

GUSTAVO REIS DE BRITO

**EFEITOS DA VARIAÇÃO DO TAMANHO AMOSTRAL E DA QUANTIDADE DE
PONTOS DE PSEUDO AUSÊNCIA E BACKGROUND EM MODELOS DE NICHOS
ECOLÓGICOS**

ASSIS

2023

GUSTAVO REIS DE BRITO

**EFEITOS DA VARIAÇÃO DO TAMANHO AMOSTRAL E DA QUANTIDADE DE
PONTOS DE PSEUDO AUSÊNCIA E BACKGROUND EM MODELOS DE NICHO
ECOLÓGICO**

Tese apresentada à Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Letras, Assis, para obtenção do título de Doutor em Biociências (Área de Conhecimento: Caracterização e Aplicação da Diversidade Biológica)

Orientador(a): Dr. Fernando Frei

ASSIS

2023

Dados Internacionais de Catalogação na Publicação (CIP)
Ana Cláudia Inocente Garcia - CRB 8/6887

B862e Brito, Gustavo Reis de
Efeitos da variação do tamanho amostral e da quantidade de pontos de pseudo ausência e *Background* em modelos de nicho ecológico / Gustavo Reis de Brito. — Assis, 2023
122 p. : il.

Tese de Doutorado - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Letras, Assis
Orientador: Dr. Fernando Frei

1. Biodiversidade - Conservação. 2. Nicho ecológico - Modelagem preditiva. 3. Distribuição de Espécies (MDE) - Modelagem preditiva. 4. Biotecnologia animal - Pesquisa qualitativa. I. Título.

CDD 577

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE GUSTAVO REIS DE BRITO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM BIOCIÊNCIAS, DA FACULDADE DE CIÊNCIAS E LETRAS - CÂMPUS DE ASSIS.

Aos 01 dias do mês de março do ano de 2023, às 14:00 horas, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de GUSTAVO REIS DE BRITO, intitulada **EFEITOS DA VARIAÇÃO DO TAMANHO AMOSTRAL E DA QUANTIDADE DE PONTOS DE PSEUDO AUSÊNCIA E BACKGROUND EM MODELOS DE NICHOS ECOLÓGICOS**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. FERNANDO FREI (Orientador(a) - Participação Presencial) do(a) Departamento de Ciências Biológicas / UNESP/FCL-Assis, Profa. Dra. MARIA CECILIA BARBOSA DE TOLEDO (Participação Virtual) do(a) Departamento de Biologia / UNITAU/Taubaté, Profa. Dra. LUISA MARIA DIELE-VIEGAS COSTA SILVA (Participação Virtual) do(a) UFBA/Bahia, Prof. Dr. ALEX BARBOSA DE MORAES (Participação Virtual) do(a) UFES/Vitória-ES, Prof. Dr. PITÁGORAS DA CONCEIÇÃO BISPO (Participação Presencial) do(a) Departamento de Ciências Biológicas / UNESP/FCL-Assis. Após a exposição pelo doutorando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: APROVADO. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.



Prof. Dr. FERNANDO FREI

Aos meus avós (in memoriam).

AGRADECIMENTOS

Ao meu orientador, professor Dr. Fernando Frei, pelas contribuições e pela confiança depositada durante todos os anos de mestrado e doutorado.

Aos meus pais, José Antonio e Daisa, pelo inquestionável apoio, paciência e compreensão durante toda essa jornada, especialmente durante os dias mais longos e trabalhosos. Sem vocês não haveria sonho. Sem vocês, não haveria nada.

Aos meus avós Nady, Ademar, Teresa, e Alcides, *in memoriam*, pela condução no caminho da vida, imensurável apoio e incentivo ao desenvolvimento pessoal e acadêmico. Embora não estejam aqui hoje, foi graças a tudo o que fizeram por mim que permitiram esse dia acontecer.

A Dr^a. Solange Bongiovanni, pela amizade em todos esses anos, pelo incentivo no aprimoramento da vida acadêmica e pessoal, e pela referência profissional e humana que se tornou.

A Celine Lopes e Hellen Ceriello, pelos anos de amizade, companheirismo, e por compartilharem dos diversos momentos da vida de pós-graduação, sempre com bom humor.

A Giovanna Eller, pelo apoio, paciência com minhas ausências e por me incentivar a prosseguir nos momentos cruciais desse caminho. Pode não parecer muito, mas seu apoio sempre foi fundamental.

Aos demais amigos, colegas, e mestres que a vida me apresentou e que contribuíram quer de forma acadêmica, quer de forma pessoal, para que esse momento chegasse, os meus sinceros agradecimentos.

A matemática sem história natural é estéril, mas a história natural sem matemática é confusa.

(MAYNARD SMITH, 1988, p. 13)

BRITO, G.R. **Efeitos da variação do tamanho amostral e da quantidade de pontos de pseudo ausência e background em Modelos de Nicho Ecológico**. 200 p. Tese (Doutorado Acadêmico em Biociências). - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Letras, Assis, 2023.

RESUMO

A biodiversidade se altera espaço e temporalmente através de fatores naturais. No entanto, ações antrópicas são consideradas como o principal fator influenciador de transformações na biota do planeta. Estudar e acompanhar as mudanças na biodiversidade se tornou essencial à ciência da Conservação, sendo potencializado pela integração de campos da biologia com ferramentas tecnológicas, levando a criação de grandes bancos de dados globais e o desenvolvimento de ferramentas como os Modelos de Distribuição de Espécies e os Modelos de Nicho Ecológico. No presente trabalho foi aplicado a Modelagem de Nicho Ecológico a três espécies de aves brasileiras, através dos algoritmos Bioclim, *Generalized Additive Models* (GAM, em inglês), *Generalized Linear Models* (GLM, em inglês), Maxent e Random Forests, de modo a testar a influência da variação do tamanho amostral e da quantidade de pseudo ausência e pontos de *background* na projeção e validação estatística dos modelos. Também foi avaliado como o modo de geração dos pontos de pseudo ausência e *background* influenciam a qualidade e significância biológica das projeções. Os resultados obtidos demonstram como a diminuição do tamanho amostral influencia diretamente a projeção final ao sub ou superestimar a espécie modelada, bem como torna dúbia a significância biológica dos modelos. A sub ou superestimação da adequabilidade ambiental nos modelos ocorre independente da espécie modelada, sendo o tamanho amostral, a quantidade de pontos de pseudo ausência e *background*, e o tipo de algoritmo utilizados fatores críticos à construção dos modelos de adequabilidade. Deste modo, o presente trabalho evidencia a necessidade do pesquisador em definir corretamente o objetivo da pesquisa ao fazer uso das técnicas de Modelagem de Nicho Ecológico, bem como em selecionar, adequadamente a sua realidade de estudo, o tipo de algoritmo e a quantidade de pontos de ocorrência, pseudo ausência ou, ainda, de *background* para que as projeções tenham validação estatística e biológica.

Palavras-chave: Modelagem de Distribuição de Espécies; SDM; Modelos Ecológicos; ENM; Tamanho da amostra; Adequabilidade Ambiental.

BRITO, G.R. **Effects of varying sample size and number of pseudo absence and background points in Ecological Niche Models**. 200 p. Thesis (PhD in Biosciences). - São Paulo State University (UNESP), School of Sciences, Humanities, and Languages, Assis, 2023.

ABSTRACT

Biodiversity changes spatially and temporally through natural factors. However, anthropic actions are considered the main factor influencing transformations in the planet's biota. Studying and tracking changes in biodiversity has become essential to conservation science, and has been enhanced by the integration of biological fields with technological tools, leading to the creation of large global databases and the development of tools such as Species Distribution Models and Ecological Niche Models. In this work we applied the Ecological Niche Modeling to three species of Brazilian birds, using the Bioclim, Generalized Additive Models (GAM), Generalized Linear Models (GLM), Maxent, and Random Forests algorithms, in order to test the influence of sample size variation and the amount of pseudo absence and background points on the projection and statistical validation of the models. It also evaluated how the generation mode of the pseudo absence and background points influence the quality and biological significance of the projections. The results obtained show how the decrease in sample size directly influences the final projection by under or overestimating the modeled suitability, as well as making the biological significance of the models dubious. The under or overestimation of environmental suitability in the models occurs regardless of the species modeled, and the sample size, the number of points of pseudo absence and background, and the type of algorithm used are critical factors in the construction of suitability models. Thus, the present work highlights the need for the researcher to correctly define the research objective when making use of Ecological Niche Modeling techniques, as well as to adequately select the type of algorithm and the number of occurrence points, pseudo absence, or, still, of background for the projections to have statistical and biological validation.

Keywords: Species Distribution Modeling; SDM; Ecological Models; ENM; Sample Size; Environmental Suitability.

LISTA DE FIGURAS

Figura 1 - Distribuição dos pontos de ocorrência obtidos na base de dados GBIF dentro das Extensões de Ocorrência para as espécies estudadas. A linha vermelha denota os limites do EOO. A versão em alta resolução se encontra no Material Suplementar 1. Fonte: autoria própria (2022). **30**

Figura 2 - Visão geral do processo de construção de um SDM / ENM, desde a obtenção dos dados biológicos e ambientais, calibração (i.e. *fitting*) do modelo e posterior avaliação. A versão em alta resolução se encontra no Material Suplementar 1. Fonte: autoria própria (2022). **31**

Figura 3 - Percentual of predicted suitable area for all three species, considering the 0.5 - 1.0 threshold of suitability (red line). All models were fitted using the first set of pseudo absence/background points (n = 1000) and projections were constrained by the EOO of the species. High resolution version is available in Suppl. Material 1. **48**

Figura 4 - Percentual of predicted suitable area for all three species, considering the 0.5 - 1.0 threshold of suitability (red line). All models were fitted using the first set of pseudo absence/background points (n = 10000) and projections were constrained by the EOO of the species. High resolution version is available in Suppl. Material 1. **49**

Figura 5 - Projected models of *Cantorchilus longirostris* based on all five algorithms considering the full spectrum of suitability. Red line represents the EOO of the species; dots represent occurrence points of each sample size used to train the models. A set of n = 1000 pseudo absence and background points were used during model fitting. **50**

Figura 6 - Projected models of *Cantorchilus longirostris* based on all five algorithms considering the threshold of 0.5 - 1.0 of suitability. Red line represents the EOO of the species; dots represent occurrence points of each sample size used to train the models. A set of n = 1000 pseudo absence and background points were used during model fitting. **50**

Figura 7 - Mean AUC (A) and TSS (B) values for 1000 pseudo absence/background points for all three species. The values represent the average value per sample size. **51**

Figura 8 - Mean AUC (A) and TSS (B) values for 10000 pseudo absence/background points for all three species. The values represent the average value per sample size. **52**

Figura 9 - Mean AUC values obtained for all five algorithms across each background / pseudo absence set for all models of *Cantorchilus longirostris*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process to generate pseudo absence and background points. The red line represents the 0.7 AUC threshold. **63**

Figura 10 - Mean AUC values obtained for all five algorithms across each background / pseudo absence set for all models of *Eupsittula cactorum*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process to generate pseudo absence and background points. The red line represents the 0.7 AUC threshold. **64**

Figura 11 - Mean AUC values obtained for all five algorithms across each background / pseudo absence set for all models of *Penelope obscura*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process to generate pseudo absence and background points. The red line represents the 0.7 AUC threshold. **65**

Figura 12 - Mean TSS values obtained for all five algorithms across each background / pseudo absence set for all models of *Cantorchilus longirostris*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process to generate pseudo absence and background points. The red line represents the 0.4 TSS threshold. **66**

Figura 13 - Mean TSS values obtained for all five algorithms across each background / pseudo absence set for all models of *Eupsittula cactorum*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process to generate pseudo absence and background points. The red line represents the 0.4 TSS threshold. **66**

Figura 15 - Mean TSS values obtained for all five algorithms across each background / pseudo absence set for all models of *Penelope obscura*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process to generate pseudo absence and background points. The red line represents the 0.4 TSS threshold. **67**

Figura 16 - Suitability map for *Penelope obscura* (A) and *Eupsittula cactorum* (B) produced by Bioclim algorithm. The models were trained using 20 and 100 presence points (sizes N3 and N2), respectively. Model A was trained 1000 background points randomly generated from the study area; Model B was trained with 250 background points sampled outside a surface range envelope draw on the study area. Red lines represents the Extent of Occurrence of the species. Black dots are the presence points used in model training. **68**

Figura 17 - Overfitted suitability map for *Eupsittula cactorum* produced by four algorithms with the highest values of AUC (AUC = 1.0) and TSS (TSS $\geq$ 0.95). The models were trained using sample size N1, with a 1000 pseudo absence and 10000 background points, respectively. Red lines represents the Extent of Occurrence of the species. Black dots are the presence points used in model training.

LISTA DE TABELAS

Tabela 1 - Total de pontos obtidos para definição do tamanho amostral principal (N1)	22
Tabela 2 - Subdivisão dos pontos de ocorrência das espécies em tamanhos amostrais com base no sorteio do conjunto principal (N1)	23
Tabela 3 - Exemplos de publicações de aplicação SDM e ENM com diferentes tamanhos amostrais obtidos em um levantamento na base de dados Web of Science (2013-2021)	24
Tabela 4 - Variáveis bioclimáticas obtidas na base de dados WorldClim, versão 2.1 e respectiva codificação	25
Tabela 5 - Subdivisão dos pontos de <i>background</i> e de pseudo ausência	27
Tabela 6 - Total de publicações e percentual de aplicação dos algoritmos mais frequentes em trabalhos de distribuição de espécies e de nicho ecológico segundo Raes e Aguirre-Gutiérrez (2018) para o período 2009-2019	28
Tabela 7 - Least correlated variables selected via VIF analysis for modeling	45
Tabela 8 - Comparison of the estimated Extent of Occurrence and IUCN Red List assessment of each species range	46
Tabela 9 - Number of occurrence points for each species, from GBIF database to the final dataset	59
Tabela 10 - Environmental variables used to train the models for all three species	60

LISTA DE ABREVIATURAS E SIGLAS

AUC	Área embaixo da Curva ROC (<i>Area under the ROC Curve</i> , em inglês)
BG	<i>Background</i> (em inglês, sem tradução)
ENM	Modelo de Nicho Ecológico (<i>Ecological Niche Model</i> , em inglês)
GAM	Modelo Aditivo Generalizado (<i>Generalized Additive Model</i> , em inglês)
GLM	Modelo Linear Generalizado (<i>Generalized Linear Model</i> , em inglês)
PA	Pseudo ausência (<i>Pseudo absence</i> , em inglês)
ROC	<i>Receiver Operator Characteristic Curve</i> (em inglês, sem tradução)
SDM	Modelo de Distribuição de Espécies (<i>Species Distribution Model</i> , em inglês)
TSS	<i>True Skill Statistic</i> (em inglês, sem tradução)

SUMÁRIO

1 INTRODUÇÃO GERAL	17
2 FUNDAMENTAÇÃO TEÓRICA	23
2.1 Estrutura da base de dados	23
2.2 Dados bióticos	24
2.3 Dados abióticos	26
2.4 Seleção de pseudo ausência e background	29
2.5 Seleção dos algoritmos	29
2.6 Estimativa da Extensão de Ocorrência das espécies	30
2.7 Modelagem, validação estatística e consenso entre modelos	31
REFERÊNCIAS	33
CAPÍTULO 1	44
Abstract	45
Introduction	45
Methods	46
Selection of species and database compilation	46
Selection of Environmental Variables	47
Selection of sample sizes	47
Modeling procedures	47
Suitable area measurement	48
Results	48
Extent of Occurrence	48
Suitable Area and Model Performance	49
Statistical Analysis	52
Discussion	53
Acknowledgments	54
References	55
CAPÍTULO 2	59
Introduction	60
Methods	61
Species and Environmental datasets	61
Selection of pseudo absence and background points	62
Modeling procedures	62
Ensembling	63
Statistical Analysis	63
Statistical Analysis	63
Model Performance	68
Discussion	70
Conclusion	71
References	71
3 CONSIDERAÇÕES FINAIS	76

REFERÊNCIAS	77
Material Suplementar 1	88
Material Suplementar 2	90

1 INTRODUÇÃO GERAL

O processo de variação espaço-temporal da biodiversidade, no que se refere à diversidade e composição de espécies, é um processo natural influenciado por fatores de larga escala, de processos ecológicos recorrentes à mudanças ambientais e biogeográficas históricas. Embora grandes mudanças na biodiversidade sejam guiadas por processos naturais, a influência antrópica se apresenta como um dos mais influentes fatores de transformação ambiental e, por consequência, de alterações na biodiversidade (PINTO-LEDEZMA; CAVENDER-BARES, 2021).

Em 2019, o relatório de avaliação global sobre biodiversidade e serviços ecossistêmicos, redigido pela Plataforma Intergovernamental de Política Científica sobre Biodiversidade e Serviços Ecossistêmicos (IPBES, em inglês), apresentou dados preocupantes quanto ao estado atual e possibilidades futuras para a biodiversidade: aproximadamente um milhão de espécies, entre animais e vegetais, encontram-se ameaçadas de extinção dentro das próximas décadas (IPBES, 2019).

De acordo com o IPBES (2019), os habitats terrestres apresentam uma redução na abundância média de espécies nativas de até 20%. A interferência humana nos ecossistemas intactos acontece, principalmente, nos trópicos, regiões com altos níveis de biodiversidade. Somente na América Latina, aproximadamente 100 milhões de hectares de floresta tropical foram perdidos para atividades de pecuária no período entre 1980 a 2000.

Diante de um cenário pessimista como apresentado pelos relatórios IPBES (2019) e pelo Painel Intergovernamental para a Mudança de Clima (IPCC, em inglês) (2019), cresce a busca por estratégias para o manejo da biota visando a conservação da biodiversidade. No entanto, desenvolver projetos nesse âmbito apresenta desafios próprios, como alocação de recursos, escolha de locais para estabelecer sítios de proteção, ou, ainda, definir as espécies que serão priorizadas no projeto (MUSCATELLO; ELITH; KUJALA, 2021).

O avanço tecnológico, em especial a informática, tem permitido a integração de informações de diferentes áreas em grandes bancos de dados virtuais, acessíveis globalmente. Jetz, McPherson e Guralnick (2012) demonstram como o uso da informática permite a estruturação de bases de dados contendo informações relativas à distribuição das espécies, item essencial para estudos de conservação.

Embora o aumento na coleta e armazenamento de dados de distribuição seja notório e favoreça o estudo das espécies, estes tendem a refletir a situação em um ponto no espaço e no tempo, não refletindo a totalidade da distribuição (RONDININI et al., 2006). Porém, ao combinar dados de observação com informações ambientais, é possível modelar a adequabilidade ambiental para uma espécie e, então, prever a ocorrência para auxiliar a tomada de decisão no que se refere à conservação (GUISAN et al., 2013).

Conhecidas como Modelos de Nicho Ecológico (ENM, em inglês), Modelos de Distribuição de Espécies (SDM, em inglês), Modelos de Adequabilidade Ambiental (HMS, em inglês), entre outras terminologias, as técnicas de modelagem que integram a distribuição de uma espécie com fatores ambientais para gerar previsões espaciais e temporais tornaram-se mais expressivas após os anos 2000 (GUISAN; ZIMMERMANN, 2000), apresentando-se como a principal ferramenta de estudo sobre a distribuição de espécies e da adequabilidade ambiental (ELITH; LEATHWICK, 2009; FRANKLIN; MILLER, 2009; GUISAN; THUILLER, 2005; PETERSON, 2011).

O termo “Modelo de Distribuição de Espécie” tornou-se sinônimo de um conjunto de técnicas que, por muitas vezes, não são aplicadas com o mesmo objetivo. Peterson e Soberón (2012) apresentam um panorama dos conceitos principais de modelagem e demonstram como essa sinonímia é encontrada na literatura: 50.8% dos trabalhos recuperados pelos autores em um levantamento bibliográfico com os termos “*species distribution model*” e “*ecological niche model*” se caracterizavam como estudos de distribuição, porém o objeto de estudo real era o nicho das espécies.

Embora SDMs e ENMs compartilhem técnicas, variáveis, algoritmos, e possam ser projetados em um mapa, o objeto de estudo e questionamentos referentes a este são, em suma, diferentes. Para que essa “diferença de objetos” seja entendida, é necessário, primeiro, entender e delimitar o que está sendo modelado (PETERSON; SOBERON, 2012).

Colwell e Rangel (2009) discutem a distribuição das espécies em duas dimensões: tanto no espaço geográfico quanto no espaço ambiental. Soberón e Peterson (2005) apresentam um *framework* que visa englobar os fatores determinantes na distribuição de espécies: o diagrama Biótico-Abiótico-Mobilidade (BAM, em inglês). Para os autores, os espaços geográfico e ambiental são, portanto, relacionados.

No espaço geográfico estarão locais de distribuição que contém fatores bióticos e abióticos necessários para o estabelecimento e manutenção da espécie, denominados no *framework* como B e A, respectivamente. O “M” se refere, portanto, a áreas acessíveis a espécie. Uma vez estabelecido a base do BAM, é possível, agora, traçar os objetos a se modelar em cada abordagem (PETERSON; SOBERON, 2012).

Para Peterson e Soberón (2012), a maioria dos trabalhos modernos que se intitulam como SDMs são, na verdade, trabalhos relativos ao nicho das espécies e, portanto, associados incorretamente com a ideia de um modelo de distribuição. Os autores apresentam que os modelos atuais são iniciados no espaço geográfico, treinados no espaço ambiental e, posteriormente, expressos (i.e., projetados) no espaço geográfico novamente.

Essa abordagem implica na comparação de áreas geográficas com áreas do espaço ambiental associadas ou não a pontos de presença para que seja traçado um gradiente de similaridade que classifique, novamente, o espaço geográfico como adequado ou inadequado para a ocorrência da espécie. Em suma, a maioria dos modelos atuais tem por alvo uma entidade no espaço ambiental (i.e., o domínio do nicho) e devem ser, portanto, classificados como ENM. Em contrapartida, um SDM estritamente deve não só estimar o nicho da espécie, mas sua capacidade de dispersão e/ou colonização do espaço modelado (PETERSON; SOBERÓN, 2012).

Em linhas gerais, um SDM ou ENM correlaciona pontos de ocorrência de uma espécie com variáveis ambientais para delinear, de forma aproximada, o conjunto de fatores ecológicos requeridos pela espécie (i.e., o nicho realizado) (GUISAN; THUILLER, 2005; GUISAN; ZIMMERMANN, 2000).

Um modelo SDM ou ENM é, usualmente, estruturado em 5 etapas: 1) conceitualização do modelo (e.g., desenho experimental, revisão de literatura, experimentação), 2) coleta dos dados de ocorrência e extração de valores das variáveis ambientais “ponto a ponto”, 3) calibração (*fitting*, em inglês) de um modelo de similaridade dos locais de ocorrência, 4) avaliação do modelo por diferentes métricas (e.g., Receiver Operating Characteristic - ROC, True Skill Statistic - TSS, Coeficiente Kappa) e 5) predição, a partir do modelo, da adequabilidade ambiental ou distribuição potencial da espécie em uma região ou, ainda, para cenários passados ou futuros (GUISAN; THUILLER; ZIMMERMANN, 2017; HIJMANS; ELITH, 2021).

Dos pontos apresentados por Guisan, Thuiller e Zimmermann (2017) e Hijmans e Elith (2021) quanto a organização de um processo de modelagem, as etapas de obtenção dos pontos de ocorrência, seleção dos algoritmos, ajuste (i.e., modelagem propriamente dita dos dados), avaliação e validação do modelo podem ser consideradas como essenciais para a qualidade das projeções finais.

Grandes bases de dados como a Global Biodiversity Information Facility – GBIF (www.gbif.org) e speciesLink (www.specieslink.net) permite o acesso a grandes quantidades de informações referentes a ocorrência de espécies em todo o globo, item essencial para o desenvolvimento de um SDM ou ENM. Todavia, as bases de dados tendem a não apresentar outra informação essencial à construção de um modelo: dados de ausência.

Independente da abordagem ecológica (i.e., nicho ou a distribuição potencial), para que seja possível ajustar o modelo é necessário que seja feita a relação entre dados de presença e ausência da espécie com as variáveis ambientais da área de estudo. Sem essa relação não é possível caracterizar estatisticamente o nicho de uma espécie e, também, projetá-lo espacialmente em um mapa (DESCOMBES et al., 2022). No entanto, comprovar a inexistência de uma espécie em uma determinada localidade é um desafio ao pesquisador, uma vez que diferentes fatores podem “mascarar a falta da espécie” no momento de coleta, especialmente em espécies migratórias ou que se movimentam com frequência pelo ambiente (MACKENZIE; ROYLE, 2005; RAES; AGUIRRE-GUTIÉRREZ, 2018). Raes e Aguirre-Gutiérrez (2018) apresentam o argumento de que “a ausência de presença não significa a presença de ausência”, reforçando a dificuldade de obter esse dado.

Essencialmente, os processos de modelagem podem ser divididos em métodos de presença e presença-ausência (BROTONS et al., 2004). Estudos comparativos (e.g., BROTONS et al., 2004; ELITH et al., 2006) demonstram que as metodologias de presença-ausência tendem a possuir performance superior aos métodos que requerem apenas presença. Todavia, devido a falta de dados de ausência real, estes são substituídos no processo de modelagem por pontos de *background* ou por pseudo ausência. O tipo de dado de “ausência” utilizado depende do tipo de algoritmo utilizado (BARBET-MASSIN et al., 2012) e, também, do objeto de pesquisa (e.g., nicho, distribuição potencial).

Todavia, a seleção do tipo de dado de “ausência” a ser utilizado na modelagem padece de um dilema análogo àquele enfrentado na nomenclatura de SDM/ENM. Sillero e Barbosa (2021) discutem como dados de pseudo ausência e *background* tendem a ser tratados como o

mesmo objeto devido a similaridades em sua concepção teórica. No entanto, estes não possuem, de fato, o mesmo significado do ponto de vista metodológico.

A diferença entre um ponto de pseudo ausência e um ponto de *background* está em sua localização. Embora ambos possam ser amostrados no mesmo espaço geográfico, a existência real encontra-se em estratos diferentes do diagrama BAM (SOBERON; PETERSON, 2005).

Gerado artificialmente, o ponto de pseudo ausência representa um local no espaço onde a espécie em estudo supostamente não ocorre, porém sem real comprovação de sua “ausência” (BARBET-MASSIN et al., 2012). Como trata-se de um substituto para a ausência real, pontos de pseudo ausência não podem ocupar o mesmo espaço geográfico que pontos de presença (i.e., não pode haver sobreposição) uma vez que isso seria contraditório.

Pontos amostrados no *background* (i.e., o espaço ambiental), por sua vez, não possuem o mesmo requerimento que a pseudo ausência: o *background* representa o espectro de condições ambientais favoráveis ou desfavoráveis à ocorrência da espécie e, portanto, um ponto amostrado no *background* ambiental poderá estar no mesmo local geográfico que um ponto de presença ou ausência/pseudo ausência (HALLGREN et al., 2019; ITURBIDE; BEDIA; GUTIÉRREZ, 2018; PHILLIPS et al., 2009).

Embora a literatura referente ao tipo de “ausência” utilizada na modelagem não seja escassa (e.g., ENGLER; GUIAN; RECHSTEINER, 2004; LOBO; JIMÉNEZ- VALVERDE; HORTAL, 2010; ZANIEWSKI; LEHMANN; OVERTON, 2002), a decisão de gerar pontos de *background* ou de pseudo ausência é uma etapa fundamental da modelagem, sendo necessário se atentar não só ao tipo de dado gerado, mas como estes foram selecionados, uma vez que dados artificiais de ausência podem impactar negativamente um SDM / ENM (BARBET-MASSIN et al., 2012; GU; SWIHART, 2004).

De igual modo, a seleção do tipo de algoritmo para ajustar o modelo é crucial. Barbet-Massin e colaboradores (2012) apresenta três categorias: tipo I (regressão), tipo II (classificação), ou tipo 3 (*machine learning*). Cada tipo de modelo exigirá 1) pontos de presença da espécie, 2) um tipo de “ausência” (i.e., pseudo ausência ou *background* ou ausência real) e 3) um algoritmo. As etapas 2 e 3 estão correlacionadas uma vez que o tipo de algoritmo requer um tipo de “ausência”.

O processo de modelagem é longo e complexo, começando antes mesmo da parte computacional. A escolha do tipo ou qual algoritmo utilizar ou, ainda, qual “ausência” incluir

no modelo, são decisões que aparentam possuir maior peso do que a delimitação da área de estudo. No entanto, o tamanho da área influenciará o processo de modelagem (SILLERO et al., 2021) quanto à quantidade de pontos de ocorrência e na captura da heterogeneidade ambiental por parte dos algoritmos.

A extensão da área de estudo impacta na percepção do espaço ambiental: áreas restritas podem não permitir o acesso a amplitude real de ocorrência da espécie (CARRETERO; SILLERO, 2016). Por outro lado, a modelagem que inclua o alcance total de distribuição da espécie permite aos algoritmos captarem a variabilidade ambiental de forma mais completa, gerando modelos mais próximos do real (BARBET-MASSIN et al., 2010; RAES, 2012; CARRETERO; SILLERO, 2016).

Torna-se, portanto, perceptível que há uma base compartilhada por SDMs e ENMs no processo de modelagem: a quantidade de pontos de presença, o tipo de ausência utilizado, e qual algoritmo fará o processamento dos dados. Falhas ou decisões equivocadas em um desses pontos (e.g., seleção da “ausência”, escolha do algoritmo) poderá levar ao enviesamento do modelo final e afetar, portanto, sua qualidade em predizer a adequabilidade do ambiental para a espécie ou a distribuição potencial.

Na avaliação estatística do modelo é possível quantificar o total de falsos positivos e falsos negativos. Um falso positivo será a predição incorreta de uma ausência ou a demarcação de um local como não adequado para a espécie quando este é, de fato, adequado. Em contrapartida, um falso negativo será o oposto: o modelo prediz uma presença falsa ou um local adequado que, na verdade, é inadequado. Para a modelagem, falsos positivos e negativos são denominados como erro de omissão e comissão, respectivamente (PETERSON; SOBERON, 2012).

Outro fator fundamental para o desenvolvimento de um modelo é o tamanho amostral representado pelos pontos de presença da espécie em estudo. Não existe um tamanho fixo ou ideal para a execução da modelagem e este tamanho dependerá das observações realizadas localmente ou disponíveis em bases de dados ou, ainda, do tipo de algoritmo utilizado (SILLERO et al., 2021).

Embora o acesso às grandes bases de dados virtuais permita ao pesquisador coletar uma quantidade considerável de ocorrências de uma espécie, muitos desses dados são duplicados ou necessitam de algum tipo de correção (e.g., remoção de centros urbanos). Raes

e Aguirre-Gutiérrez (2018) argumentam que não é possível inferir uma relação entre ocorrência e fatores ambientais a partir de poucos pontos de presença. Por outro lado, trabalhos como os de Papes e Gaubert (2007), Hernandez e colaboradores (2008) e van Proosdij e colaboradores (2016) demonstram como o tipo de algoritmo utilizado pode resultar em projeções adequadas mesmo quando o número de ocorrências é reduzido.

A variabilidade no tamanho amostral para o desenvolvimento de modelos SDMs e ENMs é evidenciada na literatura (CALL et al., 2016; SHCHEGLOVITOVA; ANDERSON, 2013; TRAN; VU, 2020), sendo possível observar modelos construídos com 3-16 pontos (VAN PROOSDIJ et al., 2016) até milhares de pontos de presença para uma única espécie (BOTELLA et al., 2018).

No entanto, é evidente na literatura que o tamanho amostral utilizado é um ponto crítico no processo de modelagem, uma vez que este influencia a acurácia dos modelos (BUISSON et al., 2010; TESSAROLO et al., 2014; THIBAUD et al., 2014). Qual o mínimo ou máximo de pontos de ocorrência a se utilizar na construção de um SDM ou de um ENM? Não há tamanho fixo, sendo necessário o pesquisador refletir com base nos demais passos da modelagem, bem como no objetivo do estudo para que tal decisão seja tomada (SILLERO et al., 2021).

Diante de um cenário de degradação ambiental e preocupação com a conservação da biodiversidade, este projeto avaliou o efeito da redução do tamanho amostral nos processos de modelagem da adequabilidade ambiental para cinco algoritmos, bem como testou a influência da quantidade de pseudo ausência e de pontos de *background* utilizados nas projeções de adequabilidade ambiental para três espécies de aves brasileiras.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Estrutura da base de dados

Os dados utilizados nos processos de modelagem foram classificados de acordo com a origem: pontos de ocorrência das espécies (i.e., dados bióticos), variáveis bioclimáticas das regiões de ocorrência (i.e., dados abióticos) e pontos de pseudo ausência e *background*.

2.2 Dados bióticos

Foram compilados os dados de ocorrência de três espécies de aves brasileiras: *Cantorchilus longirostris* Vieillot, 1819 (GBIF Secretariat, 2022), *Eupsittula cactorum* Kuhl, 1820 (GBIF Secretariat, 2022) e *Penelope obscura* Temminck, 1815 (GBIF Secretariat, 2022). A escolha das espécies foi baseada na distribuição espacial (i.e., tamanho da extensão da área de ocorrência) e na quantidade de pontos de observação disponíveis. Os dados de ocorrência foram obtidos na Global Biodiversity Information Facility - GBIF (www.gbif.org), uma base de dados global de acesso público a informações bióticas. O download, limpeza e processamento dos dados foi realizado no ambiente R 4.2.2 (R Core Team, 2022) a partir dos pacotes *rgbif* 3.7.5 (CHAMBERLAIN et al., 2023), *dplyr* 1.0.10 (WICKHAM et al., 2022) e *tidyr* 1.3.0 (WICKHAM, 2022).

Para cada espécie foi feito o download do máximo de pontos de ocorrência disponível no GBIF para cada espécie (Tabela 1). Por padrão foram eliminados dados redundantes (i.e., pontos duplicados), sem informações georreferenciadas e/ou com coordenadas incompletas, centróides municipais, bem como pontos localizados fora do Brasil ou no oceano. Para diminuir a incidência de erros no conjunto de dados, foram considerados válidas apenas observações oriundas de espécimes preservados em museu e de observação humana em campo.

Após a limpeza dos dados brutos, o conjunto de cada espécie foi analisado através do pacote *spThin* 0.2.0 (AIELLO-LAMMENS et al., 2015), sendo as ocorrências filtradas através da mensuração da distância geográfica entre cada ponto. Para tal, foi definido um raio limite de 10 km no entorno de cada presença e um número máximo de iterações do algoritmo de seleção em 100 réplicas. Isso permitiu a definição do maior tamanho amostral (i.e., tamanho N1) que contenha um conjunto de ocorrências suficientes (i.e., dados únicos e geograficamente espaçados) para modelar a espécie (Tabela 1).

Tabela 1 - Total de pontos obtidos para definição do tamanho amostral principal (N1)

Espécie	Total	Total após filtragem	Total final (N1)
<i>Cantorchilus longirostris</i>	11723	2819	645
<i>Eupsittula cactorum</i>	6004	2391	658
<i>Penelope obscura</i>	23668	5055	1131

Subsequente às etapas de limpeza e filtragem geográfica dos dados, os pontos foram sorteados aleatoriamente e sem reposição, a partir do tamanho N1, e separados em quatro tamanhos amostrais, numerados de N1 a N4, contendo ocorrências consideradas válidas para a modelagem (Tabela 2).

Tabela 2 - Subdivisão dos pontos de ocorrência das espécies em tamanhos amostrais com base no sorteio do conjunto principal (N1)

Espécie	Tamanho amostral	Total de pontos
<i>Cantorchilus longirostris</i>	N1	645
	N2	100
	N3	50
	N4	20
<i>Eupsittula cactorum</i>	N1	658
	N2	100
	N3	50
	N4	20
<i>Penelope obscura</i>	N1	1131
	N2	100
	N3	50
	N4	20

A divisão dos pontos de ocorrência em diferentes conjuntos (Tabela 2) se baseia nos diferentes tamanhos amostrais encontrados na literatura (Tabela 3). As diferenças de amostras refletem uma discussão em aberto, uma vez que não há um “tamanho ideal” e a decisão quanto ao total de pontos a se utilizar no modelo torna-se particular a cada situação (SILLERO et al., 2021).

Atingir um consenso ou, ainda, padronizar todos os tipos de modelos de distribuição ou de nicho é complexo uma vez que a quantidade de pontos utilizados em um modelo depende de fatores diversos que, muitas vezes, não estão sob o controle do pesquisador (e.g.

espécie pouco conhecida ou amostrada em campo, endemismo, estado de conservação crítico) ou, ainda, relacionam-se a configurações específicas dos algoritmos utilizados na modelagem.

Tabela 3 - Exemplos de publicações de aplicação SDM e ENM com diferentes tamanhos amostrais obtidos a partir de um levantamento na base de dados Web of Science (2013-2021)

Autor	Ano	Tamanho amostral
Shcheglovitova & Anderson	2013	6 e 9
Carvalho & Del Lama	2015	7 a 103
Spiers et al.	2018	5 a 10
Coello et al.	2021	60
Tran & Vu	2020	36
Botella et al.	2018	120 a 14278
Franco & Camps	2020	85
Jurestovsky & Andrew Joyner	2018	87
Flessner et al.	2017	102
Call et al.	2016	1578
Hemami et al.	2020	280
DeVaney	2016	719

Nota: os trabalhos de Shcheglovitova & Anderson (2013), Carvalho & Del Lama (2015) e Botella et al. (2018) fazem uso de tamanhos amostrais variados para diferentes espécies.

Características intrínsecas da programação dos algoritmos podem levar o pesquisador a utilizar um conjunto de pontos maior ou menor, dependendo, novamente, do tipo de algoritmo utilizado. Isto se torna evidente ao se analisar a literatura voltada à aplicação dos modelos de distribuição onde diversos trabalhos apontam, por exemplo, o algoritmo Maxent como satisfatório na produção de modelos robustos a partir de um conjunto de pontos reduzido (ELITH et al., 2006; PEARSON et al., 2006; WISZ et al., 2008; KUMAR & STOHLGREN, 2009; QIN et al., 2017; EL-GABBAS & DORMANN, 2018; MCCARTNEY et al., 2019; LI et al., 2020).

2.3 Dados abióticos

O conjunto de variáveis bioclimáticas (Tabela 4) foi obtido a partir da base de dados WorldClim, versão 2.1, contendo dados climáticos para o intervalo 1970 - 2000 (FICK & HIJMANS, 2017). Os arquivos originais possuem uma resolução espacial variável de 30 segundos a 10 arco-minutos, com área aproximada de 1 km² e 340 km² no equador terrestre, respectivamente, em formato GeoTiff (.tif).

Tabela 4 - Variáveis bioclimáticas obtidas na base de dados WorldClim, versão 2.1 e respectiva codificação

Código	Variável
BIO1	Temperatura média anual
BIO2	Faixa diurna média
BIO3	Isotermalidade
BIO4	Sazonalidade de temperatura
BIO5	Temperatura máxima do mês mais quente
BIO6	Temperatura mínima do mês mais frio
BIO7	Faixa anual de temperatura
BIO8	Temperatura média do trimestre mais úmido
BIO9	Temperatura média do trimestre mais seco
BIO10	Temperatura média do trimestre mais quente
BIO11	Temperatura média do trimestre mais frio
BIO12	Precipitação anual
BIO13	Precipitação do mês mais chuvoso
BIO14	Precipitação do mês mais seco
BIO15	Sazonalidade de precipitação
BIO16	Precipitação do trimestre mais úmido
BIO17	Precipitação do trimestre mais seco
BIO18	Precipitação do trimestre mais quente

Devido ao tamanho da área de amostragem ser o território brasileiro, as variáveis bioclimáticas foram amostradas em resolução de 2.5 arco-minutos, o que equivale a aproximadamente 5 km² no equador terrestre. A escolha da resolução espacial se baseou na precisão pretendida na modelagem, bem como na área de cobertura dos modelos (HIJMANS et al., 2005). Posteriormente, as variáveis bioclimáticas foram recortadas para a área de estudo através do pacote *raster* 3.6-14 (HIJMANS & ETTEN, 2023).

Anterior ao processo de modelagem, se faz necessário analisar o parâmetro de colinearidade entre as diversas variáveis que podem vir a compor o treinamento de um modelo. Entende-se por colinearidade ou multicolinearidade o efeito da correlação entre dois ou mais preditores utilizados no treinamento de um modelo (NAIMI; ARAÚJO, 2016). De acordo com Dormann e colaboradores (2013), diversos modelos estatísticos (e.g., técnicas de regressão) são suscetíveis a problemas de colinearidade, podendo enviesar sua validação estatística.

Isto posto, para cada espécie, foram cruzadas as 19 variáveis ambientais do Worldclim com os dados de observação para calcular o Fator de Inflação da Variância (VIF sigla em inglês) a partir do pacote *usdm* 1.1-18 (NAIMI et al., 2014). O cálculo do VIF se baseia em uma metodologia mais precisa do que os métodos de correlação de Spearman e Pearson, uma vez que é capaz de demonstrar a força com que uma variável é explicada pelas demais no conjunto de dados (NAIMI & ARAÚJO, 2016).

A seleção das variáveis por meio do VIF se baseia no valor de R^2 resultante da regressão de uma variável em função das outras no conjunto. Esse processo excluirá aquela com valor VIF acima do limite de corte, sendo repetido até que todas as variáveis com alta colinearidade sejam excluídas. De acordo com Chatterjee e Hadi (2006), valores de VIF acima de 10, como regra geral, indicam problemas de colinearidade no modelo. Isto posto, esse valor foi adotado como corte para o cálculo do VIF nas variáveis ambientais para as três espécies estudadas.

2.4 Seleção de pseudo ausência e background

Os dados das espécies obtidos a partir da base de dados GBIF são caracterizados apenas como dados de presença, uma vez que dados de ausência das espécies não estão disponíveis. Para efeitos de modelagem e de acordo com a metodologia empregada (i.e., tipo de modelo e algoritmo utilizado), faz-se necessário o uso de dados de ausência. Para tal, o pesquisador precisará selecionar dados de *background* ou, ainda, gerar dados de pseudo ausência (HANBERRY et al., 2012; BROTONS et al., 2004).

A seleção de *background* ou de pseudo ausência pode ser processada de diferentes formas, a depender da metodologia utilizada (ČENGÍČ et al., 2020; CHAPMAN et al., 2019; ITURBIDE et al., 2018; SENAY et al., 2013). Com base nos trabalhos de Wisz e Guisan (2009), Barbet-Massin e colaboradores (2012) e Liang e colaboradores, (2018), foram selecionadas três metodologias para a geração de pontos de *background* e pseudo ausência: seleção aleatória, seleção por exclusão geográfica ao redor dos pontos de observação e seleção por envelope climático.

