INSTITUTO LATINO-AMERICANO DE CIÊNCIAS DA VIDA E DA NATUREZA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIODIVERSIDADE NEOTROPICAL

UTILIZANDO REDES NEURAIS PROFUNDAS PARA PREDIZER ABUNDÂNCIA DE ESPÉCIES NO CONTEXTO DO GRAN CHACO

ADMIR CESAR DE OLIVEIRA JUNIOR

Foz do Iguaçu
2024

INSTITUTO LATINO-AMERICANO DE CIÊNCIAS
DA VIDA E DA NATUREZA

PROGRAMA DE PÓS-GRADUAÇÃO
EM BIODIVERSIDADE NEOTROPICAL

UTILIZANDO REDES NEURAIS PROFUNDAS PARA PREDIZER ABUNDÂNCIA DE ESPÉCIES NO CONTEXTO DO GRAN CHACO

ADMIR CESAR DE OLIVEIRA JUNIOR

Dissertação de mestrado apresentada ao Programa de Pós-Graduação Biodiversidade Neotropical, do Instituto Latino-Americano de Ciências da Vida e da Natureza, da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Mestre em Ciências Biológicas.

Orientador: Prof. Dr. Santiago José Elías Velazco

Foz do Iguaçu
2024

ADMIR CESAR DE OLIVEIRA JUNIOR

UTILIZANDO REDES NEURAIS PARA PREDIZER ABUNDÂNCIA DE ESPÉCIES NO CONTEXTO DO GRAN CHACO:

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Biodiversidade Neotropical, do Instituto Latino-Americano de Ciências da Vida e da Natureza, da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Mestre em Ciências Biológicas.

BANCA EXAMINADORA

_____

Dr. Santiago José Elías Velazco
Orientador
UNILA

_____

Dr. Michel Varajão Garey
UNILA

_____

Dr. Paulo de Marco Júnior
UFG

Foz do Iguaçu, 03 de abril de 2025

Catalogação elaborada pelo Setor de Tratamento da Informação
Catalogação de Publicação na Fonte. UNILA - BIBLIOTECA LATINO-AMERICANA - CENTRAL

O48
Oliveira Junior, Admir Cesar de.
Utilizando redes neurais profundas para predizer abundância de espécies no contexto do Gran Chaco / Admir Cesar de Oliveira Junior. - Foz do Iguaçu, 2024.
106 f.: il.

Dissertação (Mestrado) - Universidade Federal da Integração Latino-Americana. Instituto Latino-Americano de Ciências da Vida e da Natureza. Programa de Pós-Graduação em Biodiversidade Neotropical. Foz do Iguaçu-PR, 2024.
Orientador: Santiago José Elías Velazco.

1. Abundância - aspectos sociais. 2. Aprendizado do computador. 3. Gran Chaco - Biogeografia. 4. Ecologia espacial. 5. Fitogeografia - Modelos de distribuição. I. Velazco, Santiago José Elías. II. Título.

CDU 574.9

Dedico este trabalho a todos os seres deformados que habitam as fronteiras geográficas, culturais e intelectuais.

Agradecimentos

Seria impossível agradecer a todos que, de uma forma ou de outra, participaram do meu processo de conclusão deste mestrado. Este curso foi o passo seguinte na realização de um sonho de criança, que sempre compartilhei com as pessoas que amei. Estas pessoas, diferentes de outras a quem não dedicarei linhas, nunca riram dos meus sonhos. Me advertiram, mas nunca me desencorajaram. Há dias em que eu desisto dos meus objetivos. Nestes dias, estas pessoas acreditam ainda mais que eu possa alcançá-los. Diante de todas as dificuldades a mim impostas pela influência má dos signos do zodíaco, estas pessoas me motivaram a perseverar. Por isso agradeço e expresso minha profunda admiração e amor aos meus pais, Admir e Zenaide, a quem devo tudo que sou, que me proveram tudo o que foi necessário, desde a graduação até agora, mesmo enfrentando limitações financeiras, me permitindo traçar meu caminho. Agradeço aos meus três irmãos que a lua e a vida me deram. Lucas e Gustavo, que acreditam mais em mim do que eu mesmo, que me viram crescer, cresceram comigo, viveram comigo, sangraram comigo, e têm em mim seu maior e incondicional admirador. Tripaloski segue vivo e pulsa. Mateus (Edival), companheiro de viagens que ainda não ocorreram, montanhismos hipotéticos, acampamentos potenciais, jogos planejados, e esclarecedoras conversas sobre computação. Muitos foram os dias em eu decidi encarar os afazeres apenas para não os decepcionar. Obrigado, pai, mãe e irmãos, eu amo vocês. Vocês são a pedra angular da minha vida.

Neste segundo parágrafo eu agradeço aos meus professores, que se não são pais, são meus padrinhos, e a quem carrego infinita gratidão. Cada palavra, cada ideia, cada correção. Sem sua luz, eu seguiria sendo cego. Por isso eu agradeço especialmente ao meu orientador, Santiago José Elías Velazco, que, nas palavras do grande ecólogo prof. dr. Paulo de Marco Júnior, é “a melhor pessoa que já pisou na Terra”. Nunca concordei tanto com algo. Minha experiência de mestrado foi ótima, e grande parte disso é devido ao professor Santiago, que com sua inteligência e gentileza, me guiou nos momentos mais árduos, e nunca soltou minha mão. No mundo acadêmico, sou uma criança esperando para nascer, mas nascerei sob as orientações dessa figura impressionante e admirada por absolutamente todos a quem perguntei. Quem discorda não quero saber. Também agradeço as minhas eternas professoras, que hoje tenho o orgulho e o amor de chamar de colegas de trabalho, Eliane, Alessandra, Josi e Mônica. Vocês são minha raiz. Em sala de aula repito o que vocês me falavam. Muitas vezes me pego sendo vocês. E espero seguir assim. Agradeço também a todos os professores do Programa de Pós-graduação em Biodiversidade Neotropical da UNILA. Nunca imaginei que seria tão bem recebido, nem que faria tantas amizades com meus mestres. Me senti em casa nestes dois anos, e sentirei saudades quando estiver em outro lugar. Espero visitá-los um dia.

Neste terceiro parágrafo quero agradecer às instituições que fizeram parte disto tudo. À CAPES pela bolsa de estudos a mim concedida, pois sem ela não seria possível minha permanência no programa. Isto faz toda diferença para pessoas de baixa renda, como eu. Permite que nós busquemos o aperfeiçoamento profissional e intelectual. Ao Colégio Cristo Rei, de Cascavel – PR, onde eu conheci pessoas incríveis e me fiz o que viria a ser, e onde sempre encontrei trabalho quando necessitei. E, por fim, agradeço ao secretário do PPGBN, Celso Garcia Junior, que é uma pessoa, mas também é uma instituição, porque garante, consideravelmente sozinho, que este programa funcione como deveria. Nestes anos que estive no mestrado, notei que o Celso foi uma pedra angular no funcionamento do PPGBN. Obrigado por sempre ser ágil e desculpa por tantos emails enchendo o saco, Celso.

É isso. Acabou.

Depois tem mais.

“Todos os modelos estão errados, mas alguns são úteis.”

George Box (1919-2013)

“… Naquele Império, a Arte da Cartografia logrou tal perfeição que o mapa de uma única Província ocupava toda uma Cidade, e o mapa do Império, toda uma Província. Com o tempo, esses Mapas Desmedidos não satisfizeram mais e os Colégios de Cartógrafos levantaram um Mapa do Império, que tinha o tamanho do Império e coincidia pontualmente com ele. Menos dedicadas ao Estudo da Cartografia, as Gerações Seguintes entenderam que esse dilatado Mapa era Inútil e não sem Impiedade o entregaram às Inclemências do Sol e dos Invernos. Nos desertos do Oeste perduram despedaçadas Ruínas do Mapa, habitadas por Animais e por Mendigos; em todo o País não há outra relíquia das Disciplinas Geográficas.”

Jorge Luis Borges (1899-1986)

OLIVEIRA JUNIOR, Admir Cesar de. **Utilizando redes neurais profundas para predizer abundância de espécies no contexto do Gran Chaco**. 2024. 55 p. Dissertação de mestrado do Programa de Pós-Graduação em Biodiversidade Neotropical – Universidade Federal da Integração Latino-Americana, Foz do Iguaçu, 2025.

RESUMO

Os Modelos de Distribuição de Espécies (SDMs) tornaram-se importantes ferramentas, não só para a biologia da conservação como também para a ecologia e biogeografia. Entretanto, os SDM apresentam limitações por ignorarem aspectos ecológicos como tendência e distribuição espacial da abundância de espécies. Recentemente vêm sendo propostos os Modelos de Distribuição Baseados em Abundância (ADM). Como os SDM, estes são modelos correlativos, mas modelam e predizem ao longo do espaço geográfico a abundância das espécies. Entretanto, os ADM permanecem pouco desenvolvidos em comparação com os SDMs, com poucos trabalhos na literatura que explorem os ADM, em questões como algoritmos e validação, e poucos pacotes ou ferramentas desenvolvidas para suas construções. A capacidade dos ADM pode significar ganhos expressivos para a conversação biológica, resultando em melhores decisões e planejamento, especialmente em áreas ameaçadas, como o Gran Chaco. O Gran Chaco é uma região que se estende pelo norte da Argentina, oeste do Paraguai, sudeste da Bolívia e em uma pequena parte do centro-oeste Brasil. Apresenta vegetação predominantemente aberta e xeromórfica e clima semiárido e alta biodiversidade, mas possui poucas áreas protegidas e grande taxa perda de cobertura natural. Esta dissertação está estruturada em dois capítulos. No primeiro capítulo, apresenta-se o pacote para R *adm*, desenvolvido com o intuito de possibilitar fluxos de trabalho concisos e flexíveis para a construção de ADMs. Com este pacote objetiva-se alavancar o estado de desenvolvimento e a pesquisa dos ADMs fornecendo uma ferramenta útil para sua fácil construção e validação. No segundo capítulo, apresenta-se a aplicação deste pacote em um experimento que teve como objetivo avaliar e comparar o desempenho de nove algoritmos, com ênfase em diferentes tipos de Redes Neurais Artificiais, em múltiplos cenários de tratamento dos dados, com a construção de ADMs para 117 espécies arbóreas nativas do Gran Chaco. Os dados de abundância destas espécies foram compilados de inventários florestais do Brasil, Paraguai e Argentina. Os resultados do experimento mostram que decisões como o particionamento dos dados de treinamento e o balanço entre quantidade de pontos de ausência e abundância são importantes para o desempenho dos modelos. Além disso, Redes Neurais Profundas, com múltiplas camadas, tendem a ter melhor desempenho que Redes Neurais Rasas, com apenas uma camada.

Palavras-chave: Abundância. Aprendizado de máquina. Biogeografia. Ecologia espacial. Modelos de distribuição.

OLIVEIRA JUNIOR, Admir Cesar de. **Utilizando redes neurais profundas para predizer abundância de espécies no contexto do Gran Chaco**. 2024. 55 p. Master's thesis of the Graduate Program in Neotropical Biodiversity - Federal University of Latin American Integration, Foz do Iguaçu, 2024.

ABSTRACT

Species Distribution Models (SDMs) have become important tools not only for conservation biology but also for ecology and biogeography. However, SDMs have limitations because they ignore ecological aspects such as the trend and spatial distribution of species abundance. Abundance-based distribution models (ADM) have recently been proposed. Like SDMs, these are correlative models, but they model and predict the abundance of species over geographical space. However, ADMs remain underdeveloped compared to SDMs, with few papers in the literature exploring ADMs, on issues such as algorithms and validation, and few packages or tools developed for their construction. The capacity of ADMs could mean significant gains for biological conversation, resulting in better decisions and planning, especially in threatened areas such as the Gran Chaco. The Gran Chaco is a region that stretches across northern Argentina, western Paraguay, southeastern Bolivia, and a small part of Brazil. It has predominantly open, xeromorphic vegetation, a semi-arid climate, and high biodiversity, but few protected areas and a high rate of natural cover loss. This dissertation is structured in two chapters. The first chapter presents the R adm package, developed to enable concise and flexible workflows for building ADMs. Thish package aims to leverage the state of development and research of ADMs by providing a useful tool for their easy construction and validation. The second chapter presents the application of this package in an experiment that aimed to evaluate and compare the performance of nine algorithms, with an emphasis on different types of Artificial Neural Networks, in multiple data processing scenarios, with the construction of ADMs for 117 tree species native to the Gran Chaco. The abundance data for these species was compiled from forest inventories in Brazil, Paraguay, and Argentina. The results of the experiment show that decisions such as the partitioning of the training data and the balance between the number of absence and abundance points are important for the performance of the models. In addition, Deep Neural Networks with multiple layers tend to perform better than Shallow Neural Networks with only one layer.

Keywords: Distribution models. Abundance. ADM. Spatial ecology. Biogeography. Machine learning.

SUMÁRIO

1 INTRODUÇÃO GERAL

A abundância é a característica de uma população que diz respeito à quantidade de indivíduos de uma certa espécie que ocupam uma determinada área. Este tópico é de interesse dos ecólogos, desde o início da ecologia, especialmente pela importância da compreensão da abundância das espécies tanto em questões relacionados às diferentes teorias ecológicas como conservação e utilização de recursos naturais (MATTHEWS; WHITTAKER, 2015). Ainda hoje, entretanto, os fatores que determinam a abundância são discutidos, mas entende-se que se relacionem à combinação entre características ambientais bióticas e abióticas e as respostas das espécies a elas (ALVES-MARTINS et al., 2023; BORREGAARD; RAHBEK, 2010; DORIGO; BOSCUTTI; SIGURA, 2021). Um aspecto ainda em discussão da abundância é sua relação com a distribuição da espécie no espaço geográfico; i.e., os indivíduos de uma espécie não necessariamente se distribuem de maneira uniforme ao longo de toda sua área de ocorrência (BORREGAARD; RAHBEK, 2010; DE LA FUENTE et al., 2021). No centro desta discussão está a hipótese do Centro-abundante[1], que assume que a abundância de uma espécie deve ser maior no centro de distribuição geográfica e declinar em direção às bordas, acompanhando a redução de condições ótimas para a espécie (BROWN, 1984). Esta relação, entretanto, segue a ser explorada, e pesquisas recentes apresentam tanto resultados favoráveis (MARTIN; ROBINSON; BONIER, 2024; VYE et al., 2020; WALDOCK et al., 2019) quanto contrários à hipótese (DALLAS; SANTINI, 2020; NTULI et al., 2020). De forma semelhante, também se propõe a hipótese centro-abundante ecológico[2], que estabelece que a abundância de uma espécie é maior quanto mais próximo do centroide de um nicho estimado (no espaço ambiental) (OSORIO-OLVERA; SOBERÓN; FALCONI, 2019). Neste sentido, vários trabalhos verificaram uma relação positiva entre a adequabilidade ambiental e abundância local (DE LA FUENTE et al., 2021; JARNEVICH; SOFAER; ENGELSTAD, 2021; WEBER et al., 2017), enquanto outros encontraram pouca ou nenhuma relação (DALLAS; HASTINGS, 2018; LEE-YAW et al., 2022; SPORBERT et al., 2020).

Fatores como temperatura, heterogeneidade do habitat, composição química do ambiente e interações com outras espécies se mostram importantes para a abundância local de espécies de diferentes grupos biológicos (BOWLER et al., 2018). Outro aspecto importante é a sazonalidade. Trabalhos recentes demonstram que a abundância de diferentes espécies pode flutuar de forma intimamente ligada à padrões temporais de condições ambientais, como temperatura e precipitação (MTUI et al., 2022; NOVAIS et al., 2019). Entretanto, todos estes fatores podem variar, ser

[1] Do inglês "*Abundant-centre hypotesis*".
[2] Do inglês "*Ecological Abundant-centre hypotesis*".

mensurados e ter efeitos significativos sobre a abundância de uma espécie em diferentes escalas, o que é conhecido como escala de efeito. Neste sentido, trabalhos recentes demonstram que a abundância possui resposta variada a variáveis em diferentes escalas (MORAGA; MARTIN; FAHRIG, 2019; SAN-JOSÉ et al., 2019). Alguns trabalhos encontraram evidências de que variáveis em grandes escalas, a nível de paisagem, são boas preditoras da abundância de espécies, mas outros concluíram que a abundância foi mais afetada por variáveis em escalas menores (KYRÖ et al., 2018; REMM et al., 2017). Outras pesquisas ainda argumentam que esta questão depende do contexto e da espécie estudada (CATZIM et al., 2022; DORIGO; BOSCUTTI; SIGURA, 2021; GESTICH et al., 2019).

Esta discussão acerca da abundância e sua distribuição geográfica envolve a acepção "hutchisoniana" de nicho, amplamente aceita contemporaneamente, que o compreende como um hipervolume n-dimensional determinado por condições abióticas e bióticas nas quais a espécies pode manter populações viáveis (SOBERÓN; NAKAMURA, 2009). Neste sentido, a área geográfica em que ocorre uma espécie está relacionada como o "nicho realizado", que consiste nas regiões do "nicho potencial" colonizáveis por uma espécie. O "nicho potencial", por sua vez, são todas as áreas em que o "nicho fundamental" - o conjunto de todas as condições abióticas e bióticas favoráveis à espécie - está de fato presente num espaço geográfico e tempo determinado (JACKSON; OVERPECK, 2000; SOBERÓN; ARROYO-PEÑA, 2017; SOBERÓN; NAKAMURA, 2009).

Baseadas nesta compreensão, foram propostas há décadas técnicas para a construção de modelos de nicho ecológico e suas projeções no espaço geográfico (ELITH; LEATHWICK, 2009). Entre estas, os Modelos de Distribuição de Espécies (SDM) ganharam especial atenção nos últimos anos. Estes modelos correlacionam dados de presença, ausência e/ou pseudo-ausências com variáveis ambientais, projetando a adequabilidade ambiental da espécie ao longo do espaço geográfico (FRANKLIN, 2023; RATHORE; SHARMA, 2023). Entretanto, existem outras abordagens de modelagem não necessariamente correlativas, como os SDM. Os modelos mecanicistas, por exemplo, são focados em processos e visam a captura da relação entre condições ambientais e traços funcionais da espécie para projetar sua área de ocorrência (EVANS et al., 2016). Em modelos determinísticos, a distribuição da espécie é suposta como completamente determinada pelas condições iniciais e os parâmetros do modelo, sem incorporar elementos de estocasticidade (MOHD, 2022).

Embora os SDMs tenham ganhado grande atenção e aplicabilidade, sendo utilizados como importantes ferramentas na biogeografia e nos planejamentos dos esforços de conservação (FRANKLIN, 2023; RATHORE; SHARMA, 2023; SOFAER et al., 2019), eles também apresentam limitações. Uma destas é a de que os SDM, embora capazes de projetar adequabilidade

ambiental, ignoram outros aspectos ecológicos, como densidade e tendências de populações, importantes aspectos para a conservação (HASTINGS et al., 2020). Isto se torna um problema dado que, por exemplo, a população de uma espécie não necessariamente se distribui uniformemente a longo de todo o espaço geográfico ocupado (BORREGAARD; RAHBEK, 2010). Neste sentido, mudanças ambientais podem causar o declínio da população de uma espécie sem causar grandes impactos ou alterações em sua área de distribuição (HASTINGS et al., 2020). Diversas abordagens foram propostas para modelar a abundância da espécie no espaço geográfico (e.g., modelos mecanicistas como MetaRange, FALLERT; LI; CABRAL, 2025; e RangeShifter, BOCEDI et al., 2021; e correlativos, como regressões quantílicas, CADE; NOON, 2003; VILLÉN-PERÉZ et al., 2020). Recentemente vêm sendo propostos os Modelos de Distribuição Baseados em Abundância (ADM) (DREXLER; AINSWORTH, 2013; EHRLÉN; MORRIS, 2015; HILL et al., 2017; HOWARD et al., 2014; KULHANEK; LEUNG; RICCIARDI, 2011; YU; COOPER; INFANTE, 2020). Estes modelos assemelham-se aos SDM ao correlacionar dados de abundância e variáveis ambientais.

A capacidade de estimar a abundância ao longo do espaço geográfico pode ser de grande valor para a conservação das espécies (VILLÉN-PERÉZ et al., 2020). Tudo isto torna os ADM relevantes para áreas como a ecologia e conservação. Há, entretanto, poucos trabalhos na literatura, até a atualidade, tratando explicitamente destes modelos e/ou explorando extensivamente possíveis algoritmos para construí-los. Em especial, WALDOCK et al. (2022) e BOTELLA et al. (2018) conduziram experimentos comparativos com diversos algoritmos e obtiveram promissores resultados. Contudo, os ADMs ainda possuem desenvolvimento muito inferior em relação aos SDMs (WALDOCK et al., 2022).

Atualmente, devido a crescente disponibilidade de dados ecológicos e do aumento do poder computacional, a utilização de algoritmos mais complexos torna-se possível, e o ajuste de modelos ecológicos quase sempre ocorre em ambientes computacionais (GILBERT et al., 2024). A linguagem de programação estatística R (R CORE TEAM, 2024) vêm sendo cada vez mais utilizadas para ajuste e análise de modelos, revelando o papel central destas ferramentas na pesquisa em ecologia e conservação (LAI et al., 2019, 2023). Aliás, R é a principal linguagem para a construção de SDM (KASS et al., 2025). Portanto, o desenvolvimento de pacotes focados na construção de AMD em R representaria uma oportunidade para popularizar e alavancar o próprio desenvolvimento e aplicação destes modelos.

Entre os algoritmos complexos mais utilizados atualmente estão as Redes Neurais Artificiais (ANN). Estas são algoritmos que originalmente foram pensados para imitar o funcionamento do cérebro humano, e por isso consistem em redes estruturadas de neurônios interconectados; estes são unidades computacionais que executam transformações nos dados

(ALZUBAIDI et al., 2021). Nestas redes, os neurônios estão organizados em diversas camadas sucessivas. A quantidade de camadas e a quantidade de neurônios em cada camada não são parâmetros predeterminados, permitindo que redes muito grandes ou muito pequenas sejam definidas, dependendo da necessidade do usuário, e disso vem a flexibilidade das ANN (POUYANFAR et al., 2019). As ANN mais simples contam com apenas uma camada, entre a entrada e a saída dos dados, e por isso são frequentemente chamadas de Redes Neurais Rasas (NET) (PODDER et al., 2021). Com o avanço da tecnologia, tornaram-se possíveis a construção de ANN com dezenas a milhares de camadas e neurônios, surgindo o que é chamado de Redes Neurais Profundas (DNN) (ALOM et al., 2019). A variação na estrutura e funcionamento dos neurônios também tornou possível a construção de Redes Neurais Convolucionais (CNN), que são capazes de trabalhas com dados matriciais, como imagens, em vez de dados tabulares (AYENI, 2022). Todas as ANN aprendem com o próprio erro, através dos processos de *feed-foward* e *back-propagation*. No processo de aprendizagem, são imputados os dados na camada inicial e cada neurônio desta camada faz transformações numéricas nos dados e transmite seu resultado para cada neurônio da camada seguinte. Esse processo é repetido sequencialmente. Após à saída dos dados, o erro é computado e os pesos dos neurônios, que controlam as transformações numéricas, são atualizados. Este processo de treino é repetido até que o erro seja minimizado (CHAI; JIN, 2024). Atualmente, estes algoritmos vêm sendo muito utilizados na ecologia, em tarefas que vão desde a regressão de variáveis à classificação de imagens (BOROWIEC et al., 2022), mas ainda são poucos explorados no campo dos ADMs.

A aplicação de ADMs se mostra especialmente relevante em regiões com alta diversidade e perda de cobertura natural, comuns nas regiões tropicais e subtropicais (EDWARDS et al., 2019), como o Gran Chaco (Chaco seco, úmido e serrano). O Gran Chaco estende-se ao longo do norte da Argentina, oeste do Paraguai, sudeste da Bolívia e uma pequena porção do sudoeste do Mato Grosso do Sul, no Brasil (PRADO, 1993). Possui vegetação majoritariamente aberta, xeromórfica, clima semiárido, com verões de temperatura elevada e geadas no inverno (SPICHIGER et al., 2006). Alguns autores consideram a região do Gran Chaco a maior floresta seca contínua do mundo (OLSON et al., 2001) e o segundo maior bioma da América do Sul, com 1,3 milhões km² de extensão (BUCHER; HUSZAR, 1999). Apesar de possuir elevada biodiversidade (REDFORD; TABER; SIMONETTI, 1990), pouco de sua área é protegida (NORI et al., 2016), e há uma alarmante taxa de perda de cobertura natural (GASPARRI; GRAU, 2009). Por exemplo, de 2013 à 2023, a cobertura natural da região diminui ~ 4,64 milhões de hectares, dando espaço para atividades antrópicas (PROYECTO MAPBIOMAS CHACO, 2024). Portanto, a compreensão sobre a distribuição e os padrões espaciais de suas espécies nativas poderia ser benéfica para a conservação da espécie.

Este trabalho é divido em dois capítulos. No primeiro, apresenta-se o pacote para R *adm*, destinado para o ajuste, validação e predição de ADMs. No segundo, efetuou-se um experimento como o objetivo de comparar a performance de diversos algoritmos estatísticos convencionais e de aprendizagem de máquina com diferentes tratamentos de dados, utilizando-se de dados de abundância de espécies arbóreas do Gran Chaco. Com isto objetivou-se contribuir e alavancar o desenvolvimento dos ADMs, bem como incentivar futuras pesquisas sobre estes modelos, especialmente no contexto de regiões ameaçadas.

2 REFERÊNCIAS

ALOM, M. Z. et al. A State-of-the-Art Survey on Deep Learning Theory and Architectures. **Electronics**, v. 8, n. 3, p. 292, 5 mar. 2019.

ALVES-MARTINS, F. et al. **The neglected role of limiting factors in large-scale abundance patterns**. , 2 fev. 2023. Disponível em: <https://www.authorea.com/users/582375/articles/622458-the-neglected-role-of-limiting-factors-in-large-scale-abundance-patterns?commit=156c56003593d294e3e45fb4d4b1123e58d8fa27>. Acesso em: 21 jun. 2024

ALZUBAIDI, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. **Journal of Big Data**, v. 8, n. 1, p. 53, 31 mar. 2021.

AYENI, J. A. Convolutional Neural Network (CNN): The architecture and applications. **Applied Journal of Physical Science**, v. 4, n. 4, p. 42–50, 30 dez. 2022.

BOCEDI, G. et al. RangeShifter 2.0: an extended and enhanced platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. **Ecography**, v. 44, n. 10, p. 1453–1462, out. 2021.

BOROWIEC, M. L. et al. Deep learning as a tool for ecology and evolution. **Methods in Ecology and Evolution**, v. 13, n. 8, p. 1640–1660, 2022.

BORREGAARD, M. K.; RAHBEK, C. Causality of the Relationship between Geographic Distribution and Species Abundance. **The Quarterly Review of Biology**, v. 85, n. 1, p. 3–25, mar. 2010.

BOTELLA, C. et al. A Deep Learning Approach to Species Distribution Modelling. Em: JOLY, A. et al. (Eds.). **Multimedia Tools and Applications for Environmental & Biodiversity Informatics**. Cham: Springer International Publishing, 2018. p. 169–199.

BOWLER, D. E. et al. Disentangling the effects of multiple environmental drivers on population changes within communities. **Journal of Animal Ecology**, v. 87, n. 4, p. 1034–1045, jul. 2018.

