041-210X.13152

APÊNDICE

Tabela Suplementar 1: Métricas de acurácia dos modelos de nicho (esquerda), suas equações (centro) e suas descrições (direita)

descrições (direita).		
Métrica	Equação	Descrição e Fonte
Área Sob a Curva ROC (AUC)	$1/n_{i} n_{0} \sum_{i=1}^{n_{0}} \sum_{j=1}^{n_{1}} \phi(X_{i}, X_{j})$	$\phi(X, X)$ é igual a 1 se Y>X, 1/2 se Y=X, e 0 caso contrário; X_i e Y_j são o valor previsto para o site de ausência i e site de presença j (Fielding and Bell, 1997).
Área Sob a Curva <i>lift</i> (AU <i>l</i>)	$1/n_i n_r \sum_{i=1}^{n_r} \sum_{j=1}^{n_1} \phi(X_i, X_j)$	$\phi(X, X)$ é igual a 1 se Y>X, 1/2 se Y=X, e 0 caso contrário; X_i e Y_j são o valor previsto para o site de ausência i e site de presença j (Liu et al., 2013).
Jaccard	a/(c+a+b)	a corresponde aos acertos nas presenças e c corresponde aos acertos nas ausências. b corresponde aos erros nas presenças e d corresponde aos erros nas ausências da matriz de confusão.
Máxima distância da Curva <i>lift</i> (maxVD <i>l</i>)	max(TPR - PR)	TPR é a sensibilidade e PR corresponde a proporção de células da área de estudo preditas como presença (Prevalência) (Liu et al., 2013).
Pearson's D	<i>TPR(1 - PR)</i>	TPR é a sensibilidade e PR corresponde a proporção de

Symmetric Extremal Dependence Index (SEDI)	$\frac{log(TNR) - log(TPR) - log(1 - TNR) + log(1 - TPR)}{log(TNR) + log(TPR) + log(1 - TNR) + log(1 - TPR)}$	células da área de estudo preditas como presença (Prevalência) (Pearson et al., 2007). TPR é a sensibilidade e TNR a especificidade (Wunderlich et al., 2019). a corresponde aos acertos nas presenças e c corresponde aos
Sørensen	2a/(c+2a+b)	acertos nas ausências. <i>b</i> corresponde aos erros nas presenças e <i>d</i> corresponde aos erros nas ausências da matriz de confusão.
Especificidade - TNR	$\frac{b}{b+d}$	b corresponde aos erros nas presenças e d corresponde aos erros nas ausências da matriz de confusão.
Sensibilidade - TPR	$\frac{a}{a+c}$	<i>a</i> corresponde aos acertos nas presenças e <i>c</i> corresponde aos acertos nas ausências.
True Skill Statistic (TSS)	(TPR + TNR) - 1	TPR corresponde à sensibilidade e TNR corresponde à especificidade (Allouche et al., 2006).

Tabela Suplementar 2: Tabela com pacotes (esquerda) e suas respectivas versões (centro-esquerda), a função utilizada (centro-direita) e o acrônimo utilizado para para identificar os algoritmos (direita).

Packages	Version	Function	Acronym
dismo	1.3-3	bioclim	bioclim
dismo	1.3-3	domain	domain
dismo	1.3-3	mahal	maha
dismo	1.3-3	maxent	maxent
Rfast	2.0.1	dista	euclidist
adehabitatHS	0.3.15	madifa	enfa
stats	3.6.3	glm	glm
gam	1.16.1	gam	gam
mda	0.5-2	fda	fda
rpart	4.1-15	rpart	cta
gbm	2.1.8	gbm	gbm
kernlab	0.9-29	ksvm	svm
nnet	7.3-15	nnet	nnet
randomForest	4.6-14	randomForest	rf
e1071	1.7-6	naiveBayes	bayes

Tabela Suplementar 3: Tabela com variáveis utilizadas em todas as nossas análises. A esquerda estão as abreviaturas das variáveis, ao centro estão os nomes das variáveis e a direita estão as referências de cada variável.