Os pontos gerados foram subdivididos em quatro conjuntos durante o processo de seleção (Tabela 5), baseado nas recomendações de Barbet-Massin e colaboradores (2012).

Tabela 5 - Subdivisão dos pontos de *background* e de pseudo ausência

Conjunto	Total de pontos
S1	10000
S2	1000
S3	250
S4	1:1

Nota: seleção na proporção 1:1, onde para cada ponto de presença foi gerado um ponto de *background* ou de pseudo ausência.

2.5 Seleção dos algoritmos

Existem diferentes algoritmos aplicáveis aos processos de modelagem. Para seleção de quais iriam compor o estudo, foi realizado um levantamento na base de dados Web of Science no ano de 2019, para o período 2009-2019, através das palavras-chave “*species distribution model*” AND “*ecological niche model*”. Para o referido período foram obtidas 2963

publicações que se enquadraram nos processos de modelagem de distribuição de espécies ou de nicho ecológico.

O resultado foi filtrado sem distinção de área e/ou espécie estudada com base no trabalho de Raes e Aguirre-Gutiérrez (2018), onde os autores apresentam uma listagem dos algoritmos mais utilizados na modelagem de distribuição de espécies. Foi calculado, então, o percentual de aplicação dos algoritmos mais referenciados (Tabela 6) e, a partir destes, feita a seleção para o trabalho.

Tabela 6 - Total de publicações e percentual de aplicação dos algoritmos mais frequentes em trabalhos de distribuição de espécies e de nicho ecológico segundo Raes e Aguirre-Gutiérrez (2018) para o período 2009-2019

Algoritmo	Total de publicações	Percentual de aplicação
Maxent	1564	53%
Random Forests	811	27%
GLM	176	6%
GAM	174	6%
GARP	104	4%
DOMAIN	64	2%
BIOCLIM	70	2%
Total	2963	100%

Com base nos resultados bibliográficos, foram selecionados: GLM, GAM, Maxent e Random Forests para os procedimentos de modelagem das espécies, utilizando como critério de seleção o número de publicações. O algoritmo BIOCLIM também foi incluído com base na facilidade de aplicação e pelo pioneirismo da metodologia (BOOTH et al., 2013).

2.6 Estimativa da Extensão de Ocorrência das espécies

Para comparar as variações de cada modelo e avaliar os processos de modelagem, fez-se necessário obter um modelo padrão de área para cada espécie que contenha o máximo

dos pontos de distribuição. Para tal, estimou-se a área de ocorrência das espécies através do cálculo da Extensão de Ocorrência - EOO, em inglês.

Entende-se por Extensão de Ocorrência a área contida no interior do menor limite contínuo que pode ser traçado de modo a conter todos os locais de ocorrência, quer real, quer potenciais ou projetados de uma espécie (IUCN, 2012). O EOO de uma espécie pode ser traçado através de um mínimo polígono convexo (ou através de um *alpha Hull*) onde todos os locais de ocorrência estão contidos em seu interior (IUCN, 2012). Para estimar o EOO das espécies utilizou-se o pacote *ConR* 1.3.0 (DAUBY et al., 2017), sendo mapeado e plotado a Extensão de Ocorrência (Figura 1).

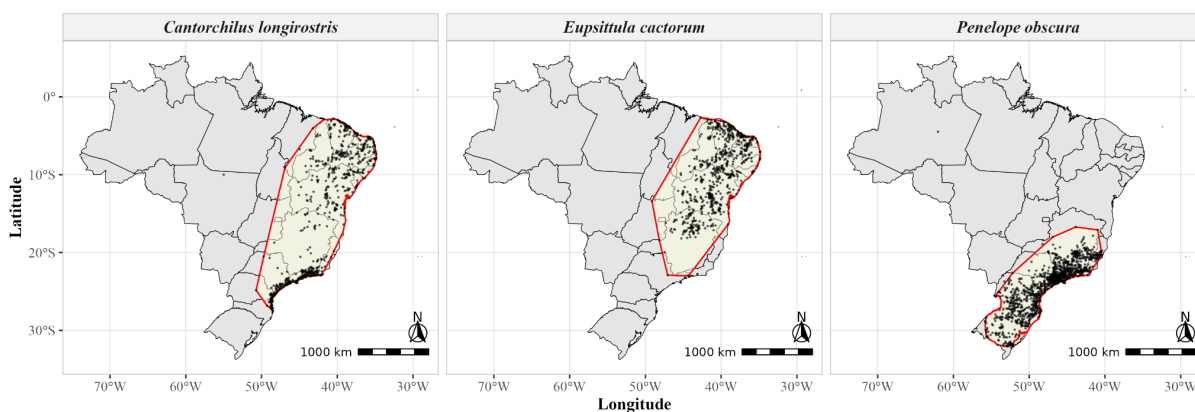


Figura 1 - Distribuição dos pontos de ocorrência obtidos na base de dados GBIF dentro das Extensões de Ocorrência para as espécies estudadas. A linha vermelha denota os limites do EOO. A versão em alta resolução se encontra no Material Suplementar 1. Fonte: autoria própria (2022).

O polígono calculado para representar o EOO das espécies, bem como o valor da área interna, foi utilizado para a comparar as áreas dos modelos gerados por cada algoritmo, combinação de tamanhos amostrais, métodos e grupos de pontos de *background* e pseudo ausência.

2.7 Modelagem, validação estatística e consenso entre modelos

Todos os modelos foram produzidos no R 4.2.2 (R CORE TEAM, 2022) através do pacote *sdm* 1.1-8 (NAIMI & ARAÚJO, 2016). Tomando por base a divisão do conjunto amostral em N1-N4 e dos pontos de *background* e pseudo ausência em S1-S4, o processo de modelagem (Figura 2) foi separado com base no tipo de amostragem de *background* e pseudo

ausência (i.e. aleatória, por exclusão geográfica, por envelope climático) e pelo algoritmo utilizado.

A calibração (i.e., *fitting* em inglês) dos modelos foi configurada de modo que 75% da amostra fosse utilizada como treinamento e 25% como teste, sendo o processo repetido 10 vezes para cada algoritmo através da técnica de *subsampling*, sem reposição dos dados a cada reamostragem. Os cinco algoritmos selecionados para o trabalho foram utilizados em sequência, em sua configuração padrão, e fazendo uso dos mesmos procedimentos de treinamento/teste descritos anteriormente.

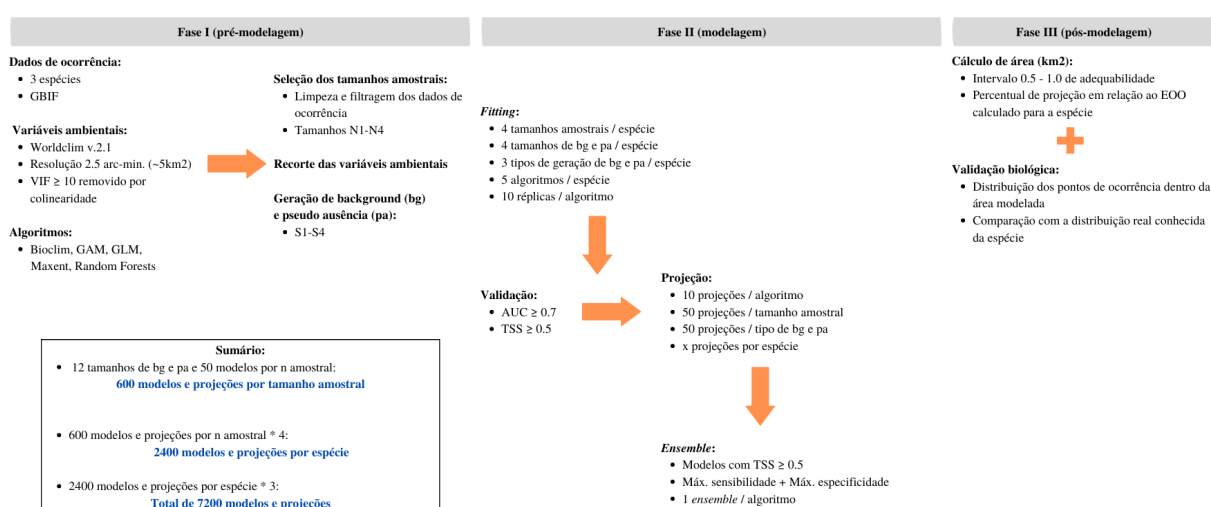


Figura 2 - Visão geral do processo de construção de um SDM / ENM, desde a obtenção dos dados biológicos e ambientais, calibração (i.e. *fitting*) do modelo e posterior avaliação. A versão em alta resolução se encontra no Material Suplementar 1. Fonte: autoria própria (2022).

A validação dos modelos se baseou no valor de ROC/AUC e TSS por serem medidas bastante utilizadas na avaliação de modelagens de distribuição de espécies e de nicho ecológico (ALLOUCHE; TSOAR; KADMON, 2006; FIELDING; BELL, 1997).

Tendo em vista que não há um único algoritmo com performance superior aos demais e, portanto, os resultados de uma modelagem podem diferir significativamente dentro do mesmo conjunto de dados (ELITH et al., 2006; THUILLER, 2004), a metodologia dos modelos de consenso (*ensemble*, em inglês) se apresenta como uma alternativa para combinar os diferentes algoritmos em uma única projeção final (MARMION et al., 2009), permitindo enxergar a tendência central dos modelos gerados pelos diferentes algoritmos.

Após a projeção individual dos modelos para cada algoritmo, foi realizada uma seleção dos melhores modelos para construção de um ensemble baseado no valor de TSS, de modo a maximizar a sensibilidade e especificidade dos modelos. Deste modo, apenas os modelos que atenderam aos critérios (i.e., máxima sensibilidade/especificidade) foram selecionados e combinados em uma única projeção final para medição das áreas.

De forma a validar a interferência do tamanho amostral no ensemble, inclui-se, também, uma medida de área modelada onde a projeção total do modelo (i.e. somatória da áreas de adequabilidade de 0% a 100%) e, também, das áreas projetadas para as probabilidades de 75% e 100%, respectivamente. Os valores projetados foram comparados por algoritmo (i.e. tamanhos amostrais e pseudo ausência por algoritmo e entre algoritmos) e com o polígono de estimação da área de ocorrência da espécie.

REFERÊNCIAS

- AIELLO-LAMMENS, Matthew E.; BORJA, Robert A.; RADOSAVLJEVIC, Aleksandar; VILELA, Bruno; ANDERSON, Robert P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. **Ecography**, [S. l.], v. 38, n. 5, p. 541–545, 2015. DOI: [10.1111/ecog.01132](https://doi.org/10.1111/ecog.01132). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01132>. Acesso em: 9 jan. 2023.
- ALLOUCHE, Omri; TSOAR, Asaf; KADMON, Ronen. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. **Journal of Applied Ecology**, [S. l.], v. 43, n. 6, p. 1223–1232, 2006. DOI: [10.1111/j.1365-2664.2006.01214.x](https://doi.org/10.1111/j.1365-2664.2006.01214.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2006.01214.x>. Acesso em: 10 jan. 2023.
- BARBET-MASSIN, Morgane; JIGUET, Frédéric; ALBERT, Cécile Hélène; THUILLER, Wilfried. Selecting pseudo-absences for species distribution models: how, where and how many?: *How to use pseudo-absences in niche modelling?* **Methods in Ecology and Evolution**, [S. l.], v. 3, n. 2, p. 327–338, 2012. DOI: [10.1111/j.2041-210X.2011.00172.x](https://doi.org/10.1111/j.2041-210X.2011.00172.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00172.x>. Acesso em: 15 fev. 2022.
- BARBET-MASSIN, Morgane; THUILLER, Wilfried; JIGUET, Frédéric. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? **Ecography**, [S. l.], v. 33, n. 5, p. 878–886, 2010. DOI: [10.1111/j.1600-0587.2010.06181.x](https://doi.org/10.1111/j.1600-0587.2010.06181.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2010.06181.x>. Acesso em: 3 jan. 2023.

BOOTH, Trevor H.; NIX, Henry A.; BUSBY, John R.; HUTCHINSON, Michael F. bioclim : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. **Diversity and Distributions**, [S. l.], v. 20, n. 1, p. 1–9, 2014. DOI: [10.1111/ddi.12144](https://doi.org/10.1111/ddi.12144). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ddi.12144>. Acesso em: 10 jan. 2023.

BOTELLA, Christophe; JOLY, Alexis; BONNET, Pierre; MONESTIEZ, Pascal; MUNOZ, François. Species distribution modeling based on the automated identification of citizen observations. **Applications in Plant Sciences**, [S. l.], v. 6, n. 2, p. e1029, 2018. DOI: [10.1002/aps3.1029](https://doi.org/10.1002/aps3.1029). Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5851560/>. Acesso em: 10 jan. 2023.

BROTONS, Lluís; THUILLER, Wilfried; ARAÚJO, Miguel B.; HIRZEL, Alexandre H. Presence-Absence versus Presence-Only Modelling Methods for Predicting Bird Habitat Suitability. **Ecography**, [S. l.], v. 27, n. 4, p. 437–448, 2004. DOI: [10.2307/3683417](https://doi.org/10.2307/3683417). Disponível em: <https://www.jstor.org/stable/3683417>. Acesso em: 4 maio. 2022.

BUISSON, Laëtitia; THUILLER, Wilfried; CASAJUS, Nicolas; LEK, Sovan; GRENOUILLET, Gaël. Uncertainty in ensemble forecasting of species distribution. **Global Change Biology**, [S. l.], v. 16, n. 4, p. 1145–1157, 2010. DOI: [10.1111/j.1365-2486.2009.02000.x](https://doi.org/10.1111/j.1365-2486.2009.02000.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2009.02000.x>. Acesso em: 25 maio. 2022.

CALL, Ashley; SUN, Yan-Xia; YU, Yan; PEARMAN, Peter B.; THOMAS, David T.; TRIGIANO, Robert N.; CARBONE, Ignazio; XIANG, Qiu-Yun Jenny. Genetic structure and post-glacial expansion of *Cornus florida* L. (Cornaceae): integrative evidence from phylogeography, population demographic history, and species distribution modeling: Phylogeography of *Cornus florida*. **Journal of Systematics and Evolution**, [S. l.], v. 54, n. 2, p. 136–151, 2016. DOI: [10.1111/jse.12171](https://doi.org/10.1111/jse.12171). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/jse.12171>. Acesso em: 24 maio. 2022.

Cantorichilus longirostris (Vieillot, 1819) in GBIF Secretariat (2022). GBIF Backbone Taxonomy. Checklist dataset <https://doi.org/10.15468/39omei> acesso via GBIF.org em: 10 jan 2023.

CARRETERO, Miguel A.; SILLERO, Neftalí. Evaluating how species niche modelling is affected by partial distributions with an empirical case. **Acta Oecologica**, [S. l.], v. 77, p. 207–216, 2016. DOI: [10.1016/j.actao.2016.08.014](https://doi.org/10.1016/j.actao.2016.08.014). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1146609X16301710>. Acesso em: 31 out. 2022.

CARVALHO, Antônio F.; DEL LAMA, Marco Antonio. Predicting priority areas for conservation from historical climate modelling: stingless bees from Atlantic Forest hotspot as a case study. **Journal of Insect Conservation**, [S. l.], v. 19, n. 3, p. 581–587, 2015. DOI: [10.1007/s10841-015-9780-7](https://doi.org/10.1007/s10841-015-9780-7). Disponível em: <https://doi.org/10.1007/s10841-015-9780-7>. Acesso em: 10 jan. 2023.

ČENGÍĆ, Mirza; ROST, Jasmijn; REMENSKA, Daniela; JANSE, Jan H.; HUIJBREGTS, Mark A. J.; SCHIPPER, Aafke M. On the importance of predictor choice, modelling technique, and number of pseudo-absences for bioclimatic envelope model performance. **Ecology and Evolution**, [S. l.], v. 10, n. 21, p. 12307–12317, 2020. DOI: [10.1002/ece3.6859](https://doi.org/10.1002/ece3.6859). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.6859>. Acesso em: 10 jan. 2023.

CHAPMAN, Daniel; PESCONT, Oliver L.; ROY, Helen E.; TANNER, Rob. Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. **Journal of Biogeography**, [S. l.], v. 46, n. 5, p. 1029–1040, 2019. DOI: [10.1111/jbi.13555](https://doi.org/10.1111/jbi.13555). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/jbi.13555>. Acesso em: 10 jan. 2023.

CHATTERJEE, Samprit; HADI, Ali S. **Regression analysis by example**. 4th ed ed. Hoboken, N.J: Wiley-Interscience, 2006.

COELLO, Alberto J.; FERNÁNDEZ-MAZUECOS, Mario; GARCÍA-VERDUGO, Carlos; VARGAS, Pablo. Phylogeographic sampling guided by species distribution modeling reveals the Quaternary history of the Mediterranean–Canarian *Cistus monspeliensis* (Cistaceae). **Journal of Systematics and Evolution**, [S. l.], v. 59, n. 2, p. 262–277, 2021. DOI: [10.1111/jse.12570](https://doi.org/10.1111/jse.12570). Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jse.12570>. Acesso em: 10 jan. 2023.

COLWELL, Robert K.; RANGEL, Thiago F. Hutchinson’s duality: The once and future niche. **Proceedings of the National Academy of Sciences**, [S. l.], v. 106, n. supplement_2, p. 19651–19658, 2009. DOI: [10.1073/pnas.0901650106](https://doi.org/10.1073/pnas.0901650106). Disponível em: <https://pnas.org/doi/full/10.1073/pnas.0901650106>. Acesso em: 3 jan. 2023.

DAUBY, Gilles et al. *ConR*: An R package to assist large-scale multispecies preliminary conservation assessments using distribution data. **Ecology and Evolution**, [S. l.], v. 7, n. 24, p. 11292–11303, 2017. DOI: [10.1002/ece3.3704](https://doi.org/10.1002/ece3.3704). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.3704>. Acesso em: 10 jan. 2023.

DESCOMBES, Patrice; CHAUVIER, Yohann; BRUN, Philipp; RIGHETTI, Damiano; WÜEST, Rafael O.; KARGER, Dirk N.; ZURELL, Damaris; ZIMMERMANN, Niklaus E. **Strategies for sampling pseudo-absences for species distribution models in complex mountainous terrain**. [s.l.] : Ecology, 2022. DOI: [10.1101/2022.03.24.485693](https://doi.org/10.1101/2022.03.24.485693). Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2022.03.24.485693>. Acesso em: 2 jan. 2023.

DEVANEY, Shannon C. Species Distribution Modeling of Deep Pelagic Eels. **Integrative and Comparative Biology**, [S. l.], v. 56, n. 4, p. 524–530, 2016. DOI: [10.1093/icb/icw032](https://doi.org/10.1093/icb/icw032). Disponível em: <https://academic.oup.com/icb/article-lookup/doi/10.1093/icb/icw032>. Acesso em: 10 jan. 2023.

DORMANN, Carsten F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, [S. l.], v. 36, n. 1, p. 27–46, 2013. DOI: [10.1111/j.1600-0587.2012.07348.x](https://doi.org/10.1111/j.1600-0587.2012.07348.x). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2012.07348.x>. Acesso em: 10 jan. 2023.

EL-GABBAS, Ahmed; DORMANN, Carsten F. Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling. **Ecology and Evolution**, [S. l.], v. 8, n. 4, p. 2196–2206, 2018. DOI: [10.1002/ece3.3834](https://doi.org/10.1002/ece3.3834). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.3834>. Acesso em: 10 jan. 2023.

ELITH, Jane et al. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, [S. l.], v. 29, n. 2, p. 129–151, 2006. DOI: [10.1111/j.2006.0906-7590.04596.x](https://doi.org/10.1111/j.2006.0906-7590.04596.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.2006.0906-7590.04596.x>. Acesso em: 15 fev. 2022.

ELITH, Jane; LEATHWICK, John R. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. **Annual Review of Ecology, Evolution, and Systematics**, [S. l.], v. 40, n. 1, p. 677–697, 2009. DOI: [10.1146/annurev.ecolsys.110308.120159](https://doi.org/10.1146/annurev.ecolsys.110308.120159). Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.110308.120159>. Acesso em: 23 abr. 2019.

ENGLER, Robin; GUISAN, Antoine; RECHSTEINER, Luca. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. **Journal of Applied Ecology**, [S. l.], v. 41, n. 2, p. 263–274, 2004. DOI: [10.1111/j.0021-8901.2004.00881.x](https://doi.org/10.1111/j.0021-8901.2004.00881.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.0021-8901.2004.00881.x>. Acesso em: 24 maio. 2022.

Eupsittula cactorum (Kuhl, 1820) in GBIF Secretariat (2022). GBIF Backbone Taxonomy. Checklist dataset <https://doi.org/10.15468/39omei> acesso via GBIF.org em: 10 jan. 2023.

FICK, Stephen E.; HIJMANS, Robert J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology**, [S. l.], v. 37, n. 12, p. 4302–4315, 2017. DOI: [10.1002/joc.5086](https://doi.org/10.1002/joc.5086). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/joc.5086>. Acesso em: 10 jan. 2023.

FIELDING, Alan H.; BELL, John F. A review of methods for the assessment of prediction errors in conservation presence/absence models. **Environmental Conservation**, [S. l.], v. 24, n. 1, p. 38–49, 1997. DOI: [10.1017/S0376892997000088](https://doi.org/10.1017/S0376892997000088). Disponível em: https://www.cambridge.org/core/product/identifier/S0376892997000088/type/journal_article. Acesso em: 10 jan. 2023.

FLESSNER, Brandon; HENRY, Mary C.; GREEN, Jerry. Species Distribution Modeling of American Beech (Fagus Grandifolia) Distribution in Southwest Ohio: **International Journal of Applied Geospatial Research**, [S. l.], v. 8, n. 3, p. 16–36, 2017. DOI: [10.4018/ijagr.2017070102](https://doi.org/10.4018/ijagr.2017070102). Disponível em:

<https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/ijagr.2017070102>.

Acesso em: 10 jan. 2023.

FRANCO, Francisco; CAMPS, Gonzalo Andrés. La aplicación de modelos de distribución de especies para la realización de inferencias arqueológicas. Una ejemplificación a partir de *Geoffroea decorticans* en el área Sudcalchaquí (Noroeste, Argentina). **InterSecciones en Antropología**, [S. l.], v. 21, n. 2, p. 131–144, 2020. DOI: [10.37176/iea.21.2.2020.489](https://doi.org/10.37176/iea.21.2.2020.489).

Disponível em:

<https://interseccionesantro.soc.unicen.edu.ar/index.php/intersecciones/article/view/489>.

Acesso em: 10 jan. 2023.

FRANKLIN, Janet; MILLER, Jennifer A. **Mapping species distributions: spatial inference and prediction**. Cambridge ; New York: Cambridge University Press, 2009.

GU, Weidong; SWIHART, Robert K. Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. **Biological Conservation**, [S. l.], v. 116, n. 2, p. 195–203, 2004. DOI: [10.1016/S0006-3207\(03\)00190-3](https://doi.org/10.1016/S0006-3207(03)00190-3). Disponível em:

<https://linkinghub.elsevier.com/retrieve/pii/S0006320703001903>. Acesso em: 15 fev. 2022.

GUISAN, Antoine et al. Predicting species distributions for conservation decisions. **Ecology Letters**, [S. l.], v. 16, n. 12, p. 1424–1435, 2013. DOI: [10.1111/ele.12189](https://doi.org/10.1111/ele.12189). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/ele.12189>. Acesso em: 10 fev. 2022.

GUISAN, Antoine; THUILLER, Wilfried. Predicting species distribution: offering more than simple habitat models. **Ecology Letters**, [S. l.], v. 8, n. 9, p. 993–1009, 2005. DOI:

[10.1111/j.1461-0248.2005.00792.x](https://doi.org/10.1111/j.1461-0248.2005.00792.x). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2005.00792.x>. Acesso em: 15 fev. 2022.

GUISAN, Antoine; THUILLER, Wilfried; ZIMMERMANN, Niklaus E. **Habitat Suitability and Distribution Models: With Applications in R**. Cambridge: Cambridge University Press, 2017. DOI: [10.1017/9781139028271](https://doi.org/10.1017/9781139028271). Disponível em:

<http://ebooks.cambridge.org/ref/id/CBO9781139028271>. Acesso em: 4 maio. 2022.

GUISAN, Antoine; ZIMMERMANN, Niklaus E. Predictive habitat distribution models in ecology. **Ecological Modelling**, [S. l.], v. 135, n. 2–3, p. 147–186, 2000. DOI:

[10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9). Disponível em:

<https://linkinghub.elsevier.com/retrieve/pii/S0304380000003549>. Acesso em: 4 maio. 2022.

HALLGREN, W.; SANTANA, F.; LOW-CHOY, S.; ZHAO, Y.; MACKEY, B. Species distribution models can be highly sensitive to algorithm configuration. **Ecological Modelling**, [S. l.], v. 408, p. 108719, 2019. DOI: [10.1016/j.ecolmodel.2019.108719](https://doi.org/10.1016/j.ecolmodel.2019.108719). Disponível em:

<https://linkinghub.elsevier.com/retrieve/pii/S0304380019302194>. Acesso em: 25 maio. 2022.

HANBERRY, Brice B.; HE, Hong S.; PALIK, Brian J. Pseudoabsence Generation Strategies for Species Distribution Models. **PLoS ONE**, [S. l.], v. 7, n. 8, p. e44486, 2012. DOI:

[10.1371/journal.pone.0044486](https://doi.org/10.1371/journal.pone.0044486). Disponível em:
<https://dx.plos.org/10.1371/journal.pone.0044486>. Acesso em: 10 jan. 2023.

HEMAMI, Mahmoud-Reza; KHOSRAVI, Rasoul; GROVES, Colin; AHMADI, Mohsen. Morphological diversity and ecological niche divergence in goitered and sand gazelles. **Ecology and Evolution**, [S. l.], v. 10, n. 20, p. 11535–11548, 2020. DOI: [10.1002/ece3.6789](https://doi.org/10.1002/ece3.6789). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.6789>. Acesso em: 10 jan. 2023.

HERNANDEZ, P. A. et al. Predicting species distributions in poorly-studied landscapes. **Biodiversity and Conservation**, [S. l.], v. 17, n. 6, p. 1353–1366, 2008. DOI: [10.1007/s10531-007-9314-z](https://doi.org/10.1007/s10531-007-9314-z). Disponível em:
<http://link.springer.com/10.1007/s10531-007-9314-z>. Acesso em: 31 out. 2022.

HIJMANS, Robert J. et al. **raster: Geographic Data Analysis and Modeling**. , 2023. Disponível em: <https://CRAN.R-project.org/package=raster>. Acesso em: 10 jan. 2023.

HIJMANS, Robert J.; CAMERON, Susan E.; PARRA, Juan L.; JONES, Peter G.; JARVIS, Andy. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, [S. l.], v. 25, n. 15, p. 1965–1978, 2005. DOI: [10.1002/joc.1276](https://doi.org/10.1002/joc.1276). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/joc.1276>. Acesso em: 10 jan. 2023.

HIJMANS, Robert J.; ELITH, Jane. **Species Distribution Models**. , 2021. Disponível em: https://rspsatial.org/raster/sdm/raster_SDM.pdf. Acesso em: 14 fev. 2022.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. **Climate Change and Land: IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems**. 1. ed. [s.l.] : Cambridge University Press, 2022. DOI: [10.1017/9781009157988](https://doi.org/10.1017/9781009157988). Disponível em:
<https://www.cambridge.org/core/product/identifier/9781009157988/type/book>. Acesso em: 3 jan. 2023.

IPBES. **Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services**. [s.l.] : Zenodo, 2019. DOI: [10.5281/ZENODO.3831673](https://doi.org/10.5281/ZENODO.3831673). Disponível em:
<https://zenodo.org/record/3831673>. Acesso em: 7 fev. 2022.

ITURBIDE, Maialen; BEDIA, Joaquín; GUTIÉRREZ, José Manuel. Background sampling and transferability of species distribution model ensembles under climate change. **Global and Planetary Change**, [S. l.], v. 166, p. 19–29, 2018. DOI: [10.1016/j.gloplacha.2018.03.008](https://doi.org/10.1016/j.gloplacha.2018.03.008). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0921818117302308>. Acesso em: 25 maio. 2022.

IUCN. **Iucn Red List Categories And Criteria: Version 3.1. Second Edition.** [s.l.] : Zenodo, 2012. DOI: [10.5281/ZENODO.15039](https://doi.org/10.5281/ZENODO.15039). Disponível em: <https://zenodo.org/record/15039>. Acesso em: 10 jan. 2023.

JETZ, Walter; MCPHERSON, Jana M.; GURALNICK, Robert P. Integrating biodiversity distribution knowledge: toward a global map of life. **Trends in Ecology & Evolution**, [S. l.], v. 27, n. 3, p. 151–159, 2012. DOI: [10.1016/j.tree.2011.09.007](https://doi.org/10.1016/j.tree.2011.09.007). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169534711002679>. Acesso em: 11 fev. 2022.

JURESTOVSKY, Derek; ANDREW JOYNER, T. Applications of species distribution modeling for palaeontological fossil detection: late Pleistocene models of Saiga (Artiodactyla: Bovidae, Saiga tatarica). **Palaeobiodiversity and Palaeoenvironments**, [S. l.], v. 98, n. 2, p. 277–285, 2018. DOI: [10.1007/s12549-017-0298-8](https://doi.org/10.1007/s12549-017-0298-8). Disponível em: <https://doi.org/10.1007/s12549-017-0298-8>. Acesso em: 10 jan. 2023.

KUMAR, Sunil; STOHLGREN, Thomas J. Maxent modeling for predicting suitable habitat for threatened and endangered tree *Canacomyrica monticola* in New Caledonia. **Journal of Ecology and Natural Environment**, [S. l.], v. 1, n. 4, p. 94–98, 2009. DOI: <https://doi.org/10.5897/JENE.9000071>. Disponível em: http://www.academicjournals.org/app/webroot/article/article1379515268_Kumar%20and%20Stohlgren.pdf.

LIANG, Wanwan; PAPEŞ, Monica; TRAN, Liem; GRANT, Jerome; WASHINGTON-ALLEN, Robert; STEWART, Scott; WIGGINS, Gregory. The effect of pseudo-absence selection method on transferability of species distribution models in the context of non-adaptive niche shift. **Ecological Modelling**, [S. l.], v. 388, p. 1–9, 2018. DOI: [10.1016/j.ecolmodel.2018.09.018](https://doi.org/10.1016/j.ecolmodel.2018.09.018). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380018303156>. Acesso em: 10 jan. 2023.

LI, Junjun; FAN, Gang; HE, Yang. Predicting the current and future distribution of three *Coptis* herbs in China under climate change conditions, using the MaxEnt model and chemical analysis. **Science of The Total Environment**, [S. l.], v. 698, p. 134141, 2020. DOI: [10.1016/j.scitotenv.2019.134141](https://doi.org/10.1016/j.scitotenv.2019.134141). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S004896971934118X>. Acesso em: 10 jan. 2023.

LOBO, Jorge M.; JIMÉNEZ-VALVERDE, Alberto; HORTAL, Joaquín. The uncertain nature of absences and their importance in species distribution modelling. **Ecography**, [S. l.], v. 33, n. 1, p. 103–114, 2010. DOI: [10.1111/j.1600-0587.2009.06039.x](https://doi.org/10.1111/j.1600-0587.2009.06039.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2009.06039.x>. Acesso em: 24 maio. 2022.

MACKENZIE, Darryl I.; ROYLE, J. Andrew. Designing occupancy studies: general advice and allocating survey effort: *Designing occupancy studies*. **Journal of Applied Ecology**, [S. l.], v. 42, n. 6, p. 1105–1114, 2005. DOI: [10.1111/j.1365-2664.2005.01098.x](https://doi.org/10.1111/j.1365-2664.2005.01098.x). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2005.01098.x>. Acesso em: 15 fev. 2022.

MARMION, Mathieu; PARVIAINEN, Miia; LUOTO, Miska; HEIKKINEN, Risto K.; THUILLER, Wilfried. Evaluation of consensus methods in predictive species distribution modelling. **Diversity and Distributions**, [S. l.], v. 15, n. 1, p. 59–69, 2009. DOI: [10.1111/j.1472-4642.2008.00491.x](https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00491.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00491.x>. Acesso em: 10 jan. 2023.

MCCARTNEY, Kevin R.; KUMAR, Sunil; SING, Sharlene E.; WARD, Sarah M. Using invaded-range species distribution modeling to estimate the potential distribution of *Linaria* species and their hybrids in the U.S. northern Rockies. **Invasive Plant Science and Management**, [S. l.], v. 12, n. 02, p. 97–111, 2019. DOI: [10.1017/inp.2019.15](https://doi.org/10.1017/inp.2019.15). Disponível em: https://www.cambridge.org/core/product/identifier/S1939729119000154/type/journal_article. Acesso em: 10 jan. 2023.

MUSCATELLO, Angela; ELITH, Jane; KUJALA, Heini. How decisions about fitting species distribution models affect conservation outcomes. **Conservation Biology**, [S. l.], v. 35, n. 4, p. 1309–1320, 2021. DOI: [10.1111/cobi.13669](https://doi.org/10.1111/cobi.13669). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/cobi.13669>. Acesso em: 10 fev. 2022.

NAIMI, Babak; ARAÚJO, Miguel B. sdm: a reproducible and extensible R platform for species distribution modelling. **Ecography**, [S. l.], v. 39, n. 4, p. 368–375, 2016. DOI: [10.1111/ecog.01881](https://doi.org/10.1111/ecog.01881). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01881>. Acesso em: 10 jan. 2023.

NAIMI, Babak; HAMM, Nicholas A. S.; GROEN, Thomas A.; SKIDMORE, Andrew K.; TOXOPEUS, Albertus G. Where is positional uncertainty a problem for species distribution modelling? **Ecography**, [S. l.], v. 37, n. 2, p. 191–203, 2014. DOI: [10.1111/j.1600-0587.2013.00205.x](https://doi.org/10.1111/j.1600-0587.2013.00205.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2013.00205.x>. Acesso em: 10 jan. 2023.

PAPES, M.; GAUBERT, P. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents: Ecological niche modelling of poorly known viverrids. **Diversity and Distributions**, [S. l.], v. 13, n. 6, p. 890–902, 2007. DOI: [10.1111/j.1472-4642.2007.00392.x](https://doi.org/10.1111/j.1472-4642.2007.00392.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2007.00392.x>. Acesso em: 31 out. 2022.

PEARSON, Richard G.; RAXWORTHY, Christopher J.; NAKAMURA, Miguel; TOWNSEND PETERSON, A. ORIGINAL ARTICLE: Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar: Predicting species distributions with low sample sizes. **Journal of Biogeography**, [S. l.], v. 34, n. 1, p. 102–117, 2006. DOI: [10.1111/j.1365-2699.2006.01594.x](https://doi.org/10.1111/j.1365-2699.2006.01594.x). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2699.2006.01594.x>. Acesso em: 10 jan. 2023.

Penelope obscura Temminck, 1815 in GBIF Secretariat (2022). GBIF Backbone Taxonomy. Checklist dataset <https://doi.org/10.15468/39omei> acesso via GBIF.org em: 10 jan. 2023.

PETERSON, A. Townsend (ORG.). **Ecological niches and geographic distributions**. Princeton, N.J: Princeton University Press, 2011.

PETERSON, A. Townsend; SOBERÓN, Jorge. Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. **Natureza & Conservação**, [S. l.], v. 10, n. 2, p. 102–107, 2012. DOI: [10.4322/natcon.2012.019](https://doi.org/10.4322/natcon.2012.019). Disponível em: <http://doi.editoracubo.com.br/10.4322/natcon.2012.019>. Acesso em: 24 maio. 2022.

PHILLIPS, Steven J.; DUDÍK, Miroslav; ELITH, Jane; GRAHAM, Catherine H.; LEHMANN, Anthony; LEATHWICK, John; FERRIER, Simon. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. **Ecological Applications**, [S. l.], v. 19, n. 1, p. 181–197, 2009. DOI: [10.1890/07-2153.1](https://doi.org/10.1890/07-2153.1). Disponível em: <http://doi.wiley.com/10.1890/07-2153.1>. Acesso em: 25 maio. 2022.

PINTO-LEDEZMA, Jesús N.; CAVENDER-BARES, Jeannine. Predicting species distributions and community composition using satellite remote sensing predictors. **Scientific Reports**, [S. l.], v. 11, n. 1, p. 16448, 2021. DOI: [10.1038/s41598-021-96047-7](https://doi.org/10.1038/s41598-021-96047-7). Disponível em: <https://www.nature.com/articles/s41598-021-96047-7>. Acesso em: 7 fev. 2022.

PROOSDIJ, André S. J.; SOSEF, Marc S. M.; WIERINGA, Jan J.; RAES, Niels. Minimum required number of specimen records to develop accurate species distribution models. **Ecography**, [S. l.], v. 39, n. 6, p. 542–552, 2016. DOI: [10.1111/ecog.01509](https://doi.org/10.1111/ecog.01509). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01509>. Acesso em: 15 fev. 2022.

QIN, Aili; LIU, Bo; GUO, Quanshui; BUSSMANN, Rainer W.; MA, Fanqiang; JIAN, Zunji; XU, Gexi; PEI, Shunxiang. Maxent modeling for predicting impacts of climate change on the potential distribution of *Thuja sutchuenensis* Franch., an extremely endangered conifer from southwestern China. **Global Ecology and Conservation**, [S. l.], v. 10, p. 139–146, 2017. DOI: [10.1016/j.gecco.2017.02.004](https://doi.org/10.1016/j.gecco.2017.02.004). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2351989416301470>. Acesso em: 10 jan. 2023.

RAES, Niels. Partial versus Full Species Distribution Models. **Natureza & Conservação**, [S. l.], v. 10, n. 2, p. 127–138, 2012. DOI: [10.4322/natcon.2012.020](https://doi.org/10.4322/natcon.2012.020). Disponível em: <http://doi.editoracubo.com.br/10.4322/natcon.2012.020>. Acesso em: 3 jan. 2023.

RAES, Niels; AGUIRRE-GUTIÉRREZ, Jesús. A Modeling Framework to Estimate and Project Species Distributions in Space and Time. Em: **Mountains, Climate and Biodiversity**. 1st. ed. Hoboken, NJ: John Wiley & Sons, 2018.

RONDININI, Carlo; WILSON, Kerrie A.; BOITANI, Luigi; GRANTHAM, Hedley; POSSINGHAM, Hugh P. Tradeoffs of different types of species occurrence data for use in systematic conservation planning: Species data for conservation planning. **Ecology Letters**,

[*S. l.*], v. 9, n. 10, p. 1136–1145, 2006. DOI: [10.1111/j.1461-0248.2006.00970.x](https://doi.org/10.1111/j.1461-0248.2006.00970.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2006.00970.x>. Acesso em: 15 fev. 2022.

SENAY, Senait D.; WORNER, Susan P.; IKEDA, Takayoshi. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. **PLoS ONE**, [*S. l.*], v. 8, n. 8, p. e71218, 2013. DOI: [10.1371/journal.pone.0071218](https://doi.org/10.1371/journal.pone.0071218). Disponível em: <https://dx.plos.org/10.1371/journal.pone.0071218>. Acesso em: 10 jan. 2023.

SHCHEGLOVITOVA, Mariya; ANDERSON, Robert P. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. **Ecological Modelling**, [*S. l.*], v. 269, p. 9–17, 2013. DOI: [10.1016/j.ecolmodel.2013.08.011](https://doi.org/10.1016/j.ecolmodel.2013.08.011). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380013004043>. Acesso em: 24 maio. 2022.

SILLERO, Neftalí; ARENAS-CASTRO, Salvador; ENRIQUEZ-URZELAI, Urtzi; VALE, Cândida Gomes; SOUSA-GUEDES, Diana; MARTÍNEZ-FREIRÍA, Fernando; REAL, Raimundo; BARBOSA, A. Márcia. Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. **Ecological Modelling**, [*S. l.*], v. 456, p. 109671, 2021. DOI: [10.1016/j.ecolmodel.2021.109671](https://doi.org/10.1016/j.ecolmodel.2021.109671). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380021002301>. Acesso em: 31 out. 2022.

SILLERO, Neftalí; BARBOSA, A. Márcia. Common mistakes in ecological niche models. **International Journal of Geographical Information Science**, [*S. l.*], v. 35, n. 2, p. 213–226, 2021. DOI: [10.1080/13658816.2020.1798968](https://doi.org/10.1080/13658816.2020.1798968). Disponível em: <https://www.tandfonline.com/doi/full/10.1080/13658816.2020.1798968>. Acesso em: 24 maio. 2022.

SOBERON, Jorge; PETERSON, A. Townsend. Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. **Biodiversity Informatics**, [*S. l.*], v. 2, n. 0, 2005. DOI: [10.17161/bi.v2i0.4](https://doi.org/10.17161/bi.v2i0.4). Disponível em: <https://journals.ku.edu/jbi/article/view/4>. Acesso em: 3 jan. 2023.

SPIERS, Joshua A.; OATHAM, Michael P.; ROSTANT, Luke V.; FARRELL, Aidan D. Applying species distribution modelling to improving conservation based decisions: a gap analysis of Trinidad and Tobago's endemic vascular plants. **Biodiversity and Conservation**, [*S. l.*], v. 27, n. 11, p. 2931–2949, 2018. DOI: [10.1007/s10531-018-1578-y](https://doi.org/10.1007/s10531-018-1578-y). Disponível em: <https://doi.org/10.1007/s10531-018-1578-y>. Acesso em: 10 jan. 2023.

TESSAROLO, Geiziane; RANGEL, Thiago F.; ARAÚJO, Miguel B.; HORTAL, Joaquín. Uncertainty associated with survey design in Species Distribution Models. **Diversity and Distributions**, [*S. l.*], v. 20, n. 11, p. 1258–1269, 2014. DOI: [10.1111/ddi.12236](https://doi.org/10.1111/ddi.12236). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ddi.12236>. Acesso em: 31 out. 2022.

THIBAUD, Emeric; PETITPIERRE, Blaise; BROENNIMANN, Olivier; DAVISON, Anthony C.; GUIBAN, Antoine. Measuring the relative effect of factors affecting species distribution model predictions. **Methods in Ecology and Evolution**, [*S. l.*], v. 5, n. 9, p.

947–955, 2014. DOI: [10.1111/2041-210X.12203](https://doi.org/10.1111/2041-210X.12203). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12203>. Acesso em: 31 out. 2022.

THUILLER, Wilfried. Patterns and uncertainties of species' range shifts under climate change: SPECIES' RANGE SHIFTS. **Global Change Biology**, [S. l.], v. 10, n. 12, p. 2020–2027, 2004. DOI: [10.1111/j.1365-2486.2004.00859.x](https://doi.org/10.1111/j.1365-2486.2004.00859.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2004.00859.x>. Acesso em: 10 jan. 2023.