BROWN, J. H. On the Relationship between Abundance and Distribution of Species. **The American Naturalist**, v. 124, n. 2, p. 255–279, ago. 1984.

BUCHER, E. H.; HUSZAR, P. C. Sustainable management of the Gran Chaco of South America: Ecological promise and economic constraints. **Journal of Environmental Management**, v. 57, n. 2, p. 99–108, out. 1999.

CADE, B. S.; NOON, B. R. A gentle introduction to quantile regression for ecologists. **Frontiers in Ecology and the Environment**, v. 1, n. 8, p. 412–420, out. 2003.

CATZIM, V. V. et al. Local and landscape correlates of coccinellid species richness, abundance, and assemblage change along a rural–urban gradient in Quintana Roo, Mexico. **Biotropica**, v. 54, n. 3, p. 776–788, maio 2022.

CHAI, Y.; JIN, L. Deep learning in data science: Theoretical foundations, practical applications, and comparative analysis. **Applied and Computational Engineering**, v. 69, n. 1, p. 1–6, 21 jun. 2024.

DALLAS, T. A.; HASTINGS, A. Habitat suitability estimated by niche models is largely unrelated to species abundance. **Global Ecology and Biogeography**, v. 27, n. 12, p. 1448–1456, dez. 2018.

DALLAS, T. A.; SANTINI, L. The influence of stochasticity, landscape structure and species traits on abundant–centre relationships. **Ecography**, v. 43, n. 9, p. 1341–1351, set. 2020.

DE LA FUENTE, A. et al. Predicting species abundance by implementing the ecological niche theory. **Ecography**, v. 44, n. 11, p. 1723–1730, nov. 2021.

DORIGO, L.; BOSCUTTI, F.; SIGURA, M. Landscape and microhabitat features determine small mammal abundance in forest patches in agricultural landscapes. **PeerJ**, v. 9, p. e12306, 16 nov. 2021.

DREXLER, M.; AINSWORTH, C. H. Generalized Additive Models Used to Predict Species Abundance in the Gulf of Mexico: An Ecosystem Modeling Tool. **PLoS ONE**, v. 8, n. 5, p. e64458, 14 maio 2013.

EDWARDS, D. P. et al. Conservation of Tropical Forests in the Anthropocene. **Current Biology**, v. 29, n. 19, p. R1008–R1020, out. 2019.

EHRLÉN, J.; MORRIS, W. F. Predicting changes in the distribution and abundance of species under environmental change. **Ecology Letters**, v. 18, n. 3, p. 303–314, mar. 2015.

ELITH, J.; LEATHWICK, J. R. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. **Annual Review of Ecology, Evolution, and Systematics**, v. 40, n. 1, p. 677–697, 1 dez. 2009.

EVANS, M. E. K. et al. Towards Process-based Range Modeling of Many Species. **Trends in Ecology & Evolution**, v. 31, n. 11, p. 860–871, nov. 2016.

FALLERT, S.; LI, L.; CABRAL, J. S. metaRange : A framework to build mechanistic range models. **Methods in Ecology and Evolution**, v. 16, n. 1, p. 49–56, jan. 2025.

FRANKLIN, J. Species distribution modelling supports the study of past, present and future biogeographies. **Journal of Biogeography**, v. 50, n. 9, p. 1533–1545, set. 2023.

GASPARRI, N. I.; GRAU, H. R. Deforestation and fragmentation of Chaco dry forest in NW Argentina (1972–2007). **Forest Ecology and Management**, v. 258, n. 6, p. 913–921, set. 2009.

GESTICH, C. C. et al. Unraveling the scales of effect of landscape structure on primate species richness and density of titi monkeys (*Callicebus nigrifrons*). **Ecological Research**, v. 34, n. 1, p. 150–159, jan. 2019.

GILBERT, N. A. et al. A century of statistical *Ecology* . **Ecology**, v. 105, n. 6, p. e4283, jun. 2024.

HASTINGS, R. A. et al. Climate Change Drives Poleward Increases and Equatorward Declines in Marine Species. **Current Biology**, v. 30, n. 8, p. 1572- 1577.e2, abr. 2020.

HILL, L. et al. Abundance distributions for tree species in Great Britain: A two-stage approach to modeling abundance using species distribution modeling and random forest. **Ecology and Evolution**, v. 7, n. 4, p. 1043–1056, fev. 2017.

HOWARD, C. et al. Improving species distribution models: the value of data on abundance. **Methods in Ecology and Evolution**, v. 5, n. 6, p. 506–513, jun. 2014.

JACKSON, S. T.; OVERPECK, J. T. Responses of plant populations and communities to environmental changes of the late Quaternary. **Paleobiology**, v. 26, n. S4, p. 194–220, 2000.

JARNEVICH, C. S.; SOFAER, H. R.; ENGELSTAD, P. Modelling presence versus abundance for invasive species risk assessment. **Diversity and Distributions**, v. 27, n. 12, p. 2454–2464, dez. 2021.

KASS, J. M. et al. Achieving higher standards in species distribution modeling by leveraging the diversity of available software. **Ecography**, v. 2025, n. 2, p. e07346, fev. 2025.

KULHANEK, S. A.; LEUNG, B.; RICCIARDI, A. Using ecological niche models to predict the abundance and impact of invasive species: application to the common carp. **Ecological Applications**, v. 21, n. 1, p. 203–213, jan. 2011.

KYRÖ, K. et al. Local habitat characteristics have a stronger effect than the surrounding urban landscape on beetle communities on green roofs. **Urban Forestry & Urban Greening**, v. 29, p. 122–130, jan. 2018.

LAI, J. et al. Evaluating the popularity of R in ecology. **Ecosphere**, v. 10, n. 1, p. e02567, jan. 2019.

LAI, J. et al. The Use of R and R Packages in Biodiversity Conservation Research. **Diversity**, v. 15, n. 12, p. 1202, 7 dez. 2023.

LEE-YAW, J. et al. Species distribution models rarely predict the biology of real populations. **Ecography**, v. 2022, n. 6, p. e05877, jun. 2022.

MARTIN, P. R.; ROBINSON, O. J.; BONIER, F. Rare edges and abundant cores: range-wide variation in abundance in North American birds. **Proceedings of the Royal Society B: Biological Sciences**, v. 291, n. 2015, p. 20231760, 31 jan. 2024.

MATTHEWS, T. J.; WHITTAKER, R. J. REVIEW: On the species abundance distribution in applied ecology and biodiversity management. **Journal of Applied Ecology**, v. 52, n. 2, p. 443–454, abr. 2015.

MOHD, M. H. Revisiting discrepancies between stochastic agent-based and deterministic models. **Community Ecology**, v. 23, n. 3, p. 453–468, out. 2022.

MORAGA, A. D.; MARTIN, A. E.; FAHRIG, L. The scale of effect of landscape context varies with the species' response variable measured. **Landscape Ecology**, v. 34, n. 4, p. 703–715, abr. 2019.

MTUI, D. T. et al. Elevational distribution of montane Afrotropical butterflies is influenced by seasonality and habitat structure. **PLOS ONE**, v. 17, n. 7, p. e0270769, 5 jul. 2022.

NORI, J. et al. Protected areas and spatial conservation priorities for endemic vertebrates of the Gran Chaco, one of the most threatened ecoregions of the world. **Diversity and Distributions**, v. 22, n. 12, p. 1212–1219, dez. 2016.

NOVAIS, S. M. A. D. et al. CHANGES IN THE INSECT HERBIVORE FAUNA AFTER THE FIRST RAINS IN A TROPICAL DRY FOREST. **Oecologia Australis**, v. 23, n. 02, p. 381–387, jun. 2019.

NTULI, N. N. et al. Rejection of the genetic implications of the “Abundant Centre Hypothesis” in marine mussels. **Scientific Reports**, v. 10, n. 1, p. 604, 17 jan. 2020.

OLSON, D. M. et al. Terrestrial Ecoregions of the World: A New Map of Life on Earth. **BioScience**, v. 51, n. 11, p. 933, 2001.

OSORIO-OLVERA, L.; SOBERÓN, J.; FALCONI, M. On population abundance and niche structure. **Ecography**, v. 42, n. 8, p. 1415–1425, ago. 2019.

PODDER, P. et al. **Artificial Neural Network for Cybersecurity: A Comprehensive Review**. arXiv, , 20 jun. 2021. Disponível em: <http://arxiv.org/abs/2107.01185>. Acesso em: 2 out. 2024

POUYANFAR, S. et al. A Survey on Deep Learning: Algorithms, Techniques, and Applications. **ACM Computing Surveys**, v. 51, n. 5, p. 1–36, 30 set. 2019.

PRADO, D. E. What is the Gran Chaco vegetation in South America? I. A review. Contribution to the study of flora and vegetation of the Chaco. **Candollea**, n. 48, p. 145–172, 1993.

PROYECTO MAPBIOMAS CHACO. **Colección 5.0 de los mapas anuales de cobertura y uso del suelo**. , 2024. Disponível em: <https://chaco.mapbiomas.org/>. Acesso em: 22 out. 2024

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2024.

RATHORE, M. K.; SHARMA, L. K. Efficacy of species distribution models (SDMs) for ecological realms to ascertain biological conservation and practices. **Biodiversity and Conservation**, v. 32, n. 10, p. 3053–3087, ago. 2023.

REDFORD, K. H.; TABER, A.; SIMONETTI, J. A. There Is More to Biodiversity than the Tropical Rain Forests. **Conservation Biology**, v. 4, n. 3, p. 328–330, set. 1990.

REMM, J. et al. Multilevel landscape utilization of the Siberian flying squirrel: Scale effects on species habitat use. **Ecology and Evolution**, v. 7, n. 20, p. 8303–8315, out. 2017.

SAN-JOSÉ, M. et al. The scale of landscape effect on seed dispersal depends on both response variables and landscape predictor. **Landscape Ecology**, v. 34, n. 5, p. 1069–1080, maio 2019.

SOBERÓN, J.; ARROYO-PEÑA, B. Are fundamental niches larger than the realized? Testing a 50-year-old prediction by Hutchinson. **PLOS ONE**, v. 12, n. 4, p. e0175138, 12 abr. 2017.

SOBERÓN, J.; NAKAMURA, M. Niches and distributional areas: Concepts, methods, and assumptions. **Proceedings of the National Academy of Sciences**, v. 106, n. supplement_2, p. 19644–19650, 17 nov. 2009.

SOFAER, H. R. et al. Development and Delivery of Species Distribution Models to Inform Decision-Making. **BioScience**, v. 69, n. 7, p. 544–557, 1 jul. 2019.

SPICHIGER, R. et al. Biogeography of the Forests of the Paraguay-Paraná Basin. Em: TOBY PENNINGTON, R.; LEWIS, G.; RATTER, J. (Eds.). **Neotropical Savannas and Seasonally Dry Forests**. Systematics Association Special Volumes. [s.l.] CRC Press, 2006. v. 20060637p. 193–211.

SPORBERT, M. et al. Testing macroecological abundance patterns: The relationship between local abundance and range size, range position and climatic suitability among European vascular plants. **Journal of Biogeography**, v. 47, n. 10, p. 2210–2222, out. 2020.

VILLÉN-PERÉZ, S. et al. Global warming will affect the maximum potential abundance of boreal plant species. **Ecography**, v. 43, n. 6, p. 801–811, jun. 2020.

VYE, S. R. et al. Patterns of abundance across geographical ranges as a predictor for responses to climate change: Evidence from UK rocky shores. **Diversity and Distributions**, v. 26, n. 10, p. 1357–1365, out. 2020.

WALDOCK, C. et al. The shape of abundance distributions across temperature gradients in reef fishes. **Ecology Letters**, v. 22, n. 4, p. 685–696, abr. 2019.

WALDOCK, C. et al. A quantitative review of abundance-based species distribution models. **Ecography**, v. 2022, n. 1, p. ecog.05694, jan. 2022.

WEBER, M. M. et al. Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? A meta-analysis. **Ecography**, v. 40, n. 7, p. 817–828, jul. 2017.

YU, H.; COOPER, A. R.; INFANTE, D. M. Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. **Ecological Modelling**, v. 432, p. 109202, set. 2020.

3 CAPÍTULO I

adm: an R package for constructing abundance-based distribution models

Abstract:

1. Abundance-based distribution models (ADM) correlate species abundance with environmental data to model and project abundance throughout space or time. This promising and still developing technique has gained significant attention in recent years.
2. Here, we present the *adm* R package developed to support the construction of ADM, including data preparation, model fitting, prediction, and model exploration. This package offers several modeling approaches (i.e., algorithms) that can be fine-tuned and customized. Models can be predicted in geographic space and explored regarding performance and response curves. Because modeling workflows in *adm* are constructed based on a combination of distinct functions and simple outputs, *adm* can be easily integrated into other packages. To illustrate this, we constructed a full modeling procedure for the shrub species *Cynophalla retusa* using *adm*.
3. To date, *adm* provides 35 functions in three categories, i) modeling: to tune, fit, and validate models with nine different algorithms, with a suite of possible model-specific hyperparameters; ii) post-modeling: to predict abundance across space and construct partial dependence plots to explore the relationships between abundance and environmental predictors; and iii) miscellaneous tools: to support the workflow in all steps, including data handling, transformations, and hyperparameter selection.
4. With *adm*, we intend to provide a flexible, straightforward, and concise toolbox for ADM construction and expect it to help users develop and leverage the promising ADM field.

Keywords: artificial neural networks, correlative models, model tuning, spatial ecology, species abundance models, species distribution models.

Introduction

Spatially explicit ecological models are an important approach in many research areas because they consider how spatially structured characteristics and constraints influence the modeled phenomena (DeAngelis & Yurek, 2017). One of the most widely used strategies for estimating species geographic distributions is species distribution model (SDM, also known as ecological niche models or habitat suitability models). This technique uses occurrence data and environmental variables to model the environmental suitability or occurrence probability of a species and predict its distribution (Franklin, 2023). Despite the importance of SDM in ecology and conservation, other important ecological aspects, such as population density and trends, are often not modeled because of the lack of data (Hastings et al., 2020).

Abundance-based Distribution Models (ADM) are similar to SDM, as both are spatially explicit correlative models; however, they model and project abundance throughout space or time by correlating species abundance and environmental data (Anadón et al., 2010; Ehrlén & Morris, 2015; Yu et al., 2020). Predictions of species abundance are useful for assessing extinction risk, estimating the effects of climate and land-use change, understanding the environmental drivers of species abundance, and performing conservation prioritization analysis (Villén-Pérez et al., 2020). Nonetheless, ADMs remain less developed than SDMs (Waldock et al., 2022), except for N-mixture models, which are appropriate for multivisit abundance data. A few R packages support ADM fitting, tuning, evaluation, and prediction, and most use only a few algorithms (Tables S1-S2). Using multiple algorithms and performing hyperparameters tuning is crucial for diversifying species abundance modeling approaches and enabling selection of best models or accounting with mode uncertainty (Qiao et al., 2015; Thuiller et al., 2019).

Here, we introduce *adm*, a new R package designed to facilitate the development of ADM workflows. *adm* is one of the only packages offering fitting, tuning, and model exploration of different modeling approaches, from Generalized Linear and Additive Models (Rigby & Stasinopoulos, 2005) to different types of Artificial Neural Networks implemented with *torch* (Falbel & Luraschi, 2024), which provides high architectural customization, native GPU (graphics processing unit) acceleration, and more complex setups. Furthermore, *adm* is structured to support modeling workflows that can be easily integrated with other R packages.

Package overview

adm was inspired by the philosophy of the *flexsdm* R package, which enables users to build flexible modeling workflows by combining user-selected functions that return widely used R objects such as *terra* SpatRaster and *tidyverse* tibbles (Velazco et al., 2022). Furthermore, *adm* works as an independent extension of *flexsdm*, offering features adapted to abundance modeling. *adm* features can be used alone or integrated with *flexsdm* features, e.g., various data partitioning approaches for model fitting and validation (k-fold, bootstrap, or environmentally and geographically structured partition), model calibration area delimitation, measuring model extrapolation and performing model truncation (Velazco et al., 2024). Building on the advances in developing conventional SDM, the integration of *adm* and *flexsdm* will improve the development of state-of-the-art ADM workflows.

Currently, *adm* provides 35 functions divided into three categories: modeling, post-modeling, and miscellaneous tools (Figure 1). Modeling abundance is more challenging than modeling presence-absence because abundance can be measured in different ways (e.g., plot

coverage percentage or absolute abundance). Therefore, *adm* is designed to handle different types of response variables by i) offering different probability distributions for different algorithms, ii) selecting suitable distributions based on response variable nature for GLM and GAM (*family_selector* function), or iii) performing data transformations with different methods (*adm_transform* function). Additionally, *adm* does not create any new R object classes; rather, most function outputs are simple R lists (i.e., heterogeneous vectors that can contain different object classes, e.g., lists, tibbles, rasters, and vectors). For instance, in the modeling functions, the outputs comprise an object of the original modeling framework and additional informative and useful tabular data. The simplicity of the *adm*'s output allows users to manipulate and explore results, making them more compatible with other packages.

Figure 1. Overview of *adm* package functions structured in modeling (fitting and validation), post-modeling (model predictions and exploratory plots), and various utilities to support modeling workflows.

Modeling functions

Selecting a modeling algorithm is a crucial step in developing a predictive model for species abundance and requires an understanding of the model assumptions and functionality. This can be a challenging task, and it is often recommended to test multiple algorithms (Qiao et al., 2015). *adm* facilitates modeling, hyperparameter tuning, and validation of nine algorithms (Table S3), grouped into two function types, denoted by *fit_abund* and *tune_abund* prefixes. Both

fit_abund and *tune_abund* validate models internally (see Table S3, full list of algorithms and hyperparameters).

fit_abund functions allow users to fit algorithms with default or user-specified hyperparameter values. However, tuning model hyperparameters influences algorithm performance and complexity and is therefore an important consideration when developing conventional SDMs (Fourcade, 2021). Choosing the optimal hyperparameter values for the data and modeling objective can significantly enhance model performance. Functions with the *tune_abund* prefix allow users to perform model tuning by using an array of model-specific hyperparameters. In *adm*, tuning is performed using a grid-search approach, i.e., by evaluating model performance under an array of possible hyperparameter combinations. To implement this approach, the user provides a data frame with hyperparameters as columns and the hyperparameter values to be tested as rows (Appendix S1). Model tuning functions then iterate through the available hyperparameter values and select the best-performing model and its associated hyperparameter values. To reduce computational time, *tune_abund* functions support parallel processing built with *parallel* (R Core Team, 2024), *doSNOW* (Microsoft & Weston, 2022a), and *foreach* (Microsoft & Weston, 2022b) packages. It is possible to evaluate models using one or more performance metrics during the model tuning process. If more than one performance metric is used, the order of the chosen metrics is important, as *model_selection* iterates sequentially through the metrics, selecting the highest quartile models of a given metric until only one model remains. Performance metrics for model evaluation are calculated using *adm_eval* (see below).

Generalized Linear and Additive Models (GLM and GAM)

GLM and GAM are expansions of Linear Models and enable the modeling of non-linear predictor-response relationships, even when the data are not normally distributed (Hastie & Tibshirani, 1986; Nelder & Wedderburn, 1972). GLM assume that the relationships between predictor and response variables are mediated by a link function that allows using different probability distributions (Nelder & Wedderburn, 1972). GAM use a link and smoothing function, which captures nonlinear relationships between response and predictor variables (Hastie & Tibshirani, 1986). In *adm*, GLM and GAM are based on the Generalized Additive Models for Location, Space, and Shape framework (GAMLSS), using the *gamlss* package (Rigby & Stasinopoulos, 2005). GAMLSS offers >100 probability distributions and the possibility of modeling any parameter that defines a family distribution (Stasinopoulos & Rigby, 2012). In *adm* users can fit GLM and GAM models using all the distribution families supported by *gamlss* (Figure S1). However, it is worth noting that the response variable must respect the family's assumptions; *tune_abund_glm and tune_abund_gam* automatically select the most suitable families if they are not provided within the user-specified grid. GLM can be parameterized with different interaction orders between explanatory variables and the degree of polynomials. For GAM, users can control the smoothness degree used in a formula. For the GLM and GAM, parameters that define a distribution (i.e., sigma, nu, and tau) can be modeled based on predictor variables.

Generalized Boosted Regression Models (GBM) and Extreme Gradient Boosting (XGB)

Boosting algorithms are machine learning algorithms that sequentially train small models, each one improving upon the errors of the previous model, which is known as 'boosting' (J. Friedman et al., 2000; J. H. Friedman, 2001, 2002). In *adm,* boosting algorithms are supported in

two different expansions and implementations of the original modeling framework (Friedman et al., 2000; Friedman, 2001, 2002), GBM and XGB, via *gbm* and *xgboost* packages, respectively (Chen et al., 2024; Greg & Developers, 2024). Both are set to use trees as boosters but have significant differences in gradient computation, hyperparameters, overfitting prevention, and regression tree construction (Chen & Guestrin, 2016) (Figures S2-S3). The user can tune hyperparameters, such as tree depth, learning rate, and distribution family (Table S3).

Random Forest (RAF)

RAF is a machine-learning algorithm based on the ensemble of multiple decision trees trained with a bootstrapped version of the original dataset and predictors subset (Breiman, 2001). RAF has been widely used in ecology and distribution modeling, generally obtaining good performance, even with small datasets (Pichler & Hartig, 2023; Valavi et al., 2022). The algorithm was implemented via *randomForest* package (Liaw & Wiener, 2002). Here, the user can set the number of trees grown in the forest and the number of predictors used for each decision tree (Table S3).

Support Vector Machine (SVM)

SVM is a machine learning algorithm that aims to define an optimal hyperplane determined by nonlinear decision boundaries that split samples into different classes within a higher-dimensional space (Salcedo-Sanz et al., 2014). SVM was implemented using *kernlab* package (Karatzoglou et al., 2004). SVM's *adm* function is set up to perform (epsilon) regressions, and the user can set the desired kernel, its parameters, and the constraint violation cost (Table S3). SVM is adapted to work with Radial Basis and Laplacian kernels, but the user can experiment with different kernels and configurations, in which case it is recommended to read *kernlab* documentation.

Artificial Neural Networks (NET), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN)

Neural Networks are systems composed of interconnected neurons capable of learning complex nonlinear data relationships. These neurons are organized into one or multiple layers called hidden layers (Alzubaidi et al., 2021). When a network features multiple serialized hidden layers, it is often called a Deep Neural Network (Alom et al., 2019). In this case, networks can be constructed with several architectures, combining multiple types of neurons, layers, and functions (Pouyanfar et al., 2019). We refer to DNN as a fully connected, feedforward, backpropagation, and artificial neuron networks, which is the most common type of deep networks (Alom et al., 2019). The neuron receives inputs, performs a weighted operation, and feeds forward a value transformed by an activation function to the next layer (Schmidhuber, 2015). When a neural network features a single hidden layer, it can be called a Shallow Neural Network (Podder et al., 2021), referred to as NET herein and in *adm* documentation. These networks function similarly to DNN, but are less computationally intensive, although they perform well (Winkler & Le, 2017). Other common structures are the Convolutional Neural Networks (CNN), which have filters or kernels that perform convolution operations across large multidimensional data matrices. This process sequentially generates "activation maps" between the layers, which allows the network to learn complex features from the data (Alzubaidi et al., 2021). Although there are no rules for their

construction, excessively deep and large neural networks of any type tend to overfit (Pichler & Hartig, 2023). In ecology, these techniques have gained significant attention, encompassing a wide range of applications, including regressions and distribution models (Borowiec et al., 2022; Pichler & Hartig, 2023).

For DNN and CNN, *adm* uses *torch* framework for R (Falbel & Luraschi, 2024). This allows the construction of highly customizable architectures, with size and number of layers defined by the user. NET is based on *nnet* package (Venables et al., 2002), which is a single-layer and less customizable option; however, it is much faster than DNN and CNN.

adm provides functions to help define the size and number of layers for constructing CNN and DNN. *generate_dnn_architecture* and the *generate_cnn_architecture* functions help to easily construct neural networks. To facilitate the tuning process, the *generate_arch_list* function builds multiple architectures with different layer configurations. To systematically sample these architectures, *select_arch_list* can be used to reduce the list of architectures, while maintaining a range of characteristics. In addition, users can manually construct a neural network using *torch* package syntax and use it within *fit_abund* and *tune_abund* functions.

Model performance metrics (adm_eval)

Model evaluation metrics are calculated using *adm_eval* function, which is implemented internally in *fit_adm* and *tune_adm* functions. *adm_eval* returns a *tibble* with results for six performance metrics calculated between observed and predicted data (Table 1), based on Waldock et al. (2022): (i) Spearman's and (ii) Pearson's correlations, (iii) Mean Absolute Error, which consists of the absolute value of the average residual, (iv) Intercept and (v) Slope of a linear model fitted with observed abundance as a function of predicted abundance, and (vi) Dispersion, calculated as the ratio between the standard deviation of predicted and observed abundance.

Table 1. Model performance metrics, acronyms, and their characteristics

Metric	Acronym	Range	Type
Spearman Correlation	corr_spear	$[-1, 1]$	Discrimination
Pearson Correlation	corr_pear	$[-1, 1]$	
Slope	slope	$(-\infty, +\infty)$	
Intercept	inter	$(-\infty, +\infty)$	
Mean Absolute Error	mae	$[0, +\infty)$	Accuracy
Dispersion	pdisp	$[0, +\infty)$	Precision

Post-modeling functions

Model prediction

In *adm*, spatial predictions for all algorithms are performed using *adm_predict*, using rasterized predictor variables as input. This function can simultaneously predict multiple models with prediction transformation, accounting for negative and scale-transformed values. Transform negative values could be valuable for algorithms that do not use a distribution (e.g., some machine learning approaches).

Partial dependence plot and partial bivariate dependence plots

Partial dependence plots allow for the exploration of marginal response curves by linearly varying the values of one predictor while maintaining other constants. In *adm*, partial dependence plots and their bivariate version can easily be constructed with *p_abund_pdp* and *p_abund_bpdp* which return a *ggplot2* object (Wickham, 2016). Both functions require only the output of *tune_* and *fit_* functions.

Miscellaneous tools

Dataset and variable manipulation

Models and predictions can be constructed by transforming response and predictor variables. To facilitate this process, *adm_transform* can scale predictor variables in rasters or response variables in a table. It can also return the values to the original scale if necessary. Because algorithms are sensitive to the number of zeros (Barbet-Massin et al., 2012; Liu et al., 2019), the *balance_dataset* can be used to perform absence data thinning by randomly selecting absences to equilibrate the number of presences and absences to a given ratio.

Additional tools

Several other *adm* functionalities can be useful in the modeling workflow. *family_selector* identifies suitable distribution families to use in GAM and GLM based on range and type of response variable (Figure S1). *model_selection* iterates over performances dataset of a given model to select the best-performing hyperparameter combination based on user-defined performance metrics. *model_selection* is implemented in each *tune_abund_* functions, but users can utilize it independently, e.g., to reselect the best hyperparameter combination based on different metrics, without the need to tune the model again. *adm_summarize* concatenates performance tables from different models into one single table. *adm_extract* retrieves values for georeferenced points from a predictor raster.