Acronym	Variable	Source
Bio 1	Annual Mean Temperature	(Fick and Hijmans, 2017)
Bio 2	Mean Diurnal Range	(Fick and Hijmans, 2017)
Bio 3	Isothermality	(Fick and Hijmans, 2017)
Bio 4	Temperature Seasonality	(Fick and Hijmans, 2017)
Bio 5	Max Temperature of Warmest Month	(Fick and Hijmans, 2017)

Bio 6	Min Temperature of Coldest Month	(Fick and Hijmans, 2017)	
Bio 7	Temperature Annual Range	(Fick and Hijmans, 2017)	
Bio 8	Mean Temperature of Wettest Quarter	(Fick and Hijmans, 2017)	
Bio 9	Mean Temperature of Driest Quarter	(Fick and Hijmans, 2017)	
Bio 10	Mean Temperature of Warmest Quarter	(Fick and Hijmans, 2017)	
Bio 11	Mean Temperature of Coldest Quarter	(Fick and Hijmans, 2017)	
Bio 12	Annual Precipitation	(Fick and Hijmans, 2017)	
Bio 13	Precipitation of Wettest Month	(Fick and Hijmans, 2017)	
Bio 14	Precipitation of Driest Month	(Fick and Hijmans, 2017)	
Bio 15	Precipitation Seasonality	(Fick and Hijmans, 2017)	
Bio 16	Precipitation of Wettest Quarter	(Fick and Hijmans, 2017)	
Bio 17	Precipitation of Driest Quarter	(Fick and Hijmans, 2017)	
Bio 18	Precipitation of Warmest Quarter	(Fick and Hijmans, 2017)	
Bio 19	Precipitation of Coldest Quarter	(Fick and Hijmans, 2017)	
Elevation	Elevation	(Amatulli et al., 2018)	
Aspect	Orientation of Slope	(Amatulli et al., 2018)	
Slope	Degree of Incline of a Surface (Amatulli et al., 2018)		

Tabela Suplementar 4: Tabela da ANOVA fatorial realizada para avaliar o potencial efeito múltiplos componentes, e das suas interações, sobre a capacidade preditiva dos modelos de nicho ecológico (ENM) e modelos de distribuição potencial de espécies (SDM). Os componentes np, bias e data correspondem aos fatores, número de pontes de ocorrência, intensidade do viés de amostragem e tipo de dados de ausência inseridos na ANOVA, respectivamente. **Effect** corresponde ao componente avalia, **DFn** são os graus de liberdade no numerador, **DFd** são os graus de liberdade no denominador, **F** é o valor da estatística e **perm_p** corresponde ao valor-p obtido por 9999 permutações.

Effect	DFn	DFd	\mathbf{F}	perm_p
(Intercept)	1	1499	46146.54	1
np	3	4497	4846.28	0.0001
bias	3	4497	1959.90	0.0001
data	1	1499	610.78	0.0001
np:bias	9	13491	126.84	0.0001
np:data	3	4497	323.51	0.0001
bias:data	3	4497	142.41	0.0001
np:bias:data	9	13491	37.72	0.0001

CONCLUSÃO DA TESE

Assim, concluímos a presente tese destacando que a Modelagem de Nicho Ecológico e Modelagem de Distribuição de Espécies constitui um conjunto de ferramentas úteis, mas que requerem aperfeiçoamentos. Os aperfeiçoamentos são desejáveis tanto no arcabouço conceitual, quanto metodológico, conforme foi o foco desta tese. Destacamos a importância de se concentrar esforços na avaliação de tais modelos para que possamos ser capazes de avaliá-los adequadamente. Desta forma, seremos capazes de selecionar e/ou ponderar adequadamente os modelos que possuem capacidade preditiva elevada.

Também destacamos a relevância de tomar cuidado ao utilizar dados enviesados, especialmente quando dados de pseudo-ausências forem utilizados. Incentivamos que dados menos enviesados e que os procedimentos de correção de viés sejam explicitamente informados.

Finalmente, destacamos a importância de trabalhos que avaliem as metodologias da Modelagem de Nicho Ecológico e Modelagem de Distribuição de Espécies. Tais trabalhos trazem à luz questões fundamentais que tenhamos modelos mais úteis.