TRAN, Dung Van; VU, Tinh Tien. Combining species distribution modeling and distance sampling to assess wildlife population size: A case study with the northern yellow-cheeked gibbon (*Nomascus annamensis*). **American Journal of Primatology**, [S. l.], v. 82, n. 9, 2020. DOI: [10.1002/ajp.23169](https://doi.org/10.1002/ajp.23169). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ajp.23169>. Acesso em: 24 maio. 2022.

WICKHAM, Hadley; FRANÇOIS, Romain; HENRY, Lionel; MÜLLER, Kirill; RSTUDIO. **dplyr: A Grammar of Data Manipulation**. , 2022. Disponível em: <https://CRAN.R-project.org/package=dplyr>. Acesso em: 10 jan. 2023.

WICKHAM, Hadley; GIRLICH, Maximilian; RSTUDIO. **tidyr: Tidy Messy Data**. , 2022. Disponível em: <https://CRAN.R-project.org/package=tidyr>. Acesso em: 10 jan. 2023.

WISZ, M. S.; HIJMANS, R. J.; LI, J.; PETERSON, A. T.; GRAHAM, C. H.; GUIBAN, A.; NCEAS PREDICTING SPECIES DISTRIBUTIONS WORKING GROUP†. Effects of sample size on the performance of species distribution models. **Diversity and Distributions**, [S. l.], v. 14, n. 5, p. 763–773, 2008. DOI: [10.1111/j.1472-4642.2008.00482.x](https://doi.org/10.1111/j.1472-4642.2008.00482.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00482.x>. Acesso em: 10 jan. 2023.

WISZ, Mary S.; GUIBAN, Antoine. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. **BMC Ecology**, [S. l.], v. 9, n. 1, p. 8, 2009. DOI: [10.1186/1472-6785-9-8](https://doi.org/10.1186/1472-6785-9-8). Disponível em: <http://bmcecol.biomedcentral.com/articles/10.1186/1472-6785-9-8>. Acesso em: 10 jan. 2023.

ZANIEWSKI, A. Elizabeth; LEHMANN, Anthony; OVERTON, Jacob McC. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. **Ecological Modelling**, [S. l.], v. 157, n. 2–3, p. 261–280, 2002. DOI: [10.1016/S0304-3800\(02\)00199-0](https://doi.org/10.1016/S0304-3800(02)00199-0). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380002001990>. Acesso em: 24 maio. 2022.

CAPÍTULO 1

THE EFFECT OF SAMPLE SIZE ON SPECIES DISTRIBUTION AND ECOLOGICAL NICHE MODELS: A CASE STUDY WITH THREE BRAZILIAN BIRDS

THE EFFECT OF SAMPLE SIZE ON SPECIES DISTRIBUTION AND ECOLOGICAL NICHE MODELS: A CASE STUDY WITH THREE BRAZILIAN BIRDS

Gustavo Reis de Brito¹ Fernando Frei²

¹Programa de Pós-Graduação em Biociências, Faculdade de Ciências e Letras, São Paulo State University, Assis, São Paulo, Avenue Dom Antônio, 2100, 190806-900, Brazil

²Department of Biology, Faculdade de Ciências e Letras, São Paulo State University, Assis, São Paulo, Avenue Dom Antônio, 2100, 190806-900, Brazil

Corresponding author: Gustavo Reis de Brito (reis.brito@unesp.br)

Abstract

The knowledge of species distribution is important for conservation planning, risk assessment, and to further ecological research. It can be derived from correlating environmental factors with occurrence records to create suitability and potential distribution predictions. This is known as Species Distribution Modeling (SDM) and Ecological Niche Modeling (ENM). Although these techniques gained popularity in recent years, there is still some confusion regarding important aspects of the modeling processes such as sample size influence in the fitting of the models. Using four different sample sizes (N1-N4) and two sets of pseudo absence and background points ($n = 1000$ and $n = 10000$) for three species of Brazilian birds, we evaluate the effect of small sample sizes on model performance and predicted suitable areas with five common SDM/ENM algorithms: Bioclim, Generalized Additive Models (GAM), Generalized Linear Models (GLM), Maximum Entropy (Maxent), and Random Forests. The reduction of the sample size to 20 points significantly affected the performance of the models, generating unreal projections that hold no biological significance when considering an interval of suitability of 0.5-1.0. AUC values for all algorithms showed poor performance for all sample sizes, with N4 showing further loss of model quality for Bioclim. The effect of sample size detected in this study can be minimized with the proper selection of algorithms that can deal with small samples, while also comparing models to a null model of the Extent of Occurrence (EOO) of the species regarding projected suitable areas for the 0.5-1.0 interval.

Keywords

sample size, ecological niche model, model evaluation, ensemble modeling

Introduction

One of the most widely used techniques to study species distribution is called Species Distribution Modeling (SDM) (Araújo and Peterson 2012). This approach combines computational and ecological concepts to devise models to explain the relationship between environmental variables and the occurrence of a species (Guisan and Zimmermann 2000, Peterson 2011, Araújo and Peterson 2012).

Correlative models are the most common approach to an SDM. This type of model aims to correlate occurrence data with environmental variables which, in turn, allows the model to predict a set of environmental factors that are relevant to the species (Guisan and Zimmermann 2000, Guisan and Thuiller 2005).

Different terminologies have been applied to the same techniques, such as Species Distribution Modeling, Ecological Niche Modeling, Bioclimatic Envelope Modeling, and Habitat Suitability Models (Elith and Leathwick 2009, Franklin and Miller 2009, Araújo and Peterson 2012). Although used as synonymous with one another, an SDM can not be classified as such when the main object being modeled is an entity in the domain of niches. Thus, many modern approaches labeled as SDMs are, in fact, Ecological Niche Models (ENM) and need to be treated as such based on the fact that both approaches (i.e., SDM and ENM) target different objects in geographic and environmental spaces (Peterson and Soberón 2012).

Despite nomenclatural confusion regarding the techniques applied, the field of SDMs/ENMs faces another problem: sample sizes. Papers such as Stockwell and Peterson (2002), Papeş and Gaubert (2007), Wisz et al. (2008), and Proosdij et al. (2016) are some examples of investigative studies regarding sample size influence in SDM/ENM methods. How many occurrence points of a species should be used for effective model calibration is, still, an open question in the literature (Sillero et al. 2021).

Biological meaningful models were produced with as low as 5 distinct occurrence records (Pearson et al. 2006, Shcheglovitova and Anderson 2013, Galante et al. 2018). Maxent (Phillips et al. 2006) is a well-known algorithm that can produce biologically significant models ranging from three to 13 presence points (van Proosdij et al. 2016), but other algorithms tend to be penalized with less than 30 observations.

Sillero et al. (2021) and Barbet-Massin et al. (2012) are important papers that deal with the implication of sample size in SDM/ENM fitting. Nevertheless, there is no consensus in the literature regarding sample size. What can be found are suggestions such as Field et al. (2012), Franklin (2010), and Stockwell and Peterson (2002).

Given the necessity to further investigate the effects of sample size on model performance and biological significance, we combined the predicted suitable area, the extension of occurrence, and traditional statistical metrics (i.e. AUC, TSS) to access species and ecological niche models performance for three Brazilian avian species.

Methods

Selection of species and database compilation

We selected three species of Brazilian birds for the modeling procedures: *Cantorchilus longirostris* (Vieillot, 1819), *Eupsittula cactorum* (Kuhl, 1820), and *Penelope obscura* (Temminck, 1815). All three species were selected based on 1) the number of occurrences and 2) their geographic distribution being constrained to Brazil. Geo-referenced presence records were obtained from the Global Biodiversity Information Facility - GBIF (www.gbif.org) via the *rgbif* 3.7.5 package (Chamberlain and Boettiger 2017); records with uncertain coordinates, absence of latitude or longitude, marked as NA (i.e. not available), and duplicates were eliminated from the data via *dplyr* 1.0.10 (Wickham et al. 2022b) and *tidyr* 1.3.0 (Wickham et al. 2022a) packages. The final datasets for all three species comprised unique and representative records: *C. longirostris* (n = 645), *E. cactorum* (n = 658), and *P. obscura* (n = 1131). All procedures were performed in R 4.2.2 (R Core Team 2022).

Selection of Environmental Variables

We obtained 19 bioclimatic variables from the WorldClim (www.worldclim.org) database at a resolution of 2.5 arc-minutes. All variables comprise version 2.1 of the WorldClim database (Fick and Hijmans 2017). A Variance Inflation Factor (VIF) test was conducted for each species to remove correlated variables to avoid over-fitting the models (Mateo et al. 2013) via the *usdm* 1.1-18 package (Naimi et al. 2014). A threshold of 10 was selected for the VIF test since variables with correlation values above it tend to overfit the model (Chatterjee and Hadi 2012). Only variables with collinearity values below the threshold were preserved for the modeling procedures. Given the three species are not from the same genus nor the same region, their ecological profile (e.g., environmental preferences) is different, which is reflected in the number of variables selected by the VIF technique (Table 1).

Table 1. The least correlated variables selected via VIF analysis for modeling

Species	Variables
<i>Cantorchilus longirostris</i>	BIO2, BIO4, BIO8, BIO13, BIO14, BIO15, BIO18, BIO19
<i>Eupsittula cactorum</i>	BIO2, BIO3, BIO4, BIO8, BIO13, BIO14, BIO15, BIO18, BIO19
<i>Penelope obscura</i>	BIO3, BIO5, BIO7, BIO8, BIO9, BIO12, BIO13, BIO18

Selection of sample sizes

To test the effect of the sample size in the fitting of an SDM/ENM, we created four sizes: N1 = the maximum number of unique points after data cleaning, N2 = 100, N3 = 50, and N4 = 20. Sizes N2-N4 were sampled from N1, with replacements to prevent repetitions. To define the number of occurrences in N2-N4 sizes, we reviewed the Web of Science database (www.webofscience.com) using “species distribution modeling” AND “ecological niche models” as keywords. We obtained a total of XX papers, from which 12 papers were selected to illustrate how sample size varies across applications of the same techniques (Suppl. material 1).

Modeling procedures

All ENMs were fitted using five widely used algorithms: Bioclim (Busby 1991), Generalized Additive Model - GAM (Hastie and Tibshirani 1986), Generalized Linear Model - GLM (McCullagh and Nelder 1998), Maxent (Phillips et al. 2006), and Random Forest (Breiman 2001). Algorithm selection was based on the work of Raes and Aguirre-Gutiérrez (2018). Modeling procedures were implemented in the *sdm* 1.1.8 package (Naimi and Araújo 2016) via the R 4.2.2 environment (R Core Team 2022), with all algorithms being used with the default configuration. Two sets of background and pseudo absence points were randomly generated from the study area, containing 1000 and 10000 points, respectively. We performed 100 runs per species (five algorithms x 10 subsampling runs x 2 sets of background and pseudo absence points), totalizing 300 model runs. The dataset was partitioned to evaluate model fitting, with 75% of the data used for training and 25% for testing. Model accuracy was analyzed via the True Skill Statistic analysis (Allouche et al. 2006) and the Area Under the

Curve (AUC) value from the Receiver Operating Characteristic (ROC) (Fielding and Bell 1997). To create the final projections, we applied the ensemble technique following Buisson et al. (2010) recommendations, defining two thresholds ($AUC \geq 0.7$ and $TSS \geq 0.4$) to exclude models with poor performance. Thus, only the best models were selected for the ensemble, yielding one single map for each algorithm.

Suitable area measurement

To assess the biological validity and the effects of sample size in model fitting, we measured 1) the total suitable area predicted (in square kilometers) and 2) areas above the 0.5 thresholds of suitability for each algorithm. Area values were calculated from the ensemble (Suppl. material 2).

Since model performance can present omission/commission errors, a minimum convex polygon was drawn to represent and constrain the Extent of Occurrence (EOO) for the three species using the *ConR* 1.3.0 package (Dauby 2020), following IUCN (2012) guidelines. Per definition, the EOO represents the area calculated from the minimum polygon that can be drawn to represent all sites of present occurrence of a species, be it known, inferred, or projected (IUCN 2012). We used the EOO polygon as a null model for the total area occupied by the species.

Suitability areas were then compared to the EOO area to assess biological viability and infer possible sample size influence in the fitting of the models by calculating the sum of predicted areas above the 0.5 thresholds of suitability. We decided to use 0.5 as the threshold for area calculation based on the hypothesis that, if the model should be used for conservation purposes, the best candidate areas for preservation should be, at least, 50% suitable for any given species to survive.

Results

Extent of Occurrence

The estimated Extent of Occurrence for each species (Suppl. Material 1) using the minimum convex polygon is consistent with the 2016 and 2018 IUCN Red List assessments (Table 2). However, differences can be noted given that all species occur beyond Brazilian frontiers and points that fall on other countries were excluded from the analysis.

Table 2. Comparison of the estimated Extent of Occurrence and IUCN Red List assessment of each species range

Species	EOO (km ²)	IUCN (km ²)
<i>Cantochilus longirostris</i>	2,330,207	2,240,000
<i>Eupsittula cactorum</i>	2,021,247	1,220,000
<i>Penelope obscura</i>	1,121,955	3,110,000

Suitable Area and Model Performance

Compared to the total area of the convex polygons (i.e. species EOO), the predicted suitable area for the 0.5-1.0 suitability interval for all five algorithms changed accordingly to sample size variation. Both sets of pseudo absence/background points also influenced the projection of the models (Figure 1, Figure 2).

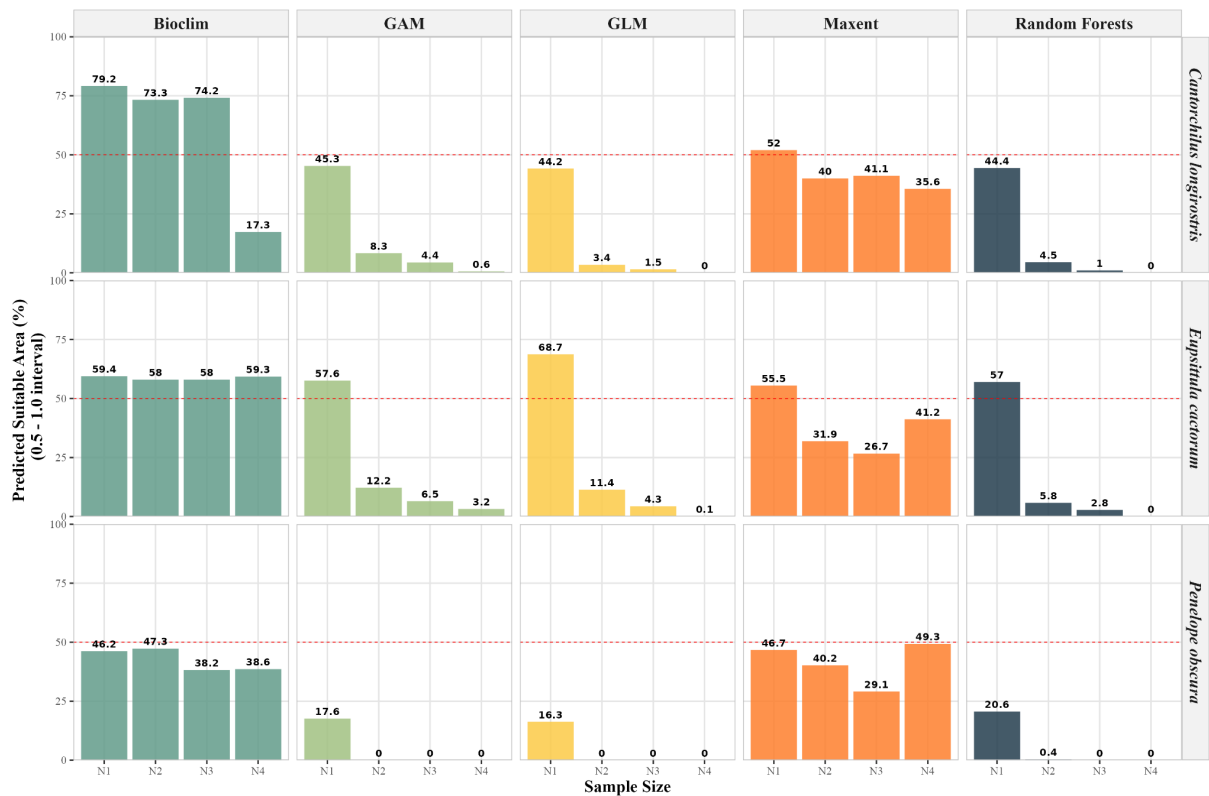


Figure 1. Percentual of predicted suitable area for all three species, considering the 0.5 - 1.0 threshold of suitability (red line). All models were fitted using the first set of pseudo absence/background points ($n = 1000$) and the EOO of the species constrained projections. A high-resolution version is available in Suppl. Material 1.

GAM, GLM, and Random Forests models showed the most noticeable area change when the sample size is reduced beyond N1. Although these models produced no area value when the suitability interval is considered for sizes N2-N4 (Figure 1, Figure 2), these algorithms were able to project suitable areas inside the EOO for all three species, although it holds little biological value due to increased omission and commission errors (Figure 3, Figure 4). All projections are available as Supplemental Material 1.

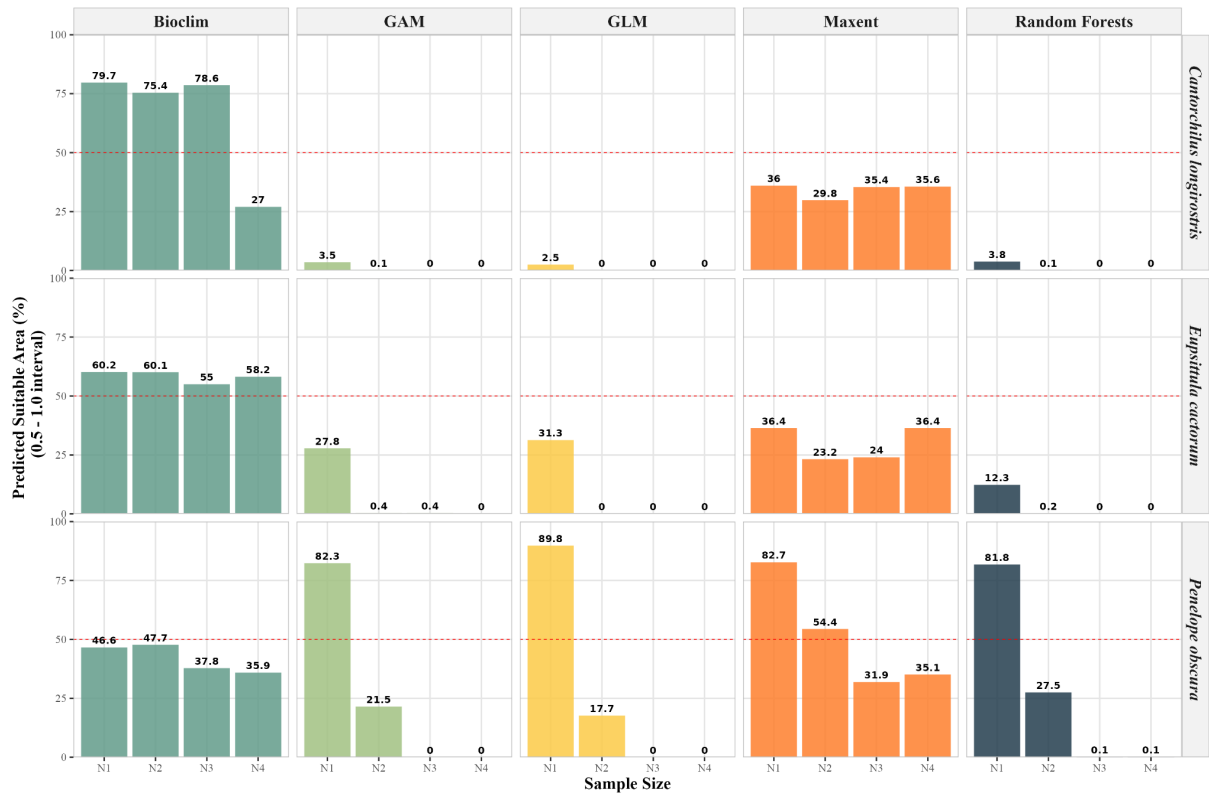


Figure 2. Percentual of predicted suitable area for all three species, considering the 0.5 - 1.0 threshold of suitability (red line). All models were fitted using the first set of pseudo absence/background points ($n = 10000$) and the EOO of the species-constrained projections. A high-resolution version is available in Suppl. Material 1.

In contrast, Bioclim and Maxent models projected more stable areas when the suitability interval is considered. Bioclim presented the most conservative approach, with a significant area retraction in N4 for *C. longirostris* and *P. obscura*; *E. cactorum* areas were stable across all four sample sizes. Maxent projections varied between sample sizes, dropping below the threshold of 0.5 starting from N2 size for all three species. For all algorithms, when sample size extremes are compared (i.e. N1 with N4), predicted suitable areas show significant reduction within the interval of 0.5 - 1.0 suitability. When all four sample sizes are considered, model performance tends to drop when departing from N1, with significant losses around N3 and N4.

Both sets of pseudo absence and background points showed influence in projected areas, with major losses being observed with $n = 10000$, except for Bioclim and Maxent models, although the biological validity of all models cannot be ascertained.

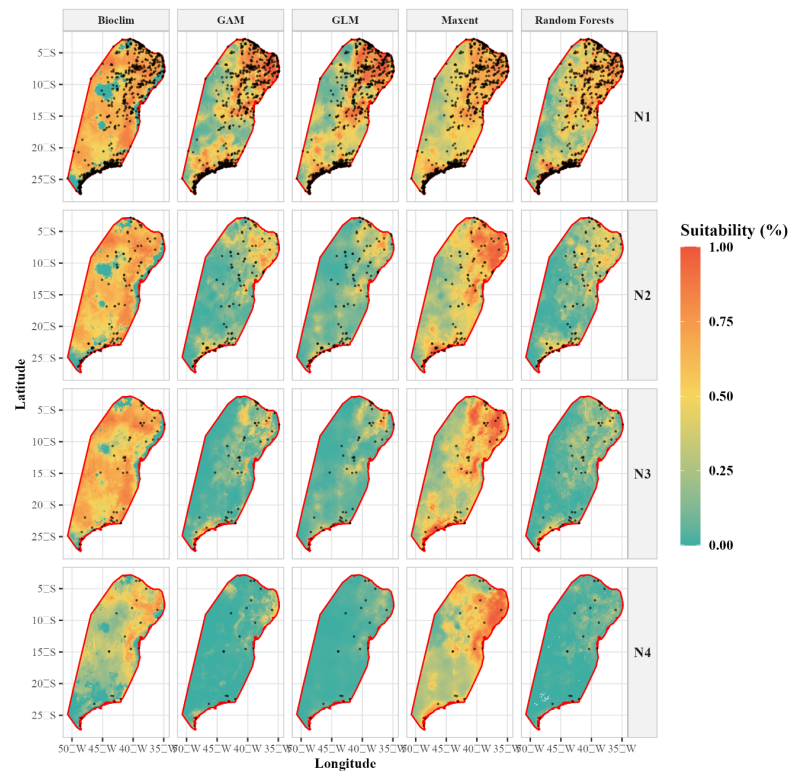


Figure 3. Projected models of *Cantorchilus longirostris* based on all five algorithms considering the full spectrum of suitability. The red line represents the EOO of the species; the dots represent the occurrence points of each sample size used to train the models. A set of $n = 1000$ pseudo absence and background points were used during model fitting.

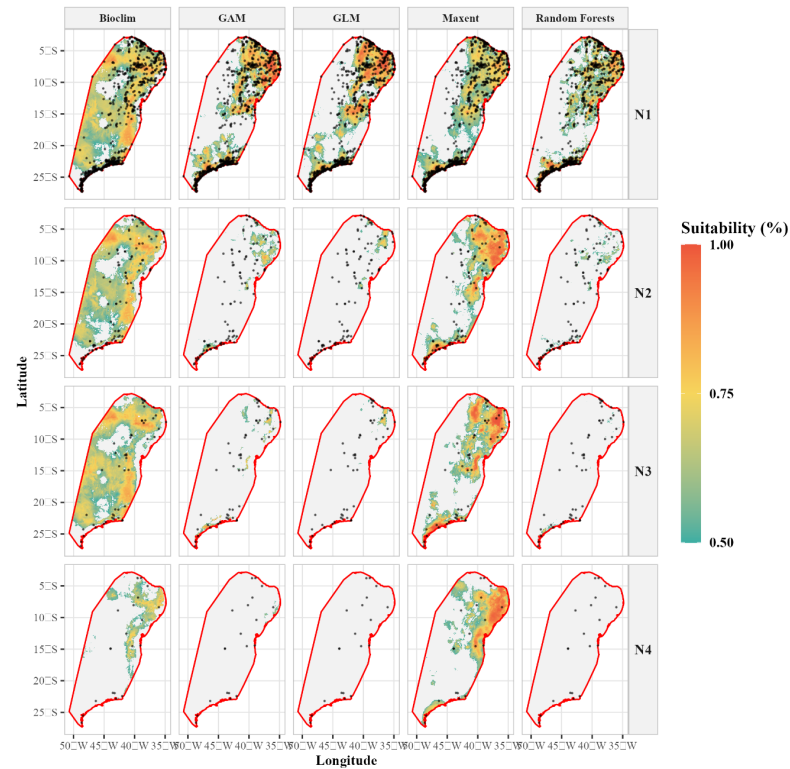


Figure 4. Projected models of *Cantorchilus longirostris* based on all five algorithms considering the threshold of 0.5 - 1.0 of suitability. The red line represents the EOO of the species; the dots represent the occurrence points of each sample size used to train the models. A set of $n = 1000$ pseudo absence and background points were used during model fitting.

As shown in Figures 3 and 4, model performance is hindered by the sample size reduction and when a threshold of suitability is defined. This behavior was observed for all three species (Suppl. material 3). During model fitting, GAM and GLM models failed when the distribution of presence and “absence” data was right-skewed (i.e. more absence points than occurrence points in the sample). This imbalance in data is further evidenced in Figure 2, where GAM, GLM, and Random Forest models produced low to zero suitability with sample sizes as N3 ($n = 50$) and N4 ($n = 20$). These models were trained with a set of pseudo absence/background of 10000, a ratio of 1:200 and 1:500, respectively.

Statistical Analysis

Mean AUC and TSS values showed no significant variation for sizes N1-N3 when comparisons were made within the same algorithm, independently of pseudo absence/background points quantity (Figure 5, Figure 6). Comparisons between algorithms show higher AUC and TSS values ($AUC > 0.7$, $TSS > 0.4$) for all algorithms except Bioclim; Random Forests presented the highest mean AUC across all three species for size N1 in both sets of pseudo absence/background points, with the lowest values being presented by Bioclim in all sizes.

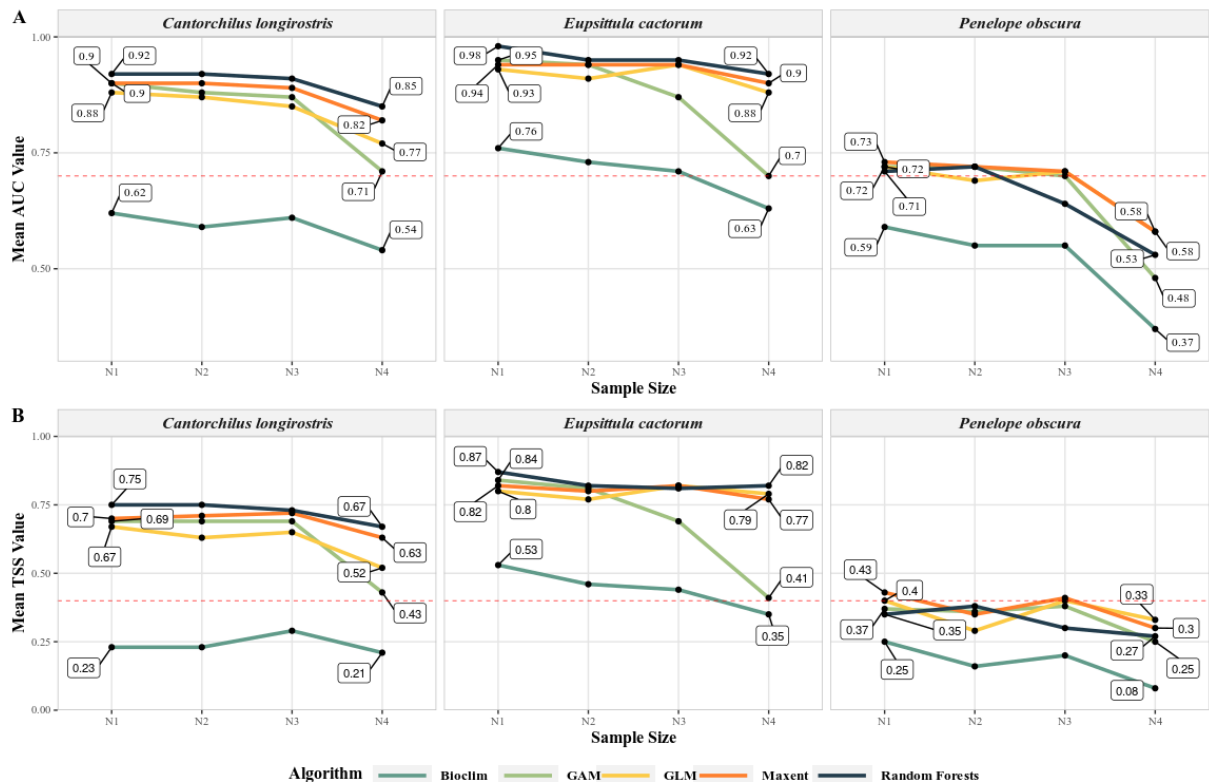


Figure 5. Mean AUC (A) and TSS (B) values for 1000 pseudo absence/background points for all three species. The values represent the average value per sample size.

Except for Bioclim, all algorithms presented significantly lower AUC and TSS values for sample size N4, with the lowest scores being registered by GAM ($AUC = 0.48$, $TSS = 0.25$).

for *Penelope obscura*) and Bioclim (AUC = 0.37, TSS = 0.08 for *Penelope obscura*), respectively, for models fitted with 1000 pseudo absence/background points (Figure 5 A and B). When the models were fitted with 10000 pseudo absence/background points, all algorithms except Bioclim presented values above the validation thresholds (AUC ≥ 0.7 and TSS ≥ 0.4) for *Cantorchilus longirostris* and *Eupsittula cactorum*; for the third species, *Penelope obscura*, all algorithms except Bioclim presented values above the validation thresholds only for sample N1, dropping in N2-N4. Bioclim scored above the thresholds only in *Eupsittula cactorum*, dropping below it in size N4.

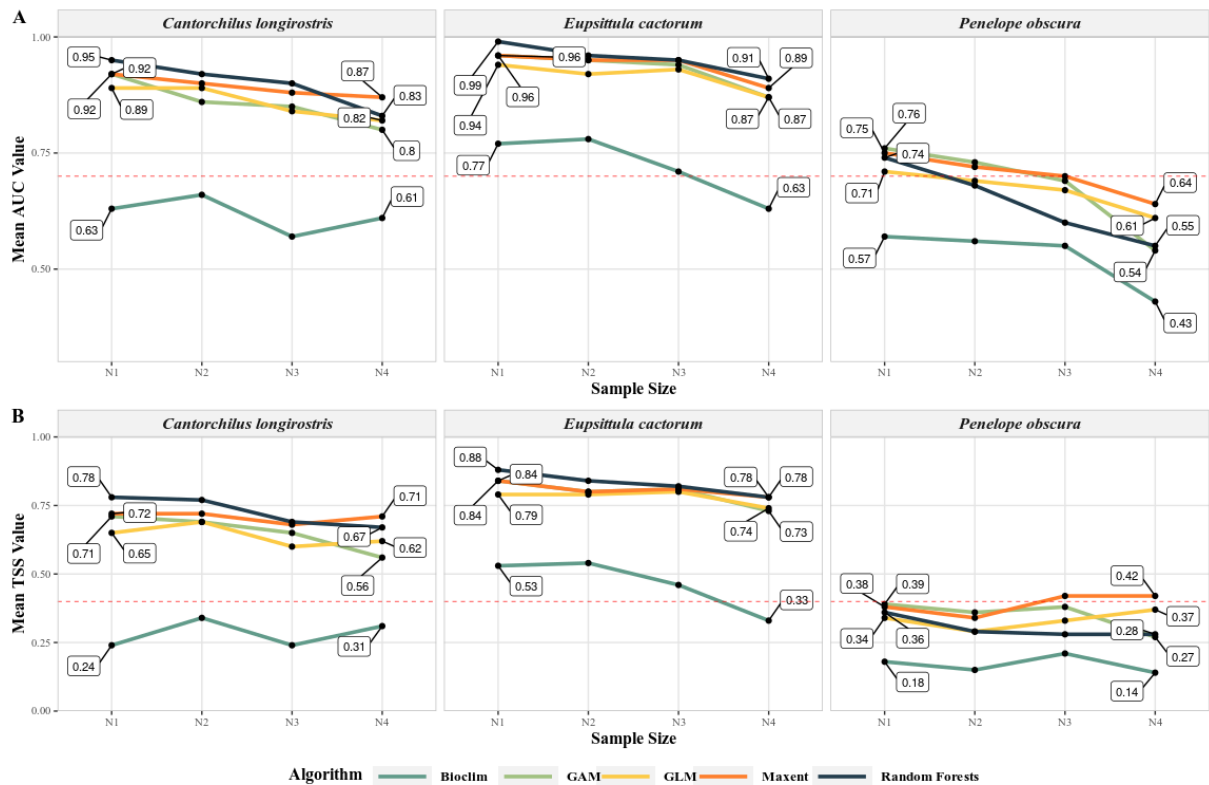


Figure 6. Mean AUC (A) and TSS (B) values for 10000 pseudo absence/background points for all three species. The values represent the average value per sample size.

Discussion

Sample size influence on model performance is well known in the literature (e.g. Buisson et al. 2010; Tessarolo et al. 2014; Thibaud et al. 2014), even though there is no consensus in the literature. Bioclim and Maxent are two common algorithms that show how sample size influences model performance: it's possible to fit a model with less than 20 occurrences but each algorithm will perform differently. Bioclim usually can predict a very large or a very small suitable area, while Maxent is well known for producing good results with samples as small as 13 (Papes and Gaubert 2007; Hernandez et al. 2008; van Proosdij et al. 2016). Although algorithm programming should be taken into account in model fitting, knowledge of the ecological traits of the species and the quality of the occurrence records are also crucial factors in model performance (Sillero et al. 2021).

While suitable areas predicted by Bioclim and Maxent models were relatively constant when considering sample sizes N1-N3, these areas showed reduction when the sample size is N4 (Figure 3). On the other hand, GAM, GLM, and Random Forests algorithms showed the most

prominent area reduction as sample size changes (Figure 1, Figure 2). This type of result is known as omission and commission errors (Peterson and Soberón 2012), which can hinder model performance, resulting in abnormal projections (Figure 3) and AUC values (Figure 4). Da Silva et al. (2020) suggest that uncertainty in model projections is not only related to algorithm programming, but to environmental variables not considered for the fitting of the models. Per Sillero et al. (2021) suggestions, our results show that to avoid these pitfalls, algorithms should be chosen based on the type of data available and the object of the model (i.e., environmental niche or the potential distribution of the species) (Peterson and Soberón 2012).

All three species showed a significant contraction of the predicted suitable area as the sample size is reduced from a maximum number of occurrences (N1) to as few as 20 (N4), which is to be expected since a species range influences environmental complexity (Sillero et al. 2021). When a threshold of suitability is defined (0.5 - 1.0), the size of predicted suitable areas can be reduced further or models can fail (Figure 3), which, for instance, can hinder the use of such models in conservation planning since habitat suitability/unsuitability is not a probability of a species occurrence or absence (Sillero et al. 2021).

For all three species, when considering all five algorithms, AUC values were high for all sample sizes except N4 (Figure 4). High AUC values are a rough estimate of model predictability of the niche or potential distribution of a species (Elith and Leathwick 2009), but it also can hide omission/commission errors since the ROC curve is well known for being influenced by the total extent where models are fitted (Lobo et al. 2008). By reducing the sample size across the study area (i.e. EOO), our results demonstrate how the algorithms see only part of the full range of environmental variation for the species, reflecting in small and not biologically significant suitable areas being projected. Carretero and Sillero (2016) state that partial distributions, represented by smaller and constrained sample sizes, can negatively affect an ENM/SDM by constraining the model to a space that does not correspond to the full extent of the species niche, which was seen in our results.

As a way to improve model quality and its biological significance, instead of choosing an algorithm by its popularity in the literature, we suggest fitting the models with at least 5 different algorithms and using no less than 50 unique presence records since performance tends to drop when occurrences are below that number. When more than one algorithm is used to model a species niche/potential distribution, we recommend the use of the ensemble technique as a way to maximize the specificity and sensibility of the models by capturing the best estimates of each algorithm into a single final projection (Marmion et al. 2009).

Regarding suitable area and/or potential distribution projections, we suggest the estimation of the Extent of Occurrence of a species and using it as a null model, comparing 1) the total predicted suitable area and 2) areas for the 0.5 - 1.0 interval to the EOO for final validation. Model evaluation can be misled by thresholding and the technique used since it can hide errors and mislead projection interpretation when high values (e.g. AUC and TSS > 0.8) are obtained. Algorithm selection, sample size, and the statistical techniques used to validate an ENM or SDM are crucial points that need to be thoroughly addressed to achieve statistical and biologically robust models.

Acknowledgments

We like to thank Ciro Branco and Pitágoras Bispo for their scientific contributions.

References

- Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. *Journal of Applied Ecology* 43: 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Araújo MB, Peterson AT (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology* 93: 1527–1539. <https://doi.org/10.1890/11-1930.1>
- Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* 3: 327–338. <https://doi.org/10.1111/j.2041-210X.2011.00172.x>
- Breiman L (2001) Random Forests. *Machine Learning* 45: 5–32. <https://doi.org/10.1023/A:1010933404324>
- Buisson L, Thuiller W, Casajus N, Lek S, Grenouillet G (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology* 16: 1145–1157. <https://doi.org/10.1111/j.1365-2486.2009.02000.x>
- Carretero MA, Sillero N (2016) Evaluating how species niche modelling is affected by partial distributions with an empirical case. *Acta Oecologica* 77: 207–216. <https://doi.org/10.1016/j.actao.2016.08.014>
- Chamberlain SA, Boettiger C (2017) R Python, and Ruby clients for GBIF species occurrence data. *PeerJ Preprints*. preprint <https://doi.org/10.7287/peerj.preprints.3304v1>
- Chatterjee S, Hadi AS (2012) *Regression analysis by example*. Fifth edition. Wiley, Hoboken, New Jersey, 393 pp.
- Da Silva FP, Fernandes-Ferreira H, Montes MA, da Silva LG (2020) Distribution modeling applied to deficient data species assessment: A case study with *Pithecopus nordestinus* (Anura, Phyllomedusidae). *Neotropical Biology and Conservation* 15: 165–175. <https://doi.org/10.3897/neotropical.15.e47426>
- Dauby G (2020) ConR: Computation of Parameters Used in Preliminary Assessment of Conservation Status. Available from: <https://CRAN.R-project.org/package=ConR> (October 23, 2022).
- Elith J, Leathwick JR (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 SE, Hijmans RJ (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>
- Field AP (2013) *Discovering statistics using IBM SPSS statistics: and sex and drugs and rock “n” roll*. 4th edition. Sage, Los Angeles, 915 pp.
- Franklin J, Miller JA (2009) *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge ; New York, 320 pp.

- Franklin J (2010) Mapping Species Distributions: Spatial Inference and Prediction. 1st ed. Cambridge University Press. <https://doi.org/10.1017/CBO9780511810602>
- Galante PJ, Alade B, Muscarella R, Jansa SA, Goodman SM, Anderson RP (2018) The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. *Ecography* 41: 726–736. <https://doi.org/10.1111/ecog.02909>
- Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147–186. [https://doi.org/10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9)
- Hastie T, Tibshirani R (1986) Generalized Additive Models. *Statistical Science* 1. <https://doi.org/10.1214/ss/1177013604>
- Hernandez PA, Franke I, Herzog SK, Pacheco V, Paniagua L, Quintana HL, Soto A, Swenson JJ, Tovar C, Valqui TH, Vargas J, Young BE (2008) Predicting species distributions in poorly-studied landscapes. *Biodiversity and Conservation* 17: 1353–1366. <https://doi.org/10.1007/s10531-007-9314-z>
- IUCN (2012) Iucn Red List Categories And Criteria: Version 3.1. Second Edition. Zenodo <https://doi.org/10.5281/ZENODO.15039>
- Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145–151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* 15: 59–69. <https://doi.org/10.1111/j.1472-4642.2008.00491.x>
- Mateo RG, Vanderpoorten A, Muñoz J, Laenen B, Désambré A (2013) Modeling Species Distributions from Heterogeneous Data for the Biogeographic Regionalization of the European Bryophyte Flora. Bond-Lamberty B (Ed.). *PLoS ONE* 8: e55648. <https://doi.org/10.1371/journal.pone.0055648>
- McCullagh P, Nelder JA (1998) Generalized linear models. 2nd ed. Chapman & Hall/CRC, Boca Raton, 511 pp.
- Naimi B, Araújo MB (2016) sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography* 39: 368–375. <https://doi.org/10.1111/ecog.01881>
- Naimi B, Hamm NAS, Groen TA, Skidmore AK, Toxopeus AG (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography* 37: 191–203. <https://doi.org/10.1111/j.1600-0587.2013.00205.x>
- Pearson RG, Raxworthy CJ, Nakamura M, Townsend Peterson A (2006) ORIGINAL ARTICLE: Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar: Predicting species distributions with low sample sizes. *Journal of Biogeography* 34: 102–117. <https://doi.org/10.1111/j.1365-2699.2006.01594.x>

- Papeş M, Gaubert P (2007) Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents: Ecological niche modelling of poorly known viverrids. *Diversity and Distributions* 13: 890–902. <https://doi.org/10.1111/j.1472-4642.2007.00392.x>
- Peterson AT (Ed.) (2011) *Ecological niches and geographic distributions*. Princeton University Press, Princeton, N.J, 314 pp.
- Peterson AT, Soberón J (2012) Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. *Natureza & Conservação* 10: 102–107. <https://doi.org/10.4322/natcon.2012.019>
- Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259. <https://doi.org/10.1016/j.ecolmodel.2005.03.026>
- Proosdij ASJ, Sosef MSM, Wieringa JJ, Raes N (2016) Minimum required number of specimen records to develop accurate species distribution models. *Ecography* 39: 542–552. <https://doi.org/10.1111/ecog.01509>
- R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Sillero N, Arenas-Castro S, Enriquez-Urzelai U, Vale CG, Sousa-Guedes D, Martínez-Freiría F, Real R, Barbosa AM (2021) Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. *Ecological Modelling* 456: 109671. <https://doi.org/10.1016/j.ecolmodel.2021.109671>
- Shcheglovitova M, Anderson RP (2013) Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling* 269: 9–17. <https://doi.org/10.1016/j.ecolmodel.2013.08.011>
- Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148: 1–13. [https://doi.org/10.1016/S0304-3800\(01\)00388-X](https://doi.org/10.1016/S0304-3800(01)00388-X)
- Tessarolo G, Rangel TF, Araújo MB, Hortal J (2014) Uncertainty associated with survey design in Species Distribution Models. Beaumont L (Ed.). *Diversity and Distributions* 20: 1258–1269. <https://doi.org/10.1111/ddi.12236>
- Thibaud E, Petitpierre B, Broennimann O, Davison AC, Guisan A (2014) Measuring the relative effect of factors affecting species distribution model predictions. O'Hara RB (Ed.). *Methods in Ecology and Evolution* 5: 947–955. <https://doi.org/10.1111/2041-210X.12203>
- Wickham H, Girlich M, RStudio (2022a) tidy: Tidy Messy Data. Available from: <https://CRAN.R-project.org/package=tidy> (October 23, 2022).
- Wickham H, François R, Henry L, Müller K, RStudio (2022b) dplyr: A Grammar of Data Manipulation. Available from: <https://CRAN.R-project.org/package=dplyr> (October 23, 2022).
- Wisn MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, NCEAS Predicting Species Distributions Working Group† (2008) Effects of sample size on the performance of species

distribution models. Diversity and Distributions 14: 763–773.
<https://doi.org/10.1111/j.1472-4642.2008.00482.x>

CAPÍTULO 2

DO PSEUDO ABSENCE AND BACKGROUND POINTS QUANTITY AND SELECTION METHODS IMPACT ON SPECIES DISTRIBUTION AND ECOLOGICAL NICHE MODELS?