Example

We illustrate the use of *adm* and its integration with *flexsdm* (Velazco et al., 2022) by modeling the abundance of *Cynophalla retusa* (Griseb.) Cornejo & Iltis (Capparaceae) (Appendix S2). It is a shrub native to northeastern Argentina, Paraguay, Bolivia, and central Brazil, and is distributed mainly in dry biomes. We compiled and standardized data from the first (1998-2002) and second (2020) national forestry surveys in Argentina (MAyDS, 2022; SAyDS, 2005), constructing an abundance (individuals/ha) and absence dataset (sites with 0 individuals/ha). Using the *adm::balance_dataset*, we balanced presence (sites with >0 individuals/hectare) and absence at a 1:1 ratio. The absences were limited to the species training area, constructed as a 200-km buffered minimum convex polygon around presence points, using *flexsdm::calib_area.* We performed a principal component analysis using *flexsdm::correct_colinvar*, with 35 climatic and edaphic variables (Table S4), and selected the first seven principal components that represented >

90% of cumulative variance as predictors (Table S5). We partitioned the dataset into three spatial blocks using *flexsdm::part_sblock*. To construct the models, we used DNN, RAF, and GLM algorithms, fitted and validated by *adm::tune_abund_dnn*, *adm::tune_abund_raf* and *adm::tune_abund_glm*, aiming to maximize Pearson's correlation and minimize MAE (Table S6). Because DNN often perform better with scaled data (LeCun et al., 1998), input response data were standardized by Z-score using *adm::adm_transform* before fitting this algorithm, and all architectures tested featured batch normalization between layers, generated with *adm::generate_arch_list*. Predictions were generated, restricted to the species calibration area, with *adm::adm_predict*. To explore how models extrapolate, we produced bivariate partial dependence plots with *adm::p_abund_bpdp*, taking as an example the first and seventh principal components (Figure 2, Figure S4-S9 for all bivariate and univariate partial dependence plots). Comprehensive functions documentation and illustrative examples are available on the *adm* website (example of the website is available in: https://figshare.com/s/0906509c8b0b38c92243).

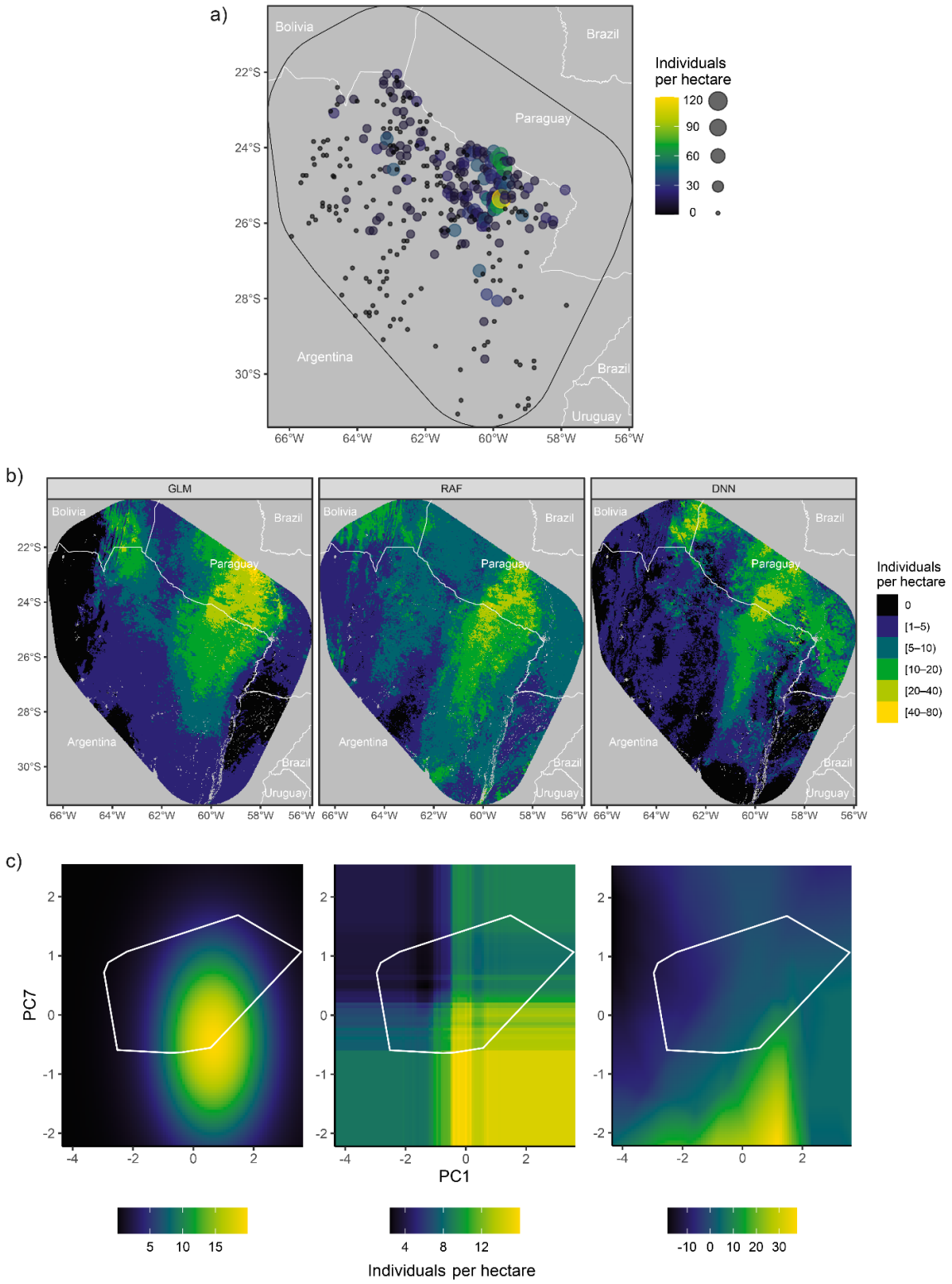


Figure 2. *Cynophalla retusa* observed abundance and ADM constructed with Generalized Linear Models, Random Forest, and Deep Neural Networks. a) Observed abundance for each sample point used in models training. Points' size is proportional to its abundance. b) Abundance maps predicted by each algorithm within training area. Predictions were classified into intervals to facilitate visualization. c) Partial bivariate dependence plots for each model. White polygon in b represents the range of environmental conditions explored by abundance data.

Conclusion

The *adm* R package provides functions to construct a full workflow to model and predict species abundance in the geographic and environmental space. We highlight the possibility of using a variety of highly customizable algorithms and provide several functions for predicting and exploring ADMs. The complete integration of *adm* with *flexsdm* creates a holistic environment for modeling conventional species' presence-absence and species' abundance, allowing users to seamlessly combine and compare both approaches. In the future, we aim to expand *adm*'s features by implementing algorithm ensembles, ensembles of small models (Breiner et al., 2015), new algorithms, variable importance, and other evaluation metrics. We expect that *adm* will help users to further develop the promising ADM field by providing a flexible, straightforward, integrated, and concise toolbox.

References

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., & Asari, V. K. (2019). A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, *8*(3), 292. https://doi.org/10.3390/electronics8030292

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Anadón, J. D., Giménez, A., & Ballestar, R. (2010). Linking local ecological knowledge and habitat modelling to predict absolute species abundance on large scales. *Biodiversity and Conservation*, *19*(5), 1443–1454. https://doi.org/10.1007/s10531-009-9774-4

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022). Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, *13*(8), 1640–1660. https://doi.org/10.1111/2041-210X.13901

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiner, F. T., Guisan, A., Bergamini, A., & Nobis, M. P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, *6*(10), 1210–1218. https://doi.org/10.1111/2041-210X.12403

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2024). *xgboost: Extreme Gradient Boosting*. https://CRAN.R-project.org/package=xgboost

DeAngelis, D. L., & Yurek, S. (2017). Spatially Explicit Modeling in Ecology: A Review. *Ecosystems*, *20*(2), 284–300. https://doi.org/10.1007/s10021-016-0066-z

Ehrlén, J., & Morris, W. F. (2015). Predicting changes in the distribution and abundance of species under environmental change. *Ecology Letters*, *18*(3), 303–314. https://doi.org/10.1111/ele.12410

Falbel, D., & Luraschi, J. (2024). *torch: Tensors and Neural Networks with "GPU" Acceleration*. https://torch.mlverse.org/docs

Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian Obama nungara. *Ecological Modelling*, *457*, 109686. https://doi.org/10.1016/j.ecolmodel.2021.109686

Franklin, J. (2023). Species distribution modelling supports the study of past, present and future biogeographies. *Journal of Biogeography*, *50*(9), 1533–1545. https://doi.org/10.1111/jbi.14617

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2). https://doi.org/10.1214/aos/1016218223

Greg, R., & Developers, G. B. M. (2024). *gbm: Generalized Boosted Regression Models*. https://CRAN.R-project.org/package=gbm

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3). https://doi.org/10.1214/ss/1177013604

Hastings, R. A., Rutterford, L. A., Freer, J. J., Collins, R. A., Simpson, S. D., & Genner, M. J. (2020). Climate Change Drives Poleward Increases and Equatorward Declines in Marine Species. *Current Biology*, *30*(8), 1572-1577.e2. https://doi.org/10.1016/j.cub.2020.02.043

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). **kernlab**—An *S4* Package for Kernel Methods in *R*. *Journal of Statistical Software*, *11*(9). https://doi.org/10.18637/jss.v011.i09

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Madsen, L., & Royle, J. A. (2023). A review of N-mixture models. *WIREs Computational Statistics*, *15*(6), e1625. https://doi.org/10.1002/wics.1625

MAySD, M. de A. y D. S. de la N. (2022). *Segundo Inventario Nacional de Bosques Nativos: Informe Nacional*. Ministerio de Ambiente y Desarrollo Sostenible de la Nación.

Microsoft, & Weston, S. (2022a). *doSNOW: Foreach Parallel Adaptor for the "snow" Package*. https://CRAN.R-project.org/package=doSNOW

Microsoft, & Weston, S. (2022b). *foreach: Provides Foreach Looping Construct*. https://CRAN.R-project.org/package=foreach

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370. https://doi.org/10.2307/2344614

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Podder, P., Bharati, S., Mondal, M. R. H., Paul, P. K., & Kose, U. (2021). *Artificial Neural Network for Cybersecurity: A Comprehensive Review* (arXiv:2107.01185). arXiv. http://arxiv.org/abs/2107.01185

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2019). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys*, *51*(5), 1–36. https://doi.org/10.1145/3234150

Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, *6*(10), 1126–1136. https://doi.org/10.1111/2041-210X.12397

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *54*(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Salcedo-Sanz, S., Rojo-Álvarez, J. L., Martínez-Ramón, M., & Camps-Valls, G. (2014). Support vector machines in engineering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(3), 234–267. https://doi.org/10.1002/widm.1125

SAyDS, S. de A. y D. S. (2005). *Primer Inventario Nacional de Bosques Nativos*. Ministerio de Salud y Ambiente de la Nación.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Stasinopoulos, M., & Rigby, R. (2012). *gamlss: Generalized Additive Models for Location Scale and Shape* (p. 5.4-22) [Dataset]. https://doi.org/10.32614/CRAN.package.gamlss

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*(1), e01486. https://doi.org/10.1002/ecm.1486

Velazco, S. J. E., Rose, M. B., De Andrade, A. F. A., Minoli, I., & Franklin, J. (2022). FLEXSDM: An R package for supporting a comprehensive and flexible species distribution modelling workflow. *Methods in Ecology and Evolution*, *13*(8), 1661–1669. https://doi.org/10.1111/2041-210X.13874

Velazco, S. J. E., Rose, M. B., De Marco, P., Regan, H. M., & Franklin, J. (2024). How far can I extrapolate my species distribution model? Exploring shape, a novel method. *Ecography*, *2024*(3), e06992. https://doi.org/10.1111/ecog.06992

Venables, W. N., Ripley, B. D., & Venables, W. N. (2002). *Modern applied statistics with S* (4th ed). Springer.

Villén-Pérez, S., Heikkinen, J., Salemaa, M., & Mäkipää, R. (2020). Global warming will affect the maximum potential abundance of boreal plant species. *Ecography*, ecog.04720. https://doi.org/10.1111/ecog.04720

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W. L., Edgar, G. J., Mouillot, D., Tjiputra, J., & Pellissier, L. (2022). A quantitative review of abundance-based species distribution models. *Ecography*, *2022*(1), ecog.05694. https://doi.org/10.1111/ecog.05694

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Winkler, D. A., & Le, T. C. (2017). Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. *Molecular Informatics*, *36*(1–2), 1600118. https://doi.org/10.1002/minf.201600118

Yu, H., Cooper, A. R., & Infante, D. M. (2020). Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. *Ecological Modelling*, *432*, 109202. https://doi.org/10.1016/j.ecolmodel.2020.109202

Supplementary Material

Appendixes

Appendix S1. Data sent and code to reproduce *Cynophalla retusa* example

Code sources, raster, and abundance database used to reproduce de modeling protocol of *Cynophalla retusa* can be downloaded from the following link (https://figshare.com/s/fde3ceccf8983a96f255).

Tables

Table S1. R packages that construct spatially explicit abundance-based models compared to *adm*.

Package	Repository	Reference
spAbund	https://doserlab.com/files/spabundance-web/reference/spabund	Doser, J. W., Finley, A. O., Kéry, M., & Zipkin, E. F. (2024). spAbundance: An R package for single-species and multi-species spatially explicit abundance models. Methods in Ecology and Evolution, 15(6), 1024–1033. https://doi.org/10.1111/2041-210X.14332
SDMaps	https://www.sovon.nl/onderzoek/methoden-en-technieken/ict-toepassingen/sdmaps	Kampichler, C., Hallmann, C., & Sierdsema, H. (2020). SDMaps: An R package for the analysis of species abundance and distribution data. Sovon Dutch Centre for Field Ornithology. https://pub.sovon.nl/pub/publicatie/20823
unmarked	https://github.com/rbchan/unmarked	Kellner, K. F., Smith, A. D., Royle, J. A., Kéry, M., Belant, J. L., & Chandler, R. B. (2023). The unmarked R package: Twelve years of advances in occurrence and abundance modelling in ecology. Methods in Ecology and Evolution, 14(6), 1408–1415. https://doi.org/10.1111/2041-210X.14123
ubms	https://github.com/biodiverse/ubms	Kellner, K. F., Fowler, N. L., Petroelje, T. R., Kautz, T. M., Beyer, D. E., & Belant, J. L. (2022). ubms: An R package for fitting hierarchical occupancy and N-mixture abundance models in a Bayesian framework. Methods in Ecology and Evolution, 13(3), 577–584. https://doi.org/10.1111/2041-210X.13777
inlabru	https://inlabru-org.github.io/inlabru/	Bachl, F. E., Lindgren, F., Borchers, D. L., & Illian, J. B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. Methods in Ecology and Evolution, 10(6), 760–766. https://doi.org/10.1111/2041-210X.13168
RISDM	https://github.com/Scott-Foster/RISDM	Foster, S. D., Peel, D., Hosack, G. R., Hoskins, A., Mitchell, D. J., Proft, K., Yang, W., Uribe-Rivera, D. E., & Froese, J. G. (2024). 'RISDM': Species distribution modelling from multiple data sources in R. Ecography, 2024(6), e06964. https://doi.org/10.1111/ecog.06964
ModEco*	https://3decology.org/	Guo, Q., & Liu, Y. (2010). ModEco: An integrated software package for ecological niche modeling. Ecography, 33(4), 637–642. https://doi.org/10.1111/j.1600-0587.2010.06416.x

DynamicSDM	https://github.com/r-a-dobson/dynamicSDM	Dobson, R., Challinor, A. J., Cheke, R. A., Jennings, S., Willis, S. G., & Dallimer, M. (2023). DYNAMICSDM: An R package for species geographical distribution and abundance modelling at high spatiotemporal resolution. Methods in Ecology and Evolution, 14(5), 1190–1199. https://doi.org/10.1111/2041-210X.14101

*ModEco is a software, but was included here to recognize the previous work.

Table S2. Features of *adm* compared with other R packages.

Features	Details	adm	spAbund	inlabru	RISDM	ModEco	unmarked	ubms	SDMaps	DynamicSDM
Response type	Single species	1	1	1	1	1	1	1	1	1
	Multi-species	0	1	0	0	0	0	0	1	0
Plots	Partial Dependence Plots	1	0	0	0	0	0	0	0	0
	Scatter plot	0	0	0	0	1	0	0	0	0
	Histogram	0	0	0	0	1	0	0	0	0
	Samples plot	0	0	1	0	0	0	0	0	0
	Mesh plot	0	0	0	1	0	0	0	0	0
	Marginal covariate effects	0	0	0	0	0	0	1	0	0
	Trace plot	0	1	0	0	0	0	1	0	0
	Residuals plot	0	0	0	1	0	0	1	0	0
	Base plot methods	0	1	0	1	0	1	1	0	0
Modelling algorithms	Generalized Additive Models	1	0	0	0	0	0	0	1	0
	Generalized Linear Models	1	1	0	0	1	0	0	1	0
	Generalized Boosted Regression	1	0	0	0	0	0	0	1	1
	Extreme Gradient Boosting	1	0	0	0	0	0	0	0	0
	Artificial Neural Networks	1	0	0	0	1	0	0	1	0
	Deep Neural Networks	1	0	0	0	0	0	0	1	0
	Convolutional Neural Networks	1	0	0	0	0	0	0	0	0
	Random Forest	1	0	0	0	0	0	0	1	0
	Support Vector Machine	1	0	0	0	1	0	0	0	0
	HDS	0	1	0	0	0	1	0	0	0
	N-mixture	0	1	0	0	0	1	0	0	0
	INLA	0	0	1	0	0	0	0	0	0
	ISDM	0	0	0	1	0	0	0	0	0
Tunned modeling algorithms	Generalized Additive Models	1	0	0	0	0	0	0	0	0
	Generalized Linear Models	1	0	0	0	0	0	0	0	0
	Generalized Boosted Regression	1	0	0	0	0	0	0	0	0
	Extreme Gradient Boosting	1	0	0	0	0	0	0	0	0
	Artificial Neural Networks	1	0	0	0	0	0	0	0	0

	Deep Neural Networks	1	0	0	0	0	0	0	0	0
	Convolutional Neural Networks	1	0	0	0	0	0	0	0	0
	Random Forest	1	0	0	0	0	0	0	0	0
	Support Vector Machine	1	0	0	0	0	0	0	0	0
	HDS	0	1	0	0	0	0	0	0	0
	N-mixture	0	1	0	0	0	0	0	0	0
	INLA	0	0	0	0	0	0	0	0	0
	ISDM	0	0	0	0	0	0	0	0	0
Performance metrics	MAE	1	0	0	*	0	0	0	1	0
	Spearman correlation	1	0	0	*	0	0	0		0
	Pearson correlation	1	0	0	*	0	0	0	1	0
	Dispersion	1	0	0	*	0	0	0	0	0
	Slope	1	0	0	*	0	0	0	0	0
	Intercept	1	0	0	*	0	0	0	0	0
	RMSE	0	0	0	*	1	0	0	1	1
	Widely Applicable Information Criterion	0	1	1	*	0	0	0	0	0
	Akaike Information Criterion	0	0	0	*	0	1	0	0	0
	Deviance Information Criterion	0	0	1	*	0	0	0	0	0
	Expected predictive accuracy (LOO)	0	0	0	*	0	0	1	0	0

* Feature that could not be verified.

Table S3. Algorithms implemented in *adm* and their respective *fit_* and *tune_* functions, hyperparameters, and source packages.

Algorithm	Function	Hyparameters	Package
Artificial Neural Networks (_ann)	fit_abund_ann	size decay	*nnet*
	tune_abund_ann		
Convolutional Neural Networks (_cnn)*	fit_abund_cnn	learning_rate n_epochs batch_size validation_patience fitting_patience	*torch, luz, torchvision*
	tune_abund_cnn		
Deep Neural Networks (_dnn)*	fit_abund_dnn		
	tune_abund_dnn		
Extreme Gradient Boosting (_xgb)	fit_abund_xgb	nrounds max_depth eta gamma colsample_bytree min_child_weight subsample objective	*xgboost*
	tune_abund_xgb		
Generalized Additive Models (_gam)	fit_abund_gam	distribution; inter	*gamlss*
	tune_abund_gam		
Generalized Linear Models (_glm)	fit_abund_glm	distribution inter poly inter_order	
	tune_abund_glm		
Generalized Boosted Regression (_gbm)	fit_abund_gbm	distribution n.trees interaction.depth n.minobsinnode shrinkage	*gbm*
	tune_abund_gbm		
Random Forests (_raf)	fit_abund_raf	mtry ntree	*randomForest*
	tune_abund_raf		
Support Vector Machines (_svm)	fit_abund_svm	kernel sigma C	*kernlab*
	tune_abund_svm		

* Number of layers and neurons not taken into account, as they are parameters of *generate_arch_list*, *generate_dnn_architecture,* and *generate_cnn_architecture*.

Table S4. Source and names of variables used to perform the Principal Component Analysis.

Source	Name
Chelsa	Annual mean temperature
	Temperature Seasonality (standard deviation ×100)
	Mean Temperature of Driest Quarter
	Mean Temperature of Warmest Quarter
	Annual precipitation
	Precipitation Seasonality (Coefficient of Variation)
	Precipitation of Wettest Quarter
	Precipitation of Driest Quarter
	Precipitation of Warmest Quarter
	Climate moisture index
	Near-surface relative humidity
	Vapor pressure deficit
	Surface downwelling shortwave radiation
SRTM	Elevation
SoilGrids*	Bulk density
	Cation exchange capacity
	Volumetric fraction of coarse fragments (> 2 mm)
	Proportion of clay particles (< 0.002 mm)
	Total nitrogen (N)
	Organic carbon density
	Organic carbon stocks
	Soil pH
	Proportion of sand particles (> 0.05/0.063 mm)
	Proportion of silt particles (≥ 0.002 mm and ≤ 0.05/0.063 mm)
	Soil organic carbon content

*Variables disponible in seven depths.

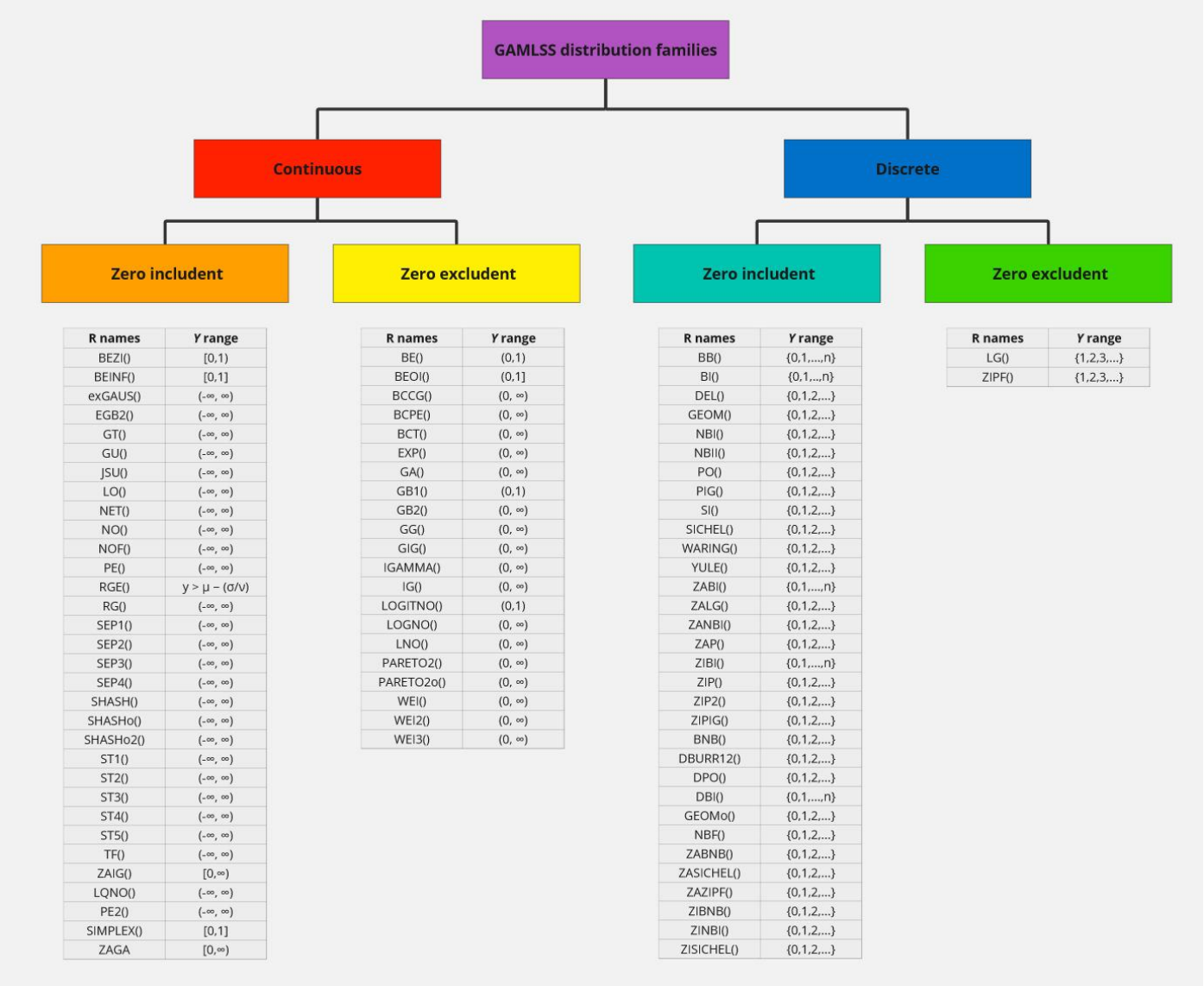
Table S5. Variance explained by principal components selected from the principal component analysis performed with climate and edaphic data.

Principal Components	Variance explained for each PC	Cumulative variance explained
1	33.44	33.44
2	26.08	59.52
3	13.12	72.64
4	8.34	80.98
5	4.84	85.82
6	3.04	88.86
7	2.04	90.90

Table S6. Performance metrics for Random Forest (RAF), Deep Neural Network (DNN), and Generalized Linear Models (GLM).

Model	MAE mean	Spearman mean	Pearson mean	Inter mean	Slope mean	PDISP mean
DNN	7.18	0.46	0.40	4.05	1.99	0.27
GLM	7.59	0.50	0.40	1.26	0.94	0.56
RAF	8.27	0.37	0.32	0.88	0.87	0.47

Figures



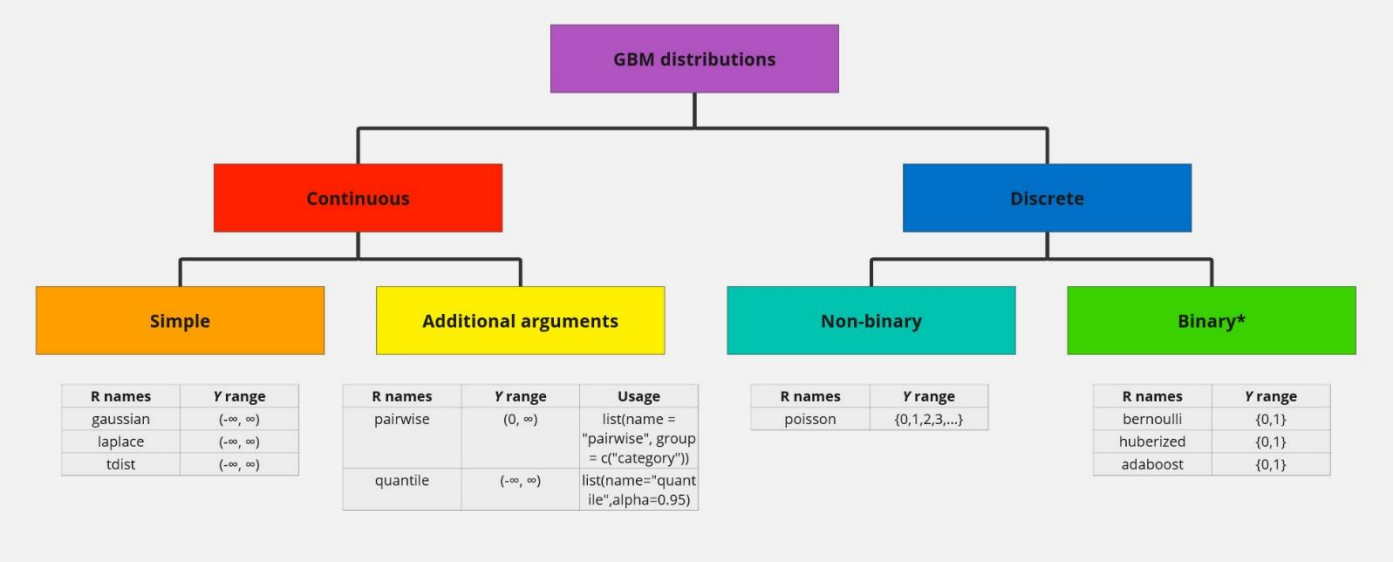
R names	Y range
BEZI()	[0,1)
BEINF()	[0,1]
exGAUS()	(-∞, ∞)
EGB2()	(-∞, ∞)
GT()	(-∞, ∞)
GU()	(-∞, ∞)
JSU()	(-∞, ∞)
LO()	(-∞, ∞)
NET()	(-∞, ∞)
NO()	(-∞, ∞)
NOF()	(-∞, ∞)
PE()	(-∞, ∞)
RGE()	y > μ − (σ/ν)
RG()	(-∞, ∞)
SEP1()	(-∞, ∞)
SEP2()	(-∞, ∞)
SEP3()	(-∞, ∞)
SEP4()	(-∞, ∞)
SHASH()	(-∞, ∞)
SHASHo()	(-∞, ∞)
SHASHo2()	(-∞, ∞)
ST1()	(-∞, ∞)
ST2()	(-∞, ∞)
ST3()	(-∞, ∞)
ST4()	(-∞, ∞)
ST5()	(-∞, ∞)
TF()	(-∞, ∞)
ZAIG()	[0,∞)
LQNO()	(-∞, ∞)
PE2()	(-∞, ∞)
SIMPLEX()	[0,1]
ZAGA	[0,∞)

R names	Y range
BE()	(0,1)
BEOI()	(0,1]
BCCG()	(0, ∞)
BCPE()	(0, ∞)
BCT()	(0, ∞)
EXP()	(0, ∞)
GA()	(0, ∞)
GB1()	(0,1)
GB2()	(0, ∞)
GG()	(0, ∞)
GIG()	(0, ∞)
IGAMMA()	(0, ∞)
IG()	(0, ∞)
LOGITNO()	(0,1)
LOGNO()	(0, ∞)
LNO()	(0, ∞)
PARETO2()	(0, ∞)
PARETO2o()	(0, ∞)
WEI()	(0, ∞)
WEI2()	(0, ∞)
WEI3()	(0, ∞)

R names	Y range
BB()	{0,1,...,n}
BI()	{0,1,..,n}
DEL()	{0,1,2,...}
GEOM()	{0,1,2,...}
NBI()	{0,1,2,...}
NBII()	{0,1,2,...}
PO()	{0,1,2,...}
PIG()	{0,1,2,...}
SI()	{0,1,2,...}
SICHEL()	{0,1,2,...}
WARING()	{0,1,2,...}
YULE()	{0,1,2,...}
ZABI()	{0,1,...,n}
ZALG()	{0,1,2,...}
ZANBI()	{0,1,2,...}
ZAP()	{0,1,2,...}
ZIBI()	{0,1,...,n}
ZIP()	{0,1,2,...}
ZIP2()	{0,1,2,...}
ZIPIG()	{0,1,2,...}
BNB()	{0,1,2,...}
DBURR12()	{0,1,2,...}
DPO()	{0,1,2,...}
DBI()	{0,1,...,n}
GEOMo()	{0,1,2,...}
NBF()	{0,1,2,...}
ZABNB()	{0,1,2,...}
ZASICHEL()	{0,1,2,...}
ZAZIPF()	{0,1,2,...}
ZIBNB()	{0,1,2,...}
ZINBI()	{0,1,2,...}
ZISICHEL()	{0,1,2,...}

R names	Y range
LG()	{1,2,3,...}
ZIPF()	{1,2,3,...}

Figure S1. Distribution families available in *gamlss* and implemented in *adm*. Other distributions may be available but were not tested. For further details, see *gamlss* and *gamlss.dist* packages help pages.



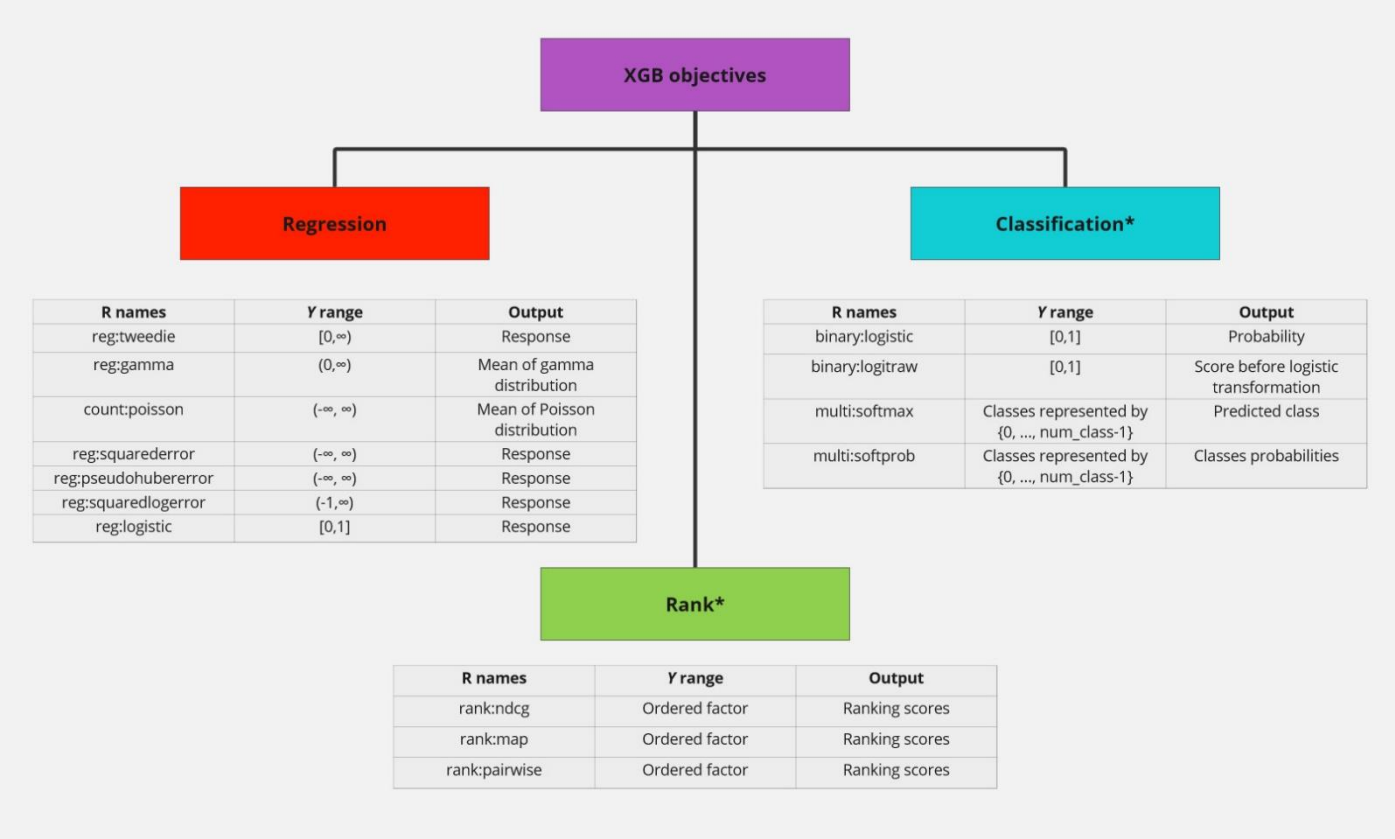
R names	*Y* range
gaussian	(-∞, ∞)
laplace	(-∞, ∞)
tdist	(-∞, ∞)

R names	*Y* range	Usage
pairwise	(0, ∞)	list(name = "pairwise", group = c("category"))
quantile	(-∞, ∞)	list(name="quantile",alpha=0.95)

R names	*Y* range
poisson	{0,1,2,3,...}

R names	*Y* range
bernoulli	{0,1}
huberized	{0,1}
adaboost	{0,1}

Figure S2. Distributions available for GBM (*gbm*).
* Not implemented in *adm*.



R names	*Y* range	Output
reg:tweedie	[0,∞)	Response
reg:gamma	(0,∞)	Mean of gamma distribution
count:poisson	(-∞, ∞)	Mean of Poisson distribution
reg:squarederror	(-∞, ∞)	Response
reg:pseudohubererror	(-∞, ∞)	Response
reg:squaredlogerror	(-1,∞)	Response
reg:logistic	[0,1]	Response

R names	*Y* range	Output
binary:logistic	[0,1]	Probability
binary:logitraw	[0,1]	Score before logistic transformation
multi:softmax	Classes represented by {0, ..., num_class-1}	Predicted class
multi:softprob	Classes represented by {0, ..., num_class-1}	Classes probabilities

R names	*Y* range	Output
rank:ndcg	Ordered factor	Ranking scores
rank:map	Ordered factor	Ranking scores
rank:pairwise	Ordered factor	Ranking scores

Figure S3. Objectives available for XGB (*xgboost*).
* Not implemented in *adm*.

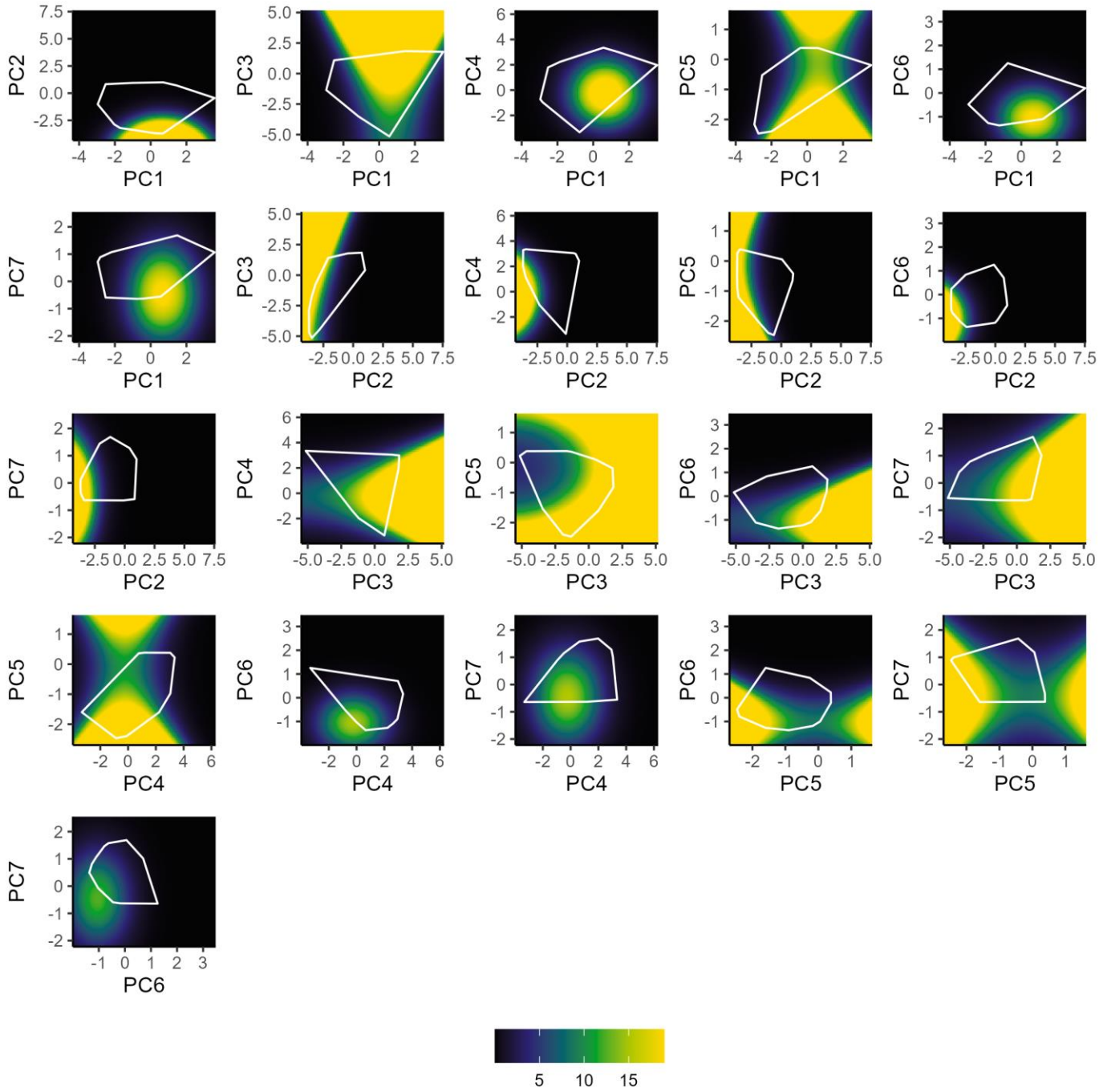


Figure S4. Bivariate partial dependence plots of a Generalized Linear Model used for modeling *Cynophalla retusa* abundance. The white polygon indicates the training conditions boundaries. For visualization purposes, color scale was truncated to approximate Figure 2.

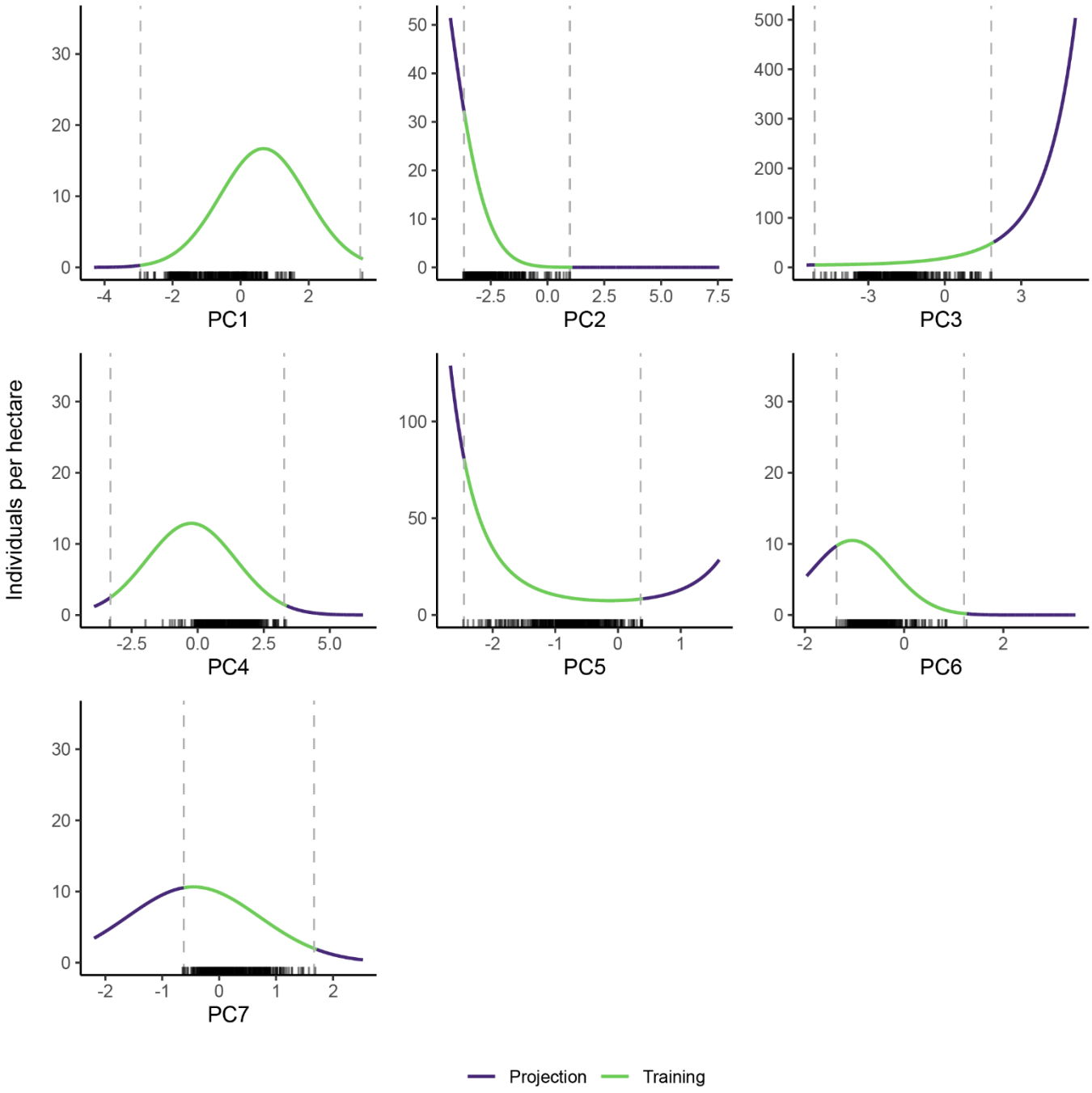


Figure S5. Univariate dependence plots of a Generalized Linear Model used for modeling *Cynophalla retusa* abundance. The green section indicates the variable values seen in training, while the purple sections indicate variable values only seen in projection.

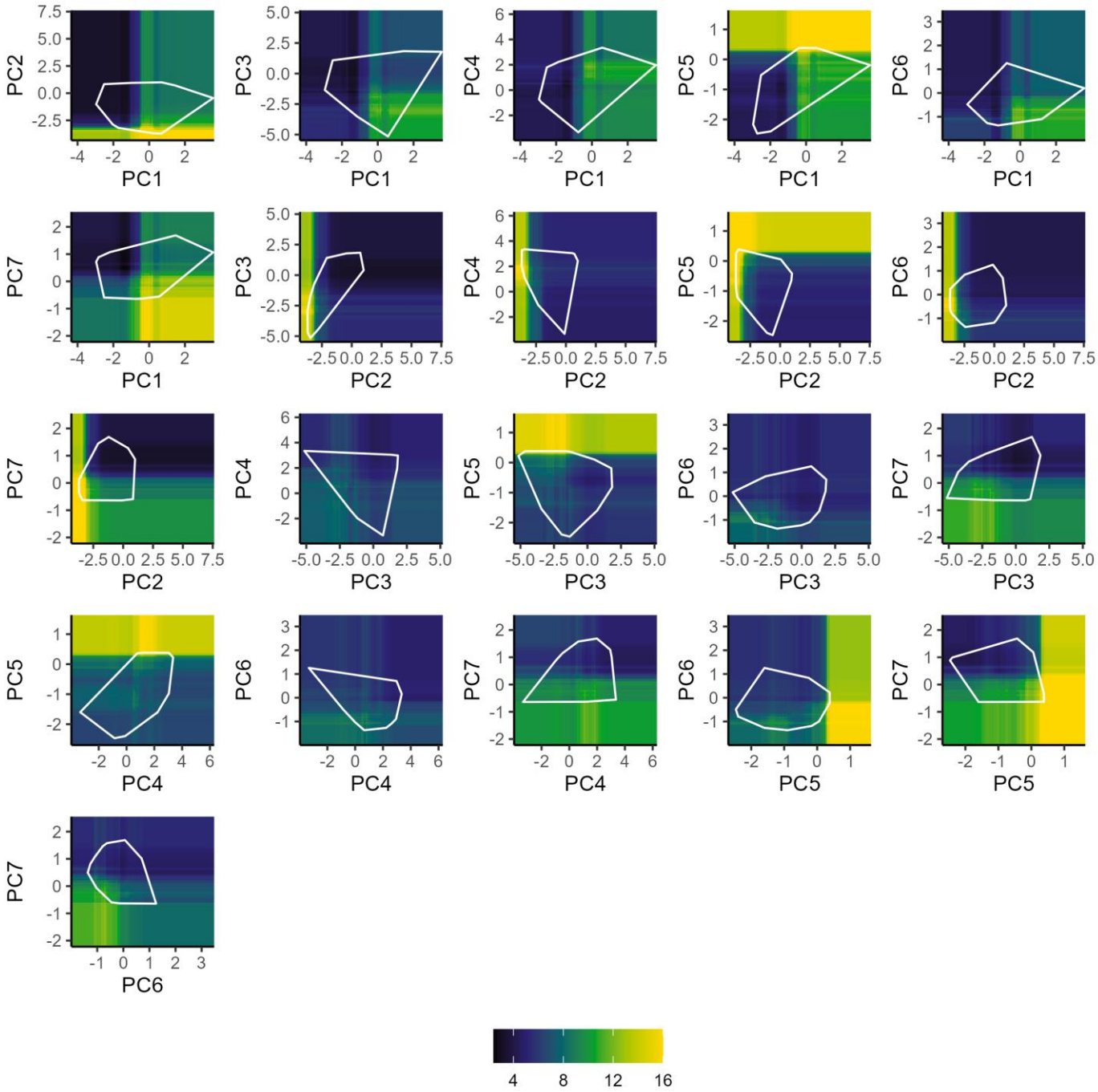


Figure S6. Bivariate partial dependence plots of a Random Forest used to model *Cynophalla retusa* abundance. The white polygon indicates the training conditions boundaries. For visualization purposes, color scale was truncated to approximate Figure 2.

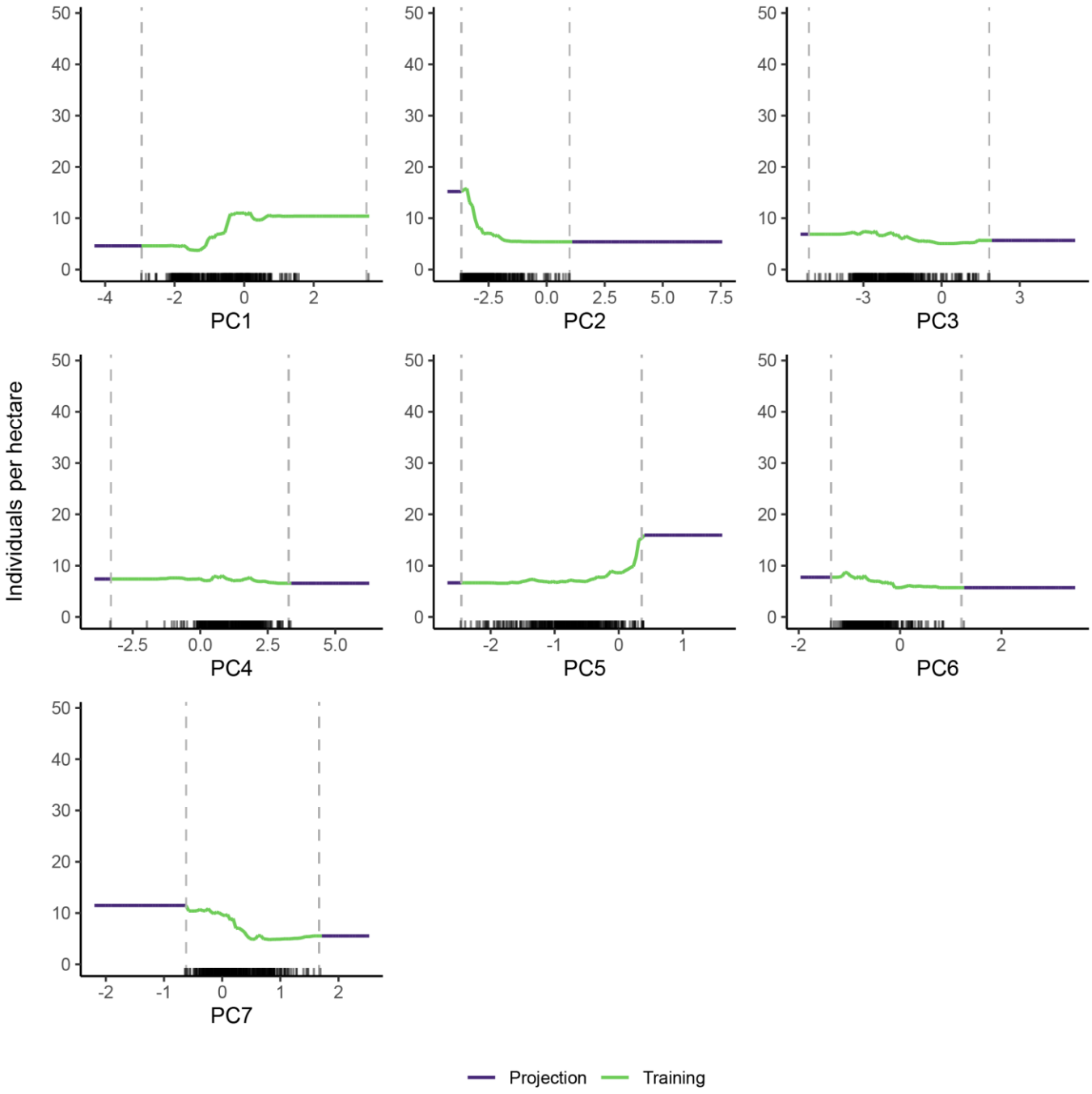


Figure S7. Univariate partial dependence plots of a Random Forest a used to model *Cynophalla retusa* abundance. The green section indicates the variable values seen in training, while the purple sections indicate variable values only seen in projection.