DO PSEUDO ABSENCE AND BACKGROUND POINTS QUANTITY AND SELECTION METHODS IMPACT ON SPECIES DISTRIBUTION AND ECOLOGICAL NICHE MODELS?

Gustavo Reis de Brito¹ Fernando Frei²

¹Programa de Pós-Graduação em Biociências, Faculdade de Ciências e Letras, São Paulo State University, Assis, São Paulo, Avenue Dom Antônio, 2100, 190806-900, Brazil

²Department of Biology, Faculdade de Ciências e Letras, São Paulo State University, Assis, São Paulo, Avenue Dom Antônio, 2100, 190806-900, Brazil

Corresponding author: Gustavo Reis de Brito (reis.brito@unesp.br)

Abstract

The understanding of species distribution is crucial for conservation planning, risk assessment, and advancing ecological research. Species Distribution Modeling (SDM) and Ecological Niche Modeling (ENM) are widely used approaches that involve correlating environmental factors with occurrence records to create predictions of species suitability and potential distribution. Despite the growing popularity of these techniques, there is still some ambiguity surrounding key aspects of the modeling process, such as the influence of pseudo absence and background data on model fitting. To test the influence of pseudo absence and background data on model performance, we selected four pseudo absence / background datasets (S1 = 10000, S2 = 1000, S3 = 250, S4 = proportional to presence data), generated via three methods (random sampling, surface range envelope, and geographical exclusion in a 10km from a presence point). As biological data for modeling we selected three species of Brazilian Birds (*Cantorchilus longirostris*, *Eupsittula cactorum*, and *Penelope obscura*). Our results show that the performance of species distribution models can be adversely affected not only by the size of the sample, but also by the quantity and type of pseudo-absence or background data used for model fitting. Our findings reveal that commonly used algorithms exhibit subpar performance when there is an imbalance between presence points and pseudo-absence or background data. Furthermore, we highlight that relying solely on statistical metrics to assess model performance can lead to misinterpretation of results, as the biological relevance of the models is crucial, given the diverse applications of SDM/ENM techniques. Therefore, we recommend a three-step approach for setting up SDM/ENM models: (1) clearly defining the aim of the project (e.g., niche modeling, potential distribution), (2) selecting appropriate algorithms based on the sample size and study area extent, and (3) evaluating model performance not only through robust statistical methods (e.g., True Skill Statistic, TSS), but also considering the biological significance of the models.

Keywords

absence data, modeling evaluation, conservation, sdm, enm

Introduction

Biodiversity changes in space and time. While this is a natural process, it can be heavily influenced by human activities (Pinto-Ledezma; Cavender-Bares 2021). Many techniques can be used to study and quantify biodiversity data (e.g. alpha and beta diversity indexes), relying not only on biological and ecological concepts but also on field sampling and data analysis. The creation of global repositories of species data such as occurrence points is a direct result of integrating Biology and Informatics (Jetz; McPherson; Guralnick 2012).

Discerning why a species is where it is is a difficult task. One approach to this question is to know what environmental factors contribute to the presence of the species (e.g. precipitation, mean temperature) and try to correlate it to occurrence data in the field. This is the concept behind the Species Distribution Model - SDM technique, one of the most widely used approaches to studying biodiversity distribution (Araújo and Peterson 2012). The combination of occurrence and environmental data allows the construction of a type of model known as a correlative model which aims to try and correlate the presence of a species and environmental characteristics where it occurs (Guisan and Zimmermann 2000, Peterson 2011, Araújo and Peterson 2012).

In theory, a basic SDM model requires three types of data: presence and absence points, and environmental variables. Although global databases such as the Global Biodiversity Information Facility - GBIF (www.gbif.org) gives access to occurrence data, often these databases lack absence information. Without it, there's no statistical possibility of characterizing the niche of a species (Descombes et al. 2022).

To prove or disprove the occurrence of a species in a location is a difficult task, given that many factors can “mask” its presence in the environment (e.g. migratory species) (Mackenzie; Royle 2005, Raes; Aguirre-Gutiérrez 2018). As Raes and Aguirre-Gutiérrez (2018) offer that “the absence of a presence does not imply the presence of absence”. Given that some algorithms require presence and absence data to model the species, while others do not, we can classify SDM/ENM techniques as presence-only and presence-absence (Brotons et al. 2004).

There are many studies that focus on pseudo absence influence on SDM models (e.g. Wisz; Guisan 2009, Barbet-Massin et al. 2012, Hanberry et al. 2012, Liang et al. 2018, Descombes et al. 2022) that sought to identify the potential influence of pseudo absence quantity and selection on SDM/ENM approaches. Although these studies (and many others) share pseudo absence as a “common ground”, they all focus on the statistical analysis as tools of assessing model performance rather than on the biological significance of such models.

Another characteristic of studies focusing on pseudo absence potential influence on model techniques is the continuation of “pseudo synonyms”: most works use the words “pseudo absence” and “background” as synonyms to name two different entities of the models. As Sillero and Barbosa (2021) point out, presence-only algorithms such as Maxent (Phillips et al. 2006) use background data to model environmental suitability but are commonly used as a presence/pseudo absence to model potential distribution (e.g. Zhang et al. 2019). Mistaking background data (i.e. the whole study area) with artificially created absence data (i.e. pseudo absence) is, also, another source of bias for SDM/ENM approaches.

In light of that, given that most approaches that study pseudo absence generation focus on statistical performance, we sought to investigate the influence of the quantity of pseudo absence and background points, and how it is generated, in Ecological Niche Models from a biological point of view. We applied ENM techniques to three species of Brazilian birds, focusing on model performance, while also analyzing model significance via statistical methods commonly used in SDM and ENM.

Methods

Species and Environmental datasets

We chose three species of Brazilian birds as the main subjects for modeling: *Cantorchilus longirostris* (Vieillot, 1819), *Eupsittula cactorum* (Kuhl, 1820), and *Penelope obscura* (Temminck, 1815). Occurrence data was downloaded from the Global Biodiversity Information Facility - GBIF (www.gbif.org) via the *dismo* 1.3-9 package (Hijmans et al. 2022). To structure the occurrence dataset for all three species we downloaded the maximum number of occurrences from GBIF and cleaned the data to remove points with uncertain coordinates, with an absence of longitude or latitude, data marked as NA (i.e. not available), and points that occurred outside of Brazil. We also eliminated duplicated points to ensure the final dataset consisted of unique presence points. Occurrence data was also trimmed to ensure no point occurred inside a radius of 10 km of a presence, resulting in a final dataset of unique and spatially distributed points for each species (Table 1). All procedures, from obtaining occurrence data to cleaning and trimming of the datasets, were achieved via *dplyr* 1.0.10 (Wickham et al. 2022b), *tidyr* 1.3.0 (Wickham et al. 2022a), and *spThin* 0.2.0 (Aiello-Lammens et al. 2015) packages, implemented in R 4.2.2 (R Core Team 2022).

Table 1. Number of occurrence points for each species, from GBIF database to the final dataset

Species	Raw Data	Points retained after data cleaning	Points retained after spatial trimming (N1)
<i>Cantorchilus longirostris</i>	11723	2819	645
<i>Eupsittula cactorum</i>	6004	2391	658
<i>Penelope obscura</i>	23668	5055	1131

Two additional datasets were sampled from size N1: N2 = 100 and N3 = 20. All sample sizes consist of unique and spatially distributed data for all three species. We decided to include three sample sizes in the modeling procedures to try and fully capture the influence of background and pseudo absence points on different sizes of occurrence data, a situation fairly common in the literature.

For the environmental dataset, we downloaded 19 bioclimatic variables from WorldClim (www.worldclim.org), version 2.1, at a spatial resolution of 2.5 arc-minutes (Fick and Hijmans 2017). Given that using too many variables can bias the models through over-fitting (Mateo et al. 2013), the environmental dataset was tested via the Variance Inflation Factor - VIF test to compute multicollinearity values and remove correlated variables. Following Chatterjee and Hadi (2012), we selected a threshold of 10 for the VIF test, removing from the dataset variables with collinearity values above it. For all three species, the final dataset comprises 8 environmental layers for *Cantorchilus longirostris* and *Penelope obscura*, and 9 layers for *Eupsittula cactorum* (Table 2).

Table 2. Environmental variables used to train the models for all three species

Species	Variables
---------	-----------

<i>Cantorichilus longirostris</i>	BIO2, BIO4, BIO8, BIO13, BIO14, BIO15, BIO18, BIO19
<i>Eupsittula cactorum</i>	BIO2, BIO3, BIO4, BIO8, BIO13, BIO14, BIO15, BIO18, BIO19
<i>Penelope obscura</i>	BIO3, BIO5, BIO7, BIO8, BIO9, BIO12, BIO13, BIO18

Selection of pseudo absence and background points

To test the influence of pseudo absence and background points quantity in species distribution and ecological niche models, we selected four sizes and three types of generation for the data. The quantity of points generated is based on the work of Barbet-Massin et al. (2012): S1 = 10000, S2 = 1000, S3 = 250, S4 = proportional to presence data (i.e. for each presence point, one point of pseudo absence or background was generated). The methods to generate pseudo absence and background points were random sampling from the study area (excluding presence points), via a surface range envelope to select random points outside an estimated suitable area, and via geographically excluding points situated at least 10km from a presence point. Pseudo absence and background points were generated via the *biomod2* 4.2-2 package (Thuiller et al. 2022), implemented in R 4.2.2 (R Core Team 2022).

Modeling procedures

To model each species, we used cleaned and trimmed datasets, containing unique and clean records: *C. longirostris* (n = 645), *E. cactorum* (n = 658), and *P. obscura* (n = 1131). Based on Raes and Aguirre-Gutiérrez (2018), we selected the five most common algorithms used in ENM and SDM procedures: Bioclim (Busby 1991), Generalized Additive Model - GAM (Hastie and Tibshirani 1986), Generalized Linear Model - GLM (McCullagh and Nelder 1998), Maxent (Phillips et al. 2006), and Random Forest (Breiman 2001). A total of 600 runs per species was performed (five algorithms x 10 subsampling runs x 4 sets of background/pseudo absence points per algorithm x 3 types of generated data per set of background/pseudo absence), totalizing 1800 model runs for all three species. The dataset of each species was partitioned to evaluate model fitting, with 75% of the data used for training and 25% for testing.

Ensembling

Given the overwhelming number of models fitted and the fact that not all models present statistical or biological significance, we decided to apply the ensemble technique, yielding a single consensus map for each algorithm. To create the ensembles we considered only the models with AUC and TSS values above 0.7 and 0.4 thresholds, respectively, following Buisson et al. (2010) recommendations. The definition of an AUC and TSS threshold is important to try and select only models that performed well during the model fitting phase.

Statistical Analysis

To assess model accuracy, we performed the True Skill Statistic - TSS analysis (Allouche et al. 2006) and calculated the Area Under the Curve - AUC value from the Receiver Operating Characteristic (ROC) (Fielding and Bell 1997). Given the number of models per algorithm,

we calculated the mean values of AUC and TSS. Model validation follows Buisson et al. (2010) recommendations, validating models above 0.7 and 0.4 thresholds, respectively.

Results

Statistical Analysis

Across all three species two patterns can be observed: 1) Bioclim models always perform below the threshold and/or below other algorithms and 2) when sample size varies, performance tends to drop as the sample reaches 20 presence points and possibly below it (Figure 1, Figure 2, Figure 3).

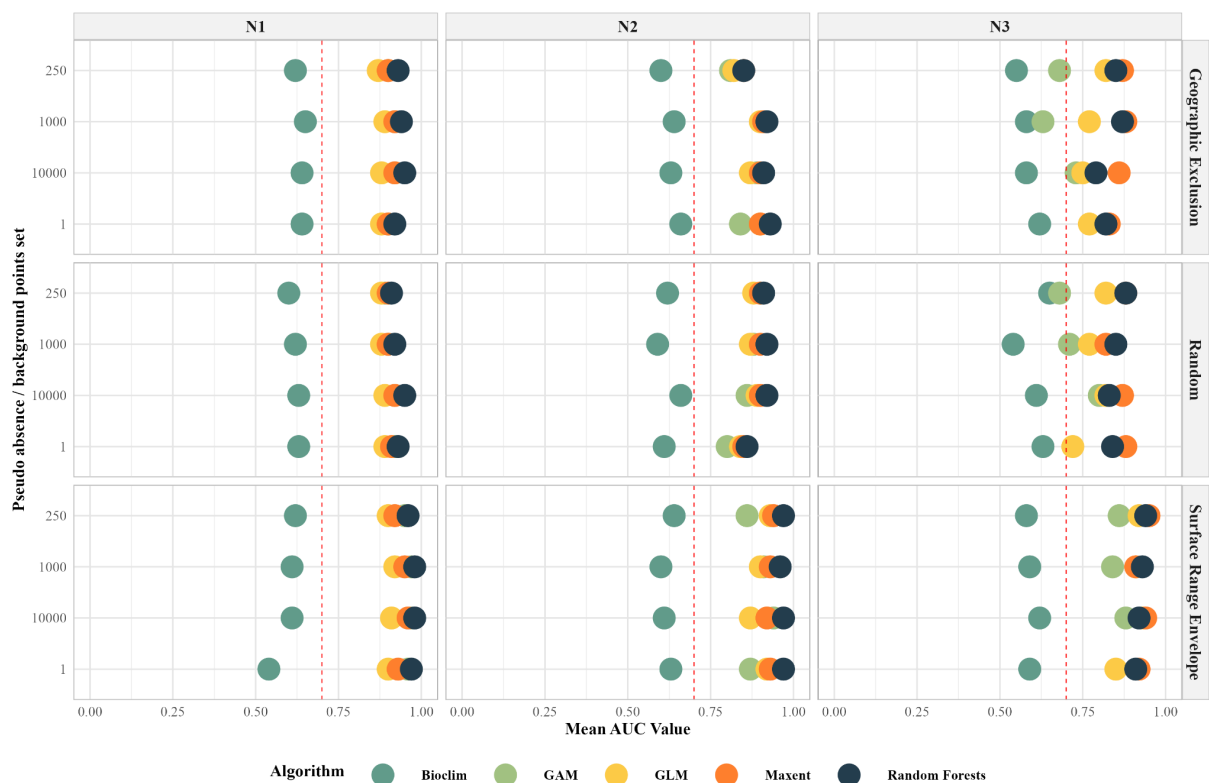


Figure 1. Mean AUC values obtained for all five algorithms across each background / pseudo absence set for all models of *Cantorchilus longirostris*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process of generating pseudo absence and background points. The red line represents the 0.7 AUC threshold.

Cantorchilus longirostris and *Eupsittula cactorum* models presented two groups of algorithms: Bioclim represents the first group, while GAM, GLM, Maxent, and Random Forests are the second one (Figure 1, Figure 2). *Eupsittula cactorum* was the species with the most uniform variation in AUC values across all three sample sizes (Figure 2), with *Cantorchilus longirostris* and *Penelope obscura* (Figure 1, Figure 3) showing variations when the sample is reduced to N3.

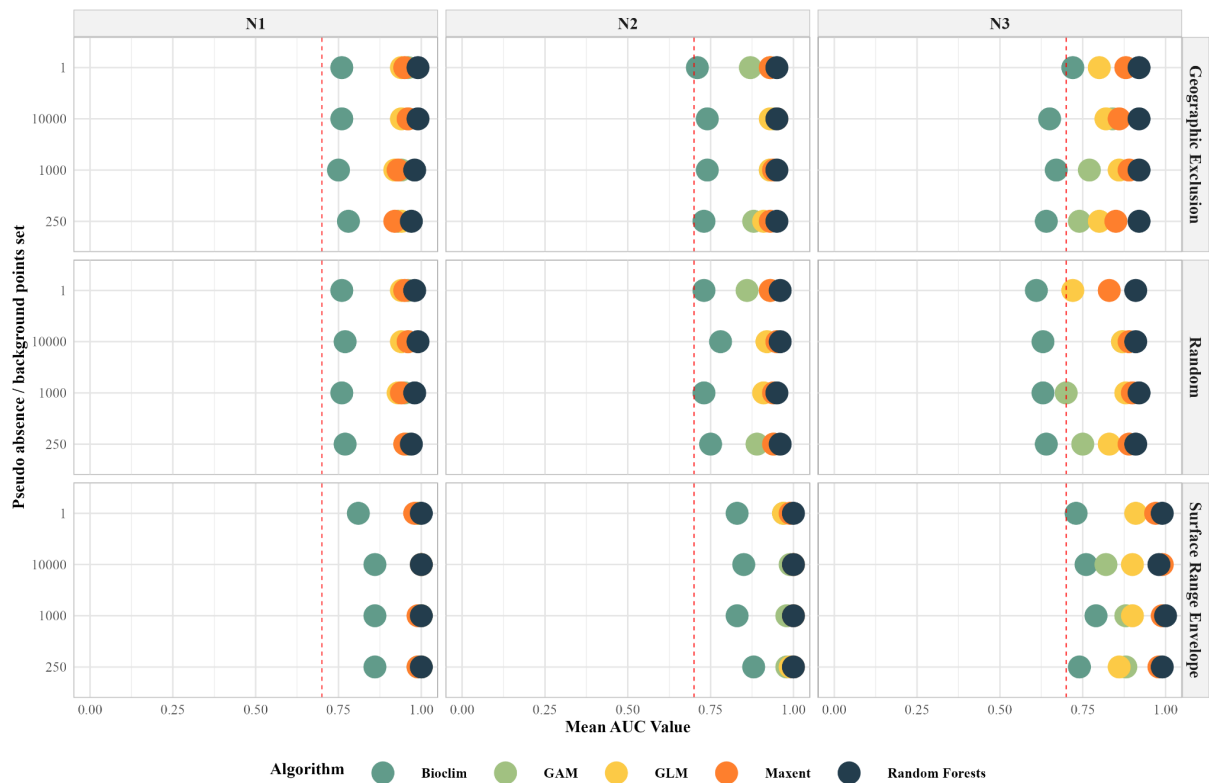


Figure 2. Mean AUC values obtained for all five algorithms across each background / pseudo absence set for all models of *Eupsittula cactorum*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process of generating pseudo absence and background points. The red line represents the 0.7 AUC threshold.

Although Bioclim tends to underperform compared to other algorithms (Figure 1, Figure 3), it produced models with performance above the 0.7 AUC threshold in *Eupsittula cactorum*, dropping below it only when the sample size was reduced to N3 (Figure 2).

Overall, model performance seems to be affected more by sample size and the set of pseudo absence/background points than by how the “absence” data was generated. Across all three species, all types of “absence” generation (i.e. Geographic Exclusion, Random, Surface Range Envelope) employed produced similar results when the AUC is used as the reference point to assess model performance (Figure 1, Figure 2, Figure 3).

When the sample size is reduced to N2 or N3 but the pseudo absence/background points set isn't (e.g. 10000 “absence” points are used to train the models), some algorithms fail to model the data (Figure 1, Figure 2, Figure 3).

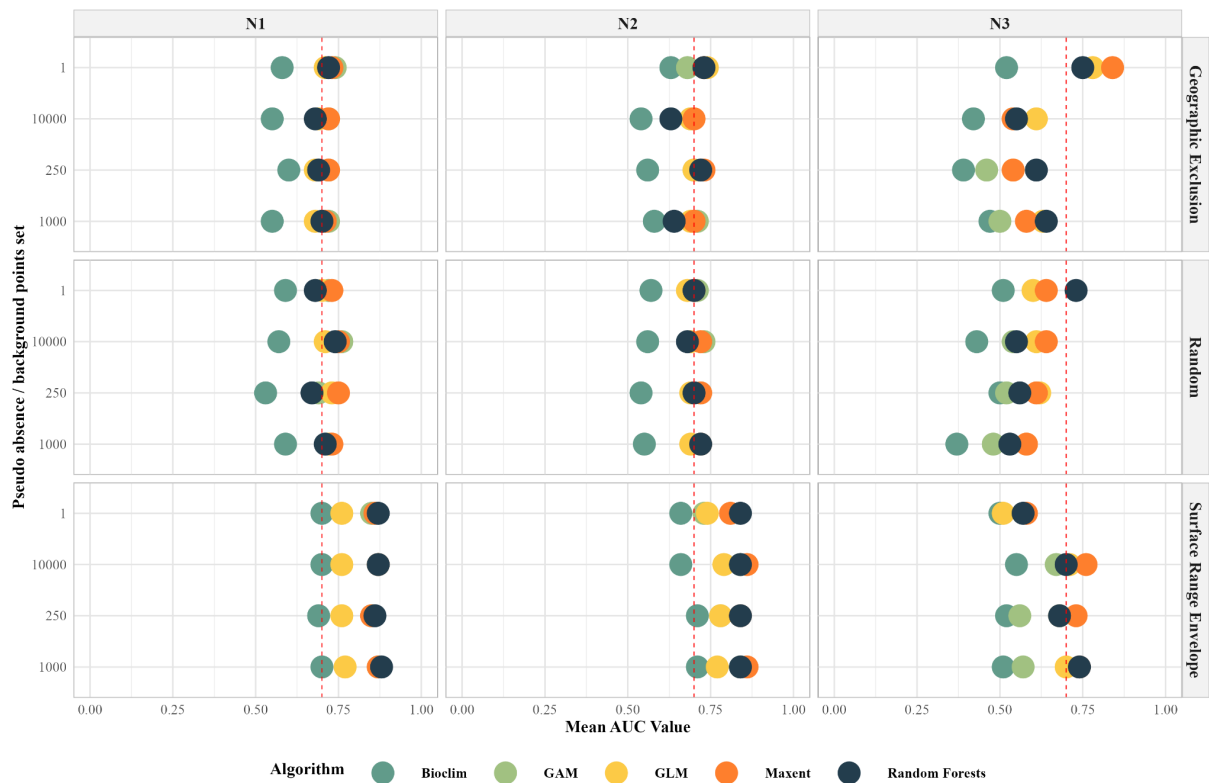


Figure 3. Mean AUC values obtained for all five algorithms across each background / pseudo absence set for all models of *Penelope obscura*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process of generating pseudo absence and background points. The red line represents the 0.7 AUC threshold.

When the model performance was assessed via the TSS values, the same patterns are repeated: Bioclim underperforms overall for all three species (Figure 4, Figure 5, Figure 6) except in *Eupsittula cactorum* (Figure 5), with two groups forming based on model performance.

Model performance under the TSS analysis presents the most stable groupings and values under sizes N1 and N2, with sample N3 presenting slightly separating GAM, GLM, Maxent, and Random Forests (Figure 4, Figure 5, Figure 6) when their values are compared across all three methods to generate “absence” data. Although size N3 showed a decrease in model performance for all three species independently of the pseudo absence/background points set, *Penelope obscura* (Figure 6) shows a further decrease in performance for all algorithms.

The same algorithms that failed to produce AUC values, failed to yield a TSS value when this analysis was applied. When combined, both analyses demonstrated that two main factors act to decrease model performance: sample size and the number of pseudo absence/background points generated to train the models.

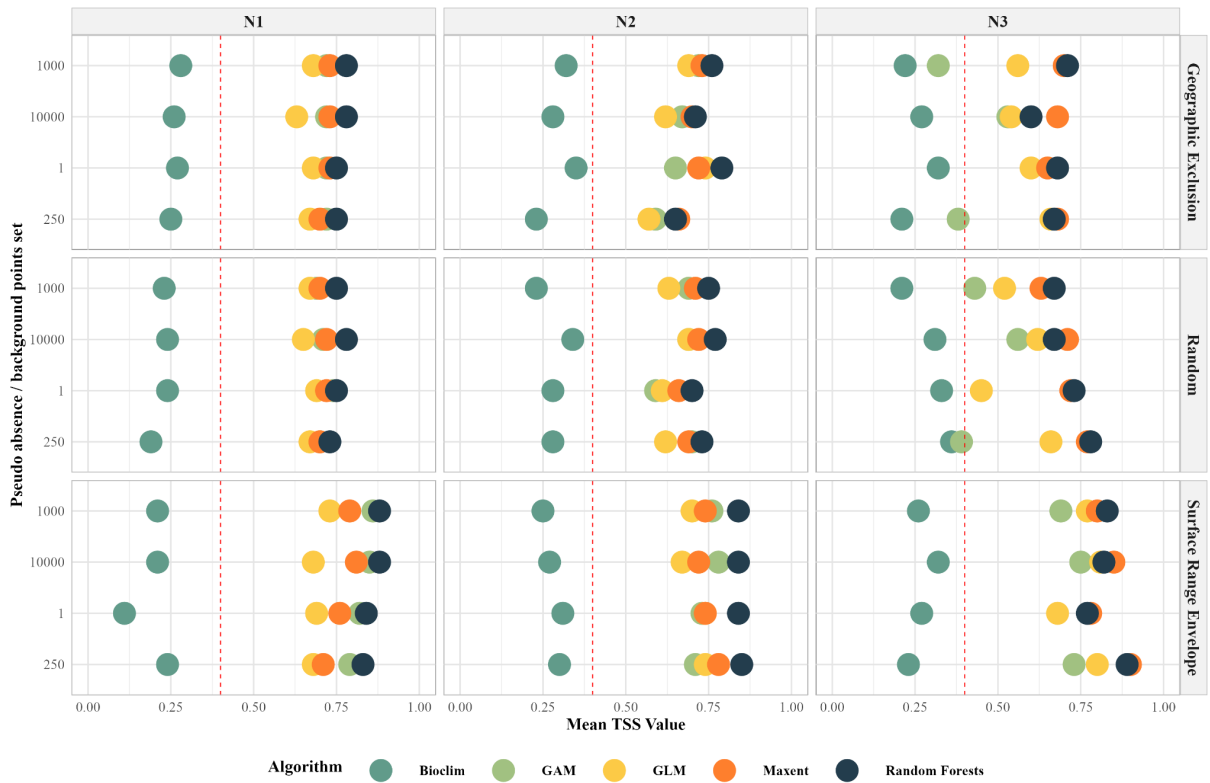


Figure 4. Mean TSS values obtained for all five algorithms across each background / pseudo absence set for all models of *Cantorchilus longirostris*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process of generating pseudo absence and background points. The red line represents the 0.4 TSS threshold.

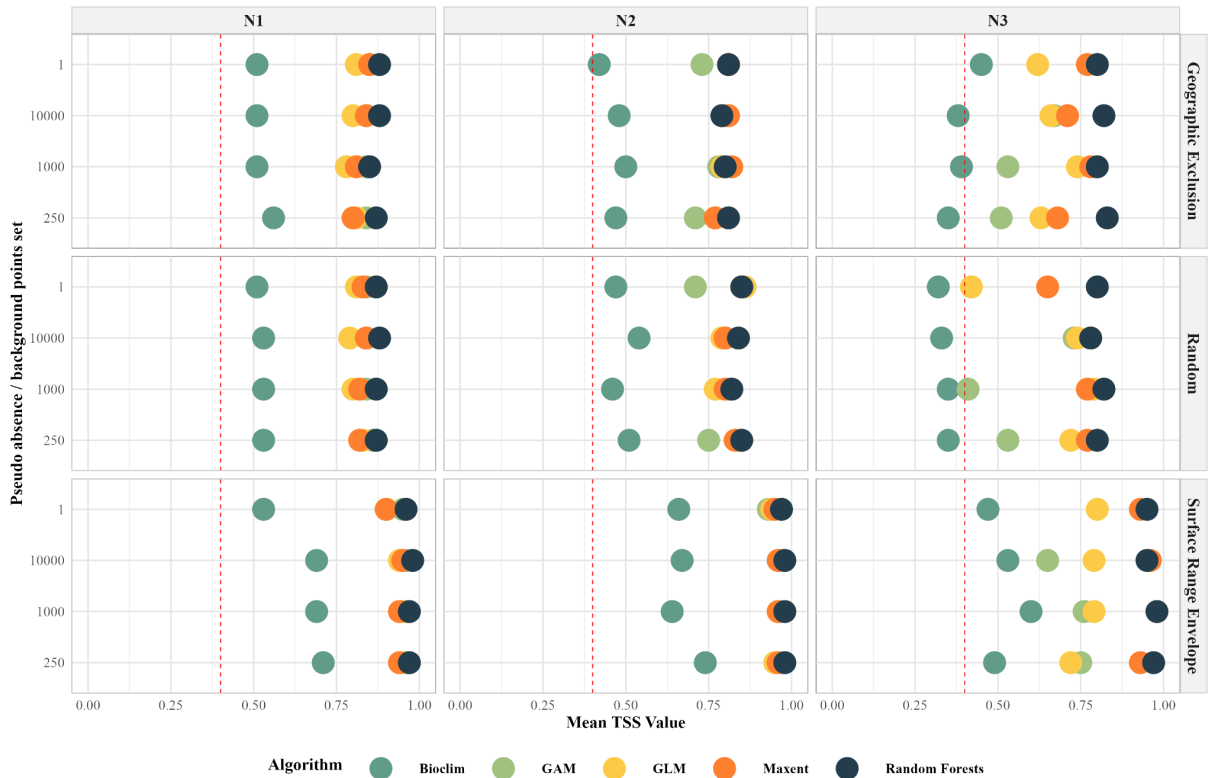


Figure 5. Mean TSS values obtained for all five algorithms across each background / pseudo absence set for all models of *Eupsittula cactorum*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process of generating pseudo absence and background points. The red line represents the 0.4 TSS threshold.



Figure 6. Mean TSS values obtained for all five algorithms across each background / pseudo absence set for all models of *Penelope obscura*. Columns represent the sample size (N1 = 645, N2 = 100, N3 = 20). Rows present the process of generating pseudo absence and background points. The red line represents the 0.4 TSS threshold.

Model Performance

Values of AUC and TSS below the thresholds ($AUC \geq 0.7$, $TSS \geq 0.4$) observed in Bioclim are reflected in the suitable area predicted by this algorithm: the lowest values of AUC ($AUC = 0.37$) and TSS ($TSS = 0.08$) are found in *Penelope obscura* when models were trained with 1000 background points, generated at random from the study area (Figure 7-A).

In contrast, Bioclim produced the highest values of AUC ($AUC = 0.88$) and TSS ($TSS = 0.74$) in *Eupsittula cactorum*, when the species was modeled using 250 background points, generated outside of a surface range envelope calculated on the study area (Figure 7-B).

The behavior of GAM, GLM, Maxent, and Random Forests when only TSS is considered as a metric of model evaluation is similar to AUC: for *Cantorchilus longirostris* and *Eupsittula cactorum*, the type of “absence” generation doesn’t seem to hold influence on model fitting (Figure 4, Figure 5); only in *Penelope obscura* one of the three methods - Surface Range Envelope - produced models above the threshold of validation across all three sample sizes (Figure 6).

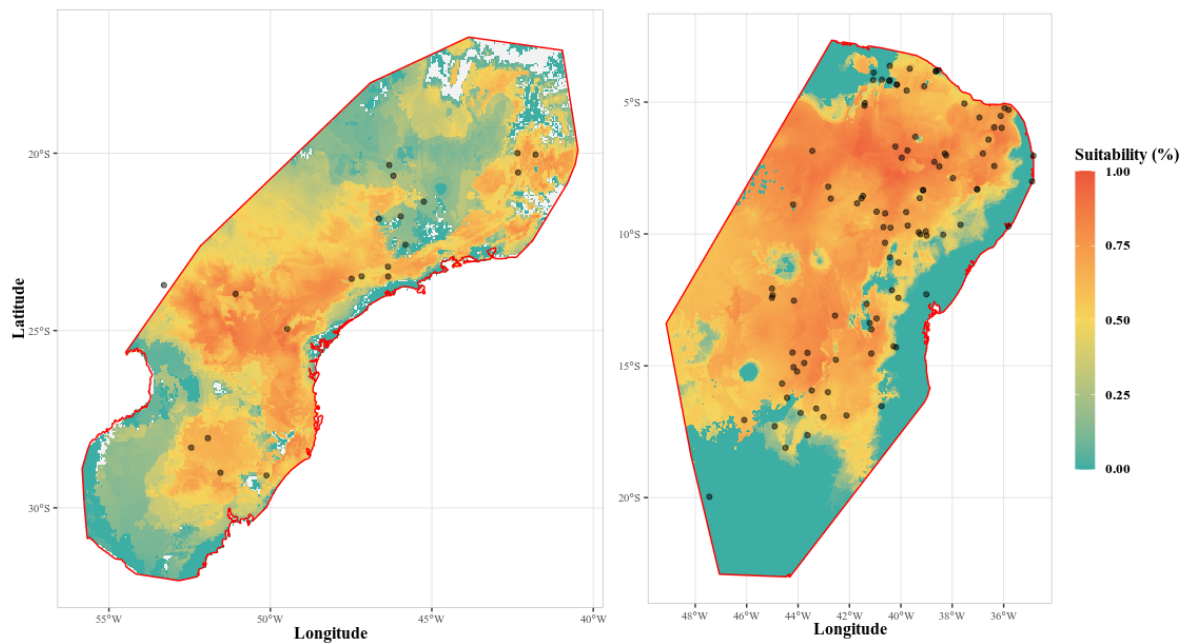


Figure 7. Suitability map for *Penelope obscura* (A) and *Eupsittula cactorum* (B) produced by Bioclim algorithm. The models were trained using 20 and 100 presence points (sizes N3 and N2), respectively. Model A was trained with 1000 background points randomly generated from the study area; Model B was trained with 250 background points sampled outside a surface range envelope drawn on the study area. Red lines represent the Extent of Occurrence of the species. Black dots are the presence points used in model training.

High AUC and TSS values were obtained for GAM, GLM, Maxent, and Random Forests algorithms in *Eupsittula cactorum* (Figure 2, Figure 5) when models were trained with 1000 pseudo absence (GAM, GLM, Random Forests) and 10000 background points (Maxent), respectively. Absence data was generated using the Surface Range Envelope selection method and the sample size used was N1 (Figure 8). Models with high values of AUC (AUC = 1.0) and TSS (TSS $\geq$ 0.95) produced overfitted maps of suitability for *Eupsittula cactorum*, holding little to no biological significance.

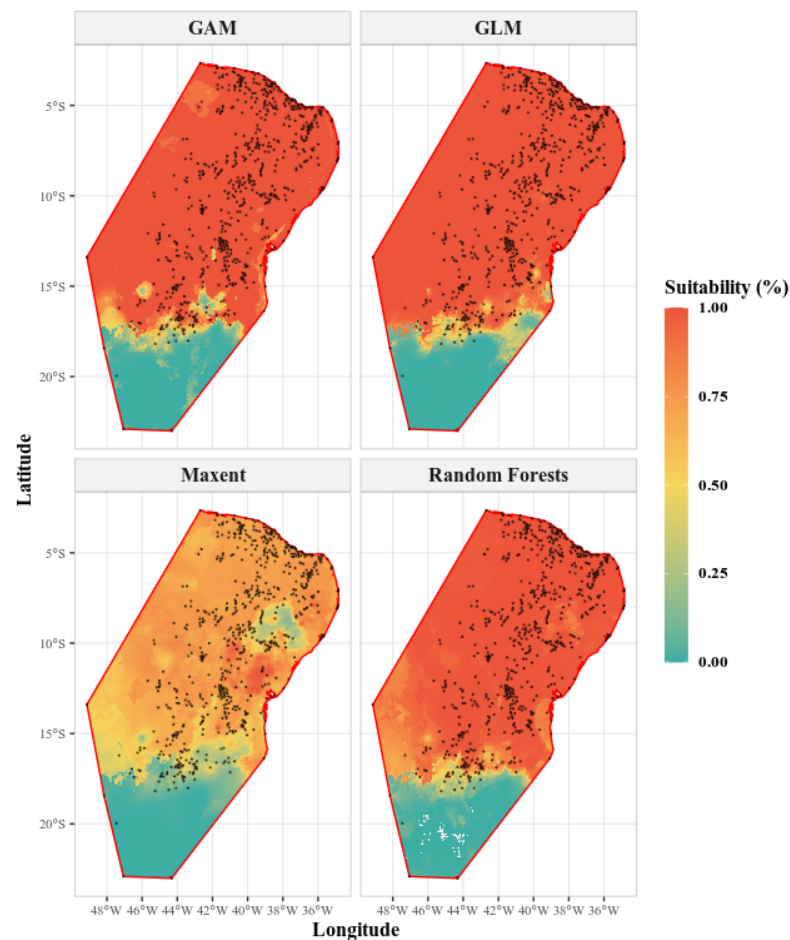


Figure 8. Overfitted suitability map for *Eupsittula cactorum* produced by four algorithms with the highest values of AUC (AUC = 1.0) and TSS (TSS $\geq$ 0.95). The models were trained using sample size N1, with a 1000 pseudo absence and 10000 background points, respectively. Red lines represent the Extent of Occurrence of the species. Black dots are the presence points used in model training.

Discussion

SDM applications can be classified as presence-only and presence-absence methods (Brotons et al. 2004). In this paper we used a combination of both methods to model the environmental suitability for three species of Brazilian birds, yielding different results when comparisons were drawn between algorithms and pseudo absence/background data generation and quantity.

When comparing algorithms, Bioclim models performed below the threshold for two of the three species. GAM, GLM, Maxent, and Random Forests models, on the contrary, performed better when a validation threshold is present. This can be explained by the fact that Bioclim is based on the calculation of a rectilinear envelope, modeling two dimensions at a time (Busby 1991), while Maxent, for instance, uses the concept of maximum entropy to derive the environmental suitability of an area (Phillips et al. 2006). Also, Maxent is considered a presence-absence method by some authors (e.g. Barbet-Massin et al. 2012), being shown that this type of model performs better than presence-only (Elith et al. 2006).

Using suitability values as a tool to access model performance can yield important insights into the biological significance of such models. The four algorithms with better performance

based on AUC and TSS metrics (i.e. GAM, GLM, Maxent, and Random Forests) produced strange suitability maps (Figure 8) when what varies is not the sample size but the number of pseudo absence/background points and how this absence data was generated.

GAM, GLM, and Random Forests practically predicted almost 90% of the study area of *Eupsittula cactorum* as highly suitable for the species (Figure 8), which is a strange result. This behavior of algorithms based on the use of pseudo absence is well known since artificially generated absence can “mask” environmental values as “not suitable” when, in fact, they are (Gu; Swihart 2004). This behavior of the algorithms can, also, be explained by the presence points being climatically and geographically biased, with important environmental variations not perceived by the algorithms (Barbet-Massin; Thuiller; Jiguet, 2010).

All five algorithms produced omission and commission errors (Peterson and Soberón 2012), under or overestimating the environmental suitability of the study area for all three species, hindering model performance and resulting in abnormal projections (Figure 7, Figure 8) and AUC (Figures 1-3) and TSS values (Figures 4-6). Biases can be introduced in models in a variety of ways, such as using datasets climatically biased or by a poor choice of algorithms, mixing background data with pseudo absence. Thus, it's possible to say that model uncertainty and abnormal projections are not only related to the algorithmic processing but to environmental variables not being considered for the fitting of the models (Barbet-Massin; Thuiller; Jiguet, 2010, Da Silva et al. 2020).

We followed the protocol proposed by Barbet-Massin et al. (2012) when pseudo absences were selected, which can be demonstrated by the resulting models showing variations when the AUC and TSS metrics were used for model assessment. The decrease in performance of all algorithms when the sample size is reduced but not the number of pseudo absences/background points shows that an imbalance between presence and absence not only masks possible suitable environment but also breaks the prevalence, which is a well-known factor that influences model performance (Barbet-Massin et al. 2012).

Conclusion

Model performance can be hindered by not only the size of the sample but also by the number and type of pseudo absence/background data used to fit the models. Our results show that common algorithms have poor performance when there is an imbalance between presence points and pseudo absence/background data. Also, we conclude that using statistics metrics as the only tool to access model performance can lead to misinterpretation of the results since biological significance is important given the many applications of SDM/ENM techniques.

Therefore, we recommend setting an SDM/ENM model by 1) delimiting the aim of the project (e.g. niche modeling, potential distribution), 2) choosing the algorithms based on the sample size and the size of the study area, and 3) evaluating model performance not only via robust statistical methods (e.g. TSS) but also by its biological significance.

References

AIELLO-LAMMENS, Matthew E.; BORJA, Robert A.; RADOSAVLJEVIC, Aleksandar; VILELA, Bruno; ANDERSON, Robert P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, [S. l.], v. 38, n. 5, p.

541–545, 2015. DOI: 10.1111/ecog.01132. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01132>. Acesso em: 9 jan. 2023.

ALLOUCHE, Omri; TSOAR, Asaf; KADMON, Ronen. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. **Journal of Applied Ecology**, [S. l.], v. 43, n. 6, p. 1223–1232, 2006. DOI: [10.1111/j.1365-2664.2006.01214.x](https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2006.01214.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2006.01214.x>. Acesso em: 10 jan. 2023.

ARAÚJO, Miguel B.; PETERSON, A. Townsend. Uses and misuses of bioclimatic envelope modeling. **Ecology**, [S. l.], v. 93, n. 7, p. 1527–1539, 2012. DOI: [10.1890/11-1930.1](https://onlinelibrary.wiley.com/doi/10.1890/11-1930.1). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1890/11-1930.1>. Acesso em: 24 out. 2022.