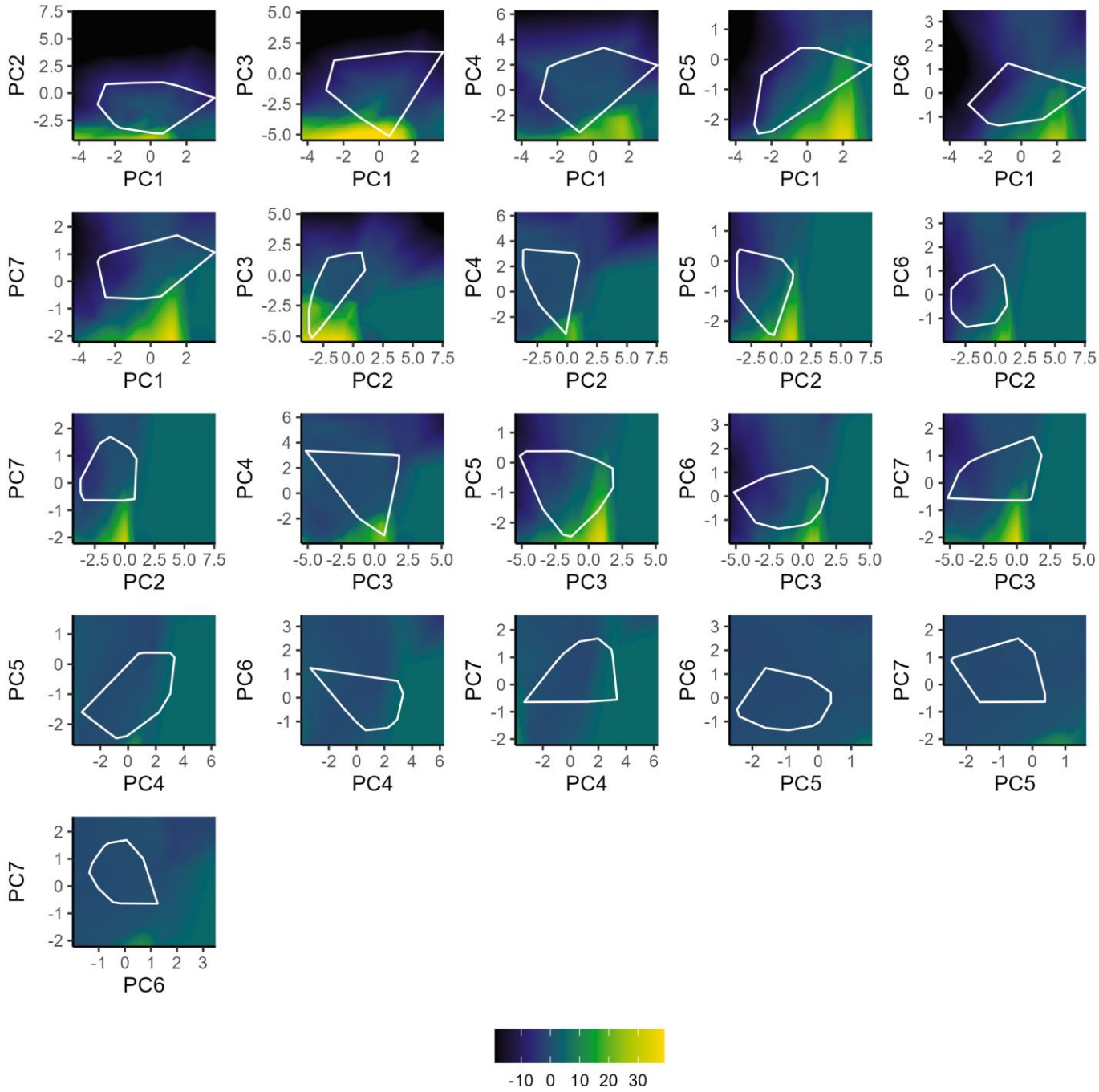


Figure S8. Bivariate partial dependence plots of a Deep Neural Network used for modeling *Cynophalla retusa* abundance. The white polygon indicates the training conditions boundaries. For visualization purposes, color scale was truncated to approximate Figure 2.

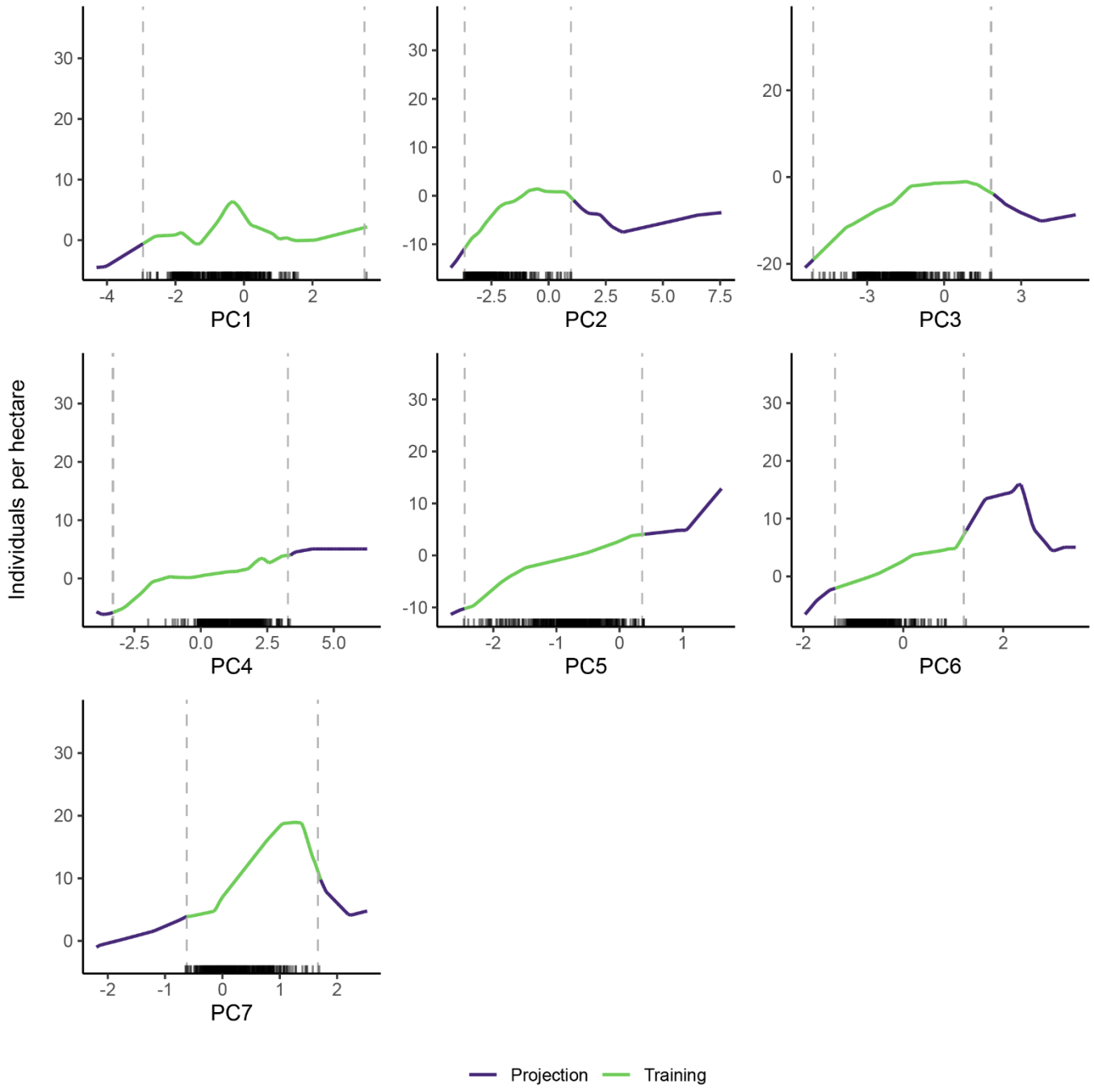


Figure S9. Univariate partial dependence plots of the Deep Neural Network used for modeling *Cynophalla retusa* abundanced. The green section indicates the variable values seen in training, while the purple sections indicate variable values only seen in projection.

4 CAPÍTULO II

Using deep learning to model species abundance in the context of Gran Chaco[3]

Introduction

Studying species distribution is important for the understanding of biogeographical and ecological patterns (Miller, 2010), and for the development of more effective conservation strategies (Guisan et al., 2013; Sofaer et al., 2019). Different approaches were proposed for estimating species distributions and their attributes (e.g., environmental suitability, abundance, traits, or genetic diversity, Číhal, 2023; Marcer et al., 2016; Pollock et al., 2018; Yu et al., 2020), being Species Distribution Models (SDM) one of the most widely used approaches. SDM correlates presence and absence data (or presence and pseudo-absences) with environmental data to model the ecological niche of species and project species environmental suitability on geographical space (Guisan et al., 2013; R. D. Holt, 2009; Soberón & Nakamura, 2009). However, despite SDM importance, this approach provides information only on environmental suitability, ignoring population aspects, like density and trends (Hastings et al., 2020).

Recently, techniques for Abundance-based Distribution Models (ADMs) have been proposed. ADM correlate species abundance with environmental data to predict abundance throughout geographic space (Ehrlén & Morris, 2015; Howard et al., 2014; Yu et al., 2020) allowing for capturing aspects of species' distribution not considered by SDM (Mi et al., 2017; Waldock et al., 2022). For instance, even if a species distributes in a region, its abudancemay distribute unequally throughout it, due to ecological and environmental factors (e.g., spatial heterogeneity, biological interactions) (Borregaard & Rahbek, 2010; Holt et al., 2004). Considering abundance variates throughout a species distribution has direct implications for species conservation planning, as distribution may remain constant even if the population is declining (Hastings et al., 2020). Moreover, while some studies have demonstrated a positive correlation between environmental suitability and local abundance (e.g., De La Fuente et al., 2021; Weber et al., 2017), others have found little or no correlation (e.g., Dallas; Hastings, 2018; Sporbert et al., 2020).

Despite the study of ADM being relevant to ecology, biogeography, and conservation, ADM are still methodologically unexplored and undeveloped compared to conventional SDM (Waldock et al., 2022). For example, modeling aspects such as sample size (Moudrý et al., 2024),

[3] Artigo formatado seguindo as normas da revista *Methods in Ecology and Evolution*.

effects of absence data on models (Ogurtsov, 2023; Stokland et al., 2011), and data partition (Roberts et al., 2017) have been explored in SDM literature for decades, while those factors effects remaining unclear for ADM. For example, is known that, for SDM, algorithms could be sensitive to the amount of absence data and how they are sampled, thus many methods have been developed to sampling and balancing absence or pseudo-absence (Barbet-Massin et al., 2012; Liu et al., 2019). Moreover, spatially structured data partitioning is assumed to allow a more rigorous model evaluation than random partitioning, such as k-fold, as it also accounts for models' spatial transferability (i.e., the capacity of a model to predict in environmental conditions or regions not used in model training) (Huang et al., 2024; Roberts et al., 2017). Both aspects, absence data amount and data partitioning have important methodological and ecological implications but are still unexplored in ADM.

Several modeling techniques have been employed for building ADM, e.g., Generalized Linear Models (GLM) (Nelder & Wedderburn, 1972), Generalized Additive Models (GAM) (Hastie & Tibshirani, 1986), Random Forest (RAF) (Breiman, 2001), Gradient Boosting Machine (GBM) (Friedman, 2001), Support Vector Machine (SVM) (Boser et al., 1992), and Artificial Neural Networks (ANN) (Rumelhart et al., 1986). However, a few examples in the literature explored and compared different algorithms performance in the ADM context. Waldock et al. (2022) evaluated the performance of different algorithms for ADM construction, excluding ANN, while Botella et al. (2018) developed ADM using ANN framework and compared it with Maximum Entropy, a common algorithm for SDM, adapted to predict species environmental suitability. Therefore, although several algorithms have been tested, ANN are still underexplored, even though they are a promising technique.

ANN have a wide variety of possible neuron organizations and connections, representing diverse structures and adaptability (Borowiec et al., 2022; Pichler & Hartig, 2023). ANN consist of interconnected nodes, in various layers or not, called neurons, which process data and transmit results to other neurons (LeCun et al., 2015). Quantity and type architecture (i.e., organizations of neurons and layers) are not predefined and must be defined (Alzubaidi et al., 2021). When the network features multiple layers between input and output (hidden layers), ANN is called a deep network (Borowiec et al., 2022). ANN composed of multiple fully connected sequential hidden layers that function with feedforward and back-propagation are often called Deep Neural Networks (DNN) (Alom et al., 2019). When the hidden layers perform convolution operations, with kernels and filters scanning a matrix, instead of tabular data as DNN, it is often called a Convolutional Neural Network (CNN) (Alzubaidi et al., 2021). Conversely, when neural networks are constructed with a single hidden-layer they could be named a Shallow Neural Network (NET). NET are smaller versions of DNN and work similarly but with much lower computational cost (Podder et al., 2021).

ANN have been widely employed in ecology for prediction and classification tasks, particularly in species and individual recognition through images and sounds, as well as forecasting responses to environmental variables, presenting good performance (Borowiec et al., 2022; Pichler & Hartig, 2023). In different distribution modeling approaches, several research used DNN (Rew et al., 2021), NET (Lek & Guégan, 1999), and CNN (Deneu et al., 2021). However, those types are still relatively underexplored in ADM.

Predicting abundance patterns is relevant in regions with high diversity and natural cover loss, common in tropical and subtropical areas (Edwards et al., 2019). The *Gran Chaco* is an ecoregion with xeromorphic vegetation that extends into northern Argentina, western Paraguay, southeastern Bolivia, and a small portion of southwestern Brazil (Prado, 1993). The *Gran Chaco* presents high biodiversity (Redford et al., 1990); however, it suffers from few protected areas (Nori et al., 2016) and an alarming rate of natural cover loss, mainly because of industrial agriculture and infrastructure expansion (Barral et al., 2020; Gasparri & Grau, 2009). To contribute to a better understanding of ADM, here we used abundance data of 117 tree species of the *Gran Chaco* to i) explore ADM modeling techniques based on nine algorithms with emphasis on different ANN forms (DNN, NET, and CNN); and ii) evaluated how algorithms perform in different data-partition and absence data quantity.

Methods

General experiment workflow

In this experiment, we used an abundance database of 117 tree species based on forest inventories from Paraguay, Argentina, and southern Brazil. ADM were constructed using nine algorithms. To evaluate how absences quantity of training data (i.e., plots with no species abundance) affects models' performance, we tested three absence amounts, i) Two-absences: two absences per presence (i.e., plots with abundance >0), ii) One-absence: one absence per presence, and iii) No-absence: only presence. To test the data partition effects, we partitioned each dataset randomly (K-folds) and geographically (Spatial blocks). Our experiment has a factorial design where all levels of three factors were combined, and species were used as experimental units totalizing 6,318 models (Figure 1).

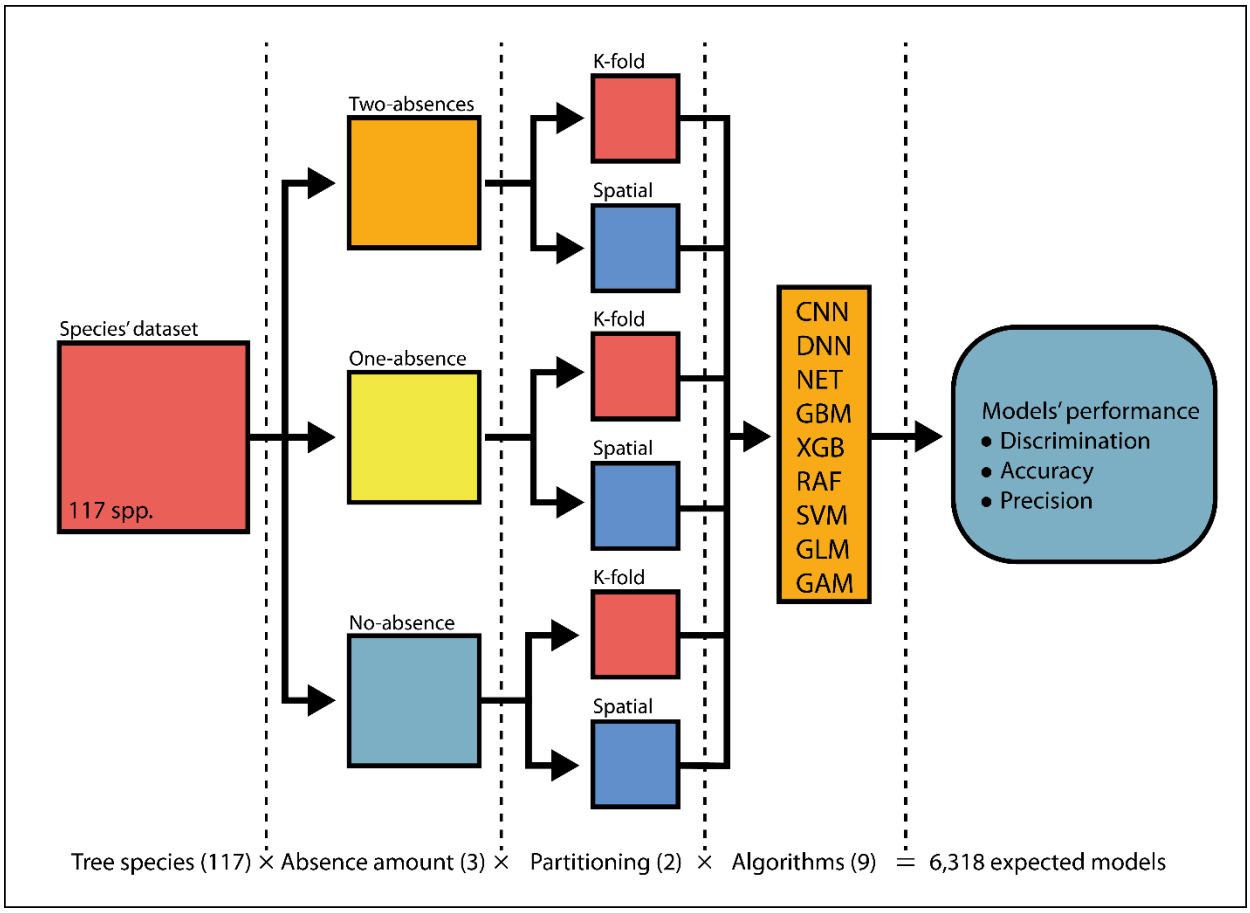


Figure 1. Overview of the experiment used to evaluate the effect of absence amount in the dataset, data partitioning, and algorithms on ADM performance.

Study area

Some consider the *Gran Chaco* as the largest continuous dry forest in the world (Olson et al., 2001) and the second-largest biome in South America, covering 1.3 million km² (Bucher & Huszar, 1999). The *Gran Chaco* (which encompasses dry, humid, and mountainous *Chaco*) is a highly diverse region, with predominantly open and xeromorphic vegetation and a semi-arid climate (mean annual precipitation ranging from 450-1200 mm), hot summers (mean temperature of the warmest month of 26-28 C°), and winter frosts (coldest month presents mean temperature of 12-17 C° and up to 28 days of frost) (Bucher, 1982). However, the region is becoming warmer and drier, due to climate changes, which also increases the risk of extreme wildfires, endangering vegetation (Feron et al., 2024). Common tree species in the region belong to the genera *Aspidosperma* (Apocynaceae), *Cereus* (Cactacea), *Copernicia* (Arecaceae), *Prosopis* (Fabaceae), *Schinopsis* (Anacardiaceae), *Tabebuia* (Bignoniaceae), *Trithrinax* (Arecaceae), and *Bulnesia* (Zygophyllaceae) (Borghetti et al., 2023; Prado, 1993).

Environmental data

To construct the ADM, climatic, edaphic, and elevation environmental variables were used. The edaphic consisted of 11 physical and chemical variables obtained from SoilGrids for seven soil depths (Poggio et al., 2021) at 250m resolution. Climatic variable consisted of 19 bioclimatic variables sourced by Chelsa v1.2 (Karger & Zimmer, 2019). Elevation was sourced by SRTM (https://srtm.csi.cgiar.org) (Table S2). Climatic and elevation were at 1km resolution. Edaphic data was upscaled to 1km by averaging 250m cell values. We calculated a pairwise Pearson's correlation matrix and for every pair of variables with correlation > |0.7|, the one with greater biological significance was kept. This process resulted in 13 climatic variables, 11 edaphic, and elevation. We used Principal component analysis to avoid any problem of model multicollinearity and reduce the number of variables. Principal component analysis was calculated based on a correlation matrix using the 25 variables (Table S2). We selected seven principal components that explained > 90% of the original data variance (Table S3) (De Marco & Nóbrega, 2018). Eigenvectors were used to calculate the scores of each derived principal component to be used as new predictors.

Abundance data

We compiled abundance data for 349 species over four forestry surveys of three countries (Paraguay: INFONA, 2024; Argentina: MAySD, 2022; SAyDS, 2005; Brazil, Santa Catarina: SFB, 2018). While our study area is limited to the *Gran Chaco*, abundance data from geographically adjacent regions was also used in ADM construction to better capture the environmental conditions species, increase abundance data, and enhance the models' performance (Sánchez-Fernández et al., 2011). Abundance was calculated to express individuals per hectare (ind/ha). Geographical coordinates of the forest inventory plots were converted into decimal degrees with WGS 84 projection (EPSG:4326). Then, all data were integrated into a unified database.

To cleanse errors and assure data quality, taxonomic and geographic cleaning protocols were performed. First, problematic coordinates were corrected. Then, species scientific names were corrected, standardized, and updated using *LCVP* R package (Freiberg et al., 2020), using Leipzig Catalog of Vascular Plants (Freiberg et al., 2020) as the taxonomic authority. Finally, we kept only the species natives to the study area and with ≥ 50 location with presences (ind/ha > 0). 232 were excluded and 117 species remained (Table S1 for species list and their summary statistics). We used *Flora Argentina y del Conosur* platform (http://www.floraargentina.edu.ar) and Tropicos Paraguay (http://legacy.tropicos.org/Project/Paraguay) to get species origin information.

Data refining and dataset construction

For each species a dataset was constructed combining plots with abundance and absences, assuming as absences plots with 0 ind/ha. Using absence data from regions that have been inaccessible to species could inflate model performance metrics and introduce artifacts to model predictions (VanDerWal et al., 2009). Moreover, well-defined calibration areas are impactful to distribution models performance (Luna et al., 2024), so in the same way that the SDM training areas geographically constrain absences (or pseudo-absence) data used to fit models, we delimited each species training area by a 100 km buffer around the species presences (i.e., ind/ha > 0) and restricted the absences within this area.

To test the impact of the absence amounts in training data, we randomly thinned absence points to be double or equal to the number of presences, i.e., Two-absence and One-absence absences amount, respectively. Finally, we removed every absence and created the No-absence dataset (Figure 2).

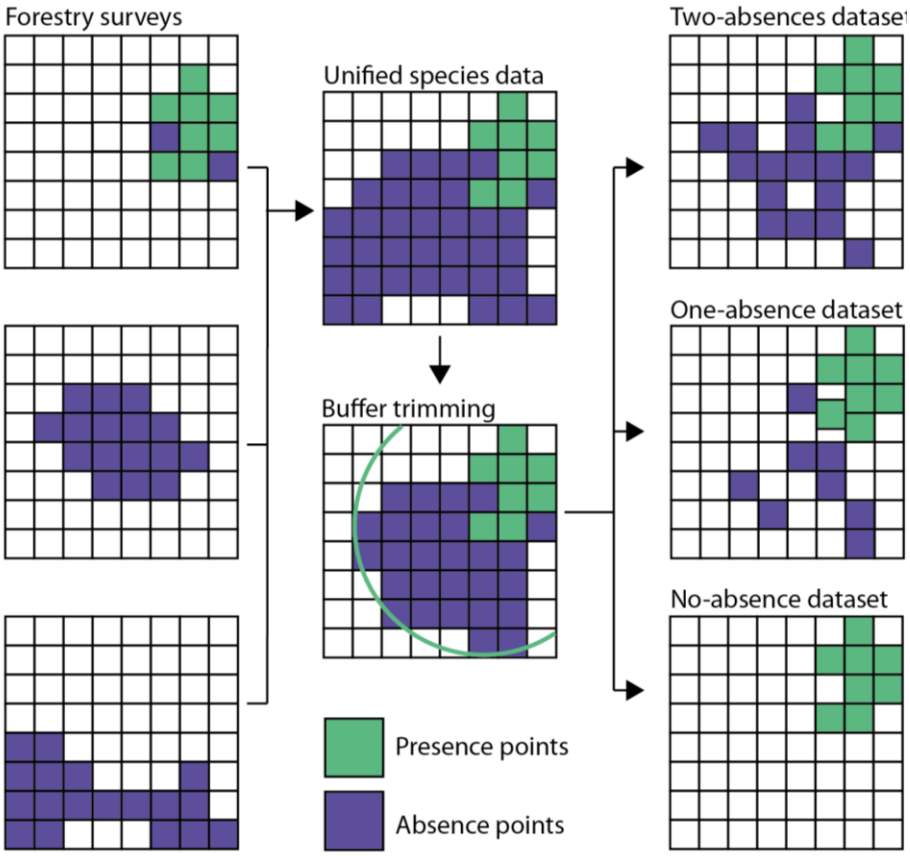


Figure 2. Representation of dataset construction for each absence amount tested.

Data partition

We used two data partitioning approaches, k-fold and spatial block cross-validation. The former is a random partitioning approach, while the second is spatially structured (Figure S1). For k-fold approach, we used 5 folds, while for spatial block cross-validation, we used three spatial blocks in a checkerboard scheme. We tested 32 block sizes and the best size was the one with the lower spatial autocorrelation, highest environmental similarity, and lower standard deviation of data amount among partitions (Velazco et al., 2019). This size selection was performed using *part_sblock* function of *flexsdm* R package (Velazco et al., 2022).

Model fitting and validation metrics

Because algorithm hyperparameters (i.e., those model parameters that are defined before model fitting) could affect model performance (Fourcade, 2021; Schratz et al., 2019), it is essential to perform tuning to select the best hyperparameters combination. We performed a grid search tuning exploring all possible combinations between predefined hyperparameter values (Table S4). We made three selections of best hyperparameter combinations for each species, selecting combinations that maximized models discrimination, accuracy, and precision, measured by Spearman's rank correlation (Spearman), Mean Absolute Error (MAE), and Dispersion (PDISP), respectively (Waldock et al. 2022).

MAE was calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where n is the total number of samples, y_i is the observed value for the i-th sample, and $\hat{y}_i$ is the predicted value for the i-th sample. MAE values range interval is inherently $[0, \infty^+)$, but values depend on the scale of the response variable. Thus, comparisons between species with different abundance ranges may be difficult or inappropriate. For this reason, we decided to normalize the MAE values, scaling it to $[0, 1]$, allowing better comparisons. To perform this, we applied the following equation:

$$MAE_{normalized} = \frac{MAE}{\max(y) - \min(y)}$$

Where $\max(y)$ and $\min(y)$ are the maximum and minimum values of the species abundance, respectively.

PDISP was based on Waldock et al. (2022) and is calculated as:

$$PDISP = \frac{\sigma(\hat{y})}{\sigma(y)}$$

Where $\sigma(\hat{y})$ is the standard deviation of predicted values and $\sigma(y)$ is the standard deviation of observed values.

A model will show higher performance in discrimination when Spearman value is closer to 1, higher performance in accuracy when MAE is closer to 0, and higher performance in precision when PDISP is closer to 1. Full ADM modeling protocol was constructed with *adm* and *flexsdm* R package. Because some models did not converge for some species, we were able to fit 6,234 models (i.e., 98.7% of expected models).

Data analysis

Experiment results were analyzed using Generalized Additive Models for Location, Scale, and Shape approach (GAMLSS, Rigby & Stasinopoulos, 2005) using the *gamlss* R package (Stasinopoulos & Rigby, 2012). GAMLSS offer the benefit of providing > 100 distribution families and allowing great flexibility by modeling explicitly any parameter that defines a distribution family regarding predictor variables (Rigby & Stasinopoulos, 2005). We constructed GLM for each performance metric (i.e., Spearman, MAE, and PDISP) using algorithms, absences amount, and partition method as predictor variables (i.e., fixed effects). To deal with the lack of independence of data, we included the species as a random effect. Several models were constructed for each response, varying in terms of family distributions, predictor interactions, and model parametrization. We tested different family distributions available in *gamlss*, exploring various parameter settings within these families to optimize model fit. Two-way interactions between predictor variables were also tested, to examine whether combining certain predictors would better explain variations in the response variable. The final model was selected based on both a visual analysis of residuals normality and homoscedasticity and Akaike Information Criterion. We then performed a post-hoc analysis by estimating means using the *emmeans* R package (Lenth, 2017) and conducting pairwise mean comparisons by HDS Tukey test between algorithms for different combinations of absence amount and partition type. Pairwise mean comparisons between partitions for different combinations of algorithms and absence amount, and between absence amount for different combinations of partition type and algorithm were also performed. Since *emmeans* does not deal with *gamlss* random effects, the models used to estimate means were refitted without random effects (Figures S2-S4 for models' residuals and formula, and Table S5 for all estimated means values).