BARBET-MASSIN, Morgane; JIGUET, Frédéric; ALBERT, Cécile Hélène; THUILLER, Wilfried. Selecting pseudo-absences for species distribution models: how, where and how many?: *How to use pseudo-absences in niche modelling?* **Methods in Ecology and Evolution**, [S. l.], v. 3, n. 2, p. 327–338, 2012. DOI: [10.1111/j.2041-210X.2011.00172.x](https://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00172.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00172.x>. Acesso em: 15 fev. 2022.

BARBET-MASSIN, Morgane; THUILLER, Wilfried; JIGUET, Frédéric. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? **Ecography**, [S. l.], v. 33, n. 5, p. 878–886, 2010. DOI: [10.1111/j.1600-0587.2010.06181.x](https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2010.06181.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2010.06181.x>. Acesso em: 3 jan. 2023.

BREIMAN, Leo. Random forests. **Machine Learning**, [S. l.], v. 45, n. 1, p. 5–32, 2001. DOI: [10.1023/A:1010933404324](http://link.springer.com/10.1023/A:1010933404324). Disponível em: <http://link.springer.com/10.1023/A:1010933404324>. Acesso em: 13 ago. 2022.

BROTONS, Lluís; THUILLER, Wilfried; ARAÚJO, Miguel B.; HIRZEL, Alexandre H. Presence-Absence versus Presence-Only Modelling Methods for Predicting Bird Habitat Suitability. **Ecography**, [S. l.], v. 27, n. 4, p. 437–448, 2004. DOI: [10.2307/3683417](https://www.jstor.org/stable/3683417). Disponível em: <https://www.jstor.org/stable/3683417>. Acesso em: 4 maio. 2022.

BUISSON, Laëtitia; THUILLER, Wilfried; CASAJUS, Nicolas; LEK, Sovan; GRENOUILLET, Gaël. Uncertainty in ensemble forecasting of species distribution. **Global Change Biology**, [S. l.], v. 16, n. 4, p. 1145–1157, 2010. DOI: [10.1111/j.1365-2486.2009.02000.x](https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2009.02000.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2009.02000.x>. Acesso em: 25 maio. 2022.

Busby, J.R. (1991a) BIOCLIM – a bioclimate analysis and prediction system. Nature conservation: cost effective biological surveys and data analysis (ed. by C.R. Margules and M.P. Austin), pp. 64–68. CSIRO, Melbourne.

CHATTERJEE, Samprit; HADI, Ali S. **Regression analysis by example**. 4th ed ed. Hoboken, N.J: Wiley-Interscience, 2006.

DESCOMBES, Patrice; CHAUVIER, Yann; BRUN, Philipp; RIGHETTI, Damiano; WÜEST, Rafael O.; KARGER, Dirk N.; ZURELL, Damaris; ZIMMERMANN, Niklaus E. **Strategies for sampling pseudo-absences for species distribution models in complex mountainous terrain**. [s.l.] : Ecology, 2022. DOI: [10.1101/2022.03.24.485693](https://doi.org/10.1101/2022.03.24.485693). Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2022.03.24.485693>. Acesso em: 2 jan. 2023.

ELITH, Jane et al. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, [S. l.], v. 29, n. 2, p. 129–151, 2006. DOI: [10.1111/j.2006.0906-7590.04596.x](https://doi.org/10.1111/j.2006.0906-7590.04596.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.2006.0906-7590.04596.x>. Acesso em: 15 fev. 2022.

FICK, Stephen E.; HIJMANS, Robert J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology**, [S. l.], v. 37, n. 12, p. 4302–4315, 2017. DOI: [10.1002/joc.5086](https://doi.org/10.1002/joc.5086). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/joc.5086>. Acesso em: 10 jan. 2023.

FIELDING, Alan H.; BELL, John F. A review of methods for the assessment of prediction errors in conservation presence/absence models. **Environmental Conservation**, [S. l.], v. 24, n. 1, p. 38–49, 1997. DOI: [10.1017/S0376892997000088](https://doi.org/10.1017/S0376892997000088). Disponível em: https://www.cambridge.org/core/product/identifier/S0376892997000088/type/journal_article. Acesso em: 10 jan. 2023.

GU, Weidong; SWIHART, Robert K. Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. **Biological Conservation**, [S. l.], v. 116, n. 2, p. 195–203, 2004. DOI: [10.1016/S0006-3207\(03\)00190-3](https://doi.org/10.1016/S0006-3207(03)00190-3). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0006320703001903>. Acesso em: 15 fev. 2022.

GUISAN, Antoine; ZIMMERMANN, Niklaus E. Predictive habitat distribution models in ecology. **Ecological Modelling**, [S. l.], v. 135, n. 2–3, p. 147–186, 2000. DOI: [10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380000003549>. Acesso em: 4 maio. 2022.

HANBERRY, Brice B.; HE, Hong S.; PALIK, Brian J. Pseudoabsence Generation Strategies for Species Distribution Models. **PLoS ONE**, [S. l.], v. 7, n. 8, p. e44486, 2012. DOI: [10.1371/journal.pone.0044486](https://doi.org/10.1371/journal.pone.0044486). Disponível em: <https://dx.plos.org/10.1371/journal.pone.0044486>. Acesso em: 10 jan. 2023.

HASTIE, Trevor; TIBSHIRANI, Robert. Generalized Additive Models. **Statistical Science**, [S. l.], v. 1, n. 3, 1986. DOI: [10.1214/ss/1177013604](https://doi.org/10.1214/ss/1177013604). Disponível em: <https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Generalized-Additive-Models/10.1214/ss/1177013604.full>. Acesso em: 13 ago. 2022.

HIJMANS, Robert J.; PHILLIPS, Steven; ELITH, John Leathwick and Jane. **dismo: Species Distribution Modeling**, 2022. Disponível em: <https://CRAN.R-project.org/package=dismo>. Acesso em: 4 fev. 2023.

JETZ, Walter; MCPHERSON, Jana M.; GURALNICK, Robert P. Integrating biodiversity distribution knowledge: toward a global map of life. **Trends in Ecology & Evolution**, [S. l.], v. 27, n. 3, p. 151–159, 2012. DOI: [10.1016/j.tree.2011.09.007](https://doi.org/10.1016/j.tree.2011.09.007). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169534711002679>. Acesso em: 11 fev. 2022.

LIANG, Wanwan; PAPEŞ, Monica; TRAN, Liem; GRANT, Jerome; WASHINGTON-ALLEN, Robert; STEWART, Scott; WIGGINS, Gregory. The effect of pseudo-absence selection method on transferability of species distribution models in the context of non-adaptive niche shift. **Ecological Modelling**, [S. l.], v. 388, p. 1–9, 2018. DOI: [10.1016/j.ecolmodel.2018.09.018](https://doi.org/10.1016/j.ecolmodel.2018.09.018). Disponível em:

<https://linkinghub.elsevier.com/retrieve/pii/S0304380018303156>. Acesso em: 10 jan. 2023.

MACKENZIE, Darryl I.; ROYLE, J. Andrew. Designing occupancy studies: general advice and allocating survey effort: *Designing occupancy studies*. **Journal of Applied Ecology**, [S. l.], v. 42, n. 6, p. 1105–1114, 2005. DOI: [10.1111/j.1365-2664.2005.01098.x](https://doi.org/10.1111/j.1365-2664.2005.01098.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2005.01098.x>. Acesso em: 15 fev. 2022.

MATEO, Rubén G.; VANDERPOORTEN, Alain; MUÑOZ, Jesús; LAENEN, Benjamin; DÉSAMORÉ, Aurélie. Modeling Species Distributions from Heterogeneous Data for the Biogeographic Regionalization of the European Bryophyte Flora. **PLoS ONE**, [S. l.], v. 8, n. 2, p. e55648, 2013. DOI: [10.1371/journal.pone.0055648](https://doi.org/10.1371/journal.pone.0055648). Disponível em: <https://dx.plos.org/10.1371/journal.pone.0055648>. Acesso em: 23 out. 2022.

MCCULLAGH, P.; NELDER, John A. **Generalized linear models**. 2nd ed ed. Boca Raton: Chapman & Hall/CRC, 1998.

PETERSON, A. Townsend (ORG.). **Ecological niches and geographic distributions**. Princeton, N.J: Princeton University Press, 2011.

PETERSON, A. Townsend; SOBERÓN, Jorge. Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. **Natureza & Conservação**, [S. l.], v. 10, n. 2, p. 102–107, 2012. DOI: [10.4322/natcon.2012.019](https://doi.org/10.4322/natcon.2012.019). Disponível em: <http://doi.editoracubo.com.br/10.4322/natcon.2012.019>. Acesso em: 24 maio. 2022.

PHILLIPS, Steven J.; ANDERSON, Robert P.; SCHAPIRE, Robert E. Maximum entropy modeling of species geographic distributions. **Ecological Modelling**, [S. l.], v. 190, n. 3–4, p. 231–259, 2006. DOI: [10.1016/j.ecolmodel.2005.03.026](https://doi.org/10.1016/j.ecolmodel.2005.03.026). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S030438000500267X>. Acesso em: 7 fev. 2023.

PINTO-LEDEZMA, Jesús N.; CAVENDER-BARES, Jeannine. Predicting species distributions and community composition using satellite remote sensing predictors. **Scientific Reports**, [S. l.], v. 11, n. 1, p. 16448, 2021. DOI: [10.1038/s41598-021-96047-7](https://doi.org/10.1038/s41598-021-96047-7). Disponível em: <https://www.nature.com/articles/s41598-021-96047-7>. Acesso em: 7 fev. 2022.

RAES, Niels; AGUIRRE-GUTIÉRREZ, Jesús. A Modeling Framework to Estimate and Project Species Distributions in Space and Time. Em: **Mountains, Climate and Biodiversity**. 1st. ed. Hoboken, NJ: John Wiley & Sons, 2018.

SILLERO, Neftalí; BARBOSA, A. Márcia. Common mistakes in ecological niche models. **International Journal of Geographical Information Science**, [S. l.], v. 35, n. 2, p. 213–226, 2021. DOI: [10.1080/13658816.2020.1798968](https://doi.org/10.1080/13658816.2020.1798968). Disponível em: <https://www.tandfonline.com/doi/full/10.1080/13658816.2020.1798968>. Acesso em: 24 maio. 2022.

THUILLER, Wilfried; GEORGES, Damien; GUEGUEN, Maya; ENGLER, Robin; BREINER, Frank; LAFOURCADE, Bruno; PATIN, Remi. **biomod2: Ensemble Platform for Species Distribution Modeling**, 2023. Disponível em: <https://CRAN.R-project.org/package=biomod2>. Acesso em: 4 fev. 2023.

WICKHAM, Hadley; FRANÇOIS, Romain; HENRY, Lionel; MÜLLER, Kirill; RSTUDIO. **dplyr: A Grammar of Data Manipulation.** , 2022. Disponível em: <https://CRAN.R-project.org/package=dplyr>. Acesso em: 10 jan. 2023.

WICKHAM, Hadley; GIRLICH, Maximilian; RSTUDIO. **tidyr: Tidy Messy Data.** , 2022. Disponível em: <https://CRAN.R-project.org/package=tidyr>. Acesso em: 10 jan. 2023.

WISZ, Mary S.; GUIBAN, Antoine. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. **BMC Ecology**, [S. l.], v. 9, n. 1, p. 8, 2009. DOI: [10.1186/1472-6785-9-8](https://doi.org/10.1186/1472-6785-9-8). Disponível em: <http://bmcecol.biomedcentral.com/articles/10.1186/1472-6785-9-8>. Acesso em: 10 jan. 2023.

ZHANG, Zhixin; XU, Shengyong; CAPINHA, César; WETERINGS, Robbie; GAO, Tianxiang. Using species distribution model to predict the impact of climate change on the potential distribution of Japanese whiting *Sillago japonica*. **Ecological Indicators**, [S. l.], v. 104, p. 333–340, 2019. DOI: [10.1016/j.ecolind.2019.05.023](https://doi.org/10.1016/j.ecolind.2019.05.023). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1470160X19303681>. Acesso em: 6 fev. 2023.

3 CONSIDERAÇÕES FINAIS

A Modelagem de Distribuição de Espécies, em conjunto com os Modelos de Nicho Ecológico demonstra-se como uma importante ferramenta para o estudo e manejo da biodiversidade. A aplicação dos modelos requer, no entanto, a tomada de decisões que podem enviesar os resultados da modelagem quando pontos fundamentais do processo não são considerados corretamente.

Isto posto, as principais conclusões do presente estudo podem ser sumarizadas como:

- O processo de modelagem, quer de nicho, quer de distribuição potencial, é longo e complexo.
- É fundamental ao pesquisador estabelecer o que está sendo modelado: a distribuição geográfica potencial da espécie ou nicho fundamental são feições diferentes dentro do mesmo espectro do diagrama BAM proposto por Soberón e Peterson (2005). Portanto, a sinonímia “Modelo de Distribuição de Espécies” x “Modelo de Nicho Ecológico” não equivale a mesma técnica, mas sim a um conjunto de técnicas compartilhadas entre processos diferentes.
- Como não há um algoritmo melhor do que os demais, a seleção de qual algoritmo utilizar para modelar a espécie precisa estar embasada não só na literatura, mas na realidade biológica da espécie e na logística/escopo do estudo. O tamanho e características abióticas da área de estudo também são pontos fundamentais nessa etapa.
- O total de pontos de presença e pseudo ausência ou *background* utilizado na calibração dos modelos são pontos críticos. A decisão de quantos pontos utilizar é fundamental para minimizar o enviesamento (i.e., surgimentos de erros de omissão e comissão) no modelo final.
- O tipo de pseudo ausência ou de pontos de *background* a ser utilizado na calibração dos modelos depende, principalmente, do tipo de algoritmo utilizado no processo. O modo como a “ausência” será gerada também é importante para que os resultados da modelagem sejam válidos estatística e biologicamente. É importante, também, conceituar adequadamente o tipo de “ausência” utilizada: pseudo ausência não é *background*.

- A validação estatística dos modelos não deve ser embasada em uma única métrica, uma vez que esta pode ser enviesada ou facilmente mascarar erros de omissão ou comissão.
- O sentido biológico tem prioridade em relação à validação estatística dos modelos.

O presente trabalho contribui para fomentar as discussões no que concerne aos processos e aplicações da Modelagem de Distribuição de Espécies, bem como dos Modelos de Nicho Ecológico. Espera-se que futuras pesquisas na área possibilitem não só o aprimoramento das técnicas, mas também uma melhor compreensão e fundamentação dos conceitos ecológicos, estatísticos e biológicos que embasam as técnicas de modelagem.

REFERÊNCIAS

- AIELLO-LAMMENS, Matthew E.; BORJA, Robert A.; RADOSAVLJEVIC, Aleksandar; VILELA, Bruno; ANDERSON, Robert P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. **Ecography**, [S. l.], v. 38, n. 5, p. 541–545, 2015. DOI: [10.1111/ecog.01132](https://doi.org/10.1111/ecog.01132). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01132>. Acesso em: 9 jan. 2023.
- ALLOUCHE, Omri; TSOAR, Asaf; KADMON, Ronen. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. **Journal of Applied Ecology**, [S. l.], v. 43, n. 6, p. 1223–1232, 2006. DOI: [10.1111/j.1365-2664.2006.01214.x](https://doi.org/10.1111/j.1365-2664.2006.01214.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2006.01214.x>. Acesso em: 10 jan. 2023.
- BARBET-MASSIN, Morgane; JIGUET, Frédéric; ALBERT, Cécile Hélène; THUILLER, Wilfried. Selecting pseudo-absences for species distribution models: how, where and how many?: *How to use pseudo-absences in niche modelling?* **Methods in Ecology and Evolution**, [S. l.], v. 3, n. 2, p. 327–338, 2012. DOI: [10.1111/j.2041-210X.2011.00172.x](https://doi.org/10.1111/j.2041-210X.2011.00172.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00172.x>. Acesso em: 15 fev. 2022.
- BARBET-MASSIN, Morgane; THUILLER, Wilfried; JIGUET, Frédéric. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? **Ecography**, [S. l.], v. 33, n. 5, p. 878–886, 2010. DOI: [10.1111/j.1600-0587.2010.06181.x](https://doi.org/10.1111/j.1600-0587.2010.06181.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2010.06181.x>. Acesso em: 3 jan. 2023.
- BOOTH, Trevor H.; NIX, Henry A.; BUSBY, John R.; HUTCHINSON, Michael F. bioclim : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. **Diversity and Distributions**, [S. l.], v. 20, n. 1, p. 1–9, 2014. DOI:

[10.1111/ddi.12144](https://onlinelibrary.wiley.com/doi/10.1111/ddi.12144). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ddi.12144>. Acesso em: 10 jan. 2023.

BOTELLA, Christophe; JOLY, Alexis; BONNET, Pierre; MONESTIEZ, Pascal; MUNOZ, François. Species distribution modeling based on the automated identification of citizen observations. **Applications in Plant Sciences**, [S. l.], v. 6, n. 2, p. e1029, 2018. DOI: [10.1002/aps3.1029](https://doi.org/10.1002/aps3.1029). Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5851560/>. Acesso em: 10 jan. 2023.

BROTONS, Lluís; THUILLER, Wilfried; ARAÚJO, Miguel B.; HIRZEL, Alexandre H. Presence-Absence versus Presence-Only Modelling Methods for Predicting Bird Habitat Suitability. **Ecography**, [S. l.], v. 27, n. 4, p. 437–448, 2004. DOI: [10.2307/3683417](https://doi.org/10.2307/3683417). Disponível em: <https://www.jstor.org/stable/3683417>. Acesso em: 4 maio. 2022.

BUISSON, Laëtitia; THUILLER, Wilfried; CASAJUS, Nicolas; LEK, Sovan; GRENOUILLET, Gaël. Uncertainty in ensemble forecasting of species distribution. **Global Change Biology**, [S. l.], v. 16, n. 4, p. 1145–1157, 2010. DOI: [10.1111/j.1365-2486.2009.02000.x](https://doi.org/10.1111/j.1365-2486.2009.02000.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2009.02000.x>. Acesso em: 25 maio. 2022.

CALL, Ashley; SUN, Yan-Xia; YU, Yan; PEARMAN, Peter B.; THOMAS, David T.; TRIGIANO, Robert N.; CARBONE, Ignazio; XIANG, Qiu-Yun Jenny. Genetic structure and post-glacial expansion of *Cornus florida* L. (Cornaceae): integrative evidence from phylogeography, population demographic history, and species distribution modeling: Phylogeography of *Cornus florida*. **Journal of Systematics and Evolution**, [S. l.], v. 54, n. 2, p. 136–151, 2016. DOI: [10.1111/jse.12171](https://doi.org/10.1111/jse.12171). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/jse.12171>. Acesso em: 24 maio. 2022.

Cantorchilus longirostris (Vieillot, 1819) in GBIF Secretariat (2022). GBIF Backbone Taxonomy. Checklist dataset <https://doi.org/10.15468/39omei> acesso via GBIF.org em: 10 jan 2023.

CARRETERO, Miguel A.; SILLERO, Neftalí. Evaluating how species niche modelling is affected by partial distributions with an empirical case. **Acta Oecologica**, [S. l.], v. 77, p. 207–216, 2016. DOI: [10.1016/j.actao.2016.08.014](https://doi.org/10.1016/j.actao.2016.08.014). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1146609X16301710>. Acesso em: 31 out. 2022.

CARVALHO, Antônio F.; DEL LAMA, Marco Antonio. Predicting priority areas for conservation from historical climate modelling: stingless bees from Atlantic Forest hotspot as a case study. **Journal of Insect Conservation**, [S. l.], v. 19, n. 3, p. 581–587, 2015. DOI: [10.1007/s10841-015-9780-7](https://doi.org/10.1007/s10841-015-9780-7). Disponível em: <https://doi.org/10.1007/s10841-015-9780-7>. Acesso em: 10 jan. 2023.

ČENGÍĆ, Mirza; ROST, Jasmijn; REMENSKA, Daniela; JANSE, Jan H.; HUIJBREGTS, Mark A. J.; SCHIPPER, Aafke M. On the importance of predictor choice, modelling technique, and number of pseudo-absences for bioclimatic envelope model performance.

Ecology and Evolution, [S. l.], v. 10, n. 21, p. 12307–12317, 2020. DOI: [10.1002/ece3.6859](https://doi.org/10.1002/ece3.6859). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.6859>. Acesso em: 10 jan. 2023.

CHAPMAN, Daniel; PESSCOTT, Oliver L.; ROY, Helen E.; TANNER, Rob. Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. **Journal of Biogeography**, [S. l.], v. 46, n. 5, p. 1029–1040, 2019. DOI: [10.1111/jbi.13555](https://doi.org/10.1111/jbi.13555). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/jbi.13555>. Acesso em: 10 jan. 2023.

CHATTERJEE, Samprit; HADI, Ali S. **Regression analysis by example**. 4th ed ed. Hoboken, N.J: Wiley-Interscience, 2006.

COELLO, Alberto J.; FERNÁNDEZ-MAZUECOS, Mario; GARCÍA-VERDUGO, Carlos; VARGAS, Pablo. Phylogeographic sampling guided by species distribution modeling reveals the Quaternary history of the Mediterranean–Canarian *Cistus monspeliensis* (Cistaceae). **Journal of Systematics and Evolution**, [S. l.], v. 59, n. 2, p. 262–277, 2021. DOI: [10.1111/jse.12570](https://doi.org/10.1111/jse.12570). Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jse.12570>. Acesso em: 10 jan. 2023.

COLWELL, Robert K.; RANGEL, Thiago F. Hutchinson’s duality: The once and future niche. **Proceedings of the National Academy of Sciences**, [S. l.], v. 106, n. supplement_2, p. 19651–19658, 2009. DOI: [10.1073/pnas.0901650106](https://doi.org/10.1073/pnas.0901650106). Disponível em: <https://pnas.org/doi/full/10.1073/pnas.0901650106>. Acesso em: 3 jan. 2023.

DAUBY, Gilles et al. *ConR*: An R package to assist large-scale multispecies preliminary conservation assessments using distribution data. **Ecology and Evolution**, [S. l.], v. 7, n. 24, p. 11292–11303, 2017. DOI: [10.1002/ece3.3704](https://doi.org/10.1002/ece3.3704). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.3704>. Acesso em: 10 jan. 2023.

DESCOMBES, Patrice; CHAUVIER, Yohann; BRUN, Philipp; RIGHETTI, Damiano; WÜEST, Rafael O.; KARGER, Dirk N.; ZURELL, Damaris; ZIMMERMANN, Niklaus E. **Strategies for sampling pseudo-absences for species distribution models in complex mountainous terrain**. [s.l.] : Ecology, 2022. DOI: [10.1101/2022.03.24.485693](https://doi.org/10.1101/2022.03.24.485693). Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2022.03.24.485693>. Acesso em: 2 jan. 2023.

DEVANEY, Shannon C. Species Distribution Modeling of Deep Pelagic Eels. **Integrative and Comparative Biology**, [S. l.], v. 56, n. 4, p. 524–530, 2016. DOI: [10.1093/icb/icw032](https://doi.org/10.1093/icb/icw032). Disponível em: <https://academic.oup.com/icb/article-lookup/doi/10.1093/icb/icw032>. Acesso em: 10 jan. 2023.

DORMANN, Carsten F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, [S. l.], v. 36, n. 1, p. 27–46, 2013. DOI: [10.1111/j.1600-0587.2012.07348.x](https://doi.org/10.1111/j.1600-0587.2012.07348.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2012.07348.x>. Acesso em: 10 jan. 2023.

EL-GABBAS, Ahmed; DORMANN, Carsten F. Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling. **Ecology and Evolution**, [S. l.], v. 8, n. 4, p. 2196–2206, 2018. DOI: [10.1002/ece3.3834](https://doi.org/10.1002/ece3.3834). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.3834>. Acesso em: 10 jan. 2023.

ELITH, Jane et al. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, [S. l.], v. 29, n. 2, p. 129–151, 2006. DOI: [10.1111/j.2006.0906-7590.04596.x](https://doi.org/10.1111/j.2006.0906-7590.04596.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.2006.0906-7590.04596.x>. Acesso em: 15 fev. 2022.

ELITH, Jane; LEATHWICK, John R. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. **Annual Review of Ecology, Evolution, and Systematics**, [S. l.], v. 40, n. 1, p. 677–697, 2009. DOI: [10.1146/annurev.ecolsys.110308.120159](https://doi.org/10.1146/annurev.ecolsys.110308.120159). Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.110308.120159>. Acesso em: 23 abr. 2019.

ENGLER, Robin; GUIBAN, Antoine; RECHSTEINER, Luca. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. **Journal of Applied Ecology**, [S. l.], v. 41, n. 2, p. 263–274, 2004. DOI: [10.1111/j.0021-8901.2004.00881.x](https://doi.org/10.1111/j.0021-8901.2004.00881.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.0021-8901.2004.00881.x>. Acesso em: 24 maio. 2022.

Eupsittula cactorum (Kuhl, 1820) in GBIF Secretariat (2022). GBIF Backbone Taxonomy. Checklist dataset <https://doi.org/10.15468/39omei> acesso via GBIF.org em: 10 jan. 2023.

FICK, Stephen E.; HIJMANS, Robert J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology**, [S. l.], v. 37, n. 12, p. 4302–4315, 2017. DOI: [10.1002/joc.5086](https://doi.org/10.1002/joc.5086). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/joc.5086>. Acesso em: 10 jan. 2023.

FIELDING, Alan H.; BELL, John F. A review of methods for the assessment of prediction errors in conservation presence/absence models. **Environmental Conservation**, [S. l.], v. 24, n. 1, p. 38–49, 1997. DOI: [10.1017/S0376892997000088](https://doi.org/10.1017/S0376892997000088). Disponível em: https://www.cambridge.org/core/product/identifier/S0376892997000088/type/journal_article. Acesso em: 10 jan. 2023.

FLESSNER, Brandon; HENRY, Mary C.; GREEN, Jerry. Species Distribution Modeling of American Beech (*Fagus Grandifolia*) Distribution in Southwest Ohio: **International Journal of Applied Geospatial Research**, [S. l.], v. 8, n. 3, p. 16–36, 2017. DOI: [10.4018/ijagr.2017070102](https://doi.org/10.4018/ijagr.2017070102). Disponível em: <https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/ijagr.2017070102>. Acesso em: 10 jan. 2023.

FRANCO, Francisco; CAMPS, Gonzalo Andrés. La aplicación de modelos de distribución de especies para la realización de inferencias arqueológicas. Una ejemplificación a partir de *Geoffroea decorticans* en el área Sudcalchaquí (Noroeste, Argentina). **InterSecciones en Antropología**, [S. l.], v. 21, n. 2, p. 131–144, 2020. DOI: [10.37176/iea.21.2.2020.489](https://doi.org/10.37176/iea.21.2.2020.489).

Disponível em:

<https://interseccionesantro.soc.unicen.edu.ar/index.php/intersecciones/article/view/489>.

Acesso em: 10 jan. 2023.

FRANKLIN, Janet; MILLER, Jennifer A. **Mapping species distributions: spatial inference and prediction**. Cambridge ; New York: Cambridge University Press, 2009.

GU, Weidong; SWIHART, Robert K. Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. **Biological Conservation**, [S. l.], v. 116, n. 2, p. 195–203, 2004. DOI: [10.1016/S0006-3207\(03\)00190-3](https://doi.org/10.1016/S0006-3207(03)00190-3). Disponível em:

<https://linkinghub.elsevier.com/retrieve/pii/S0006320703001903>. Acesso em: 15 fev. 2022.

GUISAN, Antoine et al. Predicting species distributions for conservation decisions. **Ecology Letters**, [S. l.], v. 16, n. 12, p. 1424–1435, 2013. DOI: [10.1111/ele.12189](https://doi.org/10.1111/ele.12189). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/ele.12189>. Acesso em: 10 fev. 2022.

GUISAN, Antoine; THUILLER, Wilfried. Predicting species distribution: offering more than simple habitat models. **Ecology Letters**, [S. l.], v. 8, n. 9, p. 993–1009, 2005. DOI:

[10.1111/j.1461-0248.2005.00792.x](https://doi.org/10.1111/j.1461-0248.2005.00792.x). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2005.00792.x>. Acesso em: 15 fev. 2022.

GUISAN, Antoine; THUILLER, Wilfried; ZIMMERMANN, Niklaus E. **Habitat Suitability and Distribution Models: With Applications in R**. Cambridge: Cambridge University Press, 2017. DOI: [10.1017/9781139028271](https://doi.org/10.1017/9781139028271). Disponível em:

<http://ebooks.cambridge.org/ref/id/CBO9781139028271>. Acesso em: 4 maio. 2022.

GUISAN, Antoine; ZIMMERMANN, Niklaus E. Predictive habitat distribution models in ecology. **Ecological Modelling**, [S. l.], v. 135, n. 2–3, p. 147–186, 2000. DOI:

[10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9). Disponível em:

<https://linkinghub.elsevier.com/retrieve/pii/S0304380000003549>. Acesso em: 4 maio. 2022.

HALLGREN, W.; SANTANA, F.; LOW-CHOY, S.; ZHAO, Y.; MACKEY, B. Species distribution models can be highly sensitive to algorithm configuration. **Ecological Modelling**, [S. l.], v. 408, p. 108719, 2019. DOI: [10.1016/j.ecolmodel.2019.108719](https://doi.org/10.1016/j.ecolmodel.2019.108719). Disponível em:

<https://linkinghub.elsevier.com/retrieve/pii/S0304380019302194>. Acesso em: 25 maio. 2022.

HANBERRY, Brice B.; HE, Hong S.; PALIK, Brian J. Pseudoabsence Generation Strategies for Species Distribution Models. **PLoS ONE**, [S. l.], v. 7, n. 8, p. e44486, 2012. DOI:

[10.1371/journal.pone.0044486](https://doi.org/10.1371/journal.pone.0044486). Disponível em:

<https://dx.plos.org/10.1371/journal.pone.0044486>. Acesso em: 10 jan. 2023.

HEMAMI, Mahmoud-Reza; KHOSRAVI, Rasoul; GROVES, Colin; AHMADI, Mohsen. Morphological diversity and ecological niche divergence in goitered and sand gazelles. **Ecology and Evolution**, [S. l.], v. 10, n. 20, p. 11535–11548, 2020. DOI: [10.1002/ece3.6789](https://doi.org/10.1002/ece3.6789). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ece3.6789>. Acesso em: 10 jan. 2023.

HERNANDEZ, P. A. et al. Predicting species distributions in poorly-studied landscapes. **Biodiversity and Conservation**, [S. l.], v. 17, n. 6, p. 1353–1366, 2008. DOI: [10.1007/s10531-007-9314-z](https://doi.org/10.1007/s10531-007-9314-z). Disponível em: <http://link.springer.com/10.1007/s10531-007-9314-z>. Acesso em: 31 out. 2022.

HIJMANS, Robert J. et al. **raster: Geographic Data Analysis and Modeling**. , 2023. Disponível em: <https://CRAN.R-project.org/package=raster>. Acesso em: 10 jan. 2023.

HIJMANS, Robert J.; CAMERON, Susan E.; PARRA, Juan L.; JONES, Peter G.; JARVIS, Andy. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, [S. l.], v. 25, n. 15, p. 1965–1978, 2005. DOI: [10.1002/joc.1276](https://doi.org/10.1002/joc.1276). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/joc.1276>. Acesso em: 10 jan. 2023.

HIJMANS, Robert J.; ELITH, Jane. **Species Distribution Models**. , 2021. Disponível em: https://rsatial.org/raster/sdm/raster_SDM.pdf. Acesso em: 14 fev. 2022.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. **Climate Change and Land: IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems**. 1. ed. [s.l.] : Cambridge University Press, 2022. DOI: [10.1017/9781009157988](https://doi.org/10.1017/9781009157988). Disponível em: <https://www.cambridge.org/core/product/identifier/9781009157988/type/book>. Acesso em: 3 jan. 2023.

IPBES. **Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services**. [s.l.] : Zenodo, 2019. DOI: [10.5281/ZENODO.3831673](https://doi.org/10.5281/ZENODO.3831673). Disponível em: <https://zenodo.org/record/3831673>. Acesso em: 7 fev. 2022.

ITURBIDE, Maialen; BEDIA, Joaquín; GUTIÉRREZ, José Manuel. Background sampling and transferability of species distribution model ensembles under climate change. **Global and Planetary Change**, [S. l.], v. 166, p. 19–29, 2018. DOI: [10.1016/j.gloplacha.2018.03.008](https://doi.org/10.1016/j.gloplacha.2018.03.008). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0921818117302308>. Acesso em: 25 maio. 2022.

IUCN. **Iucn Red List Categories And Criteria: Version 3.1. Second Edition**. [s.l.] : Zenodo, 2012. DOI: [10.5281/ZENODO.15039](https://doi.org/10.5281/ZENODO.15039). Disponível em: <https://zenodo.org/record/15039>. Acesso em: 10 jan. 2023.

JETZ, Walter; MCPHERSON, Jana M.; GURALNICK, Robert P. Integrating biodiversity distribution knowledge: toward a global map of life. **Trends in Ecology & Evolution**, [S. l.], v. 27, n. 3, p. 151–159, 2012. DOI: [10.1016/j.tree.2011.09.007](https://doi.org/10.1016/j.tree.2011.09.007). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169534711002679>. Acesso em: 11 fev. 2022.

JURESTOVSKY, Derek; ANDREW JOYNER, T. Applications of species distribution modeling for palaeontological fossil detection: late Pleistocene models of Saiga (Artiodactyla: Bovidae, Saiga tatarica). **Palaeobiodiversity and Palaeoenvironments**, [S. l.], v. 98, n. 2, p. 277–285, 2018. DOI: [10.1007/s12549-017-0298-8](https://doi.org/10.1007/s12549-017-0298-8). Disponível em: <https://doi.org/10.1007/s12549-017-0298-8>. Acesso em: 10 jan. 2023.

KUMAR, Sunil; STOHLGREN, Thomas J. Maxent modeling for predicting suitable habitat for threatened and endangered tree *Canacomyrica monticola* in New Caledonia. **Journal of Ecology and Natural Environment**, [S. l.], v. 1, n. 4, p. 94–98, 2009. DOI: <https://doi.org/10.5897/JENE.9000071>. Disponível em: http://www.academicjournals.org/app/webroot/article/article1379515268_Kumar%20and%20Stohlgren.pdf.

LIANG, Wanwan; PAPEŞ, Monica; TRAN, Liem; GRANT, Jerome; WASHINGTON-ALLEN, Robert; STEWART, Scott; WIGGINS, Gregory. The effect of pseudo-absence selection method on transferability of species distribution models in the context of non-adaptive niche shift. **Ecological Modelling**, [S. l.], v. 388, p. 1–9, 2018. DOI: [10.1016/j.ecolmodel.2018.09.018](https://doi.org/10.1016/j.ecolmodel.2018.09.018). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380018303156>. Acesso em: 10 jan. 2023.

LI, Junjun; FAN, Gang; HE, Yang. Predicting the current and future distribution of three *Coptis* herbs in China under climate change conditions, using the MaxEnt model and chemical analysis. **Science of The Total Environment**, [S. l.], v. 698, p. 134141, 2020. DOI: [10.1016/j.scitotenv.2019.134141](https://doi.org/10.1016/j.scitotenv.2019.134141). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S004896971934118X>. Acesso em: 10 jan. 2023.

LOBO, Jorge M.; JIMÉNEZ-VALVERDE, Alberto; HORTAL, Joaquín. The uncertain nature of absences and their importance in species distribution modelling. **Ecography**, [S. l.], v. 33, n. 1, p. 103–114, 2010. DOI: [10.1111/j.1600-0587.2009.06039.x](https://doi.org/10.1111/j.1600-0587.2009.06039.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2009.06039.x>. Acesso em: 24 maio. 2022.

MACKENZIE, Darryl I.; ROYLE, J. Andrew. Designing occupancy studies: general advice and allocating survey effort: *Designing occupancy studies*. **Journal of Applied Ecology**, [S. l.], v. 42, n. 6, p. 1105–1114, 2005. DOI: [10.1111/j.1365-2664.2005.01098.x](https://doi.org/10.1111/j.1365-2664.2005.01098.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2005.01098.x>. Acesso em: 15 fev. 2022.

MARMION, Mathieu; PARVIAINEN, Miia; LUOTO, Miska; HEIKKINEN, Risto K.; THUILLER, Wilfried. Evaluation of consensus methods in predictive species distribution modelling. **Diversity and Distributions**, [S. l.], v. 15, n. 1, p. 59–69, 2009. DOI: [10.1111/j.1472-4642.2008.00491.x](https://doi.org/10.1111/j.1472-4642.2008.00491.x). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00491.x>. Acesso em: 10 jan. 2023.

MAYNARD SMITH, John. **Games, sex, and evolution**. New York: Harvester-Wheatsheaf, 1988.

MCCARTNEY, Kevin R.; KUMAR, Sunil; SING, Sharlene E.; WARD, Sarah M. Using invaded-range species distribution modeling to estimate the potential distribution of *Linaria* species and their hybrids in the U.S. northern Rockies. **Invasive Plant Science and Management**, [S. l.], v. 12, n. 02, p. 97–111, 2019. DOI: [10.1017/inp.2019.15](https://doi.org/10.1017/inp.2019.15). Disponível em:

https://www.cambridge.org/core/product/identifier/S1939729119000154/type/journal_article. Acesso em: 10 jan. 2023.

MUSCATELLO, Angela; ELITH, Jane; KUJALA, Heini. How decisions about fitting species distribution models affect conservation outcomes. **Conservation Biology**, [S. l.], v. 35, n. 4, p. 1309–1320, 2021. DOI: [10.1111/cobi.13669](https://doi.org/10.1111/cobi.13669). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/cobi.13669>. Acesso em: 10 fev. 2022.

NAIMI, Babak; ARAÚJO, Miguel B. sdm: a reproducible and extensible R platform for species distribution modelling. **Ecography**, [S. l.], v. 39, n. 4, p. 368–375, 2016. DOI: [10.1111/ecog.01881](https://doi.org/10.1111/ecog.01881). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01881>. Acesso em: 10 jan. 2023.

NAIMI, Babak; HAMM, Nicholas A. S.; GROEN, Thomas A.; SKIDMORE, Andrew K.; TOXOPEUS, Albertus G. Where is positional uncertainty a problem for species distribution modelling? **Ecography**, [S. l.], v. 37, n. 2, p. 191–203, 2014. DOI: [10.1111/j.1600-0587.2013.00205.x](https://doi.org/10.1111/j.1600-0587.2013.00205.x). Disponível em:

<https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2013.00205.x>. Acesso em: 10 jan. 2023.

PAPEȘ, M.; GAUBERT, P. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents: Ecological niche modelling of poorly known viverrids. **Diversity and Distributions**, [S. l.], v. 13, n. 6, p. 890–902, 2007. DOI: [10.1111/j.1472-4642.2007.00392.x](https://doi.org/10.1111/j.1472-4642.2007.00392.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2007.00392.x>. Acesso em: 31 out. 2022.

PEARSON, Richard G.; RAXWORTHY, Christopher J.; NAKAMURA, Miguel; TOWNSEND PETERSON, A. ORIGINAL ARTICLE: Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar: Predicting species distributions with low sample sizes. **Journal of Biogeography**, [S. l.], v. 34, n. 1, p. 102–117, 2006. DOI: [10.1111/j.1365-2699.2006.01594.x](https://doi.org/10.1111/j.1365-2699.2006.01594.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2699.2006.01594.x>. Acesso em: 10 jan. 2023.

Penelope obscura Temminck, 1815 in GBIF Secretariat (2022). GBIF Backbone Taxonomy. Checklist dataset <https://doi.org/10.15468/39omei> acesso via GBIF.org em: 10 jan. 2023.

PETERSON, A. Townsend (ORG.). **Ecological niches and geographic distributions**. Princeton, N.J: Princeton University Press, 2011.

PETERSON, A. Townsend; SOBERÓN, Jorge. Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. **Natureza & Conservação**, [S. l.], v. 10, n. 2, p. 102–107, 2012. DOI: [10.4322/natcon.2012.019](https://doi.org/10.4322/natcon.2012.019). Disponível em: <http://doi.editoracubo.com.br/10.4322/natcon.2012.019>. Acesso em: 24 maio. 2022.

PHILLIPS, Steven J.; DUDÍK, Miroslav; ELITH, Jane; GRAHAM, Catherine H.; LEHMANN, Anthony; LEATHWICK, John; FERRIER, Simon. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. **Ecological Applications**, [S. l.], v. 19, n. 1, p. 181–197, 2009. DOI: [10.1890/07-2153.1](https://doi.org/10.1890/07-2153.1). Disponível em: <http://doi.wiley.com/10.1890/07-2153.1>. Acesso em: 25 maio. 2022.

PINTO-LEDEZMA, Jesús N.; CAVENDER-BARES, Jeannine. Predicting species distributions and community composition using satellite remote sensing predictors. **Scientific Reports**, [S. l.], v. 11, n. 1, p. 16448, 2021. DOI: [10.1038/s41598-021-96047-7](https://doi.org/10.1038/s41598-021-96047-7). Disponível em: <https://www.nature.com/articles/s41598-021-96047-7>. Acesso em: 7 fev. 2022.

PROOSDIJ, André S. J.; SOSEF, Marc S. M.; WIERINGA, Jan J.; RAES, Niels. Minimum required number of specimen records to develop accurate species distribution models. **Ecography**, [S. l.], v. 39, n. 6, p. 542–552, 2016. DOI: [10.1111/ecog.01509](https://doi.org/10.1111/ecog.01509). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ecog.01509>. Acesso em: 15 fev. 2022.

QIN, Aili; LIU, Bo; GUO, Quanshui; BUSSMANN, Rainer W.; MA, Fanqiang; JIAN, Zunji; XU, Gexi; PEI, Shunxiang. Maxent modeling for predicting impacts of climate change on the potential distribution of *Thuja sutchuenensis* Franch., an extremely endangered conifer from southwestern China. **Global Ecology and Conservation**, [S. l.], v. 10, p. 139–146, 2017. DOI: [10.1016/j.gecco.2017.02.004](https://doi.org/10.1016/j.gecco.2017.02.004). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2351989416301470>. Acesso em: 10 jan. 2023.

RAES, Niels. Partial versus Full Species Distribution Models. **Natureza & Conservação**, [S. l.], v. 10, n. 2, p. 127–138, 2012. DOI: [10.4322/natcon.2012.020](https://doi.org/10.4322/natcon.2012.020). Disponível em: <http://doi.editoracubo.com.br/10.4322/natcon.2012.020>. Acesso em: 3 jan. 2023.

RAES, Niels; AGUIRRE-GUTIÉRREZ, Jesús. A Modeling Framework to Estimate and Project Species Distributions in Space and Time. Em: **Mountains, Climate and Biodiversity**. 1st. ed. Hoboken, NJ: John Wiley & Sons, 2018.

RONDININI, Carlo; WILSON, Kerrie A.; BOITANI, Luigi; GRANTHAM, Hedley; POSSINGHAM, Hugh P. Tradeoffs of different types of species occurrence data for use in systematic conservation planning: Species data for conservation planning. **Ecology Letters**, [S. l.], v. 9, n. 10, p. 1136–1145, 2006. DOI: [10.1111/j.1461-0248.2006.00970.x](https://doi.org/10.1111/j.1461-0248.2006.00970.x). Disponível

em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2006.00970.x>. Acesso em: 15 fev. 2022.

SENAY, Senait D.; WORNER, Susan P.; IKEDA, Takayoshi. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. **PLoS ONE**, [S. l.], v. 8, n. 8, p. e71218, 2013. DOI: [10.1371/journal.pone.0071218](https://doi.org/10.1371/journal.pone.0071218). Disponível em: <https://dx.plos.org/10.1371/journal.pone.0071218>. Acesso em: 10 jan. 2023.

SHCHEGLOVITOVA, Mariya; ANDERSON, Robert P. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. **Ecological Modelling**, [S. l.], v. 269, p. 9–17, 2013. DOI: [10.1016/j.ecolmodel.2013.08.011](https://doi.org/10.1016/j.ecolmodel.2013.08.011). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380013004043>. Acesso em: 24 maio. 2022.

SILLERO, Neftalí; ARENAS-CASTRO, Salvador; ENRIQUEZ-URZELAI, Urtzi; VALE, Cândida Gomes; SOUSA-GUEDES, Diana; MARTÍNEZ-FREIRÍA, Fernando; REAL, Raimundo; BARBOSA, A. Márcia. Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. **Ecological Modelling**, [S. l.], v. 456, p. 109671, 2021. DOI: [10.1016/j.ecolmodel.2021.109671](https://doi.org/10.1016/j.ecolmodel.2021.109671). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380021002301>. Acesso em: 31 out. 2022.

SILLERO, Neftalí; BARBOSA, A. Márcia. Common mistakes in ecological niche models. **International Journal of Geographical Information Science**, [S. l.], v. 35, n. 2, p. 213–226, 2021. DOI: [10.1080/13658816.2020.1798968](https://doi.org/10.1080/13658816.2020.1798968). Disponível em: <https://www.tandfonline.com/doi/full/10.1080/13658816.2020.1798968>. Acesso em: 24 maio. 2022.

SOBERON, Jorge; PETERSON, A. Townsend. Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. **Biodiversity Informatics**, [S. l.], v. 2, n. 0, 2005. DOI: [10.17161/bi.v2i0.4](https://doi.org/10.17161/bi.v2i0.4). Disponível em: <https://journals.ku.edu/jbi/article/view/4>. Acesso em: 3 jan. 2023.

SPIERS, Joshua A.; OATHAM, Michael P.; ROSTANT, Luke V.; FARRELL, Aidan D. Applying species distribution modelling to improving conservation based decisions: a gap analysis of Trinidad and Tobago's endemic vascular plants. **Biodiversity and Conservation**, [S. l.], v. 27, n. 11, p. 2931–2949, 2018. DOI: [10.1007/s10531-018-1578-y](https://doi.org/10.1007/s10531-018-1578-y). Disponível em: <https://doi.org/10.1007/s10531-018-1578-y>. Acesso em: 10 jan. 2023.

TESSAROLO, Geiziane; RANGEL, Thiago F.; ARAÚJO, Miguel B.; HORTAL, Joaquín. Uncertainty associated with survey design in Species Distribution Models. **Diversity and Distributions**, [S. l.], v. 20, n. 11, p. 1258–1269, 2014. DOI: [10.1111/ddi.12236](https://doi.org/10.1111/ddi.12236). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/ddi.12236>. Acesso em: 31 out. 2022.

THIBAUD, Emeric; PETITPIERRE, Blaise; BROENNIMANN, Olivier; DAVISON, Anthony C.; GUISAN, Antoine. Measuring the relative effect of factors affecting species distribution model predictions. **Methods in Ecology and Evolution**, [S. l.], v. 5, n. 9, p.

947–955, 2014. DOI: [10.1111/2041-210X.12203](https://doi.org/10.1111/2041-210X.12203). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12203>. Acesso em: 31 out. 2022.

THUILLER, Wilfried. Patterns and uncertainties of species' range shifts under climate change: SPECIES' RANGE SHIFTS. **Global Change Biology**, [S. l.], v. 10, n. 12, p. 2020–2027, 2004. DOI: [10.1111/j.1365-2486.2004.00859.x](https://doi.org/10.1111/j.1365-2486.2004.00859.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2004.00859.x>. Acesso em: 10 jan. 2023.

TRAN, Dung Van; VU, Tinh Tien. Combining species distribution modeling and distance sampling to assess wildlife population size: A case study with the northern yellow-cheeked gibbon (*Nomascus annamensis*). **American Journal of Primatology**, [S. l.], v. 82, n. 9, 2020. DOI: [10.1002/ajp.23169](https://doi.org/10.1002/ajp.23169). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ajp.23169>. Acesso em: 24 maio. 2022.

WICKHAM, Hadley; FRANÇOIS, Romain; HENRY, Lionel; MÜLLER, Kirill; RSTUDIO. **dplyr: A Grammar of Data Manipulation**. , 2022. Disponível em: <https://CRAN.R-project.org/package=dplyr>. Acesso em: 10 jan. 2023.

WICKHAM, Hadley; GIRLICH, Maximilian; RSTUDIO. **tidyr: Tidy Messy Data**. , 2022. Disponível em: <https://CRAN.R-project.org/package=tidyr>. Acesso em: 10 jan. 2023.

WISZ, M. S.; HIJMANS, R. J.; LI, J.; PETERSON, A. T.; GRAHAM, C. H.; GUIBAN, A.; NCEAS PREDICTING SPECIES DISTRIBUTIONS WORKING GROUP†. Effects of sample size on the performance of species distribution models. **Diversity and Distributions**, [S. l.], v. 14, n. 5, p. 763–773, 2008. DOI: [10.1111/j.1472-4642.2008.00482.x](https://doi.org/10.1111/j.1472-4642.2008.00482.x). Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2008.00482.x>. Acesso em: 10 jan. 2023.

WISZ, Mary S.; GUIBAN, Antoine. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. **BMC Ecology**, [S. l.], v. 9, n. 1, p. 8, 2009. DOI: [10.1186/1472-6785-9-8](https://doi.org/10.1186/1472-6785-9-8). Disponível em: <http://bmcecol.biomedcentral.com/articles/10.1186/1472-6785-9-8>. Acesso em: 10 jan. 2023.

ZANIEWSKI, A. Elizabeth; LEHMANN, Anthony; OVERTON, Jacob McC. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. **Ecological Modelling**, [S. l.], v. 157, n. 2–3, p. 261–280, 2002. DOI: [10.1016/S0304-3800\(02\)00199-0](https://doi.org/10.1016/S0304-3800(02)00199-0). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304380002001990>. Acesso em: 24 maio. 2022.

Material Suplementar 1

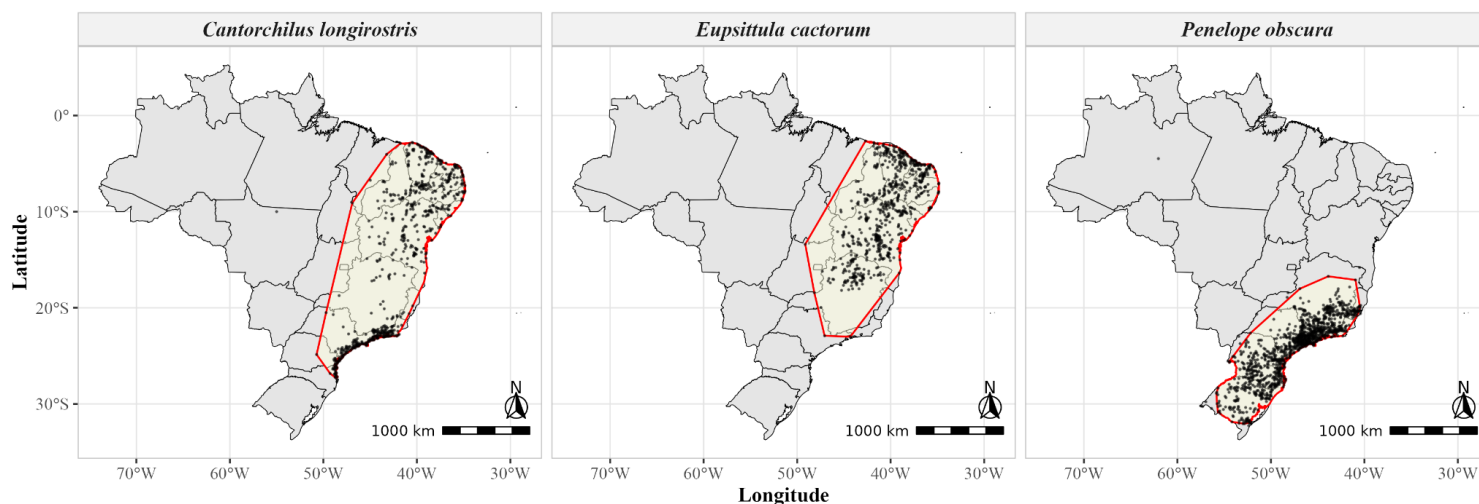


Figura S1 - Distribuição dos pontos de ocorrência obtidos na base de dados GBIF dentro das Extensões de Ocorrência para as espécies estudadas. A linha vermelha denota os limites do EOO. Fonte: autoria própria (2022).

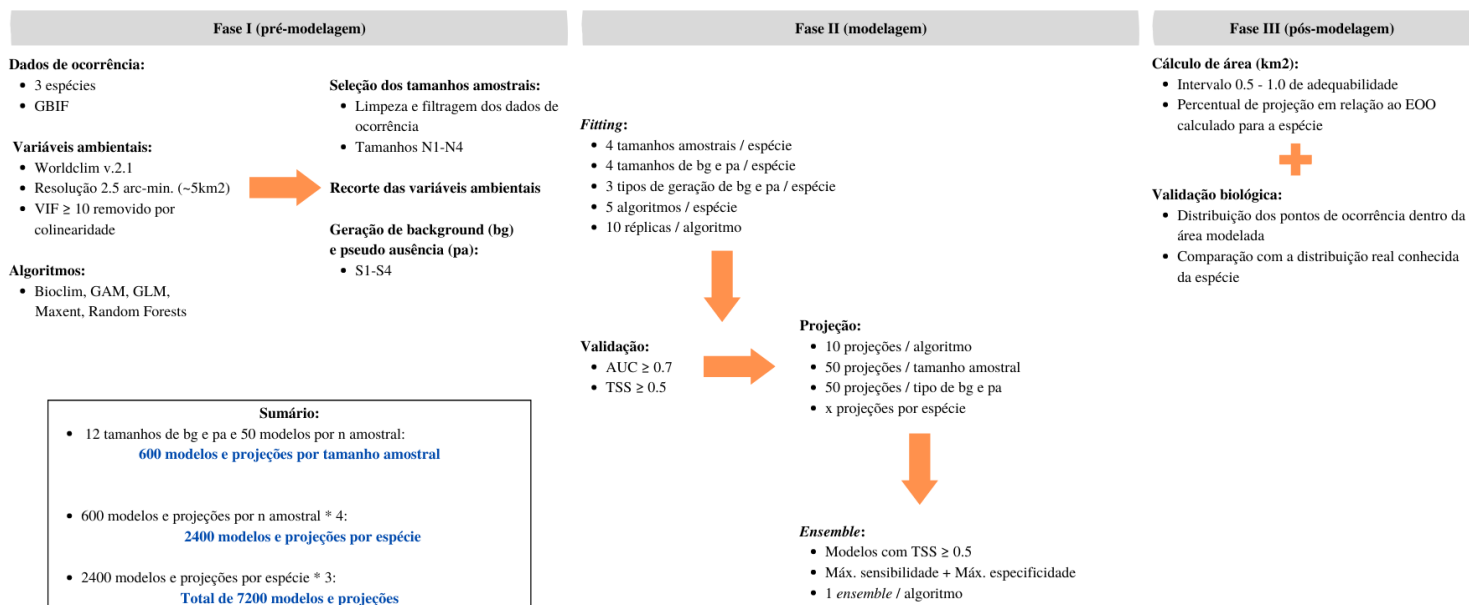


Figura S2 - Visão geral do processo de construção de um SDM / ENM, desde a obtenção dos dados biológicos e ambientais, calibração (i.e. *fitting*) do modelo e posterior avaliação. Fonte: autoria própria (2022).

Material Suplementar 2

THE EFFECT OF SAMPLE SIZE ON SPECIES DISTRIBUTION AND ECOLOGICAL NICHE MODELS: A CASE STUDY WITH THREE BRAZILIAN BIRDS

Gustavo Reis de Brito¹ Fernando Frei²

¹Programa de Pós-Graduação em Biociências, Faculdade de Ciências e Letras, São Paulo State University, Assis, São Paulo, Avenue Dom Antônio, 2100, 190806-900, Brazil

²Department of Biology, Faculdade de Ciências e Letras, São Paulo State University, Assis, São Paulo, Avenue Dom Antônio, 2100, 190806-900, Brazil

Corresponding author: Gustavo Reis de Brito (reis.brito@unesp.br)

Attention: to access the R scripts used to conduct all analysis, please send an email to reis.brito@unesp.br with the subject “R Code Request: sample size effect on SDM and ENM paper”.

Table S1. Papers sampled from a Web of Science review for the 2013 - 2021 period showing different sample sizes where SDM and ENM approaches were applied to study biodiversity

Author(s)	Year	Sample Size
Shcheglovitova & Anderson	2013	6 and 9
Carvalho & Del Lama	2015	7 to 103
Spiers et al.	2018	5 to 10
Coello et al.	2021	60
Tran & Vu	2020	36
Botella et al.	2018	120 to 14278
Franco & Camps	2020	85
Jurestovsky & Andrew Joyner	2018	87
Flessner et al.	2017	102
Call et al.	2016	1578
Hemami et al.	2020	280

DeVaney	2016	719
---------	------	-----

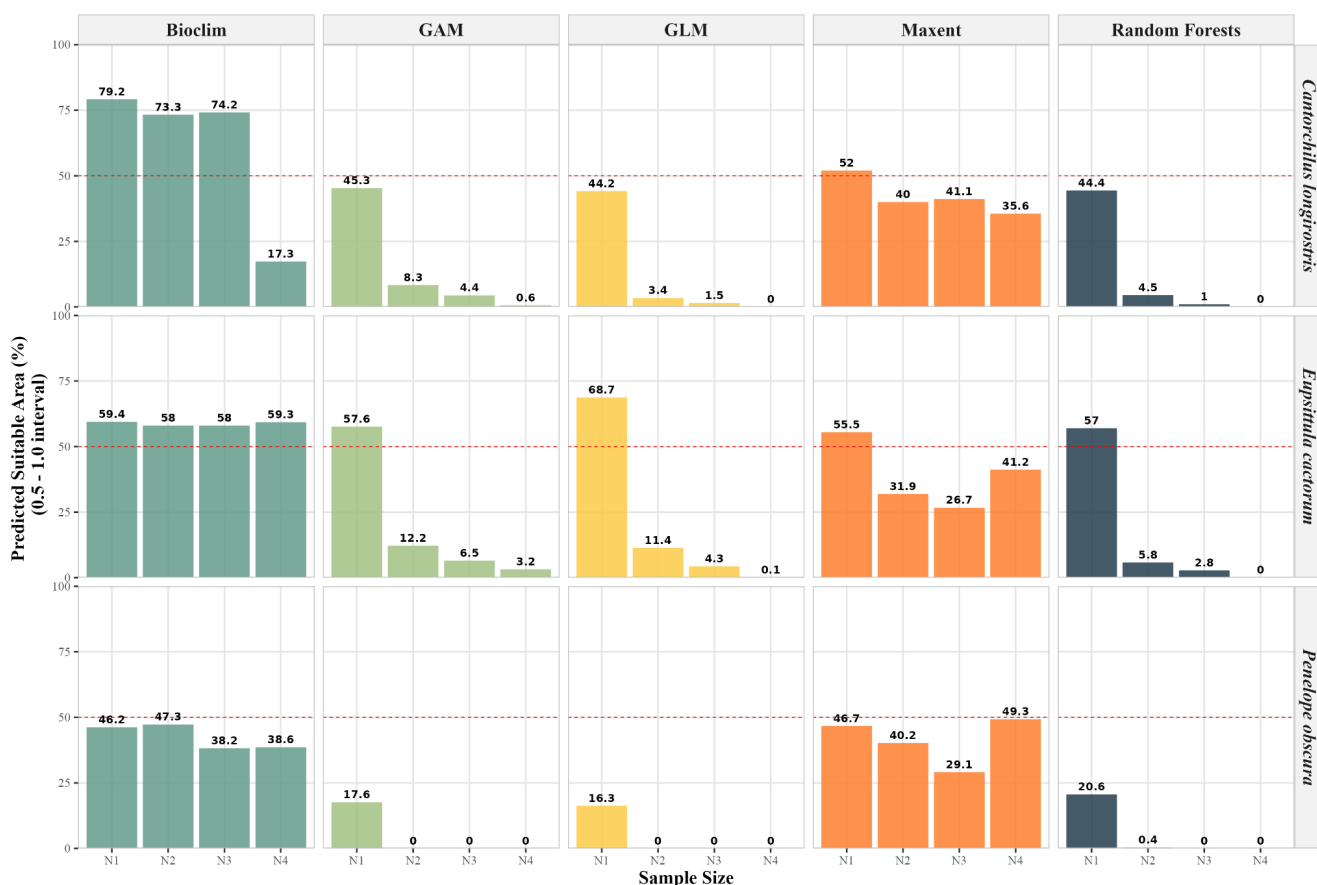


Figure S1. Percentual of predicted suitable area for all three species, considering the 0.5 - 1.0 threshold of suitability (red line). All models were fitted using the first set of pseudo absence/background points ($n = 1000$) and projections were constrained by the EOO of the species.

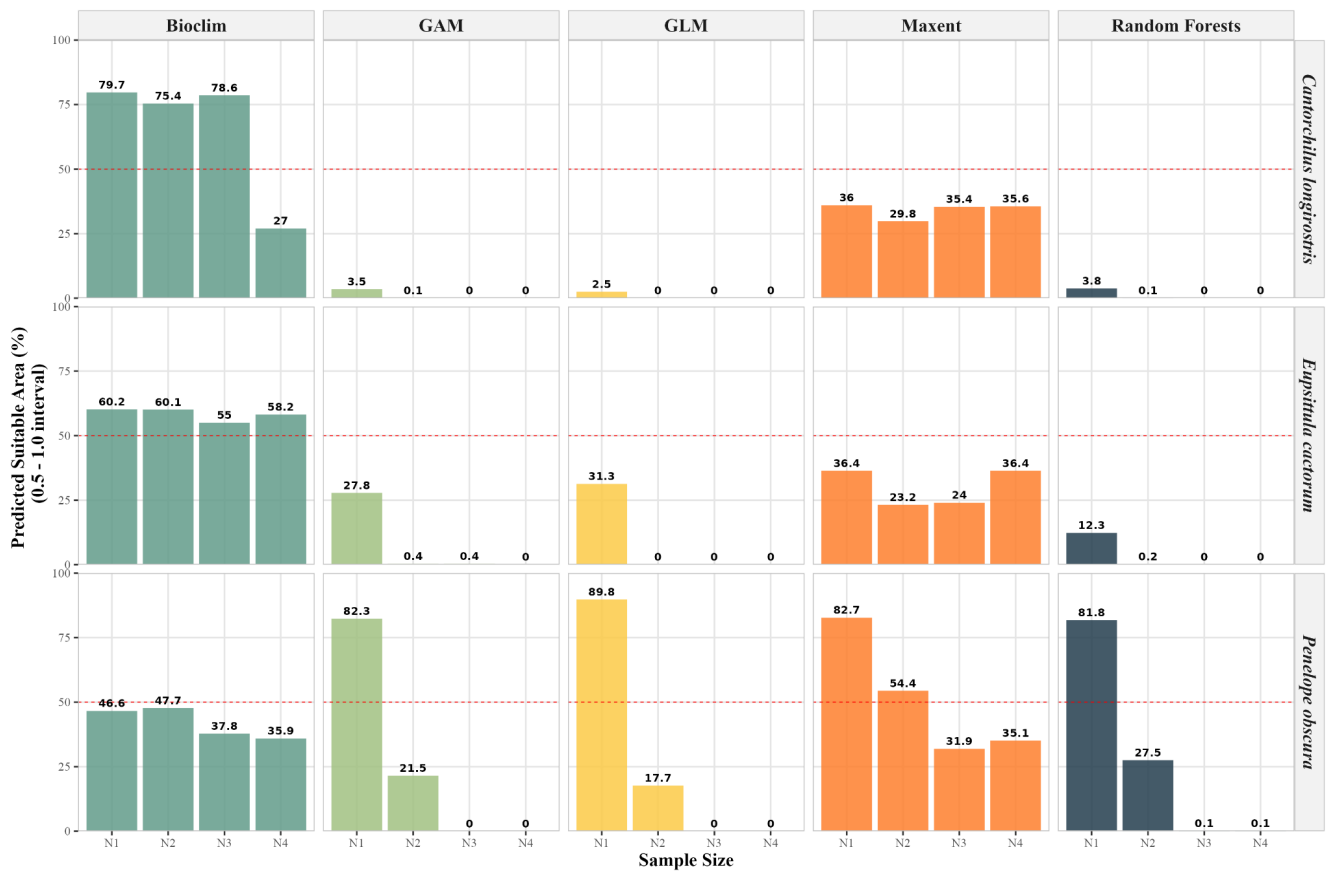


Figure S2. Percentual of predicted suitable area for all three species, considering the 0.5 - 1.0 threshold of suitability (red line). All models were fitted using the first set of pseudo absence/background points ($n = 10000$) and projections were constrained by the EOO of the species.

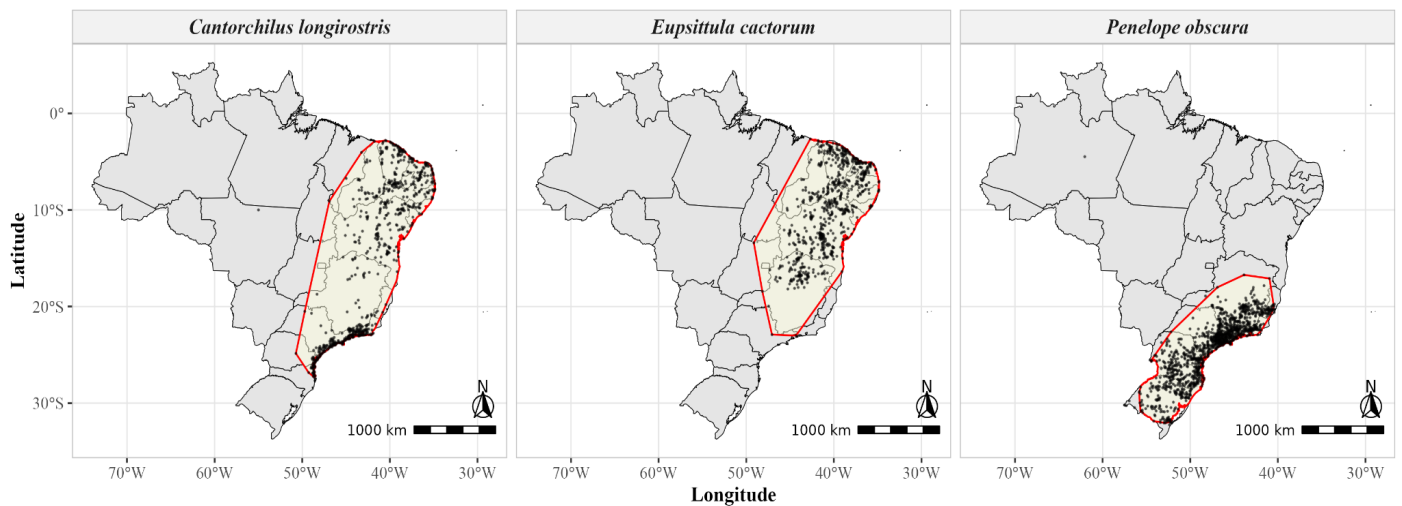


Figure S3. Extent of Occurrence calculated for all three species (red line). Black dots represent samples used in model fitting.

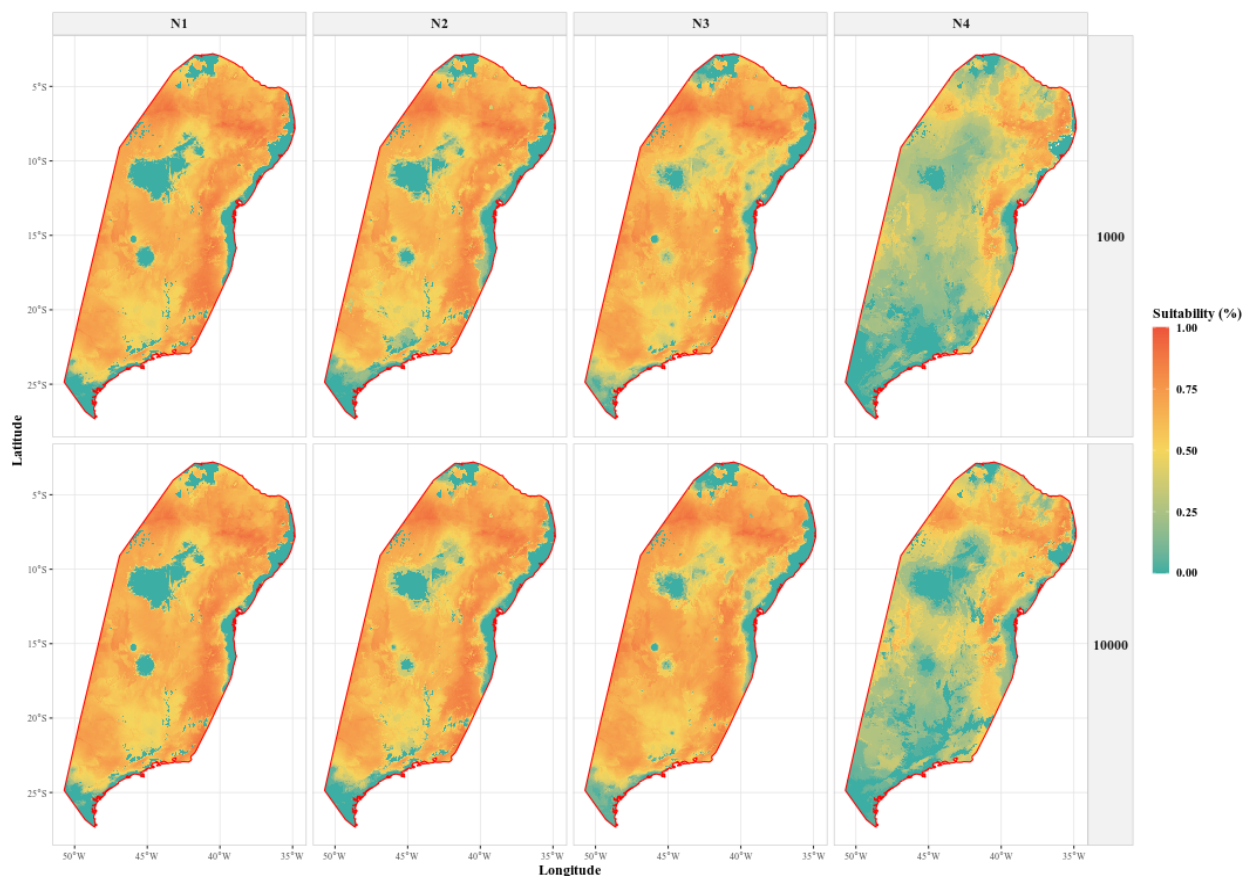


Figure S4. Full suitability Bioclim projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

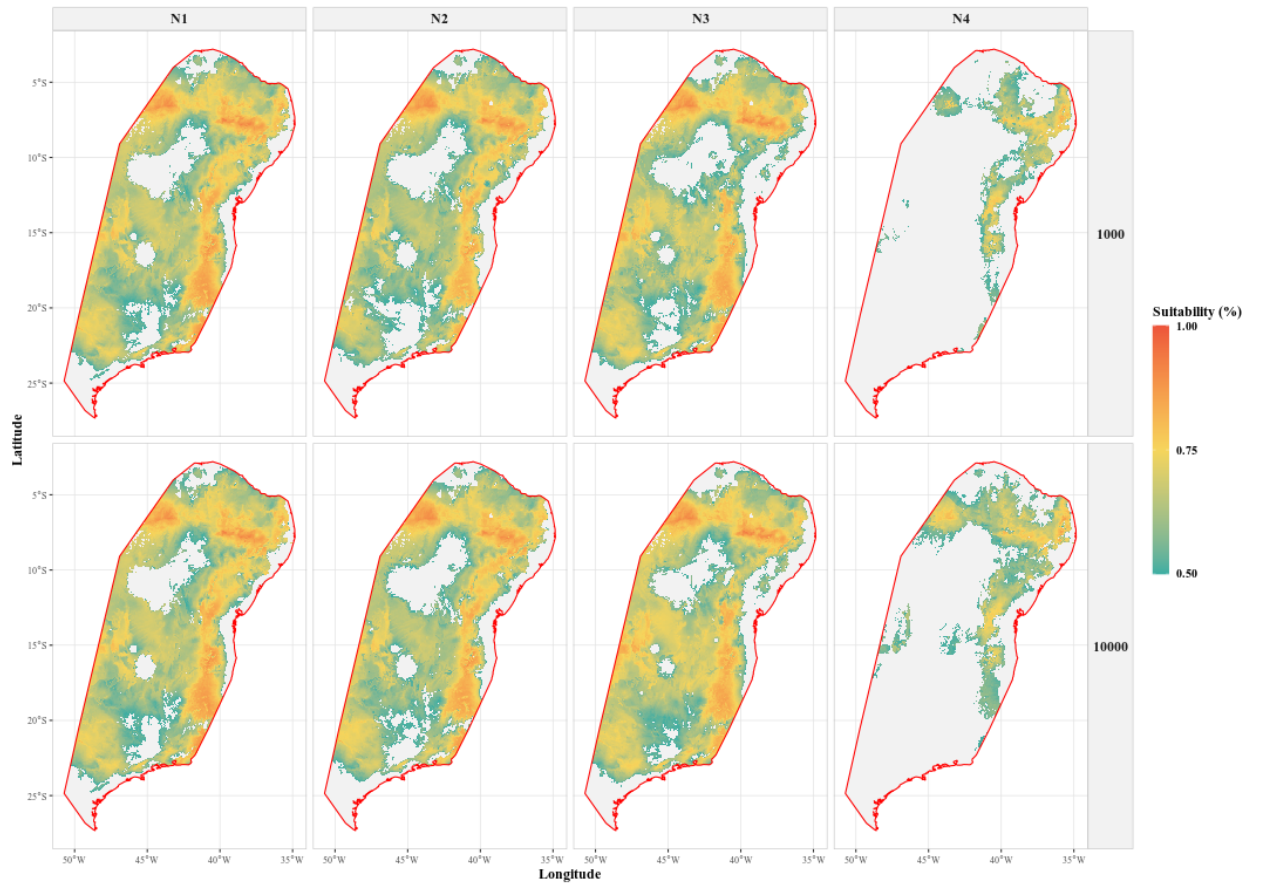


Figure S5. Suitability interval (0.5-1) Bioclim projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

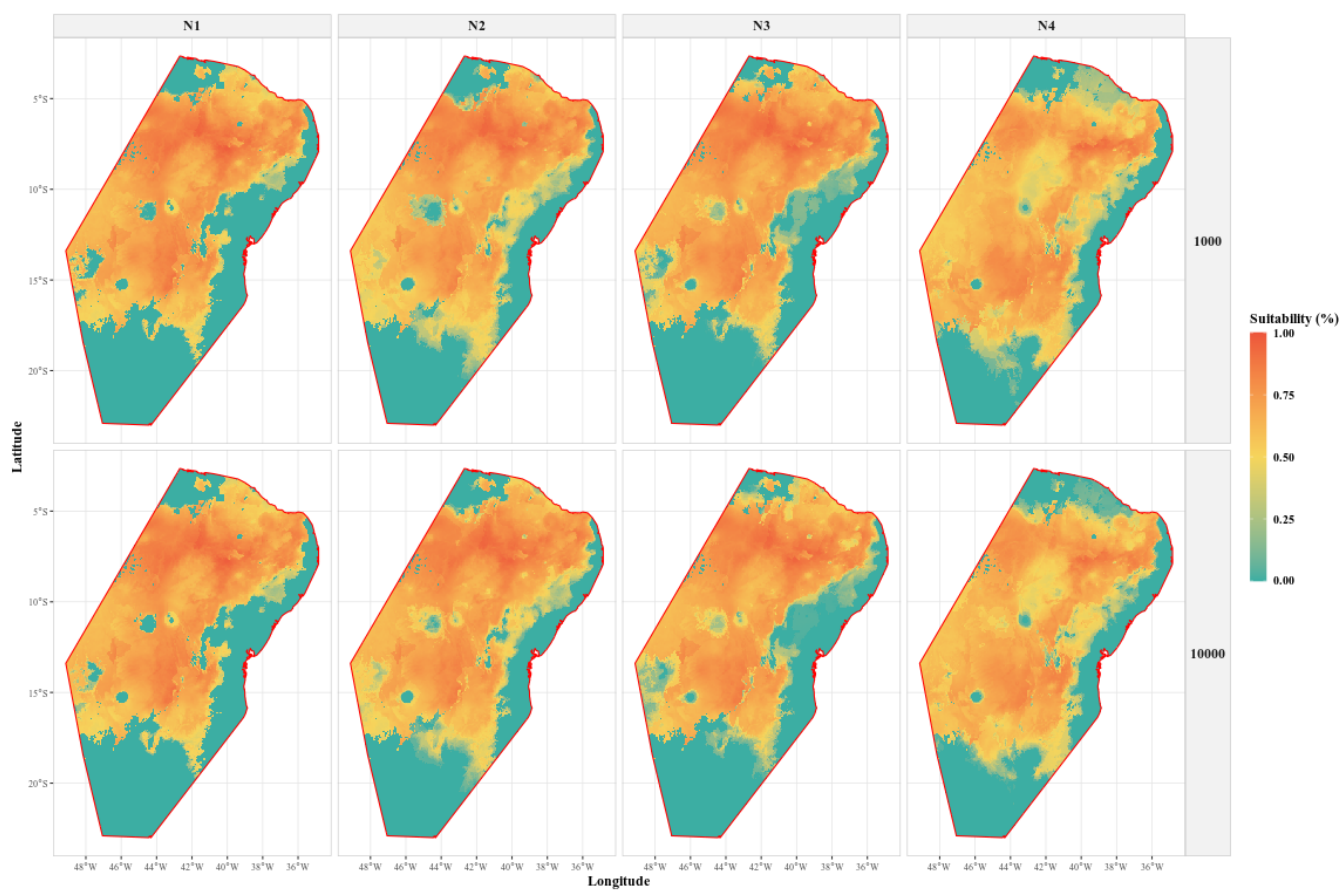


Figure S6. Full suitability Bioclim projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

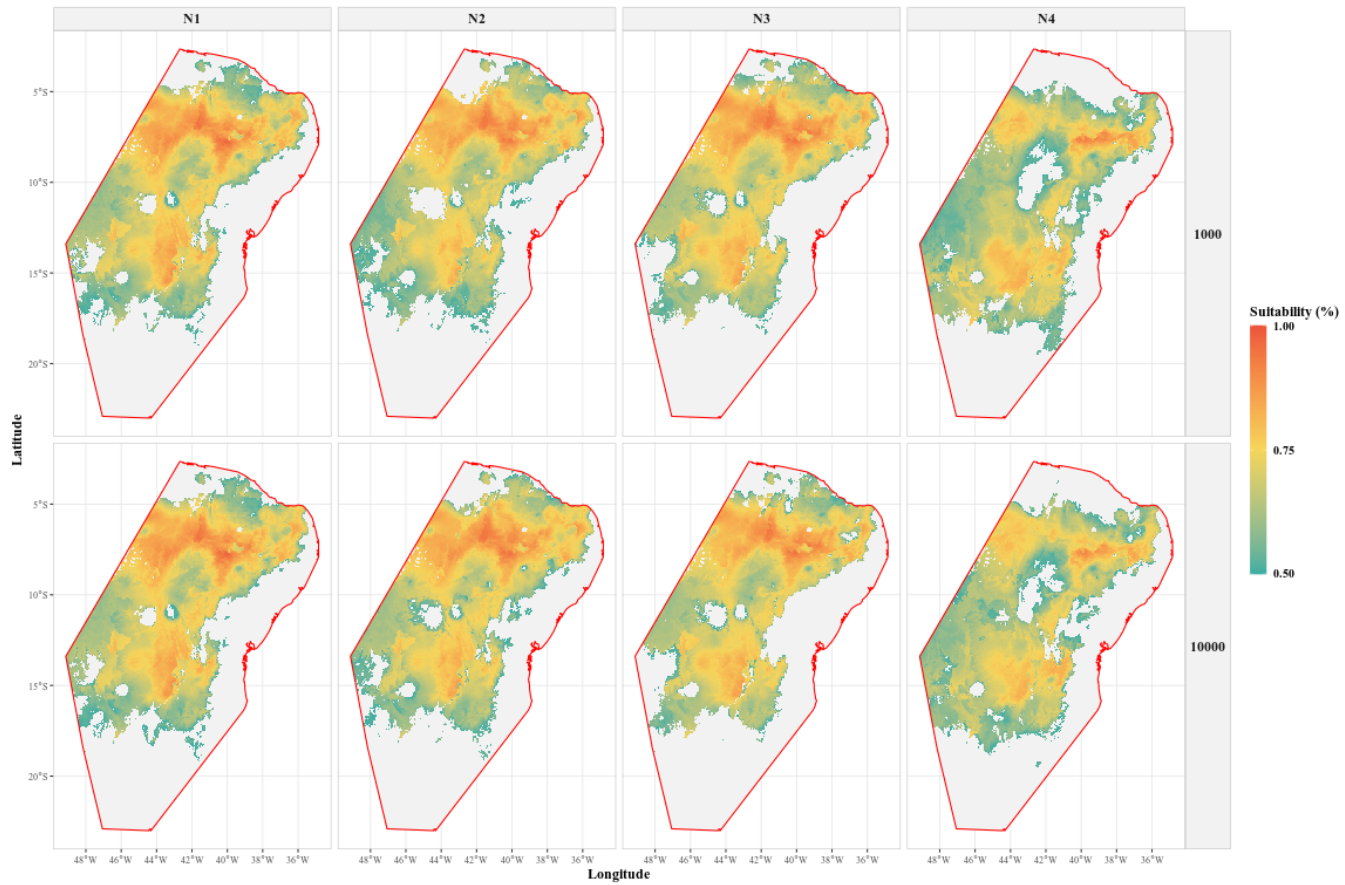


Figure S7. Suitability interval (0.5-1) Bioclim projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

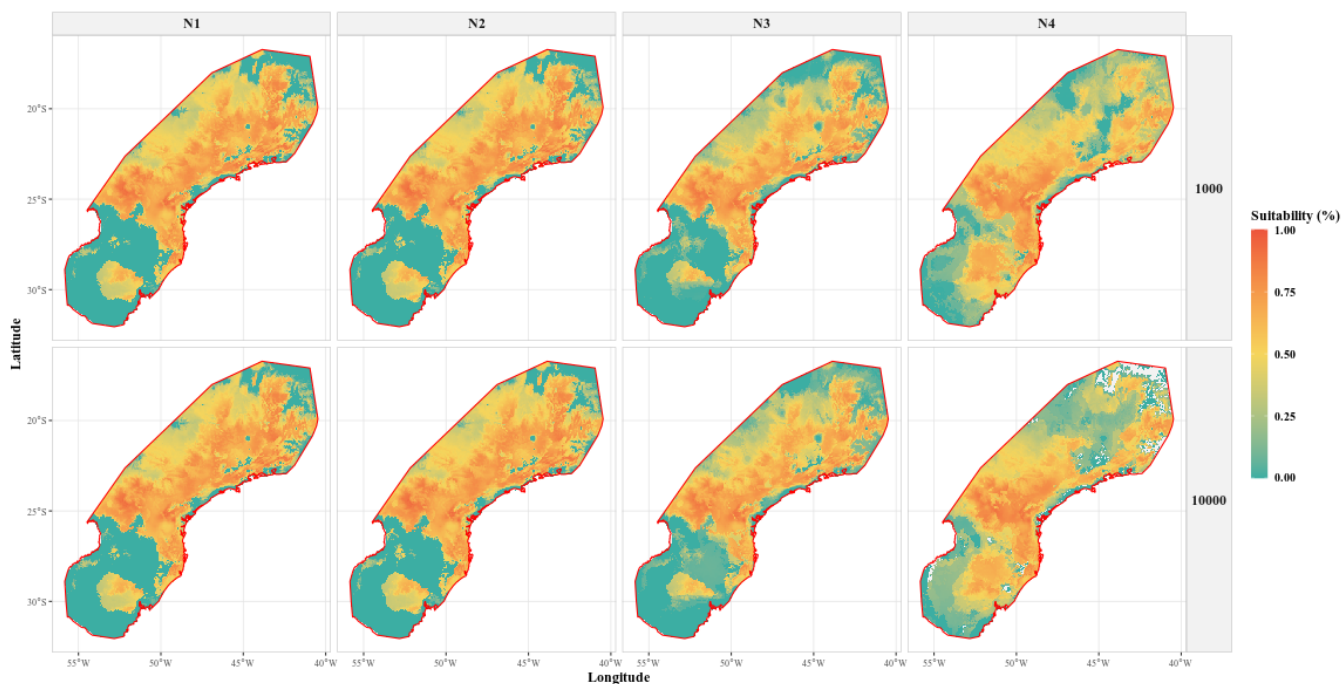


Figure S8. Full suitability Bioclim projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