Results

For discrimination, in all treatments, DNN, GBM, and XGB outperformed other algorithms while CNN was the worst (Figure 3). CNN was the only algorithm for which using no absences in the dataset increased its performance, but it was considerably low, compared to the others (Figure 3). Overall, the best-performing algorithm was DNN, but XGB often scored similarly to DNN for discrimination. The GBM, GAM, and GLM models tended to perform similarly, often being not different from XGB. SVM performed a little worse than other algorithms, but often better than RAF (Figure 3), while CNN was the worst algorithm. Focusing on ANN approaches, DNN had the best performance in every treatment, followed by NET and CNN (Figure 3). Regarding the frequency of best algorithms for discrimination, we found DNN was the best model regarding discrimination for all absence amounts in spatial partition and for No-absence in K-fold partition (Figure S5); while XGB was the best algorithm for One-absence and Two-absence for K-fold partition (Figure S5).

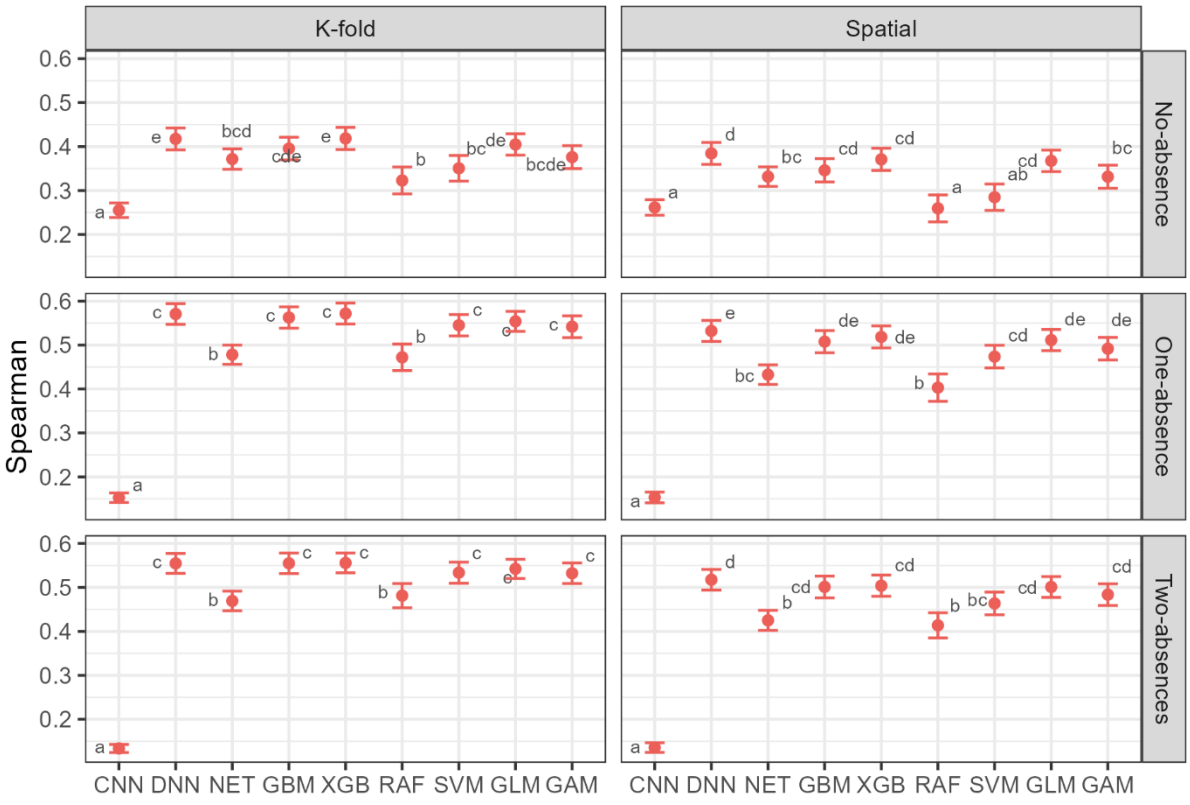


Figure 3. Post-hoc analysis of model discrimination based on Spearman correlation metric for different algorithms, partition type (columns), and absence amount (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

For accuracy, DNN and XGB were the best-performing algorithms, but SVM, GBM, GLM, and GAM performed similarly (Figure 4). NET and RAF were the worst-performing algorithms in all treatments. CNN tended to perform better than RAF and NET, and better than the other algorithms. Regarding the ANN, DNN presented the highest accuracy, followed by CNN and NET (Figure 4). DNN had the highest accuracy for most species in all treatments (Figure S6).

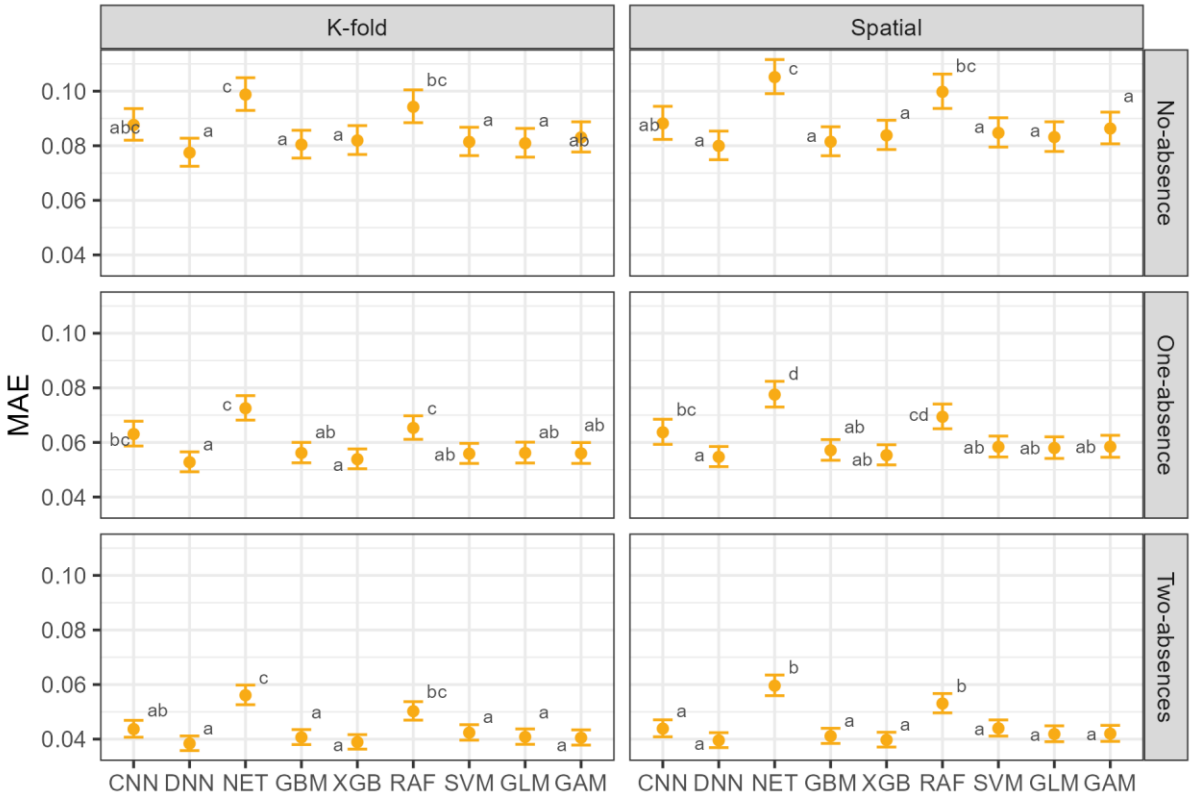


Figure 4. Post-hoc analysis of model accuracy based on MAE (normalized) metric for different algorithms, partition type (columns), and absence amount (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

For precision, algorithms tended to score much similarly among treatments (Figure 5). CNN was the worst algorithm in every treatment, except for No-absence. RAF often just outperformed CNN, being statistically different from all other algorithms. SVM tended to perform better than RAF and CNN, but worse than the other algorithms (Figure 5). NET and XGB had the highest precision for all treatments, followed by DNN and GLM. NET had the highest number of species with the highest precision for all treatments (Figure S7).

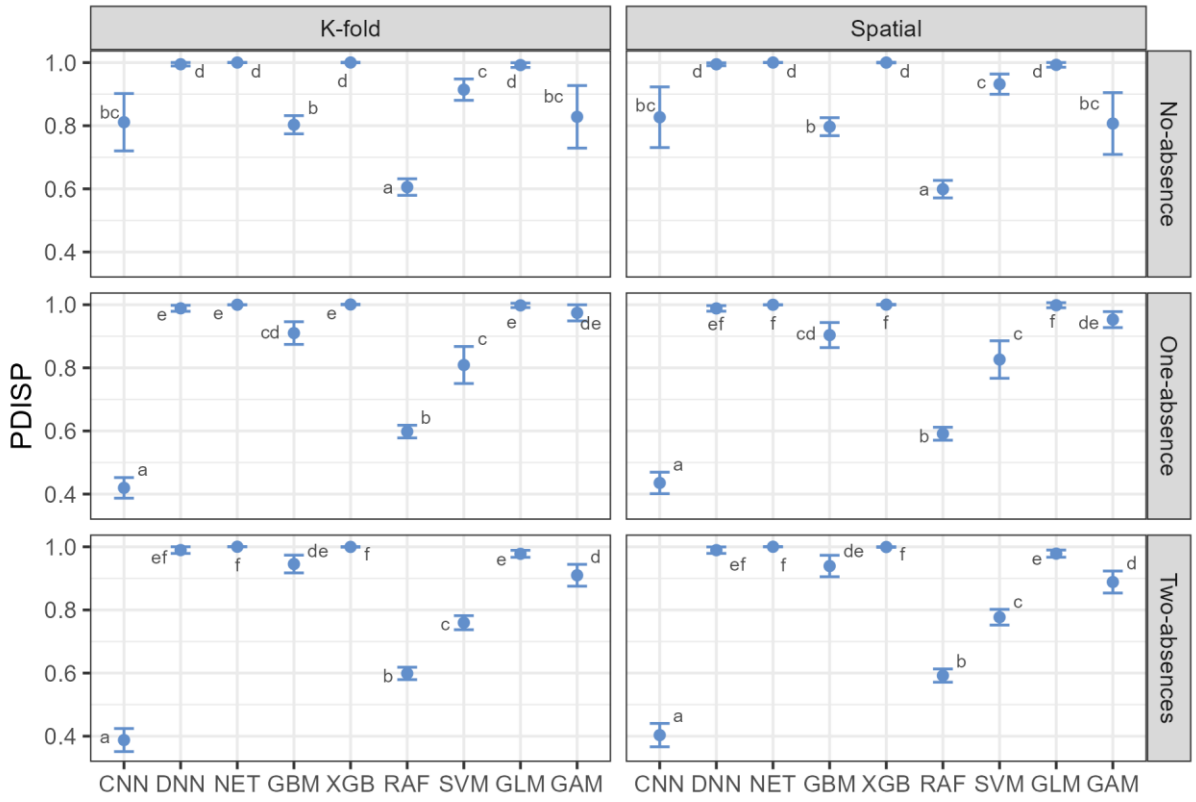


Figure 5. Post-hoc analysis of model precision based on PDISP metric for different algorithms, partition type (columns), and absence amount (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

Regarding data partitioning, we found that algorithms tended to have similar performance in spatial partition than in random partition in terms of discrimination, accuracy, and precision (Figures S8-S10). When algorithms are compared regarding absence amount, we found that performance metrics had different tendencies. For discrimination, algorithms performed similarly in One-absence and Two-absences, but the performance in No-absence was the worst for all algorithms (Figure S11). For accuracy, algorithms tended to score better in Two-absence dataset, and relatively higher in One-absence. No-absence was again the worst condition for all algorithms (Figure S12). Variation among absences amount had a small effect on precision, being the performances algorithm-dependent, with low impact of treatments (Figure S13). Thus, results indicate an effect of absence amount and data partitioning in discrimination and accuracy, but no clear tendency was found for precision.

Discussion

In our study, we aimed to the predictive performance of different ADM algorithms with emphasis on ANN and how they are affected by different absences amounts and data partitioning using 117 tree species from the *Gran Chaco*. We found that using no absence points can be detrimental to models' discrimination and accuracy for all algorithms, while impacts on models' precision seemed to be algorithm-dependent. Spatial partitioning presented a little lower performance than random partitioning. Regarding ANN algorithms, we found mixed results, with DNN scoring similarly or better in discrimination, accuracy, and precision when compared to other algorithms, NET had a good precision score, but low performance for discrimination and accuracy, and CNN had an overall low performance in all metrics. XGB tended to perform similarly to DNN in all metrics and better than the other algorithms. The following best algorithms were GBM, GLM, and GAM which performed similarly in all metrics. CNN, RAF, and SVM tended to be the worst-performing algorithms for all metrics.

Few works in the literature conducted extensive experiments on algorithm performance when building ADM (Ngor et al., 2023; Waldock et al., 2022). Waldock et al. (2022) used GAM, GLM, RAF, and GBM algorithms and observed that most models performed poorly, although some presented good discrimination (Sperman's and Pearson's correlation), accuracy (MAE), and precision metrics (Waldock et al., 2022). These authors found that RAF was the best-performing algorithm. Similarly, NGOR et al. (2023) evaluated the performance of ANN, GLM, RAF, and SVM to predict fish abundance and found that they tended to perform well in accuracy (Root Mean Squared Error – RMSE – and MAE) and discrimination (R^2), with RAF being the best algorithm. Other previous work also found good performance for RAF (Catucci et al., 2025; Mi et al., 2017). Catucci et al. (2025) predicted the red shrimps *Aristeus antennatus* (Aristeidae) and *Aristaeomorpha foliacea* (Aristeidae) biomass in the Mediterranean Sea using RAF and found good discrimination ($R^2 \approx 0.63$ and 0.74 for *A. antennatus* and *A. foliacea*, respectively) for both models. RAF also performed well for accuracy ($RMSE \approx 1.86$ and 1.04 for for *A. antennatus* and *A. foliacea*, respectively). Mi et al. (2017) constructed ADM for *Otis tarda dybowskii* (Otididae) using RAF and the algorithm performed well in both discrimination ($R^2 \approx 0.84$) and accuracy metrics ($RMSE \approx 26.54$). Several research reported that RAF had a good performance (Catucci et al., 2025; Ngor et al., 2023; Waldock et al., 2022), which is contrary to our findings. This could indicate that RAF had a lower capacity in modeling the underlying complex variables interactions, a characteristic of species abundance and distribution relationship (Borregaard & Rahbek, 2010), when compared to other algorithms, especially DNN, known for the potential in modeling highly complex ecological data (Borowiec et al., 2022).

GAM and GLM are other recurrent algorithms for ADM (Kroetz et al., 2025; Yen et al., 2004; Young & Carr, 2015). Young & Carr (2015) used GAM to predict abundance of seven fish species and got mixed performances of discrimination, with Pearson's correlation ranging from 0.26 to 0.60. Yen et al., (2004) used GLM to predict *Brachyramphus marmoratus* (Alcidae) density and evaluated the algorithm accuracy through a backfitting approach. The authors found that GLM underperformed Classification and Regression Trees (algorithm not included in our study) (Yen et al., 2004). Similarly, in other research modeling multiple species, GAM and GLM performed well but underperformed other algorithms (Ngor et al., 2023; Waldock et al., 2022). Kroetz et al. (2025) used Boosted Regression Trees (equivalent to our GBM) to predict sawfish abundance and concluded that the algorithm performed well in discrimination and accuracy.

Few works used XGB, GBM and SVM (Elmendorf & Moore, 2008; García-Gómez et al., 2023; Martín et al., 2021). García-Gómez et al. (2023) modeled fish larvae abundance with XGB and found that the algorithm performed well in accuracy (Root Mean Log Error ≅ 1.85) ; however, it consistently underestimated the species abundance. Elmendorf & Moore (2008) used an equivalent to GBM to construct ADM for 100 plant species concluding that GBM tends to have good discrimination (Spearman ranging from -0.52 to 0.61). Martín et al., (2021) predicted abundance of *Puffinus mauretanicus* (Procellariidae) with GBM, XGB and SVM, and found that SVM tended to underperform both XGB and GBM in terms of accuracy (RMSE) and discrimination (R^2). These results are consistent with ours, as our GAM and GLM models tended to have a medium performance, and GBM and XGB performed better, compared to the other algorithms (Figure 3).

Regarding ANN approaches, previous research reported ANN generally have good performance (Elmendorf & Moore, 2008; Ngor et al., 2023; Yen et al., 2004). Rocha et al. (2017) constructed a large DNN experiment, testing many architectures and predicting abundance for different macroalgae species, and found that DNN performed well in accuracy (Mean Squared Error ranging 0.018-0.060) and discrimination (R^2 ranging 0.29-0.83). Botella et al. (2018) explored DNN, CNN, and NET performance to predict plant species abundance, and found that all algorithms had good accuracy (RMSE from 2.20 to 3.98), but DNN and CNN outperformed NET. Consistent with our findings, other research found a low performance for NET (Elmendorf & Moore, 2008; Ngor et al., 2023), especially when compared to DNN. Contrary to our result, CNN presented good performance in several works (Botella et al., 2018; Deneu et al., 2021). The underperformance of CNN could be due to the simplicity of our architectures tested, which were smaller in the number of neurons and hidden-layers used in other works (Deneu et al., 2021, 2021; Hu et al., 2025). This reveals the necessity of proper planning in CNN architecture when constructing ADM. CNN are particularly more complex to fit than DNN and NET because they

work based on image-like matrices, which greatly increases the network parameters amount, the quantity of data needed, and the computational time (Alom et al., 2019; Alzubaidi et al., 2021; Martínez-González et al., 2020). These factors might make DNN and NET better options than CNN for many scenarios, as they can perform similarly or better than CNN. However, DNN and NET cannot interpret complex spatial features as CNN (Deneu et al., 2021). The good performance of DNN observed here and in previous work (Botella et al., 2018; Rocha et al., 2017; Yen et al., 2004) highlights the potential of DNN to construct ADM.

Based on our findings and those found in previous works, research performed different analysis, used several algorithms, data treatments, and transformations, with no single best algorithm. This aligned with the idea that there is no ideal algorithm to deal with all modeling conditions; in short, there is no silver bullet (Qiao et al., 2015). Previous research used a high variety of model performance metrics which denotes the need for further research to detect performance metrics suitable to ADM, something that was already explored in SDM (Abrego & Ovaskainen, 2023; Fourcade et al., 2018). We highlight that ADM performance should be evaluated with metrics related to discrimination, accuracy, and precision, as these measures carry important ecological implications that could limit or enable a model's use, depending on the objectives. Models' discrimination describes how good a model is in distinguishing low and high abundance (Norberg et al., 2019; Waldock et al., 2022). A model with low discrimination can produce erroneous and misleading spatial projections of the species abundance. Models' accuracy regards models' capacity to predict values close to the observed abundance (Waldock et al., 2022) indicating a model under or overpredicts abundance. This error could be especially dangerous when modeling rare species, as it could falsely predict large abundance in regions with few individuals, or absence in highly dense sites. Models' precision measures a model capacity to predict values with variation closer to the species' observed abundance variation (Waldock et al., 2022). Thus, a low precision represents a model that fails to capture the actual variability in species abundance, leading to predictions that are too scattered or too aggregated. This can result in unreliable spatial projections, where predicted abundance fluctuates excessively, even across regions where the observed abundance is stable. These three characteristics combined could inform if a model is realistic and reliable, thus better projecting the species' abundance across geographical space.

Absence amount

The use of absence or pseudo-absence data have been extensively explored in SDM literature (e.g., Barbet-Massin et al., 2012; Liu et al., 2019; Lobo et al., 2010); however, there is a knowledge gap for ADM, thus the question of how many absence points to use and how to sample

remains unexplored. Our findings showed that algorithm discrimination and accuracy decreased when the dataset had no absence. For discrimination, One-absence and Two-absence absence amounts tended to produce very similar models, with much higher discrimination than No-absence (Figure S11). For accuracy, Two-absence more clearly produced models with less error than One-absence and No-absence (Figure S12). This likely happened because including absence in the dataset concentrated the abundance data near zero (Figure S14). Data with these characteristics represents species more likely to have low abundance in more sites than high abundance, through a geographical region, what was observed in the literature with tree species (Murphy et al., 2006). Thus, the dataset accounting for species absences could produce models better at discriminating low and high abundance sites and in predicting accurate abundance values, because the data itself is more realistic. Otherwise, for precision, algorithms were affected differently by the absence amount, with no clear general tendency. For ANN, DNN, and NET scored near to 1 in PDISP in every treatment, showing the reliability of these algorithms in terms of precision. CNN scored better without absences in precision. This implies that the algorithm chosen to build ADM matters in terms of precision, and the modeler should be aware of algorithms' functioning and capabilities. Further explorations about the impact of the absences amount are needed to better clarify questions about the optimal amount and sampling for ADM construction.

Data partition

In this work, we compared models' performance under random (K-fold) and spatially structured data partition. We found that for most treatments and metrics, spatial partition had slightly lower or equal performance than K-fold. Differences in data partition performances are known in the literature (Gonzalez et al., 2011; Newbold et al., 2010; Roberts et al., 2017). Spatially structured partitions tend to test model spatial transferability more rigorously than random approaches (Roberts et al., 2017), i.e., the model capacity to predict correctly in regions not used during model training (Franklin, 2023; Roberts et al., 2017). For that reason, partitioning approach is meaningful for ADM extrapolation, because abundance prediction could be unbounded (i.e., no specific lower or upper limits can be predicted), contrary to SDM which typically predicts suitability between 0-1. While some algorithms could only predict positive values (e.g., a GLM fitted with Poisson distribution), others could predict negative values, especially those that are distribution-independent (e.g., DNN). Inaccurate ADM predictions could produce meaningless predictions with extremely high values, fragrantly above ecosystem carrying capacity, or inappropriate negative values. This could compromise the useability of ADM for many purposes, especially for species conservation, as they become misleading about species abundance in geographical space. Model extrapolation issue was more addressed in SDM literature (Elith et al.,

2010; Velazco et al., 2024); however, ADM field lacks experiments and development methods for addressing extrapolation problems by, for instance, correcting or rescaling model output. This highlights the necessity of further research on ADM extrapolation.

How good are DNN, NET, and CNN compared to other algorithms and is it worth the effort?

In our experiment, DNN clearly outperformed CNN and NET, being overall the best-performing algorithms. DNN consistently was the best model for most species in discrimination and accuracy, while NET was the best in precision in all treatments (Figures S5-S7, Figures S15-17). These results show the potential of ANN for ADM construction. However, these algorithms could become computationally expensive (Chen et al., 2023). As XGB often had similar performances in many situations as DNN, the question of whether it is worth the effort to use DNN to construct ADM or not must be addressed, and we identify three major topics in this discussion.

(i) ANN often take longer to train and predict than other algorithms, and the computational time and resources needed to train an ANN extensively increase with the size of the network and dataset (Caluña et al., 2020). The computational effort could be a disincentive for researchers from developing regions with a lack of financial support and resource limitations, like the *Gran Chaco* (Ocampo-Ariza et al., 2023; Silveira et al., 2023). This raises a contradiction, because in times of environmental crisis (Cowie et al., 2022) the most endangered regions, that would benefit from the most cutting-edge models, often have the least computational, financial, and technical resources.

(ii) ANN offers much more flexibility than other more conventional algorithms, due to the large number of possible architectures (Pichler & Hartig, 2023). This allows, for instance, the construction of multi-species predictions, complex model ensembles, combinations of different architectures, and the use of pre-trained networks. This flexibility allied with the ANN ability to model non-linear complex relationships provides researchers with a powerful tool to model species distribution, while enabling further advances (Botella et al., 2018). However, the planning and manipulation of ANN architectures could demand more expertise. For instance, our CNN and DNN architectures were built with *torch* (Falbel & Luraschi, 2024), which demanded the coding of the network structure. Our low performance of CNN shows that inadequately constructed ANN could produce poor models. Still, this is true for every algorithm, and one should be aware of its functioning and assumptions when constructing distribution models (Qiao et al., 2015). However, the use of ANN for many purposes is rising in ecology, and with it, the abundance of information and guides about them (Pichler & Hartig, 2023).

(iii) Because of the complexity, nonlinearity, and high-dimensional parameters of ANN, make them less interpretable than other algorithms, such as linear models (GLM and GAM) which could provide clearer relationships between predictors and abundance. (Ryo et al., 2021). The use of models to get accurate species distribution predictions to novel conditions or unsampled areas has become a common task (Andrella et al., 2023).However, they are still useful and important for explaining species distribution, e.g., to test hypotheses or evaluate the predictor variables importance (Araújo et al., 2019; Elith & Leathwick, 2009). In addition, many techniques were recently proposed in a subfield called "explainable Artifical Inteligence" (xAI). The xAI method aims to open the "black-box" of complex machine learning models to make them more interpretable (e.g., SHapley Additive explanation and Local Interpretable Model-Agnostic Explanations) (Prasad et al., 2023; Ryo et al., 2021).

About ADM nomenclature

In the literature, it is common to encounter spatially explicit correlative distribution models based on abundance data, i.e., models that have any expression of abundance as a response variable correlated with other predictor variables, what was referred as ADM, "SDM fitted with abundance data", or "Species Abundance Model" (e.g., García-Gómez et al., 2023; Kroetz et al., 2025). In fact, the ADM denomination seems to be recent in ecology (Waldock et al., 2022), and we reinforce to formalize the use of this term. We reckon that ADM derives from SDM and both use correlative models, but they produce different predictions. For instance, the relationship between abundance and environmental suitability (SDM predictions) is still not entirely clear in ecological literature, thus both should not be taken as equivalent (Osorio-Olvera et al., 2019). Moreover, SDM often fails to explain or predict local abundance (Brambilla et al., 2024). Producing predictions of different kinds, ADM and SDM must be evaluated with different metrics, adequate to their purpose, i.e., there is little sense in evaluating an ADM with True Skill Statistics, common metrics for SDM. Therefore, we argue that may be beneficial for the ADM research field to popularize the "abundance-based distribution models" denomination instead of other names that make references to SDM or that could be confused with models with other purposes, such species abundance distributions (SAD, which rank species abundance at community level Golestani & Gras, 2013).

What is the future for ADM?

The field of ADM is in ongoing construction, yet underdeveloped when compared to the SDM field (Waldock et al., 2022), highlighting the need for further research methodological

development of ADM. For example, here we addressed the effect of absence amount in the models' performance, but the methods of sampling absences for ADM were not yet explored. Recent results found that ensemble modeling approach produced better results than single algorithms (Ngor et al., 2023), thus more experiments with ensembles approach for ADM are needed. Further development of Deep Learning algorithms is possible, by experimenting with more robust, intricate, and ecologically informed architectures (see the ones of Botella et al., 2018; Deneu et al., 2021) that could deliver multi-species models based on community data (Elmendorf & Moore, 2008).

Limitations

Few large-scale experiments have compared multiple algorithms for ADM construction. For this reason, and because many studies used different metrics, it became difficult to directly compare our results with previous research. Because of the number of species modeled and the large number of fitted models, it became impractical to analyze or tune models with many more details, e.g., for each species. This, along with resources and technical limitations, impacted the construction of CNN and DNN models, as we tested fewer and simpler architectures compared to previous works.