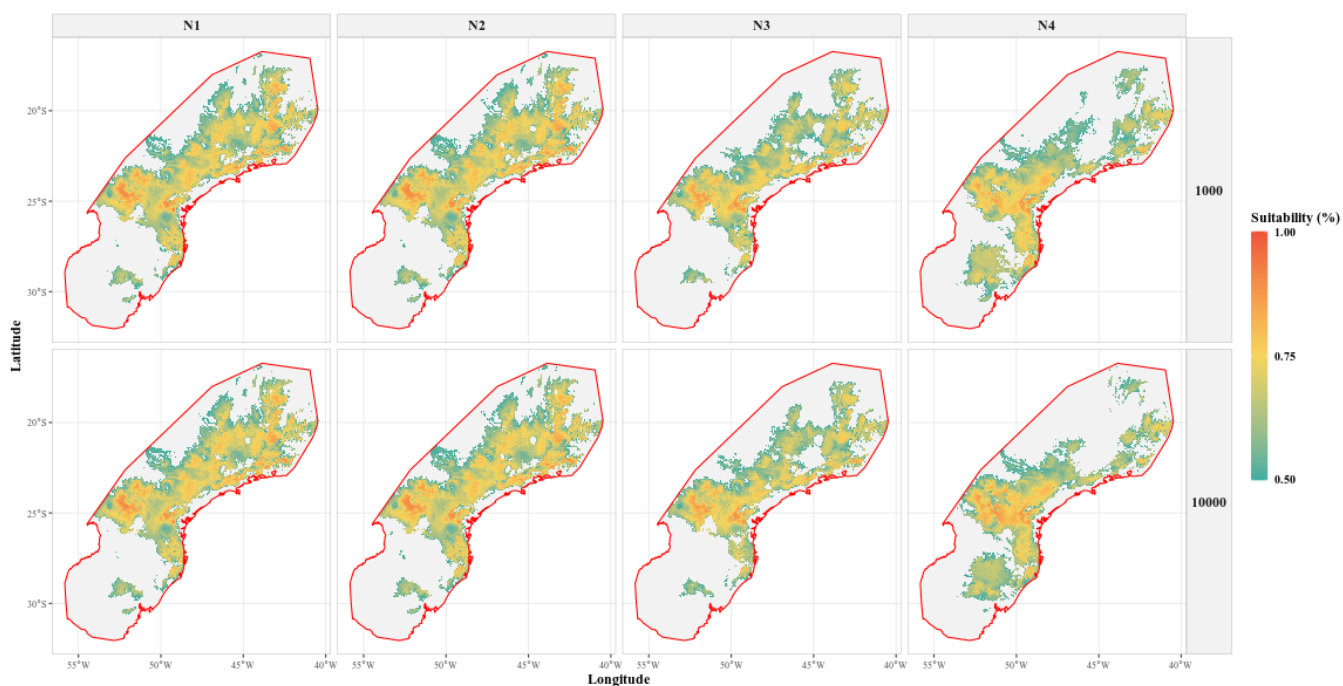


Figure S9. Suitability interval (0.5-1) Bioclim projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

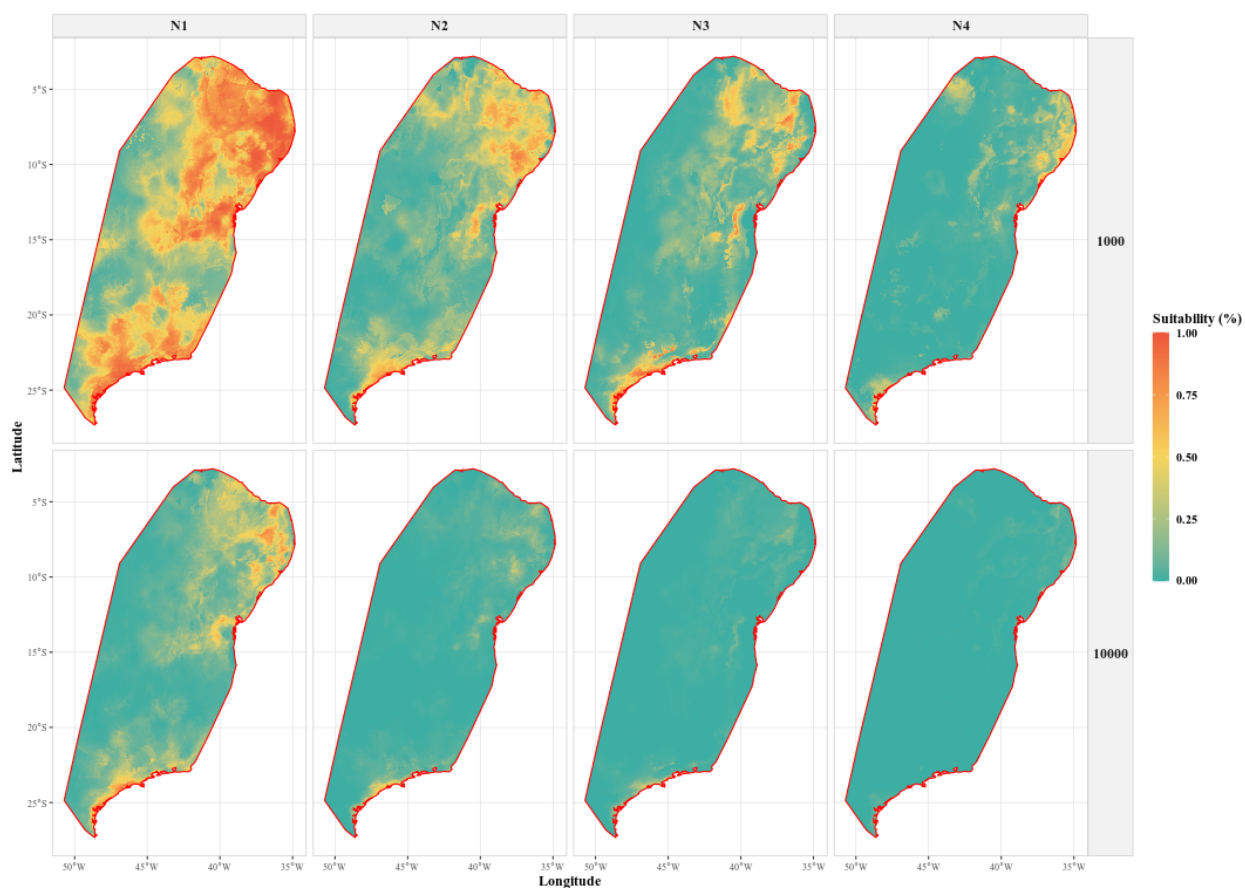


Figure S10. Full suitability GAM projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

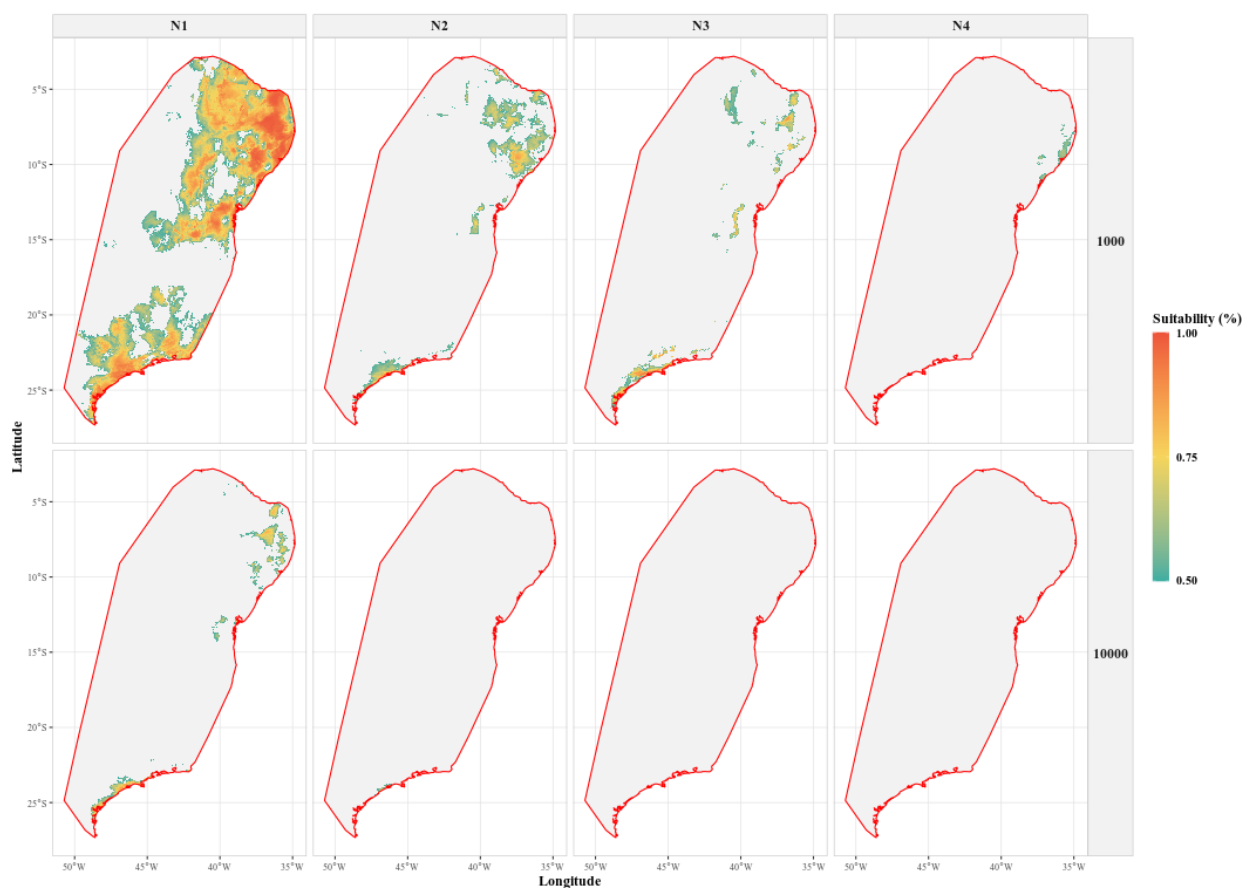


Figure S11. Suitability interval (0.5-1) GAM projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

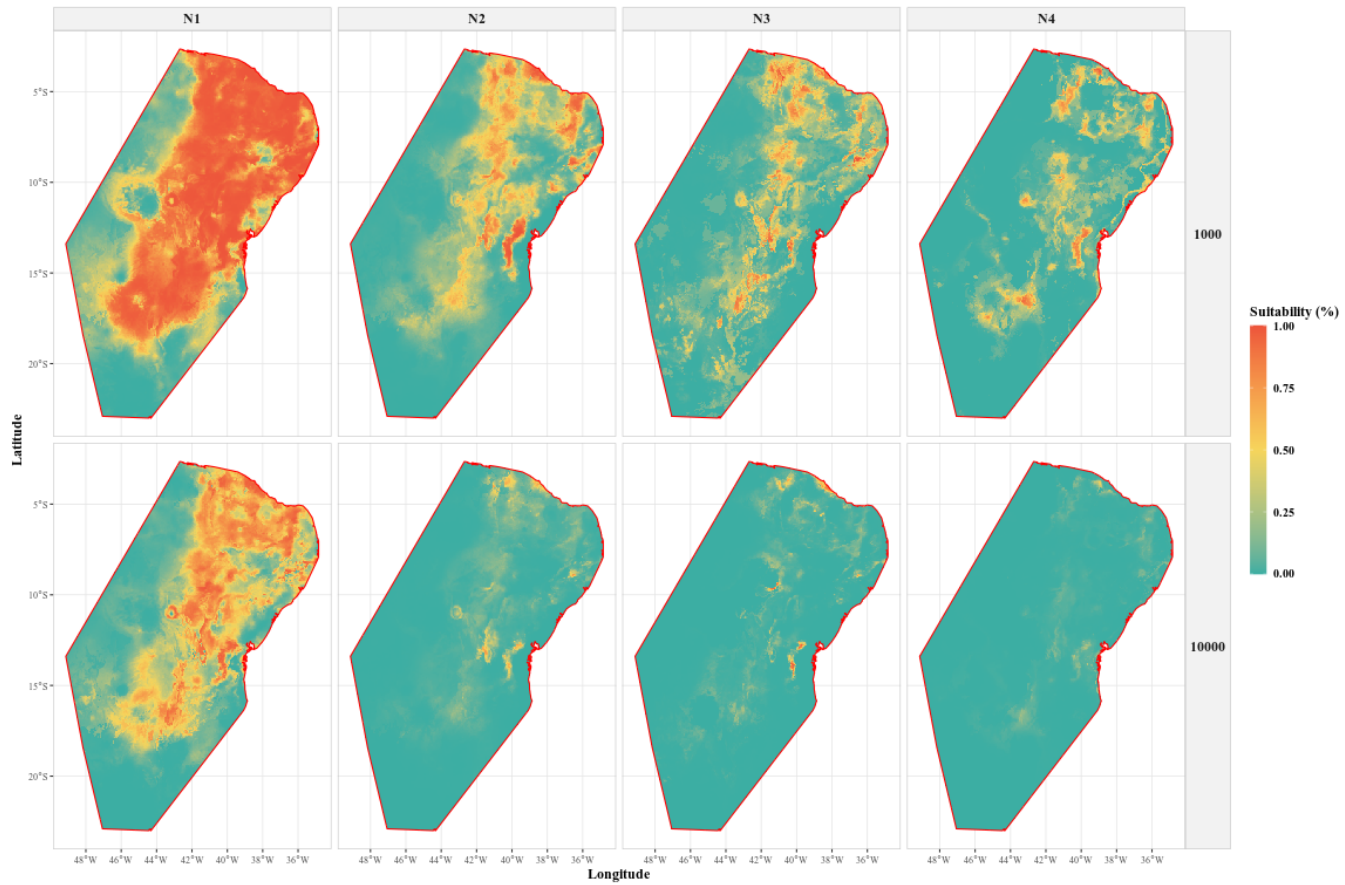


Figure S12. Full suitability GAM projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

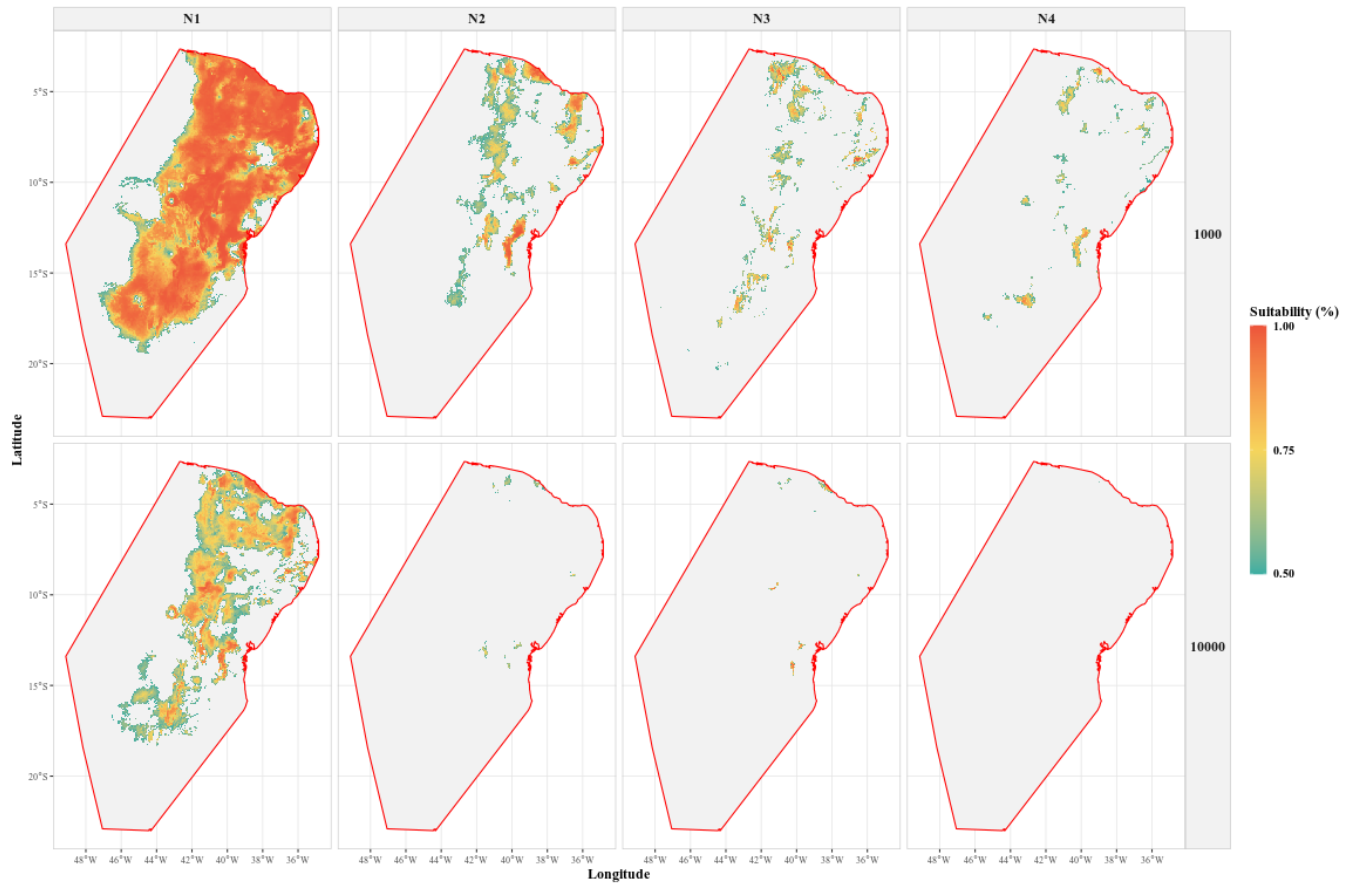


Figure S13. Suitability interval (0.5-1) GAM projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

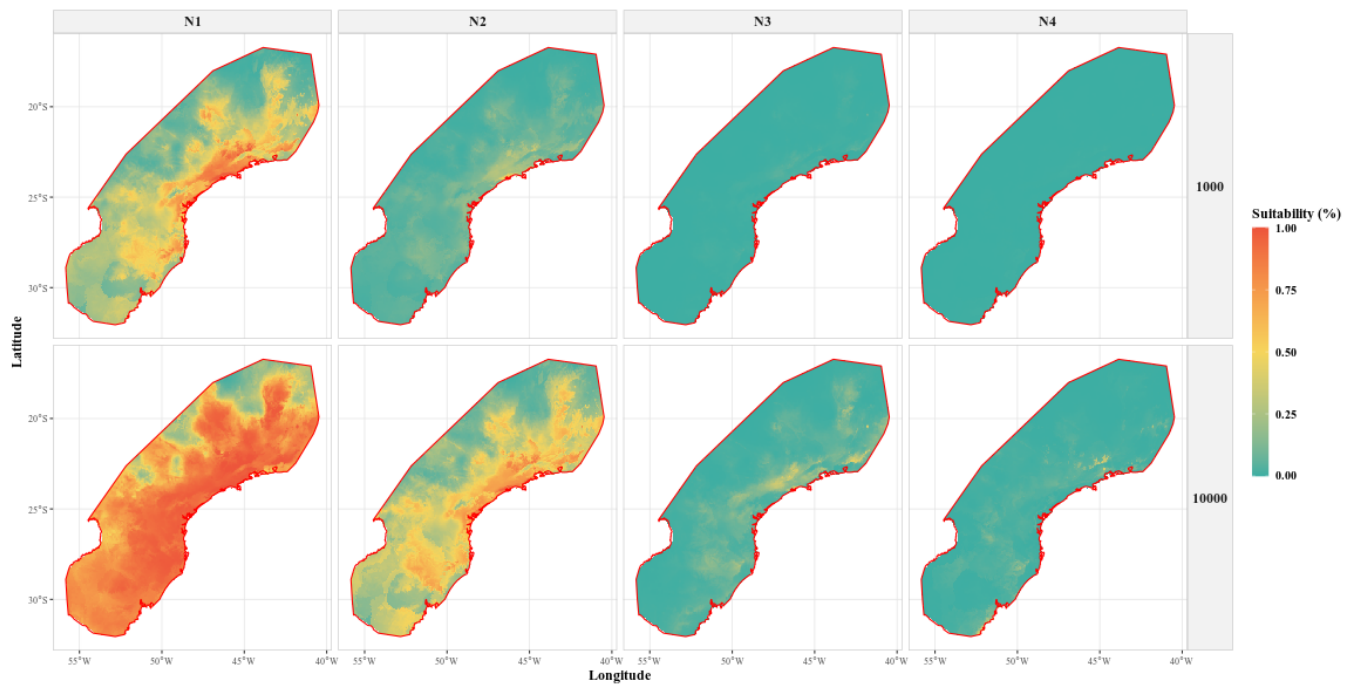


Figure S14. Full suitability GAM projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

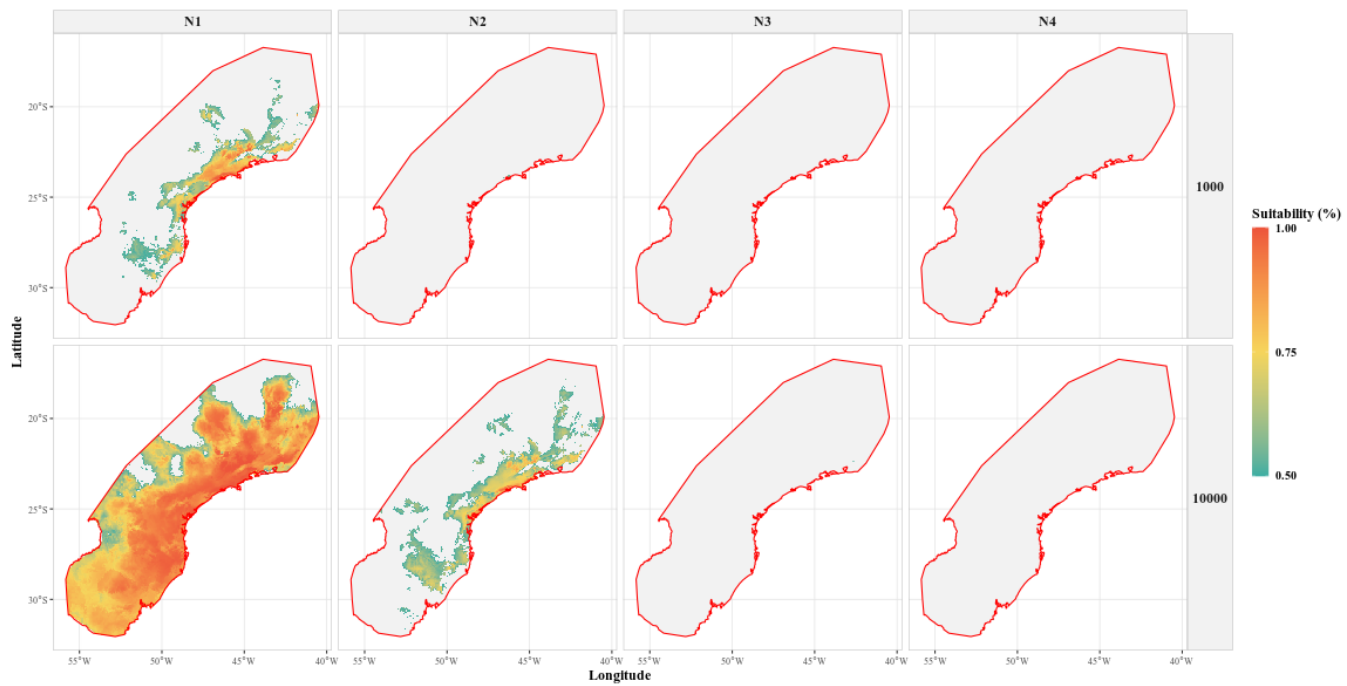


Figure S15. Suitability interval (0.5-1) GAM projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

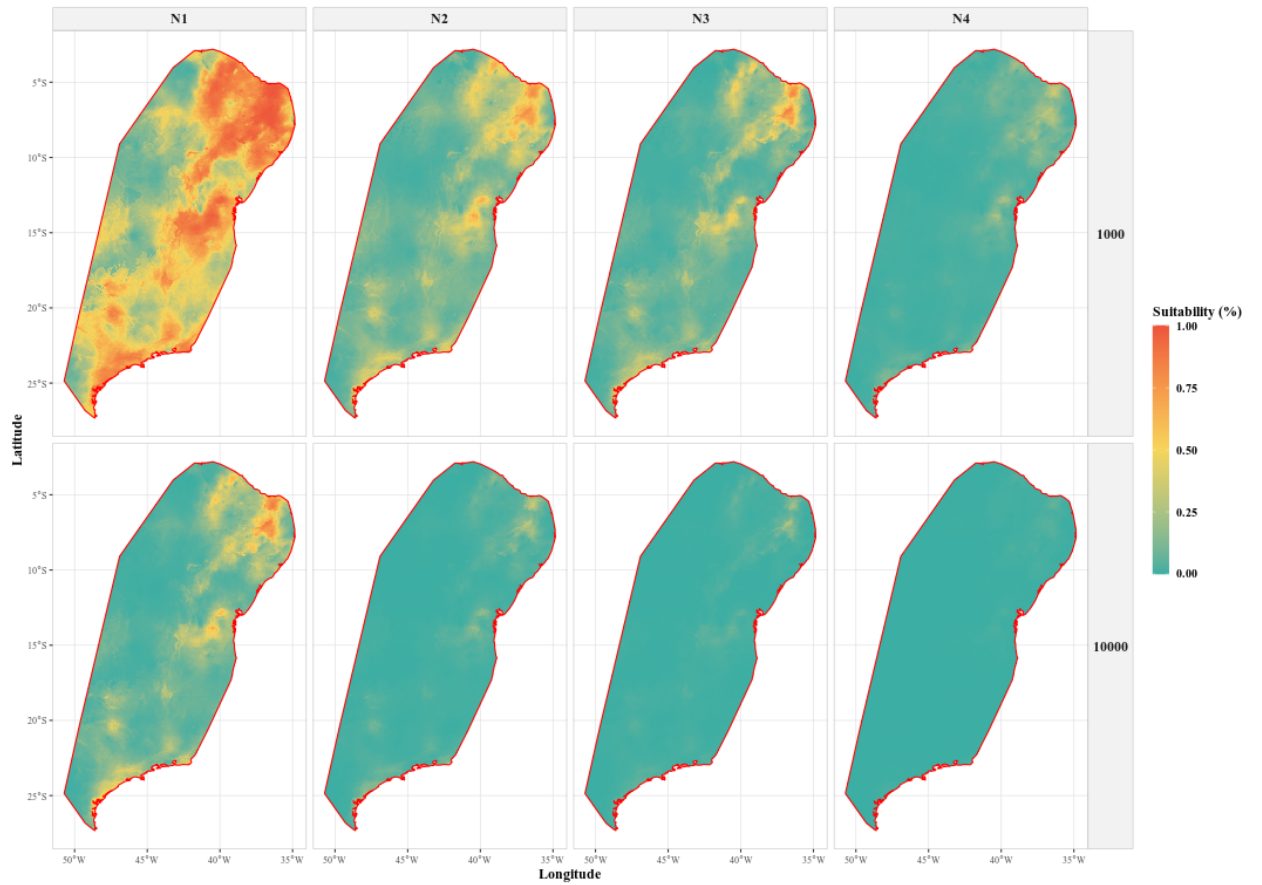


Figure S16. Full suitability GLM projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

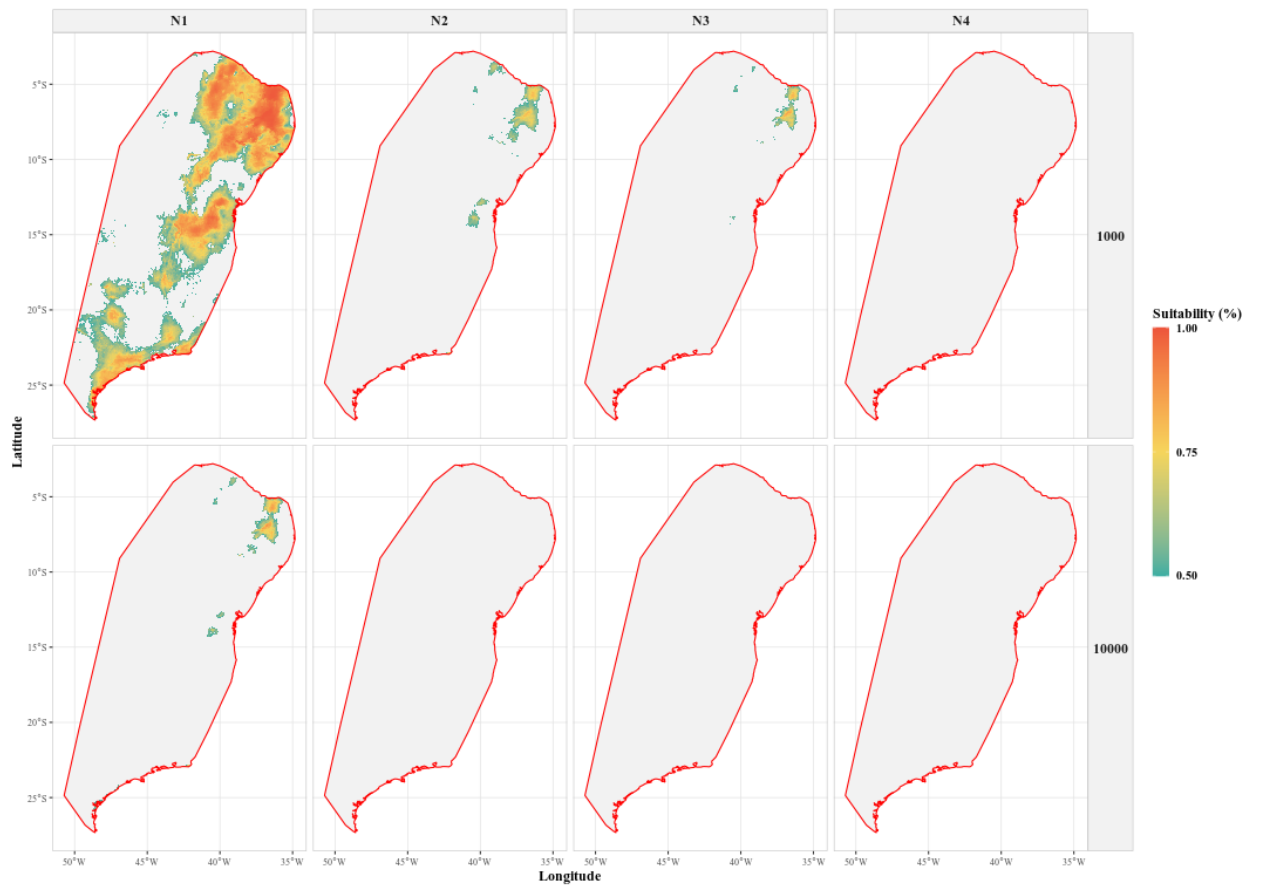


Figure S17. Suitability interval (0.5-1) GLM projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

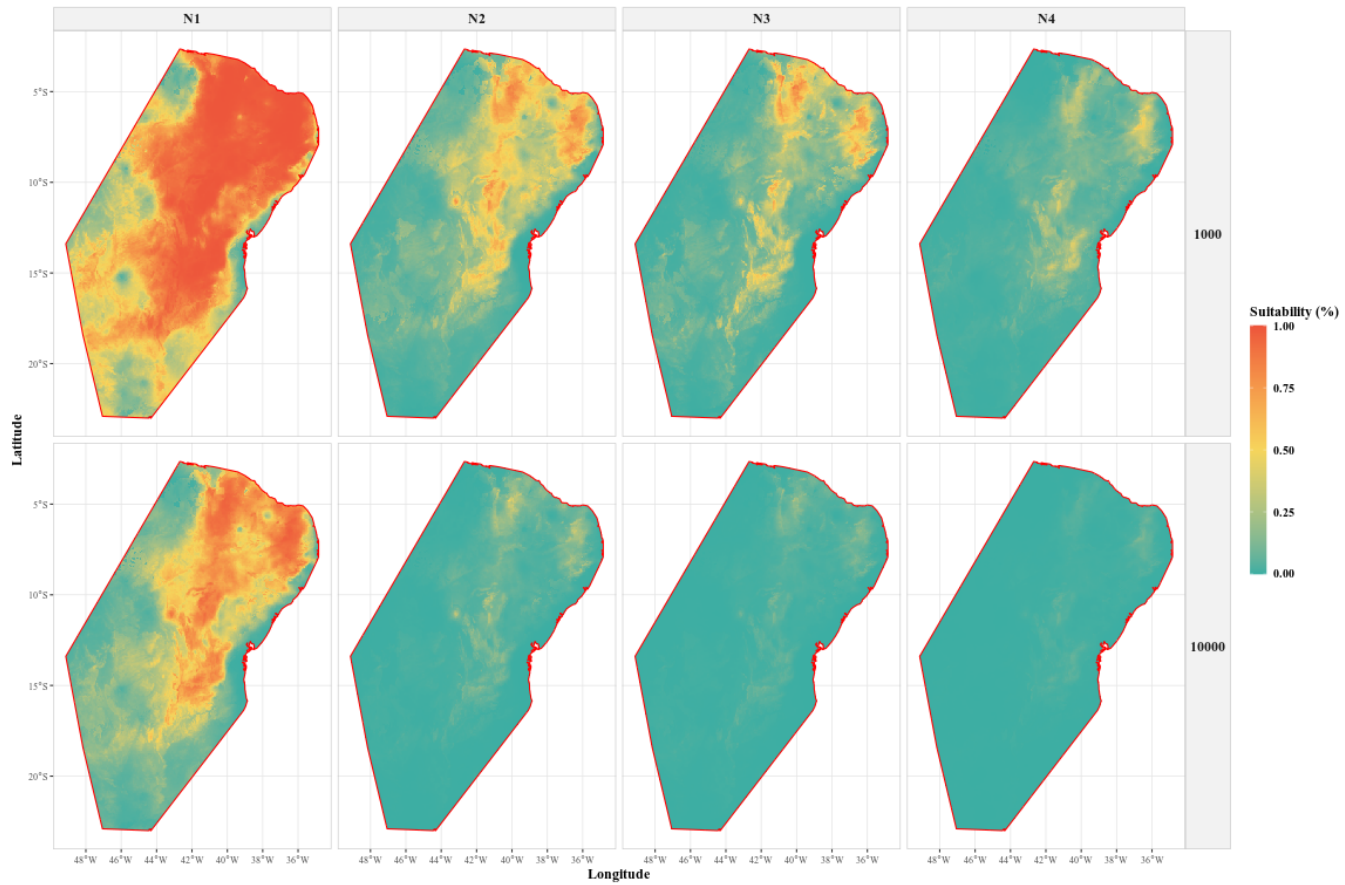


Figure S18. Full suitability GLM projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

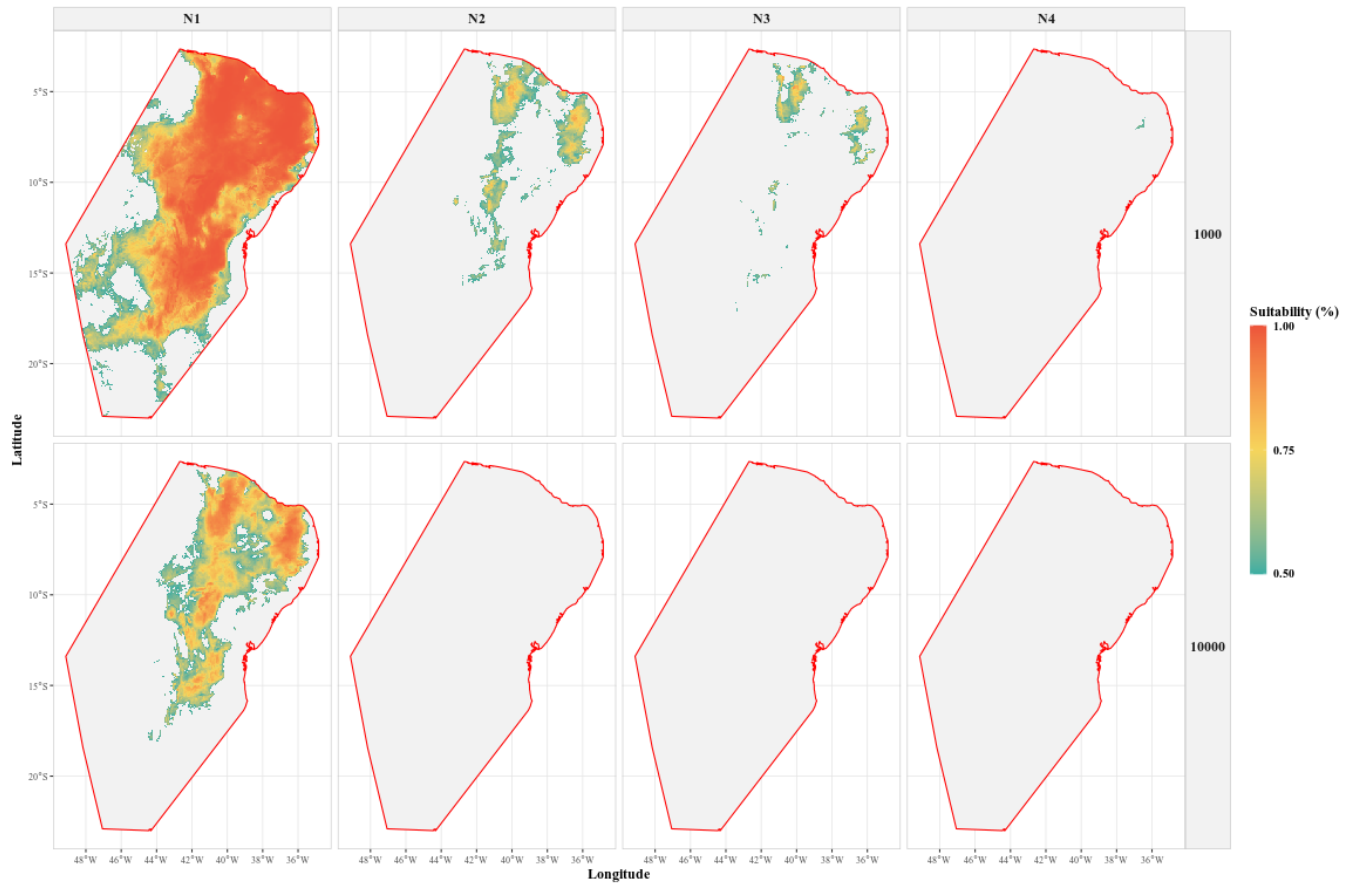


Figure S19. Suitability interval (0.5-1) GLM projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

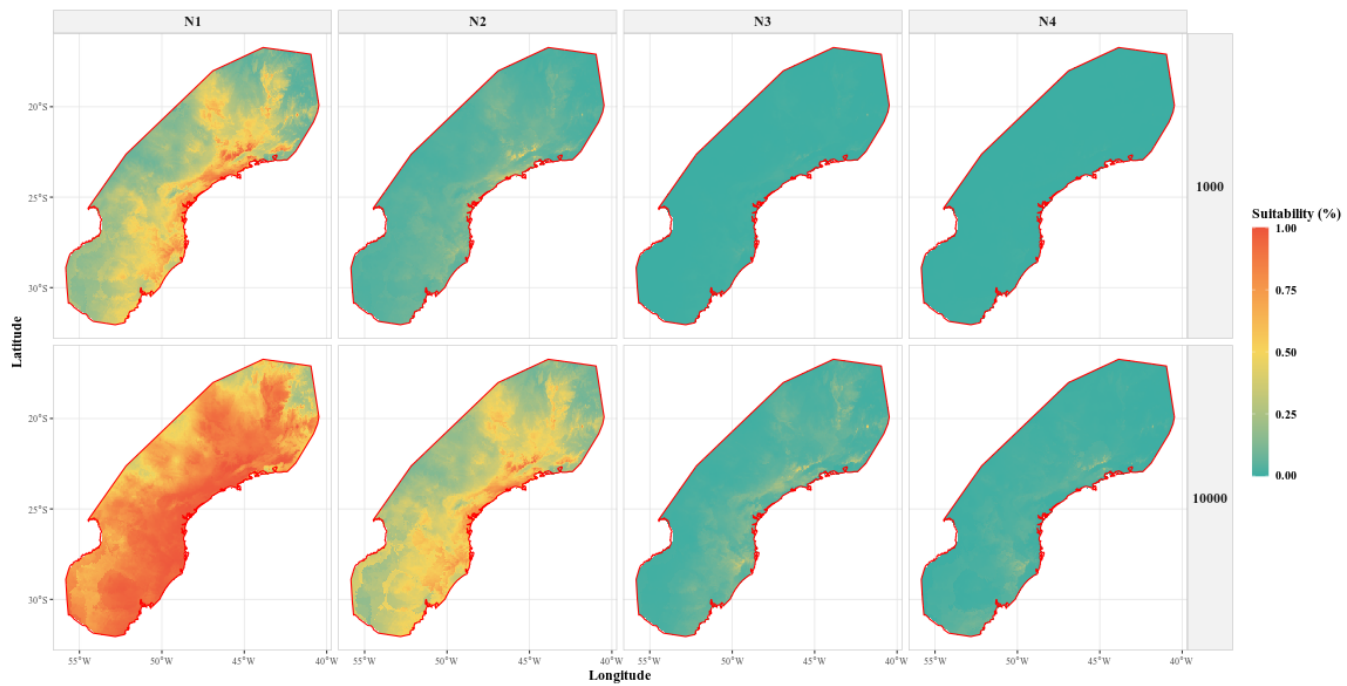


Figure S20. Full suitability GLM projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

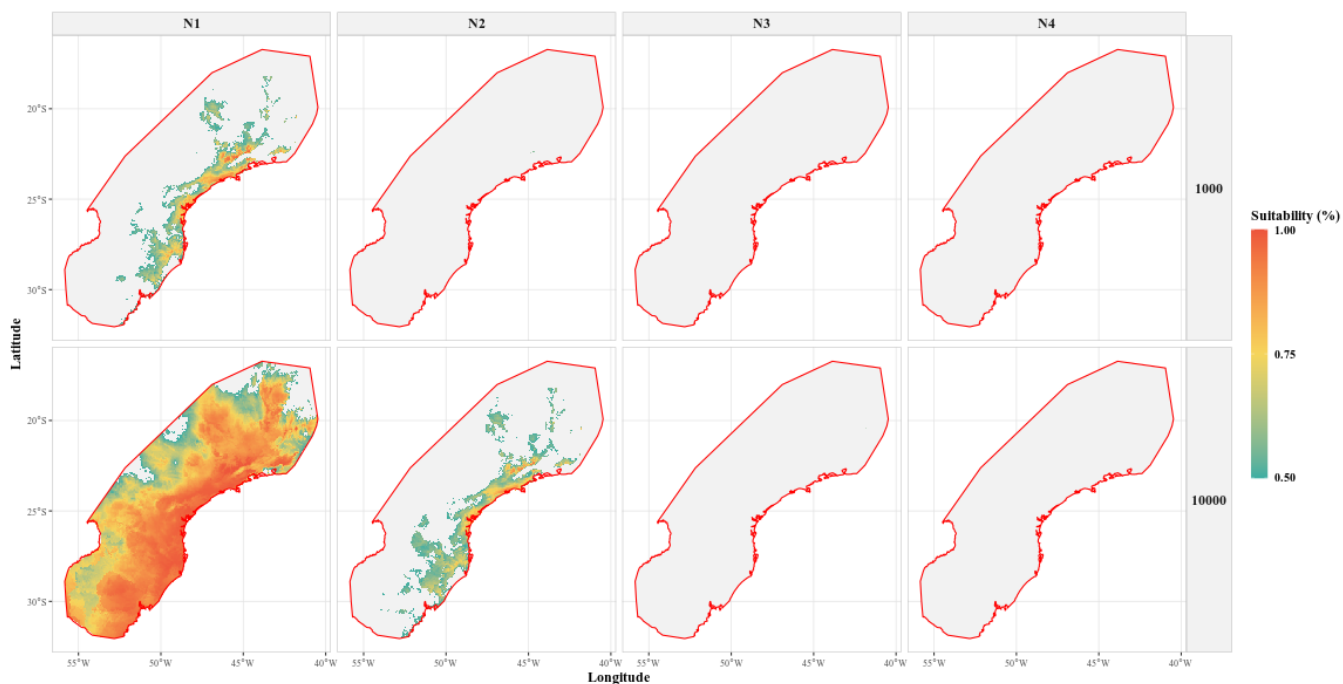


Figure S21. Suitability interval (0.5-1) GLM projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