Conclusion

As far as we know, this work is one of the first efforts to comparatively analyze the performance of Deep Learning algorithms for the construction of ADM and the impacts of absence amount and data partitioning on it. We evaluated models' discrimination, accuracy, and precision capabilities, and found that DNN was the better performing algorithm, but XGB often performed as well as DNN. GBM, GLM, and GAM had a medium performance, compared to the other algorithms, and NET, RAF, and SVM were more often the worst-performing. CNN had the worst results probably because of the simplicity of the architectures tested. We concluded (1) that the use of absence data is beneficial to all algorithms, but more research is needed to clarify even further this question, (2) that the spatially structured data partitioning produced less heterogenous results between algorithms and validates model in terms of model transferability, important for models that could make prediction unboundedly, and (3) that DNN is a viable and effective algorithm for ADM, and should be considered as a powerful tool for ADM construction. We highlight though that the algorithm selection and model evaluation metrics must be done considering the modeling purpose and ecological implications. Finally, we emphasize the need for

further research in the ADM to explore yet unclear questions to develop this field (e.g., model extrapolation, effect of calibration area, model ensembles, predictors interactions).

References

Abrego, N., & Ovaskainen, O. (2023). Evaluating the predictive performance of presence–absence models: Why can the same model appear excellent or poor? *Ecology and Evolution*, *13*(12), e10784. https://doi.org/10.1002/ece3.10784

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., & Asari, V. K. (2019). A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, *8*(3), 292. https://doi.org/10.3390/electronics8030292

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Andrella, G. C., Koch, I., & Velazco, S. J. E. (2023). Considering spatial constraints to identify areas for new species sampling: A species-specific prioritization approach. *Biological Conservation*, *288*, 110379. https://doi.org/10.1016/j.biocon.2023.110379

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Barral, M. P., Villarino, S., Levers, C., Baumann, M., Kuemmerle, T., & Mastrangelo, M. (2020). Widespread and major losses in multiple ecosystem services as a result of agricultural expansion in the Argentine Chaco. *Journal of Applied Ecology*, *57*(12), 2485–2498. https://doi.org/10.1111/1365-2664.13740

Borghetti, F., Barbosa, E., Ribeiro, L., Ribeiro, J. F., Maciel, E., & Walter, B. M. T. (2023). Fitogeografia das savanas sul-americanas. *Heringeriana*, *17*, e918014. https://doi.org/10.17648/heringeriana.v17i1.918014

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022). Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, *13*(8), 1640–1660. https://doi.org/10.1111/2041-210X.13901

Borregaard, M. K., & Rahbek, C. (2010). Causality of the Relationship between Geographic Distribution and Species Abundance. *The Quarterly Review of Biology*, *85*(1), 3–25. https://doi.org/10.1086/650265

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

Botella, C., Joly, A., Bonnet, P., Monestiez, P., & Munoz, F. (2018). A Deep Learning Approach to Species Distribution Modelling. In A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, & P. Bonnet (Eds.), *Multimedia Tools and Applications for Environmental & Biodiversity Informatics* (pp. 169–199). Springer International Publishing. https://doi.org/10.1007/978-3-319-76445-0_10

Brambilla, M., Bazzi, G., & Ilahiane, L. (2024). The effectiveness of species distribution models in predicting local abundance depends on model grain size. *Ecology*, *105*(2), e4224. https://doi.org/10.1002/ecy.4224

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bucher, E. H. (1982). Chaco and Caatinga—South American Arid Savannas, Woodlands and Thickets. In B. J. Huntley & B. H. Walker (Eds.), *Ecology of Tropical Savannas* (Vol. 42, pp. 48–79). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-68786-0_4

Bucher, E. H., & Huszar, P. C. (1999). Sustainable management of the Gran Chaco of South America: Ecological promise and economic constraints. *Journal of Environmental Management*, *57*(2), 99–108. https://doi.org/10.1006/jema.1999.0290

Caluña, G., Guachi-Guachi, L., & Brito, R. (2020). Convolutional Neural Networks for Automatic Classification of Diseased Leaves: The Impact of Dataset Size and Fine-Tuning. In O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, C. M. Torre, & Y. Karaca (Eds.), *Computational Science and Its Applications – ICCSA 2020* (Vol. 12249, pp. 951–966). Springer International Publishing. https://doi.org/10.1007/978-3-030-58799-4_68

Catucci, E., Panzeri, D., Libralato, S., Cossarini, G., Garofalo, G., Maina, I., Kavadas, S., Quattrocchi, F., Cipriano, G., Carlucci, R., Vitale, S., Mytilineou, C., Fiorentino, F., & Russo, T. (2025). Modeling the spatial distribution and abundance of deep-water red shrimps in the Mediterranean Sea: A machine learning approach. *Fisheries Research*, *281*, 107257. https://doi.org/10.1016/j.fishres.2024.107257

Chen, C.-H., Lai, J.-P., Chang, Y.-M., Lai, C.-J., & Pai, P.-F. (2023). A Study of Optimization in Deep Neural Networks for Regression. *Electronics*, *12*(14), 3071. https://doi.org/10.3390/electronics12143071

Číhal, L. (2023). *From Bioclimatic Envelopes to Machine Learning: A Journey Through the History, Present, and Future of Species Distribution Modeling With Practical Tips for Use and Notes to Bryophytes*. https://doi.org/10.20944/preprints202304.0367.v1

Cowie, R. H., Bouchet, P., & Fontaine, B. (2022). The Sixth Mass Extinction: Fact, fiction or speculation? *Biological Reviews*, *97*(2), 640–663. https://doi.org/10.1111/brv.12816

Dallas, T. A., & Hastings, A. (2018). Habitat suitability estimated by niche models is largely unrelated to species abundance. *Global Ecology and Biogeography*, *27*(12), 1448–1456. https://doi.org/10.1111/geb.12820

De La Fuente, A., Hirsch, B. T., Cernusak, L. A., & Williams, S. E. (2021). Predicting species abundance by implementing the ecological niche theory. *Ecography*, *44*(11), 1723–1730. https://doi.org/10.1111/ecog.05776

De Marco, P., & Nóbrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLOS ONE*, *13*(9), e0202403. https://doi.org/10.1371/journal.pone.0202403

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., & Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the

environment. *PLOS Computational Biology*, *17*(4), e1008856. https://doi.org/10.1371/journal.pcbi.1008856

Edwards, D. P., Socolar, J. B., Mills, S. C., Burivalova, Z., Koh, L. P., & Wilcove, D. S. (2019). Conservation of Tropical Forests in the Anthropocene. *Current Biology*, *29*(19), R1008–R1020. https://doi.org/10.1016/j.cub.2019.08.026

Ehrlén, J., & Morris, W. F. (2015). Predicting changes in the distribution and abundance of species under environmental change. *Ecology Letters*, *18*(3), 303–314. https://doi.org/10.1111/ele.12410

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species: *The art of modelling range-shifting species*. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elmendorf, S. C., & Moore, K. A. (2008). Use of Community-Composition Data to Predict the Fecundity and Abundance of Species. *Conservation Biology*, *22*(6), 1523–1532. https://doi.org/10.1111/j.1523-1739.2008.01051.x

Falbel, D., & Luraschi, J. (2024). *torch: Tensors and Neural Networks with "GPU" Acceleration*. https://torch.mlverse.org/docs

Feron, S., Cordero, R. R., Damiani, A., MacDonell, S., Pizarro, J., Goubanova, K., Valenzuela, R., Wang, C., Rester, L., & Beaulieu, A. (2024). South America is becoming warmer, drier, and more flammable. *Communications Earth & Environment*, *5*(1), 501. https://doi.org/10.1038/s43247-024-01654-7

Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian Obama nungara. *Ecological Modelling*, *457*, 109686. https://doi.org/10.1016/j.ecolmodel.2021.109686

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. https://doi.org/10.1111/geb.12684

Franklin, J. (2023). Species distribution modelling supports the study of past, present and future biogeographies. *Journal of Biogeography*, *50*(9), 1533–1545. https://doi.org/10.1111/jbi.14617

Freiberg, M., Winter, M., Gentile, A., Zizka, A., Muellner-Riehl, A. N., Weigelt, A., & Wirth, C. (2020). *The Leipzig Catalogue of Vascular Plants (LCVP) – An improved taxonomic reference list for all known vascular plants*. https://doi.org/10.1101/2020.05.08.077149

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

García-Gómez, R. E., Aceves-Medina, G., Villalobos, H., Jiménez Rosenberg, S. P. A., & Durazo, R. (2023). Predictive performance from abundance distribution models of Vinciguerria lucetia larvae in the southern portion of the California current system using XGBOOST. *Deep Sea Research Part II: Topical Studies in Oceanography*, *212*, 105336. https://doi.org/10.1016/j.dsr2.2023.105336

Gasparri, N. I., & Grau, H. R. (2009). Deforestation and fragmentation of Chaco dry forest in NW Argentina (1972–2007). *Forest Ecology and Management*, *258*(6), 913–921. https://doi.org/10.1016/j.foreco.2009.02.024

Golestani, A., & Gras, R. (2013). A New Species Abundance Distribution Model Based on Model Combination. *The International Journal of Biostatistics*, *9*(1). https://doi.org/10.1515/ijb-2012-0033

Gonzalez, S. C., Soto-Centeno, J. A., & Reed, D. L. (2011). Population distribution models: Species distributions are better modeled using biologically relevant data partitions. *BMC Ecology*, *11*(1), 20. https://doi.org/10.1186/1472-6785-11-20

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435. https://doi.org/10.1111/ele.12189

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3). https://doi.org/10.1214/ss/1177013604

Hastings, R. A., Rutterford, L. A., Freer, J. J., Collins, R. A., Simpson, S. D., & Genner, M. J. (2020). Climate Change Drives Poleward Increases and Equatorward Declines in Marine Species. *Current Biology*, *30*(8), 1572-1577.e2. https://doi.org/10.1016/j.cub.2020.02.043

Holt, A. R., Warren, P. H., & Gaston, K. J. (2004). The importance of habitat heterogeneity, biotic interactions and dispersal in abundance–occupancy relationships. *Journal of Animal Ecology*, *73*(5), 841–851. https://doi.org/10.1111/j.0021-8790.2004.00862.x

Holt, R. D. (2009). Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences*, *106*(supplement_2), 19659–19665. https://doi.org/10.1073/pnas.0905137106

Howard, C., Stephens, P. A., Pearce-Higgins, J. W., Gregory, R. D., & Willis, S. G. (2014). Improving species distribution models: The value of data on abundance. *Methods in Ecology and Evolution*, *5*(6), 506–513. https://doi.org/10.1111/2041-210X.12184

Hu, Y., Si-Moussi, S., & Thuiller, W. (2025). Introduction to deep learning methods for multi-species predictions. *Methods in Ecology and Evolution*, *16*(1), 228–246. https://doi.org/10.1111/2041-210X.14466

Huang, H., Zhang, Z., Bede-Fazekas, Á., Mammola, S., Gu, J., Zhou, J., Qu, J., & Lin, Q. (2024). Cross-validation matters in species distribution models: A case study with goatfish species. *Ecography*, e07354. https://doi.org/10.1111/ecog.07354

INFONA, I. F. N. (2024). *Instituto Forestal Nacional – INFONA*. https://infona.gov.py/

Karger, D. N., & Zimmer, H. C. (2019). *Climatologies at high resolution for the earth land surface areas CHELSA V1.2: Technical specification*.

Kroetz, A. M., Dedman, S., & Carlson, J. K. (2025). Predictive Modeling of Juvenile Smalltooth Sawfish Habitats: Challenges and Opportunities for Conservation. *Ecology and Evolution*, *15*(1), e70592. https://doi.org/10.1002/ece3.70592

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lek, S., & Guégan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, *120*(2–3), 65–73. https://doi.org/10.1016/S0304-3800(99)00092-7

Lenth, R. V. (2017). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (p. 1.10.4) [Dataset]. https://doi.org/10.32614/CRAN.package.emmeans

Liu, C., Newell, G., & White, M. (2019). The effect of sample size on the accuracy of species distribution models: Considering both presences and pseudo-absences or background sites. *Ecography*, *42*(3), 535–548. https://doi.org/10.1111/ecog.03188

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

Luna, S., Peña-Peniche, A., & Mendoza-Alfaro, R. (2024). Species distribution model accuracy is strongly influenced by the choice of calibration area. *Biodiversity Informatics*, *18*. https://doi.org/10.17161/bi.v18i.22655

Marcer, A., Méndez-Vigo, B., Alonso-Blanco, C., & Picó, F. X. (2016). Tackling intraspecific genetic structure in distribution models better reflects species geographical range. *Ecology and Evolution*, *6*(7), 2084–2097. https://doi.org/10.1002/ece3.2010

Martín, B., González–Arias, J., & Vicente–Vírseda, J. A. (2021). Machine learning as a successful approach for predicting complex spatio–temporal patterns in animal species abundance. *Animal Biodiversity and Conservation*, 289–301. https://doi.org/10.32800/abc.2021.44.0289

Martínez-González, A., Villamizar, M., Canévet, O., & Odobez, J.-M. (2020). Efficient Convolutional Neural Networks for Depth-Based Multi-Person Pose Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(11), 4207–4221. https://doi.org/10.1109/TCSVT.2019.2952779

MAySD, M. de A. y D. S. de la N. (2022). *Segundo Inventario Nacional de Bosques Nativos: Informe Nacional*. Ministerio de Ambiente y Desarrollo Sostenible de la Nación.

Mi, C., Huettmann, F., Sun, R., & Guo, Y. (2017). Combining occurrence and abundance distribution models for the conservation of the Great Bustard. *PeerJ*, *5*, e4160. https://doi.org/10.7717/peerj.4160

Miller, J. (2010). Species Distribution Modeling. *Geography Compass*, *4*(6), 490–509. https://doi.org/10.1111/j.1749-8198.2010.00351.x

Moudrý, V., Bazzichetto, M., Remelgado, R., Devillers, R., Lenoir, J., Mateo, R. G., Lembrechts, J. J., Sillero, N., Lecours, V., Cord, A. F., Barták, V., Balej, P., Rocchini, D., Torresani, M., Arenas-Castro, S., Man, M., Prajzlerová, D., Gdulová, K., Prošek, J., … Šímová, P. (2024). Optimising occurrence data in species distribution models: Sample size, positional uncertainty, and sampling bias matter. *Ecography*, e07294. https://doi.org/10.1111/ecog.07294

Murphy, H. T., VanDerWal, J., & Lovett-Doust, J. (2006). Distribution of abundance across the range in eastern North American trees. *Global Ecology and Biogeography*, *15*(1), 63–71. https://doi.org/10.1111/j.1466-822X.2006.00194.x

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370. https://doi.org/10.2307/2344614

Newbold, T., Reader, T., El-Gabbas, A., Berg, W., Shohdi, W. M., Zalat, S., El Din, S. B., & Gilbert, F. (2010). Testing the accuracy of species distribution models using species records from a new field survey. *Oikos*, *119*(8), 1326–1334. https://doi.org/10.1111/j.1600-0706.2009.18295.x

Ngor, P. B., Uy, S., Sor, R., Chan, B., Holway, J., Null, S. E., So, N., Grenouillet, G., Chandra, S., Hogan, Z. S., & Lek, S. (2023). Predicting fish species richness and abundance in the Lower Mekong Basin. *Frontiers in Ecology and Evolution*, *11*, 1131142. https://doi.org/10.3389/fevo.2023.1131142

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., … Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*(3), e01370. https://doi.org/10.1002/ecm.1370

Nori, J., Torres, R., Lescano, J. N., Cordier, J. M., Periago, M. E., & Baldo, D. (2016). Protected areas and spatial conservation priorities for endemic vertebrates of the Gran Chaco, one of the most threatened ecoregions of the world. *Diversity and Distributions*, *22*(12), 1212–1219. https://doi.org/10.1111/ddi.12497

Ocampo-Ariza, C., Toledo-Hernández, M., Librán-Embid, F., Armenteras, D., Vansynghel, J., Raveloaritiana, E., Arimond, I., Angulo-Rubiano, A., Tscharntke, T., Ramírez-Castañeda, V., Wurz, A., Marcacci, G., Anders, M., Urbina-Cardona, J. N., De Vos, A., Devy, S., Westphal, C., Toomey, A., Sheherazade, … Maas, B. (2023). Global South leadership towards inclusive tropical ecology and conservation. *Perspectives in Ecology and Conservation*, *21*(1), 17–24. https://doi.org/10.1016/j.pecon.2023.01.002

Ogurtsov, S. S. (2023). Absence of "Absences": The Engler–Hengl Approach in Species Distribution Modeling. *Biology Bulletin*, *50*(S2), S140–S155. https://doi.org/10.1134/S1062359023605311

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, *51*(11), 933. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Osorio-Olvera, L., Soberón, J., & Falconi, M. (2019). On population abundance and niche structure. *Ecography*, *42*(8), 1415–1425. https://doi.org/10.1111/ecog.04442

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Podder, P., Bharati, S., Mondal, M. R. H., Paul, P. K., & Kose, U. (2021). *Artificial Neural Network for Cybersecurity: A Comprehensive Review* (arXiv:2107.01185). arXiv. http://arxiv.org/abs/2107.01185

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, *7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

Pollock, L. J., Kelly, L. T., Thomas, F. M., Soe, P., Morris, W. K., White, M., & Vesk, P. A. (2018). Combining functional traits, the environment and multiple surveys to understand semi-arid tree distributions. *Journal of Vegetation Science*, *29*(6), 967–977. https://doi.org/10.1111/jvs.12686

Prado, D. E. (1993). What is the Gran Chaco vegetation in South America? I. A review. Contribution to the study of flora and vegetation of the Chaco. *Candollea*, *48*, 145–172.

Prasad, A., Peters, M., Matthews, S., & Iverson, L. (2023). Unpacking the 'black box': Improving ecological interpretation of regression-based models. *Diversity and Distributions*, *29*(7), 926–945. https://doi.org/10.1111/ddi.13707

Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, *6*(10), 1126–1136. https://doi.org/10.1111/2041-210X.12397

Redford, K. H., Taber, A., & Simonetti, J. A. (1990). There Is More to Biodiversity than the Tropical Rain Forests. *Conservation Biology*, *4*(3), 328–330. https://doi.org/10.1111/j.1523-1739.1990.tb00296.x

Rew, J., Cho, Y., & Hwang, E. (2021). A Robust Prediction Model for Species Distribution Using Bagging Ensembles with Deep Neural Networks. *Remote Sensing*, *13*(8), 1495. https://doi.org/10.3390/rs13081495

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *54*(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Rocha, J. C., Peres, C. K., Buzzo, J. L. L., De Souza, V., Krause, E. A., Bispo, P. C., Frei, F., Costa, L. S. M., & Branco, C. C. Z. (2017). Modeling the species richness and abundance of lotic macroalgae based on habitat characteristics by artificial neural networks: A potentially useful tool for stream biomonitoring programs. *Journal of Applied Phycology*, *29*(4), 2145–2153. https://doi.org/10.1007/s10811-017-1107-5

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199–205. https://doi.org/10.1111/ecog.05360

Sánchez-Fernández, D., Lobo, J. M., & Hernández-Manrique, O. L. (2011). Species distribution models that do not incorporate global data misrepresent potential distributions: A case study using Iberian diving beetles: Regional data misrepresent potential distributions. *Diversity and Distributions*, *17*(1), 163–171. https://doi.org/10.1111/j.1472-4642.2010.00716.x

SAyDS, S. de A. y D. S. (2005). *Primer Inventario Nacional de Bosques Nativos*. Ministerio de Salud y Ambiente de la Nación.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

SFB, S. F. B. (2018). *Inventário Florestal Nacional: Principais resultados: Santa Catarina*. MMA. https://www.florestal.gov.br/publicacoes

Silveira, F. A. O., Fuzessy, L., Phartyal, S. S., Dayrell, R. L. C., Vandelook, F., Vázquez-Ramírez, J., Tavşanoğlu, Ç., Abedi, M., Naidoo, S., Acosta-Rojas, D. C., Chen, S.-C., Cruz-Tejada, D. M., Jayasuryia, G., Ordóñez-Parra, C. A., & Saatkamp, A. (2023). Overcoming major barriers in seed ecology research in developing countries. *Seed Science Research*, *33*(3), 172–181. https://doi.org/10.1017/S0960258523000181

Soberón, J., & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, *106*(supplement_2), 19644–19650. https://doi.org/10.1073/pnas.0901637106

Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., Edwards, T. C., Guala, G. F., Howard, T. G., Morisette, J. T., & Hamilton, H. (2019). Development and Delivery of Species Distribution Models to Inform Decision-Making. *BioScience*, *69*(7), 544–557. https://doi.org/10.1093/biosci/biz045

Sporbert, M., Keil, P., Seidler, G., Bruelheide, H., Jandt, U., Aćić, S., Biurrun, I., Campos, J. A., Čarni, A., Chytrý, M., Ćušterevska, R., Dengler, J., Golub, V., Jansen, F., Kuzemko, A., Lenoir, J., Marcenò, C., Moeslund, J. E., Pérez-Haase, A., … Welk, E. (2020). Testing macroecological abundance patterns: The relationship between local abundance and range size, range position and climatic suitability among European vascular plants. *Journal of Biogeography*, *47*(10), 2210–2222. https://doi.org/10.1111/jbi.13926

Stasinopoulos, M., & Rigby, R. (2012). *gamlss: Generalized Additive Models for Location Scale and Shape* (p. 5.4-22) [Dataset]. https://doi.org/10.32614/CRAN.package.gamlss

Stokland, J. N., Halvorsen, R., & Støa, B. (2011). Species distribution modelling—Effect of design and sample size of pseudo-absence observations. *Ecological Modelling*, *222*(11), 1800–1809. https://doi.org/10.1016/j.ecolmodel.2011.02.025

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, *220*(4), 589–594. https://doi.org/10.1016/j.ecolmodel.2008.11.010

Velazco, S. J. E., Rose, M. B., De Andrade, A. F. A., Minoli, I., & Franklin, J. (2022). flexsdm: An r package for supporting a comprehensive and flexible species distribution modelling workflow. *Methods in Ecology and Evolution*, *13*(8), 1661–1669. https://doi.org/10.1111/2041-210X.13874

Velazco, S. J. E., Rose, M. B., De Marco, P., Regan, H. M., & Franklin, J. (2024). How far can I extrapolate my species distribution model? Exploring shape, a novel method. *Ecography*, *2024*(3), e06992. https://doi.org/10.1111/ecog.06992

Velazco, S. J. E., Villalobos, F., Galvão, F., & De Marco Júnior, P. (2019). A dark scenario for Cerrado plant species: Effects of future climate, land use and protected areas ineffectiveness. *Diversity and Distributions*, *25*(4), 660–673. https://doi.org/10.1111/ddi.12886

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W. L., Edgar, G. J., Mouillot, D., Tjiputra, J., & Pellissier, L. (2022). A quantitative review of abundance-based species distribution models. *Ecography*, *2022*(1), ecog.05694. https://doi.org/10.1111/ecog.05694

Weber, M. M., Stevens, R. D., Diniz-Filho, J. A. F., & Grelle, C. E. V. (2017). Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? A meta-analysis. *Ecography*, *40*(7), 817–828. https://doi.org/10.1111/ecog.02125

Yen, P. P. W., Huettmann, F., & Cooke, F. (2004). A large-scale model for the at-sea distribution and abundance of Marbled Murrelets (Brachyramphus marmoratus) during the breeding season in coastal British Columbia, Canada. *Ecological Modelling*, *171*(4), 395–413. https://doi.org/10.1016/j.ecolmodel.2003.07.006

Young, M., & Carr, M. H. (2015). Application of species distribution models to explain and predict the distribution, abundance and assemblage structure of nearshore temperate reef fishes. *Diversity and Distributions*, *21*(12), 1428–1440. https://doi.org/10.1111/ddi.12378

Yu, H., Cooper, A. R., & Infante, D. M. (2020). Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. *Ecological Modelling*, *432*, 109202. https://doi.org/10.1016/j.ecolmodel.2020.109202

Supplementary Material

Tables

Table S1. List of the 117 species used in the experiment and summary statistics about their abundance data. Statistics were computed only considering the occurrence points.