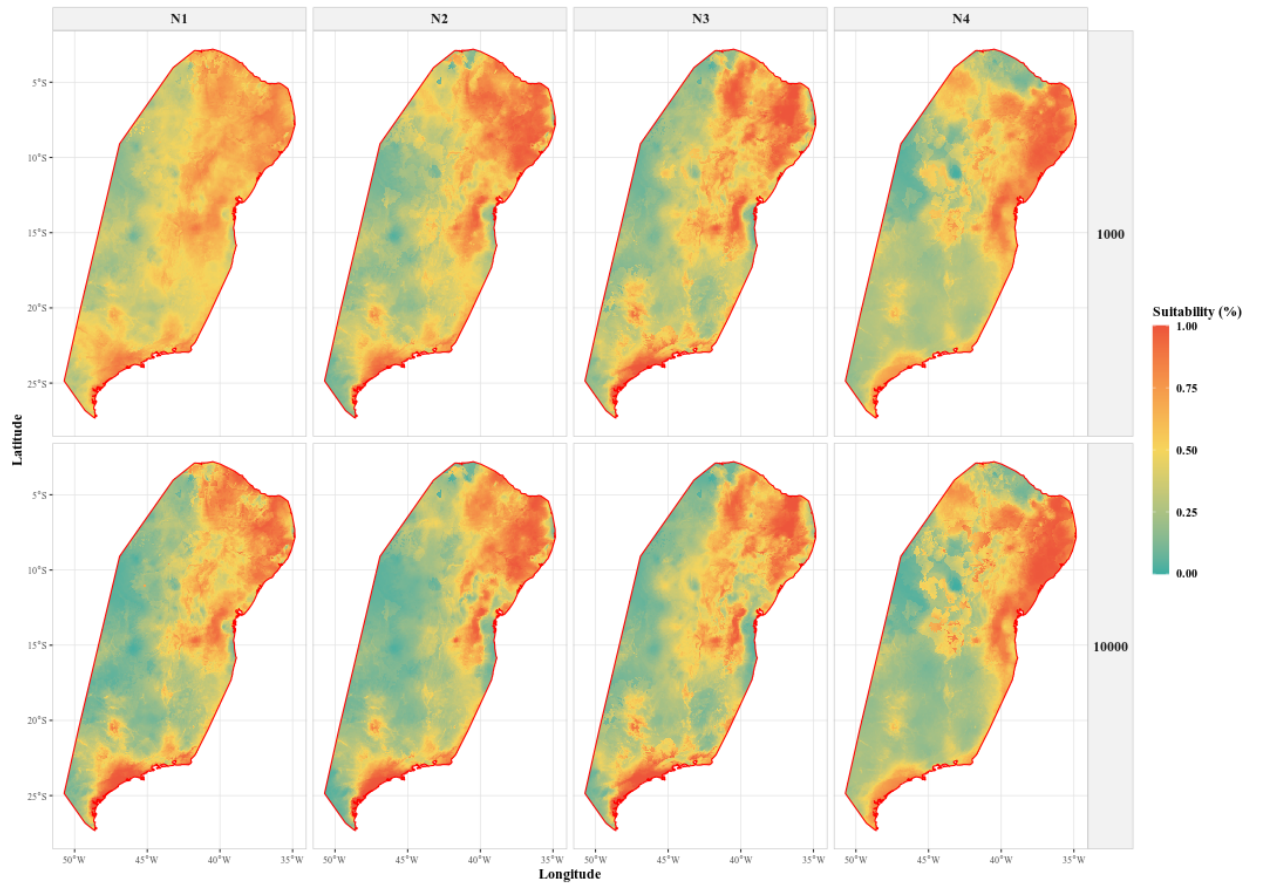


Figure S22. Full suitability Maxent projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

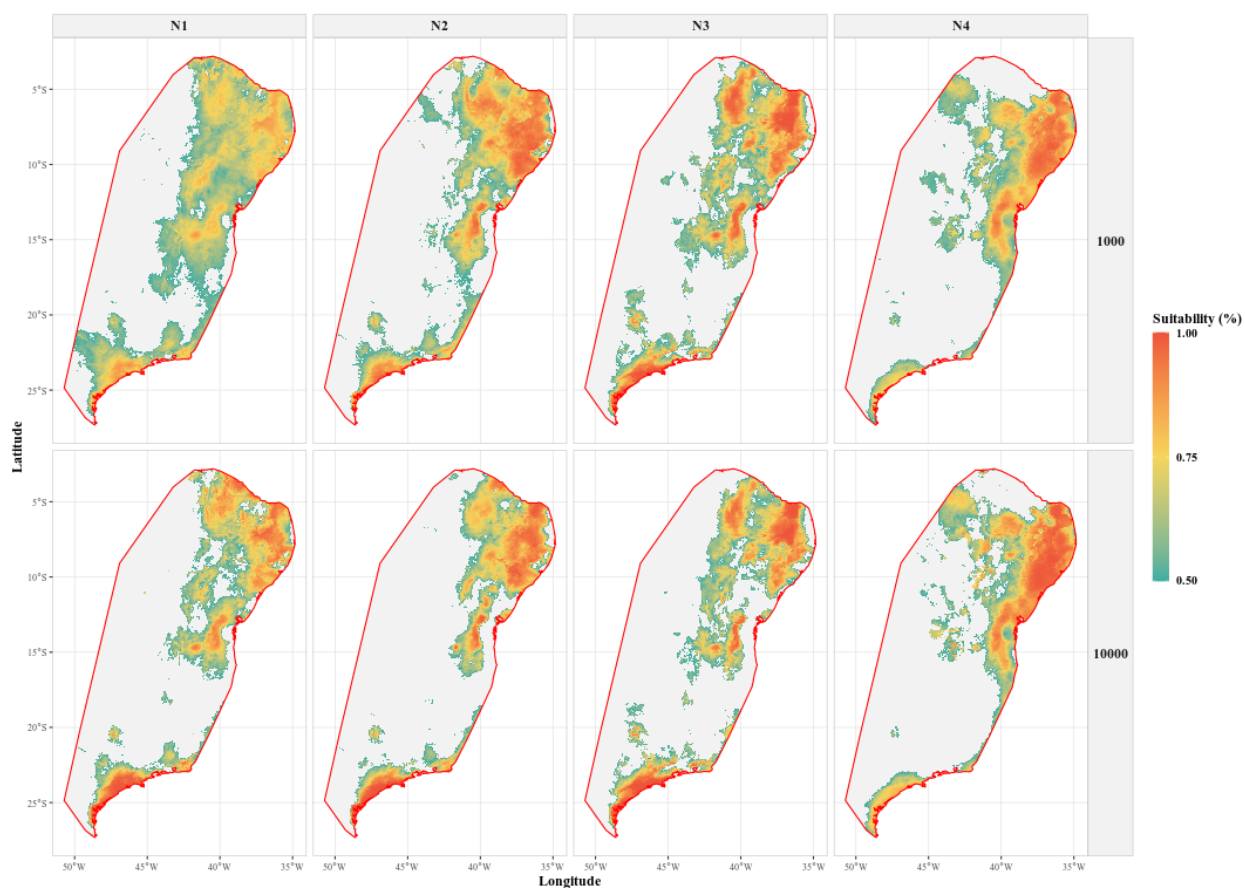


Figure S23. Suitability interval (0.5-1) Maxent projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

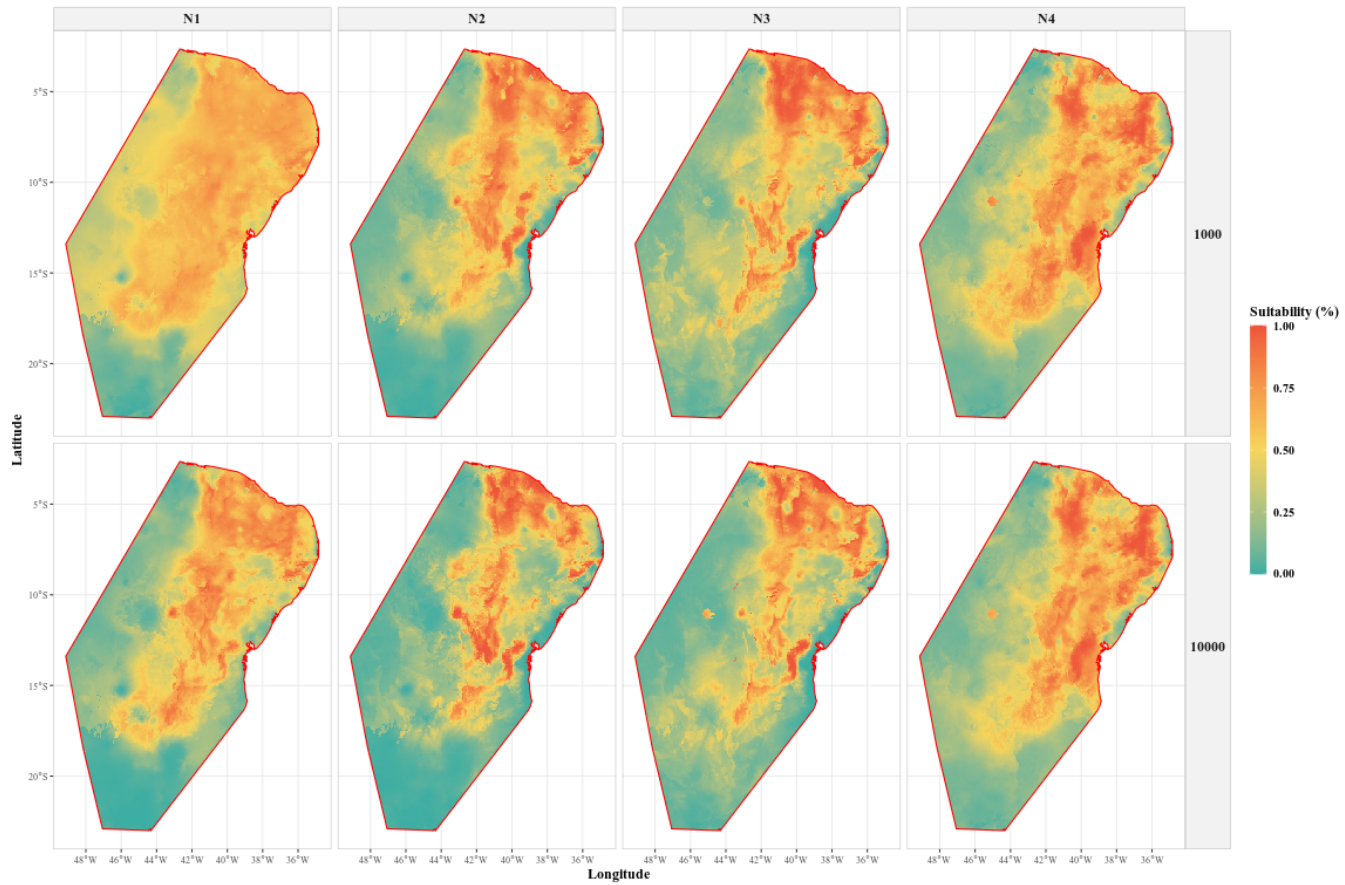


Figure S24. Full suitability Maxent projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

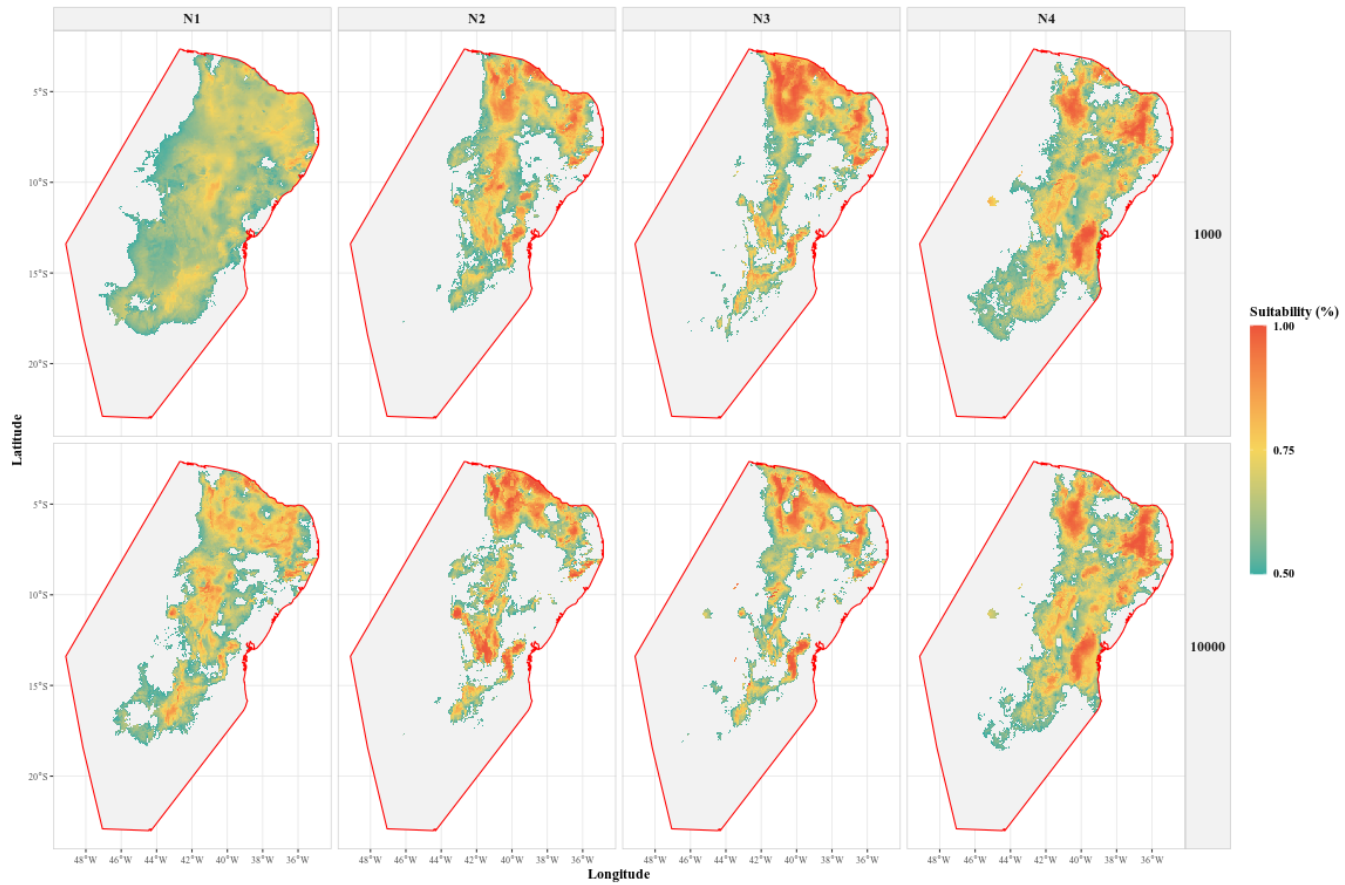


Figure S25. Suitability interval (0.5-1) Maxent projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

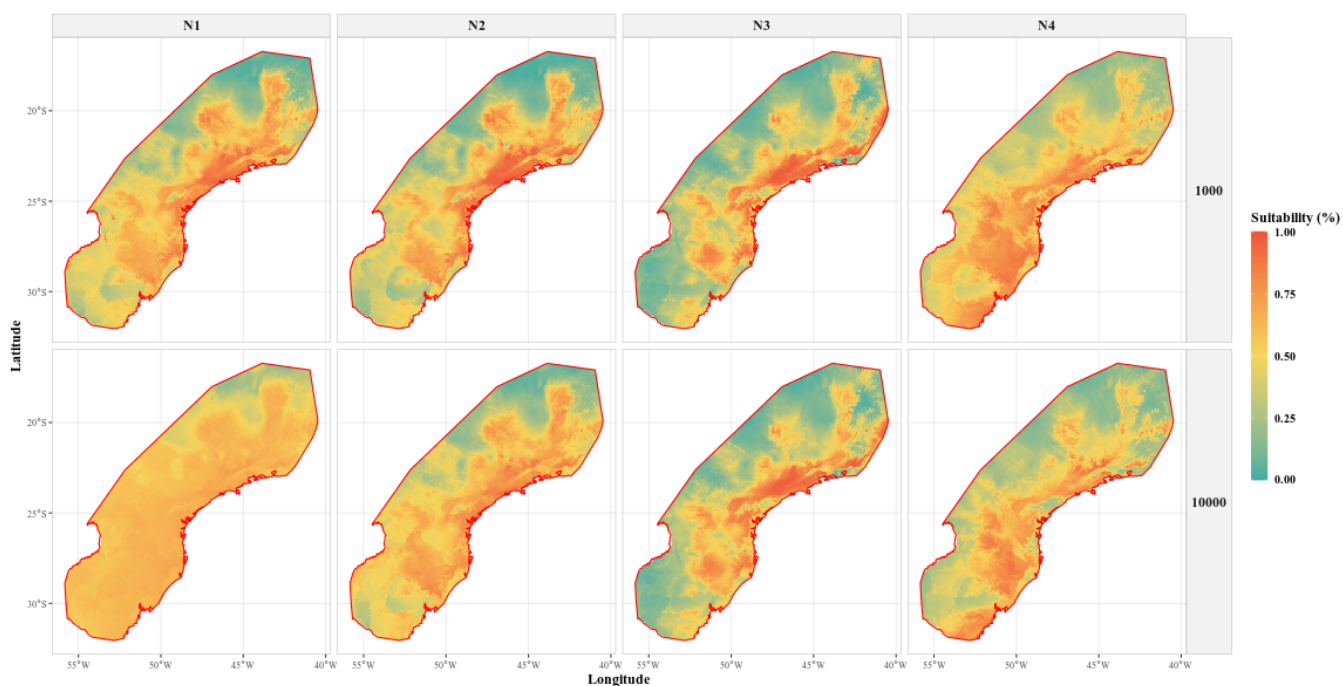


Figure S26. Full suitability Maxent projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

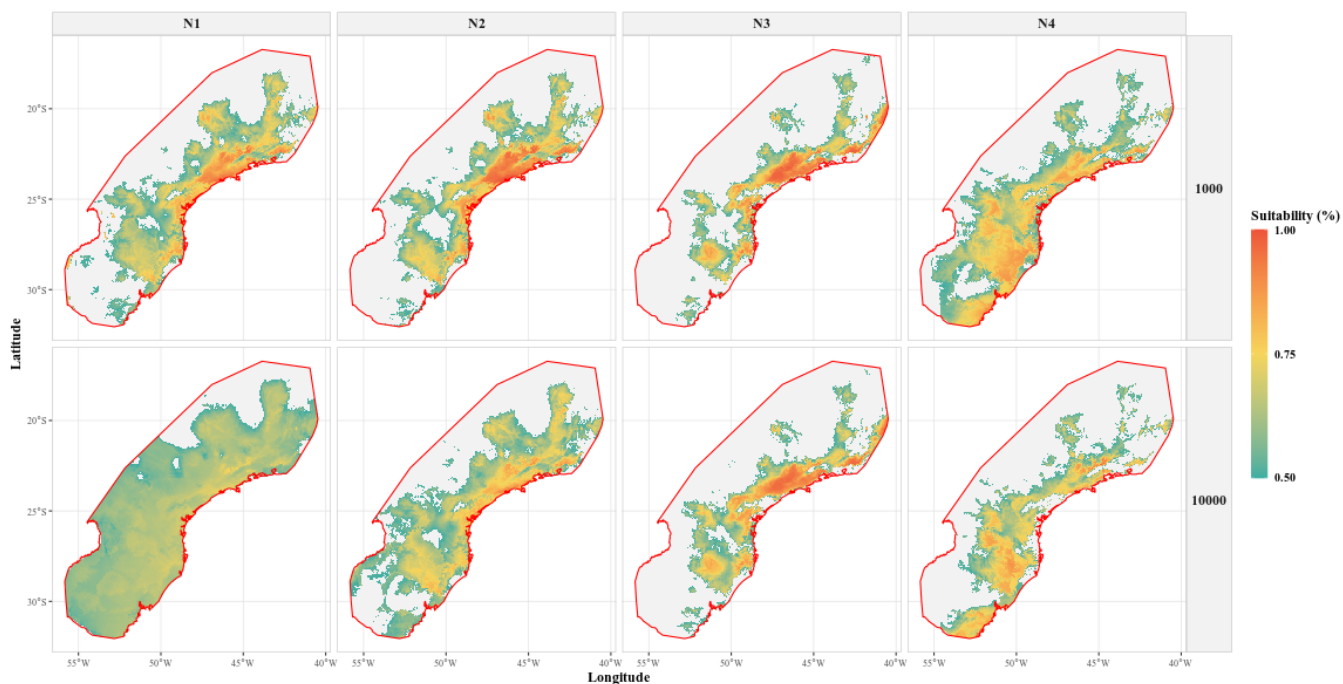


Figure S27. Suitability interval (0.5-1) Maxent projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

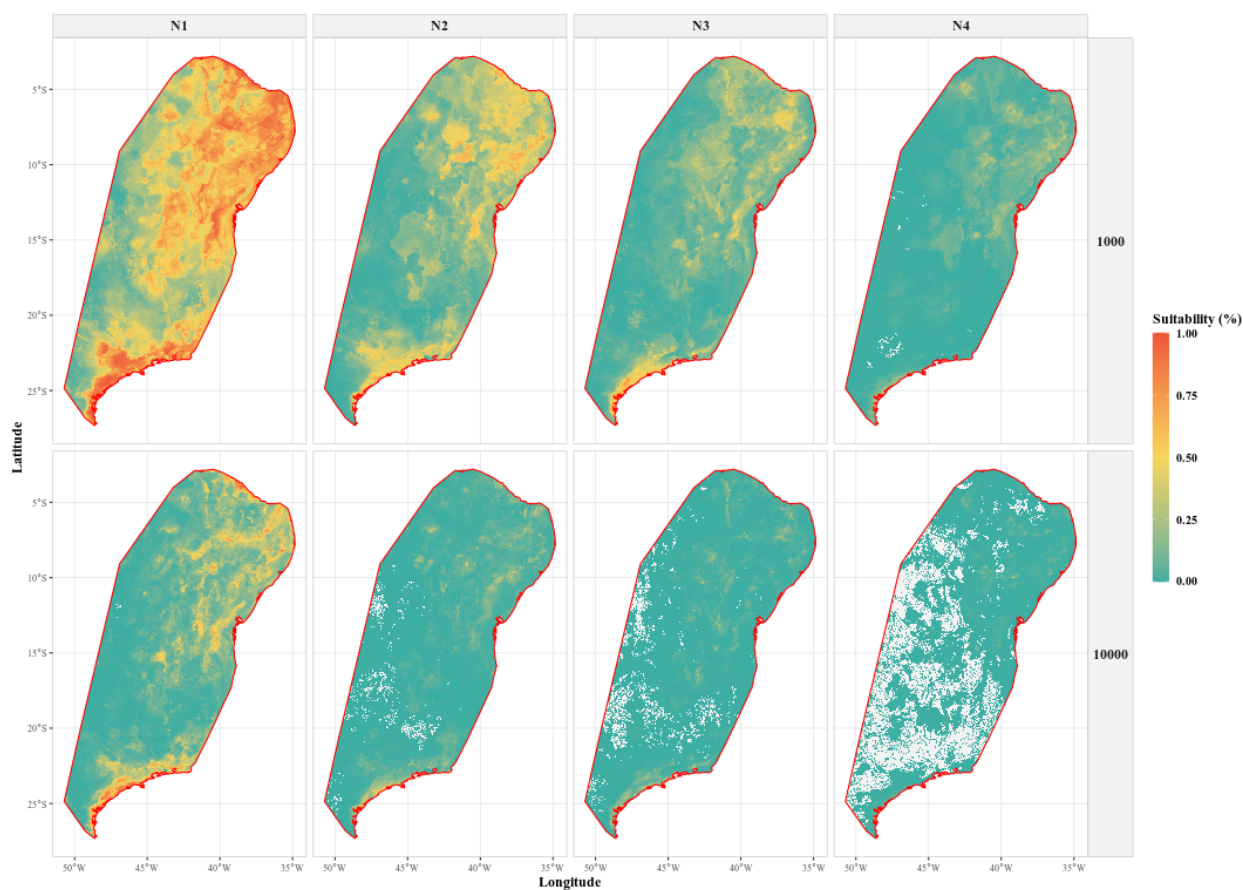


Figure S28. Full suitability Random Forests projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

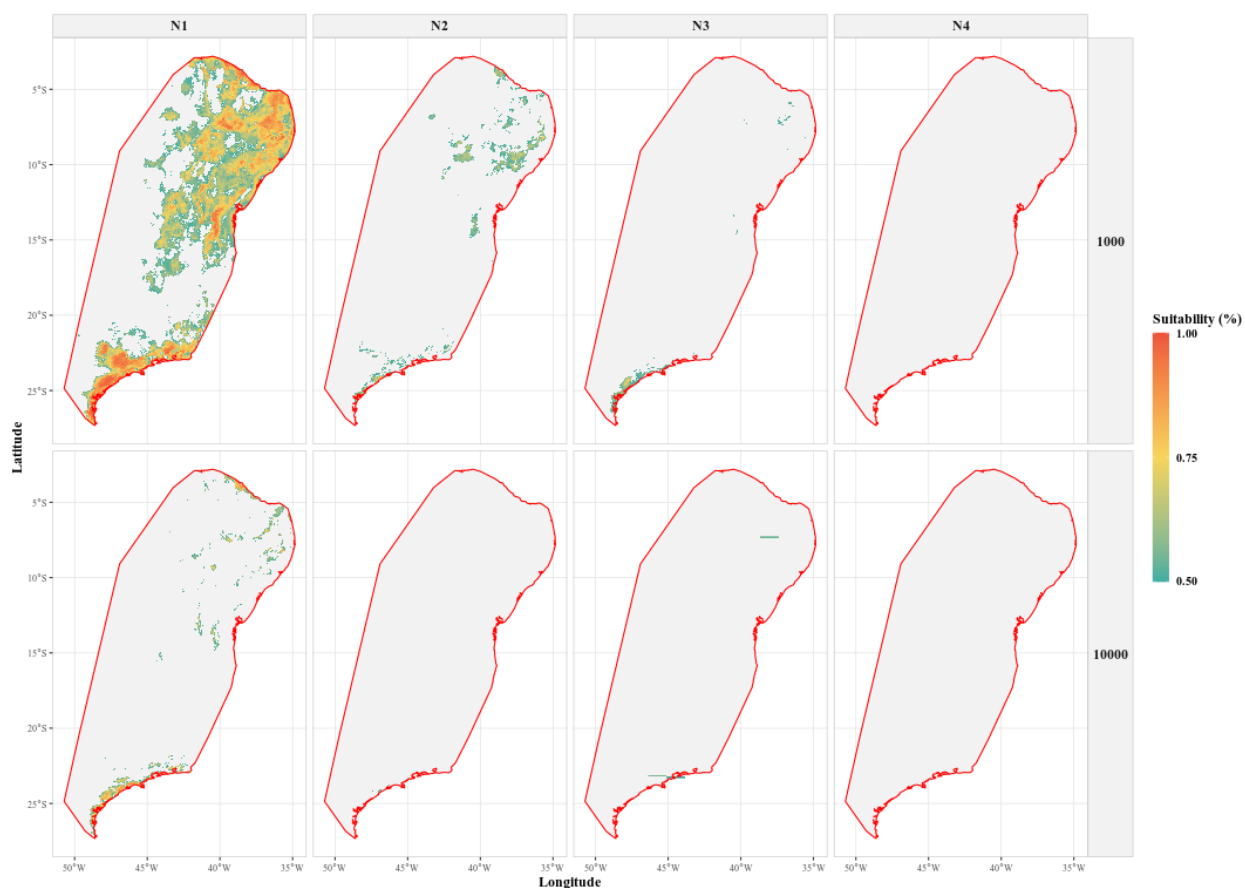


Figure S29. Suitability interval (0.5-1) Random Forests projection for *Cantorchilus longirostris* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

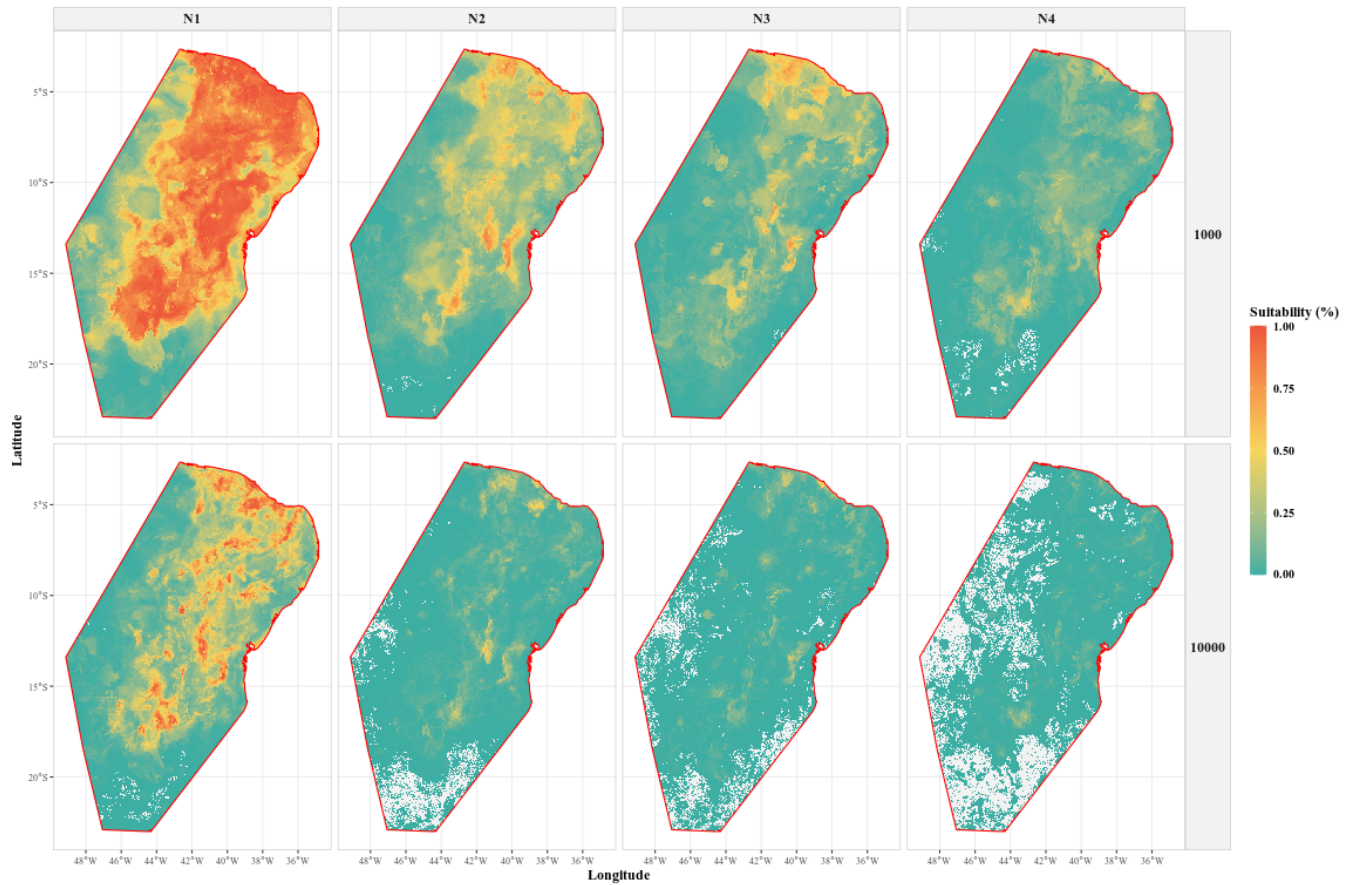


Figure S30. Full suitability Random Forests projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

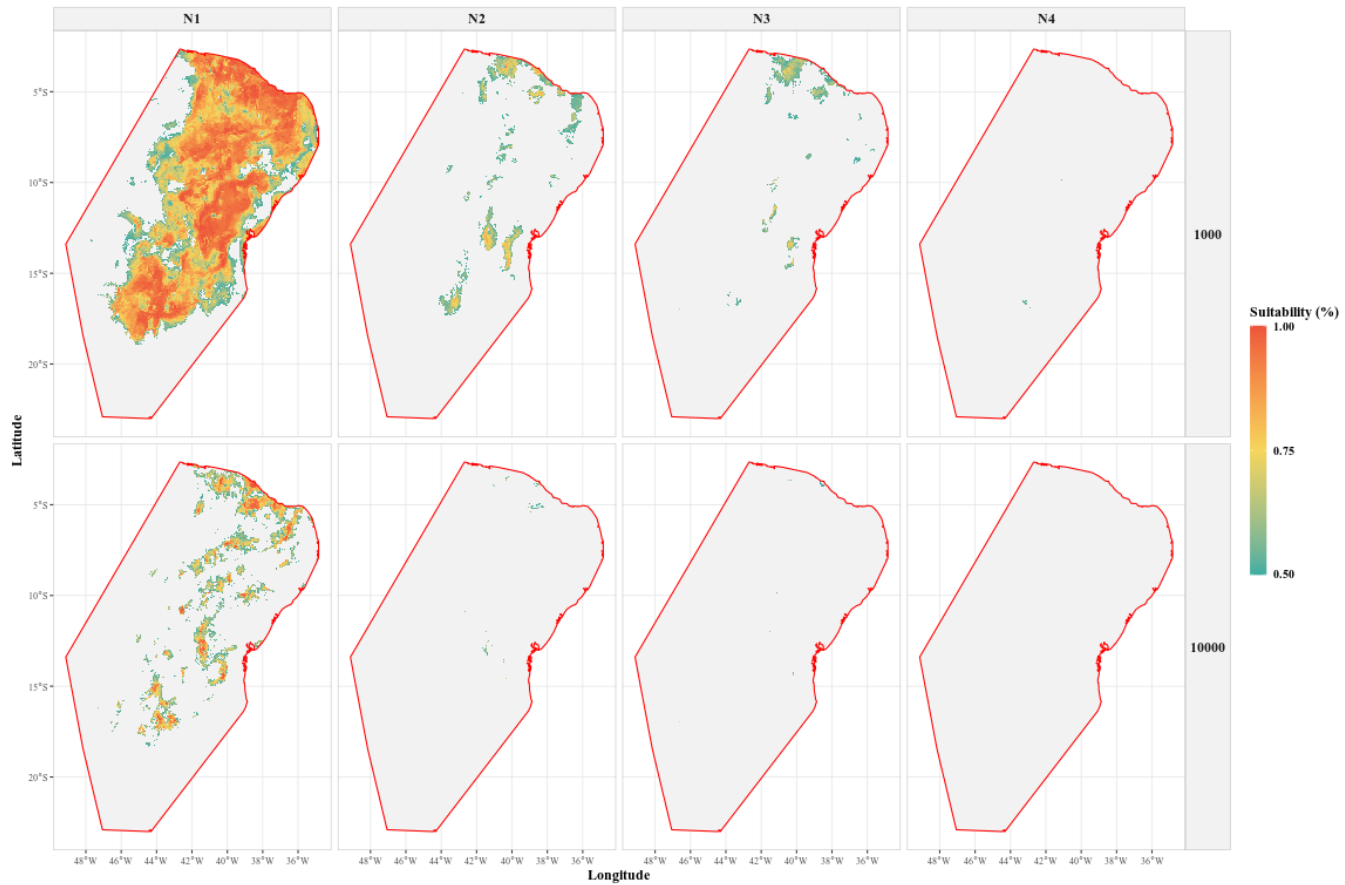


Figure S31. Suitability interval (0.5-1) Random Forests projection for *Eupsittula cactorum* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

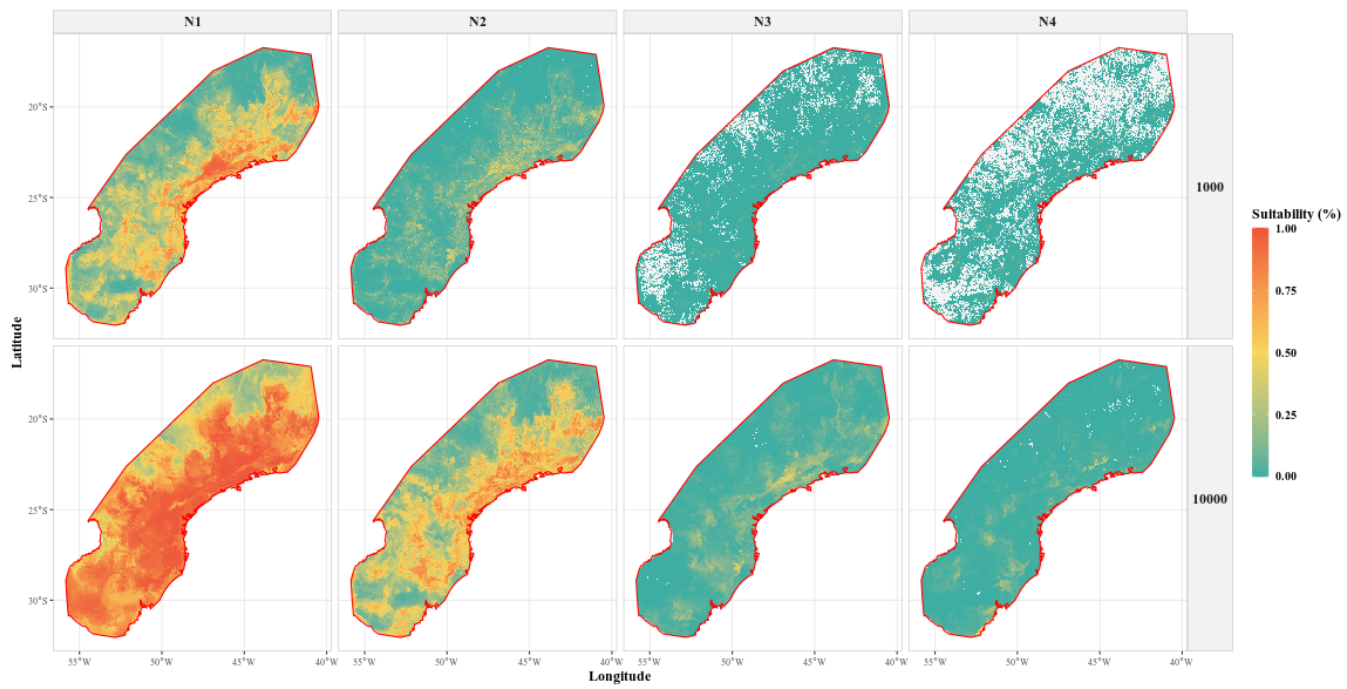


Figure S32. Full suitability Random Forests projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).

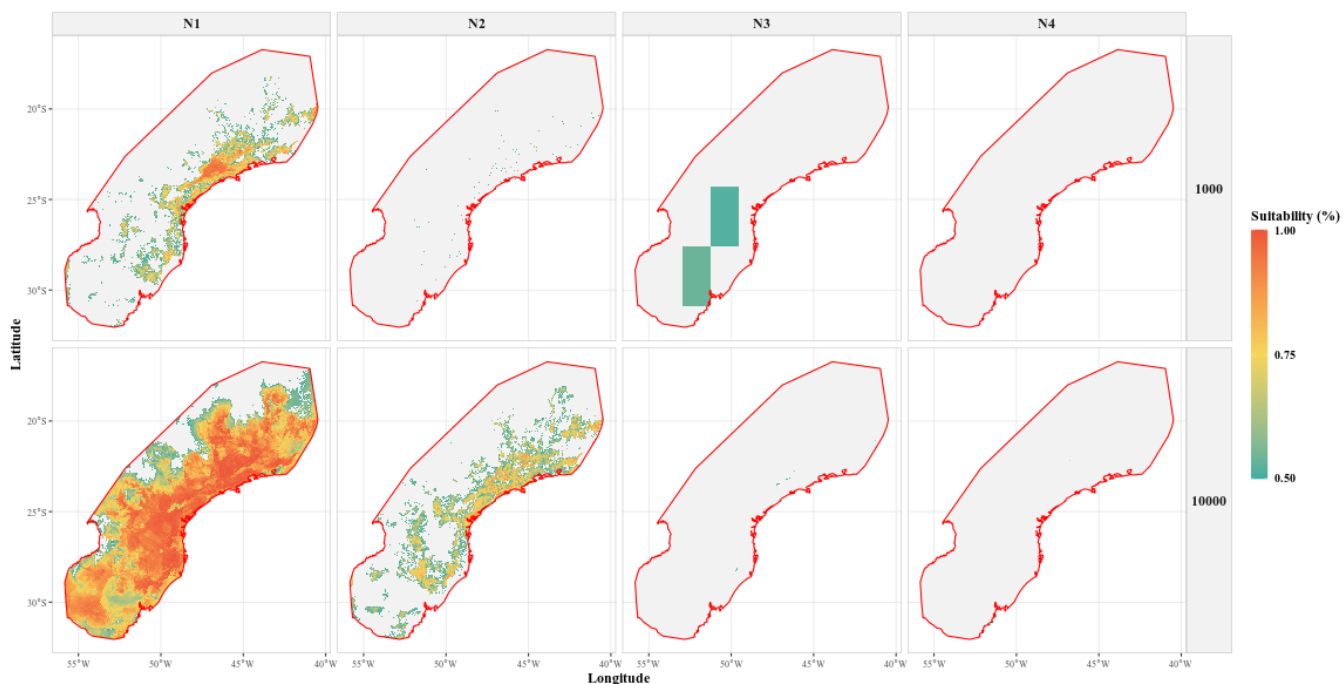


Figure S33. Suitability interval (0.5-1) Random Forests projection for *Penelope obscura* considering all sample sizes and the two pseudo absence/background points sets. Models are constrained by the Extent of Occurrence of the species (red line).