Species	Occurrence points (ind/ha > 0)	Abundance range	Mean abundance (μ)	Median abundance	Variance (σ^2)	Standard deviation (σ)	Coefficient of variation ($\frac{\sigma}{\mu}$)
Aspidosperma quebracho-blanco Schltdl.	1,849	[2.78, 400]	45.12	30	1,788.13	42.29	0.94
Sarcomphalus mistol (Griseb.) Hauenschild	1,369	[2.78, 443.06]	44.99	30	2,208.55	47	1.04
Schinopsis lorentzii (Griseb.) Engl.	1,034	[2.78, 410]	34.82	20	1,071.89	32.74	0.94
Neltuma nigra (Griseb.) C.E.Hughes & G.P.Lewis	991	[2.78, 350]	41.99	30	1,908.43	43.69	1.04
Senegalia praecox (Griseb.) Seigler & Ebinger	731	[6.25, 464.58]	43.47	30	2,548.24	50.48	1.16
Tabebuia nodosa (Griseb.) Griseb.	710	[2.78, 477.08]	49.19	27.08	3,569.57	59.75	1.21
Salta triflora (Griseb.) Adr.Sanchez	637	[2.78, 1733.33]	110.43	40	38,607.96	196.49	1.78
Libidibia paraguariensis (D.Parodi) G.P.Lewis	620	[2.78, 247.22]	25.65	19.72	720.11	26.83	1.05
Parkinsonia praecox (Ruiz & Pav.) Hawkins	525	[2.78, 190]	21.99	10	418.33	20.45	0.93
Cordia americana (L.) Gottschling & J.S.Mill.	473	[2.5, 285.29]	33.67	20	1,444.25	38	1.13
Sideroxylon obtusifolium (Roem. & Schult.) T.D.Penn.	465	[2.78, 186.86]	30.14	20	857.56	29.28	0.97
Nectandra angustifolia (Schrad.) Nees & Mart.	455	[2.5, 207.65]	31.21	20	900.83	30.01	0.96
Allophylus edulis (A.St.-Hil., A.Juss. & Cambess.) Radlk.	399	[2.5, 612.16]	28.3	10	2,767.65	52.61	1.86
Neltuma alba (Griseb.) C.E.Hughes & G.P.Lewis	366	[2.78, 250]	33.98	20	1,245.49	35.29	1.04
Neltuma ruscifolia (Griseb.) C.E.Hughes & G.P.Lewis	365	[2.78, 510]	87.96	60	8,200.65	90.56	1.03
Ruprechtia laxiflora Meisn.	340	[2.5, 300]	33.86	17.32	2,048.20	45.26	1.34
Ocotea puberula (Rich.) Nees	330	[2.5, 216.08]	27.79	12.5	1,374.96	37.08	1.33
Muellera campestris (Mart. ex Benth.) M.J.Silva & A.M.G.Azevedo	320	[2.5, 267.36]	22.85	10	1,140.10	33.77	1.48
Plectrocarpa sarmientoi (Lorentz ex Griseb.) Christenh. & Byng	318	[2.78, 420]	54.37	30	3,297.53	57.42	1.06
Cabralea canjerana (Vell.) Mart.	310	[2.5, 177.65]	15.64	10	345.69	18.59	1.19
Syagrus romanzoffiana (Cham.) Glassman	307	[2.5, 500.69]	22.7	10	1,627.41	40.34	1.78

Species	Occurrence points (ind/ha > 0)	Abundance range	Mean abundance (μ)	Median abundance	Variance (σ^2)	Standard deviation (σ)	Coefficient of variation ($\frac{\sigma}{\mu}$)
Ceiba chodatii (Hassl.) Ravenna	306	[2.78, 147.22]	16.8	10	296.83	17.23	1.03
Cupania vernalis Cambess.	304	[2.5, 235.29]	24.16	10.7	1,240.46	35.22	1.46
Neltuma flexuosa (DC.) C.E.Hughes & G.P.Lewis	293	[10, 360]	49.49	30	2,619.19	51.18	1.03
Neltuma kuntzei (Harms) C.E.Hughes & G.P.Lewis	291	[2.78, 270]	28.13	20	939.91	30.66	1.09
Bougainvillea praecox Griseb.	288	[2.78, 230]	29.78	20	1,063.09	32.61	1.1
Nectandra lanceolata Nees & Mart.	287	[2.5, 159.03]	19.74	12.5	467.39	21.62	1.1
Geoffroea decorticans (Gillies ex Hook. & Arn.) Burkart	284	[10, 800]	62.01	30	8,621.44	92.85	1.5
Morisonia retusa (Griseb.) Christenh. & Byng	274	[6.25, 781.25]	55.79	20	9,207.17	95.95	1.72
Pisonia zapallo Griseb.	273	[2.5, 375.29]	28.06	10	1,598.51	39.98	1.42
Celtis iguanaea (Jacq.) Sarg.	266	[2.5, 210]	32.12	20	1,372.06	37.04	1.15
Anadenanthera colubrina (Vell.) Brenan	261	[2.78, 1115.1]	81.13	31.25	16,915.19	130.06	1.6
Chrysophyllum gonocarpum (Mart. & Eichler) Engl.	261	[2.5, 236.08]	24.8	11.11	844.26	29.06	1.17
Phyllostylon rhamnoides (J.Poiss.) Taub.	258	[2.78, 619.8]	73.34	30	9,917.53	99.59	1.36
Luehea divaricata Mart.	252	[2.5, 178.43]	25.07	10	962.31	31.02	1.24
Casearia sylvestris Sw.	242	[2.5, 470.59]	20.48	10	1,644.00	40.55	1.98
Campomanesia xanthocarpa (Mart.) O.Berg	236	[0.88, 489.58]	17.35	10	1,517.80	38.96	2.25
Schinopsis balansae Engl.	235	[2.78, 390.28]	45.09	20	3,340.02	57.79	1.28
Morisonia speciosa (Griseb.) Christenh. & Byng	231	[6.25, 200]	23.04	10	659.29	25.68	1.11
Gleditsia amorphoides (Griseb.) Taub.	228	[2.5, 290]	43.46	30	1,936.41	44	1.01
Balfourodendron riedelianum (Engl.) Engl.	224	[2.5, 173.61]	22.13	10	779.6	27.92	1.26
Copernicia alba Morong	220	[2.78, 1270]	272.5	195	64,605.75	254.18	0.93
Vachellia caven (Molina) Seigler & Ebinger	210	[6.25, 350]	48.87	30	3,091.88	55.6	1.14
Parapiptadenia rigida (Benth.) Brenan	206	[2.5, 229.86]	19.33	10	854.52	29.23	1.51
Ximenia americana L.	199	[2.78, 110]	18.43	10	338.84	18.41	1
Chrysophyllum marginatum (Hook. & Arn.) Radlk.	191	[2.5, 865.28]	22.62	10	4,151.91	64.44	2.85
Sapium glandulosum (L.) Morong	189	[2.5, 310]	12.31	5.35	853.91	29.22	2.37
Zanthoxylum rhoifolium Lam.	185	[2.5, 250]	14	5.87	900	30	2.14

Species	Occurrence points (ind/ha > 0)	Abundance range	Mean abundance (μ)	Median abundance	Variance (σ^2)	Standard deviation (σ)	Coefficient of variation ($\frac{\sigma}{\mu}$)
Aspidosperma australe Müll.Arg.	181	[2.5, 98.43]	13.21	10	196.33	14.01	1.06
Vachellia aroma (Gillies ex Hook. & Arn.) Seigler & Ebinger	181	[6.25, 274.51]	27.82	20	938.95	30.64	1.1
Myrsine coriacea (Sw.) R.Br.	173	[2.5, 135]	12.97	7.5	271.74	16.48	1.27
Cordia trichotoma (Vell.) Arráb. ex Steud.	171	[2.5, 196.86]	17.04	10	573.46	23.95	1.41
Holocalyx balansae Micheli	171	[2.5, 80]	15.56	10	155.77	12.48	0.8
Myrocarpus frondosus Allemão	171	[2.5, 87.5]	11.07	10	147.05	12.13	1.1
Calycophyllum multiflorum Griseb.	155	[2.78, 297.22]	54.43	26	3,619.55	60.16	1.11
Phytolacca dioica L.	154	[2.5, 127.65]	10.7	10	144.87	12.04	1.13
Parapiptadenia excelsa (Griseb.) Burkart	149	[10, 326.08]	40.73	20	2,192.89	46.83	1.15
Morisonia salicifolia (Griseb.) Christenh. & Byng	142	[6.25, 402.08]	31.78	10	3,008.96	54.85	1.73
Senegalia gilliesii (Steud.) Seigler & Ebinger	139	[10, 150]	19.06	10	264.59	16.27	0.85
Castela coccinea Griseb.	138	[10, 112.5]	15.73	10	131.22	11.46	0.73
Schinus longifolia (Lindl.) Speg.	138	[10, 120]	25.43	20	417.56	20.43	0.8
Myrcianthes pungens (O.Berg) D.Legrand	136	[2.5, 500]	56.51	20	6,852.10	82.78	1.46
Annona emarginata (Schltdl.) H.Rainer	135	[2.5, 202.78]	11.62	5.78	436.89	20.9	1.8
Achatocarpus praecox Griseb.	133	[2.5, 156.25]	27.48	20	844.44	29.06	1.06
Sebastiania klotzschiana (Müll.Arg.) Müll.Arg.	133	[2.5, 400]	34.98	10	3,628.39	60.24	1.72
Mimozyganthus carinatus (Griseb.) Burkart	132	[2.78, 150]	24.56	10	712.69	26.7	1.09
Ruprechtia apetala Wedd.	132	[10, 1118.82]	60.12	30	12,994.21	113.99	1.9
Strombocarpa torquata (Lag.) Hutch. ex C.E.Hughes & G.P.Lewis	132	[10, 120]	18.11	10	319.9	17.89	0.99
Aspidosperma triternatum Rojas Acosta	130	[2.78, 200]	27.02	12.15	1,036.47	32.19	1.19
Eugenia uniflora L.	129	[2.5, 420]	30.94	10	3,745.07	61.2	1.98
Helietta apiculata Benth.	127	[2.5, 334.51]	35.88	12.5	3,358.89	57.96	1.62
Maclura tinctoria (L.) D.Don ex G.Don	127	[2.5, 152.08]	17.93	10	560.77	23.68	1.32
Sebastiania brasiliensis Spreng.	127	[2.5, 310]	28.29	12.5	1,285.87	35.86	1.27
Neltuma affinis (Spreng.) C.E.Hughes & G.P.Lewis	126	[2.78, 290]	56.26	40	2,722.48	52.18	0.93
Machaerium oblongifolium Vogel	114	[2.5, 97.44]	15.42	10	318.83	17.86	1.16

Species	Occurrence points (ind/ha > 0)	Abundance range	Mean abundance (μ)	Median abundance	Variance (σ^2)	Standard deviation (σ)	Coefficient of variation ($\frac{\sigma}{\mu}$)
Jacaratia spinosa (Aubl.) A.DC.	110	[2.5, 88.43]	15.88	10	222.79	14.93	0.94
Scutia buxifolia Reissek	101	[2.5, 402.16]	45.39	20	4,450.71	66.71	1.47
Terminalia triflora (Griseb.) Lillo	101	[2.78, 149.31]	28.9	19.44	917.05	30.28	1.05
Blepharocalyx salicifolius (Kunth) O.Berg	100	[2.5, 296.86]	30.89	10	2,586.10	50.85	1.65
Ficus luschnathiana (Miq.) Miq.	99	[2.5, 88.43]	8.59	5.41	118.13	10.87	1.27
Myrcianthes cisplatensis (Cambess.) O.Berg	96	[10, 280]	46.65	30	2,222.45	47.14	1.01
Acanthosyris falcata Griseb.	92	[2.78, 380]	25.9	10	1,978.16	44.48	1.72
Enterolobium contortisiliquum (Vell.) Morong	92	[2.5, 118.43]	16.51	10	268.74	16.39	0.99
Trichilia claussenii C.DC.	92	[2.5, 147.65]	24.67	10.59	764.09	27.64	1.12
Schinus fasciculata (Griseb.) I.M.Johnst.	88	[10, 120]	23.29	10	578.44	24.05	1.03
Jodina rhombifolia (Hook. & Arn.) Reissek	86	[2.5, 40]	13.52	10	45.26	6.73	0.5
Neltuma caldenia (Burkart) C.E.Hughes & G.P.Lewis	86	[10, 770]	197.44	175	22,095.78	148.65	0.75
Bougainvillea stipitata Griseb.	85	[10, 110]	27.17	20	438.86	20.95	0.77
Vitex megapotamica (Spreng.) Moldenke	85	[2.5, 78.43]	8.71	5	118.59	10.89	1.25
Dahlstedtia muehlbergiana (Hassl.) M.J.Silva & A.M.G.Azevedo	83	[2.5, 78.43]	14.21	10	168.46	12.98	0.91
Peltophorum dubium (Spreng.) Taub.	82	[2.5, 547.22]	26.93	10	4,036.93	63.54	2.36
Trichilia catigua A.Juss.	81	[2.5, 200]	34.29	18.75	1,821.01	42.67	1.24
Jacaranda micrantha Cham.	80	[2.5, 70]	11.32	8.23	165.58	12.87	1.14
Trema micranthum (L.) Blume	80	[2.5, 166.86]	14.68	5	773.87	27.82	1.9
Diplokeleba floribunda N.E.Br.	79	[6.25, 186.11]	37.63	24.31	1,068.02	32.68	0.87
Zanthoxylum petiolare A.St.-Hil. & Tul.	78	[2.5, 80]	15.65	10	311.49	17.65	1.13
Erythrina falcata Benth.	76	[2.5, 50]	10.48	10	70.52	8.4	0.8
Myracrodruon urundeuva M. Allemão	76	[2.78, 193.75]	26.22	12.5	1,005.46	31.71	1.21
Zanthoxylum fagara (L.) Sarg.	74	[2.5, 160]	18.92	10	660.28	25.7	1.36
Myracrodruon balansae (Engl.) Santin	73	[2.78, 210]	42.77	30	1,642.97	40.53	0.95
Sapium haematospermum Müll.Arg.	70	[2.78, 324.51]	37.92	20	3,072.70	55.43	1.46
Bougainvillea campanulata Heimerl	69	[2.78, 262.5]	63.47	46.53	3,786.97	61.54	0.97

Species	Occurrence points (ind/ha > 0)	Abundance range	Mean abundance (μ)	Median abundance	Variance (σ^2)	Standard deviation (σ)	Coefficient of variation ($\frac{\sigma}{\mu}$)
Ceiba speciosa (A.St.-Hil., A.Juss. & Cambess.) Ravenna	69	[2.5, 34.03]	9.84	10	33.07	5.75	0.58
Pseudalbizzia niopoides (Spruce ex Benth.) E.J.M.Koenen & Duno	66	[2.5, 81.25]	13.55	10	197.26	14.04	1.04
Trithrinax schizophylla Drude	66	[6.25, 840]	134.85	75	30,619.53	174.98	1.3
Didymopanax morototoni (Aubl.) Decne. & Planch.	65	[2.5, 59.22]	7.44	5	66.48	8.15	1.1
Urera baccifera (L.) Gaudich. ex Wedd.	60	[2.5, 271.53]	25.99	10.13	1,844.51	42.95	1.65
Plinia rivularis (Cambess.) Rotman	58	[2.5, 247.22]	39.27	20	2,563.16	50.63	1.29
Handroanthus impetiginosus (Mart. ex DC.) Mattos	57	[5.56, 78.43]	19.61	10	303.28	17.41	0.89
Ocotea porphyria (Griseb.) van der Werff	55	[10, 246.86]	40.17	20	2,111.26	45.95	1.14
Cordia ecalyculata Vell.	54	[2.5, 50]	13.71	6.25	219.86	14.83	1.08
Neltuma elata (Burkart) C.E.Hughes & G.P.Lewis	54	[10, 80]	17.22	10	186.73	13.66	0.79
Pterogyne nitens Tul.	53	[2.78, 78.43]	17.32	10	264.25	16.26	0.94
Cedrela balansae C.DC.	52	[10, 156.86]	27.89	12.5	755.33	27.48	0.99
Handroanthus heptaphyllus (Vell.) Mattos	52	[2.5, 70]	14.46	10	212.79	14.59	1.01
Juglans australis Griseb.	52	[10, 137.65]	39.66	20	1,145.79	33.85	0.85
Tabernaemontana catharinensis A.DC.	50	[2.5, 362.94]	22.79	10	2,888.77	53.75	2.36

Table S2. Source and names of variables used to perform the Principal Component Analysis.

Source	Name
Chelsa	Annual mean temperature
	Temperature Seasonality (standard deviation ×100)
	Mean Temperature of Driest Quarter
	Mean Temperature of Warmest Quarter
	Annual precipitation
	Precipitation Seasonality (Coefficient of Variation)
	Precipitation of Wettest Quarter
	Precipitation of Driest Quarter
	Precipitation of Warmest Quarter
	Climate moisture index
	Near-surface relative humidity
	Vapor pressure deficit
	Surface downwelling shortwave radiation
SRTM	Elevation
SoilGrids*	Bulk density
	Cation exchange capacity
	Volumetric fraction of coarse fragments (> 2 mm)
	Proportion of clay particles (< 0.002 mm)
	Total nitrogen (N)
	Organic carbon density
	Organic carbon stocks
	Soil pH
	Proportion of sand particles (> 0.05/0.063 mm)
	Proportion of silt particles (≥ 0.002 mm and ≤ 0.05/0.063 mm)
	Soil organic carbon content

*Variables disponible in seven depths.

Table S3. Variance explained by principal components selected from the principal component analysis performed with climate, topographic, and edaphic data.

Principal Components	Variance explained for each PC	Cumulative variance explained
1	33.44	33.44
2	26.08	59.52
3	13.11	72.64
4	8.34	80.98
5	4.84	85.82
6	3.04	88.86
7	2.04	90.91

Table S4. Hyperparameter values tested for each algorithm in model tunning process. The number of combinations tested is a multiplication of the number of values tested for each hyperparameter.

Model	Hyperparameter	Range tested	Combinations tested
RAF	mtry	{2, 3, ..., 7}	60
	ntree	{100, 300, ..., 1900}	
GBM	n.trees	{100, 200, 400, 600, 800}	600
	interaction.depth	{1, 2, 3, ..., 10}	
	n.minobsinnode	{5, 10, 15, 20}	
	shrinkage	{0.001, 0.01, 0.1}	
XGB	nrounds	{100, 300}	432
	max_depth	{4, 6, 8}	
	eta	{0.2, 0.5}	
	gamma	{1, 5, 10}	
	colsample_bytree	{0.5, 1}	
	min_child_weight	{0.5, 1, 2}	
	subsample	{0.5, 1}	
	objective	reg:tweedie	
SVM	C	{1, 3, 5, ..., 49}	50
	sigma	"automatic"	
	kernel	{"rbfdot", "laplacedot"}	
GLM	poly	{1, 2}	68
	inter_order	{0, 1}	
	distribution	Every suitable distribution*	
GAM	inter	"automatic"	17
	distribution	Every suitable distribution*	
DNN	batch_size	32	128
	learning_rate	{0.005, 0.001}	
	n_epochs	{100,200}	
	n_layers	{2, 3, 4}	
	n_neurons	{7, 14, 21}	
CNN	batch_size	32	108**
	learning_rate	{0.005, 0.001, 0.01}	
	n_epochs	200	
	validation_patience	{5, 10}	
	fitting_patience	{5, 10}	
	number_of_conv_layers (convolutional)	{3, 4, 5}	
	conv_layers_size	{7, 14, 28}	
	number_of_fc_layers	1	
	fc_layers_size	14	
NET	size	{7, 12, 17, ..., 97}	190
	decay	{0, 0.1, 0.2, ..., 0.9}	

*For GAM and GLM, suitable distributions for the data were achieved using *adm::family_selector*, which return a set 17 probability distributions, all of them tested: {"NO", "NOF", "RG", "TF", "ZAIG", "LQNO", "DEL", "PIG", "WARING", "YULE", "ZALG", "ZIP", "BNB", "DBURR12", "ZIBNB", "LO", "PO"}.

**Nine architectures random combining number of layers and layer size were tested. Therefore, this number comes from $batch\_size \cdot learning\_rate \cdot n\_epochs \cdot validation\_patience \cdot fitting\_patience \cdot architectures = 1 \cdot 3 \cdot 1 \cdot 2 \cdot 2 \cdot 9 = 4 \cdot 9 = 108$

Table S5. Estimated means by post-hoc analysis of model discrimination (Spearman correlation), accuracy (MAE), and precision (PDISP), for different algorithms, partition type, and absence amount. Bold values denote the highest performance value for each partition and absence amount treatment.

Performance metrics	Algorithms	No-absence		One-absence		Two-absences	
		K-fold	Spatial	K-fold	Spatial	K-fold	Spatial
Spearman	CNN	0.255	0.261	0.152	0.153	0.134	0.136
	DNN	0.418	**0.384**	0.571	**0.532**	0.555	**0.517**
	NET	0.372	0.332	0.478	0.433	0.469	0.425
	GBM	0.396	0.346	0.563	0.508	0.555	0.501
	XGB	**0.419**	0.371	**0.572**	0.518	**0.556**	0.504
	RAF	0.323	0.259	0.472	0.403	0.481	0.414
	SVM	0.351	0.285	0.545	0.474	0.534	0.464
	GLM	0.405	0.368	0.554	0.511	0.542	0.501
	GAM	0.376	0.331	0.542	0.492	0.532	0.484
MAE	CNN	0.088	0.088	0.063	0.064	0.044	0.044
	DNN	**0.077**	**0.080**	**0.053**	**0.055**	**0.038**	**0.040**
	NET	0.099	0.105	0.073	0.078	0.056	0.060
	GBM	0.080	0.081	0.056	0.057	0.041	0.041
	XGB	0.082	0.084	0.054	**0.055**	0.039	**0.040**
	RAF	0.094	0.100	0.065	0.069	0.050	0.053
	SVM	0.081	0.085	0.056	0.058	0.042	0.044
	GLM	0.081	0.083	0.056	0.058	0.041	0.042
	GAM	0.083	0.086	0.056	0.058	0.041	0.042
PDISP	CNN	0.811	0.827	0.420	0.435	0.388	0.403
	DNN	0.995	0.995	0.989	0.988	0.990	0.989
	NET	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	GBM	0.803	0.797	0.910	0.904	0.946	0.939
	XGB	**1.000**	**1.000**	1.001	**1.000**	**1.000**	0.999
	RAF	0.606	0.599	0.598	0.591	0.599	0.592
	SVM	0.914	0.932	0.809	0.826	0.760	0.777
	GLM	0.992	0.993	0.998	0.998	0.978	0.978
	GAM	0.828	0.807	0.974	0.953	0.910	0.888

Figures

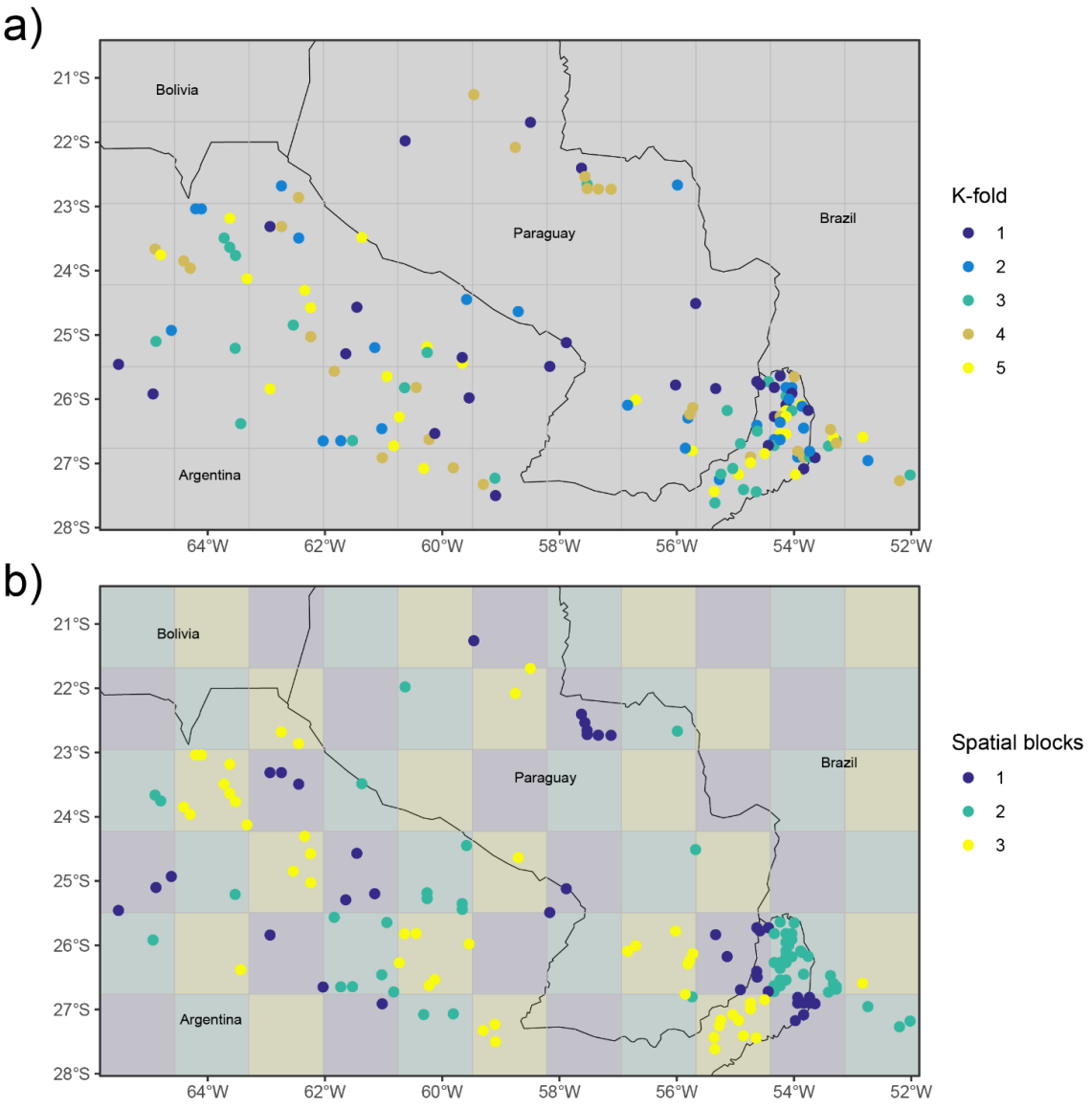


Figure S1. Data partitioning example with *Ceiba speciosa* (A.St.-Hil.) Ravenna points. a) K-fold random partition. b) Spatially structured block partition. For both, numbers and color refer to which fold the points belong.

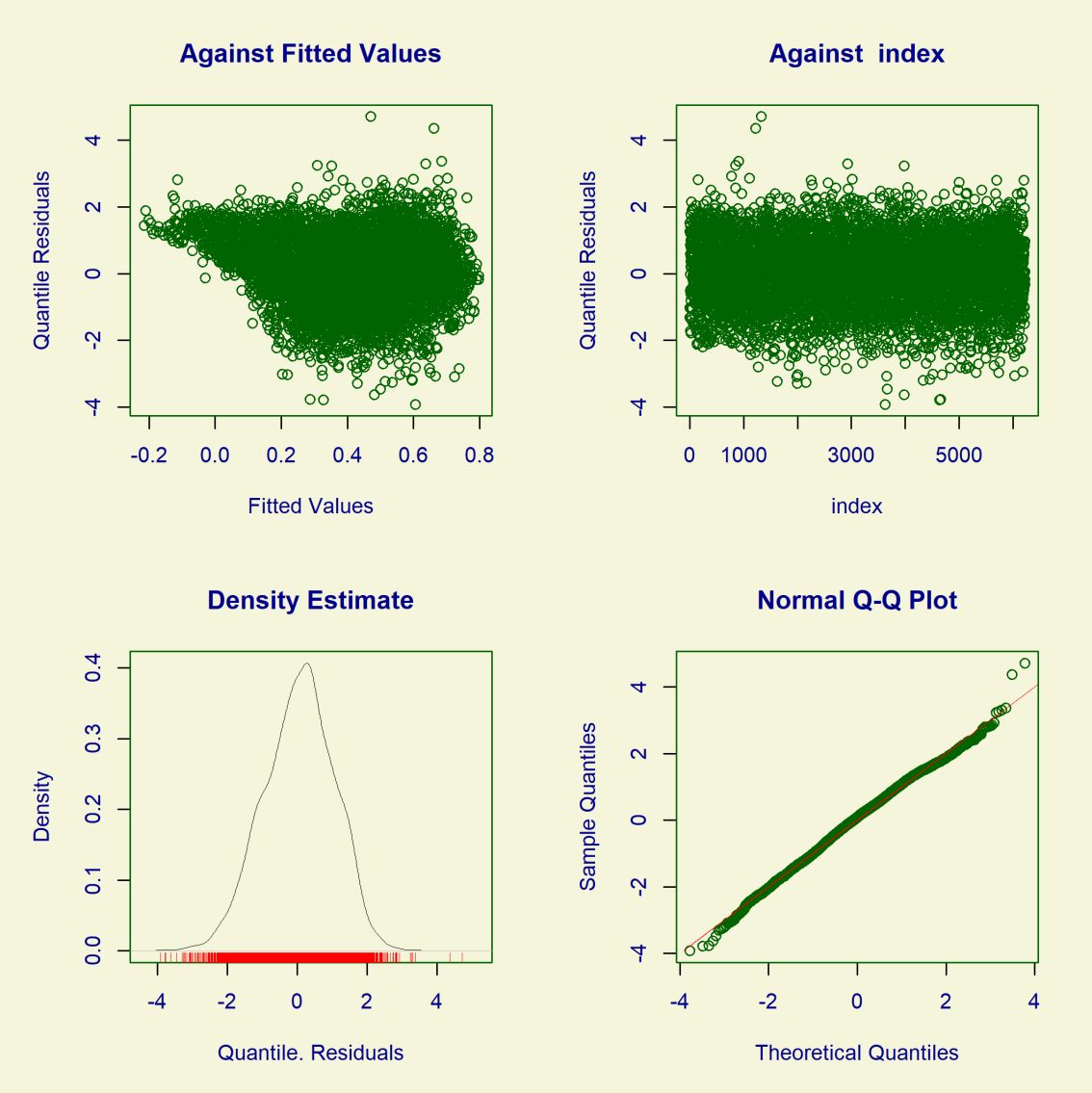


Figure S2. Residuals of GAMLSS model used for evaluating the performance ADM based Spearman's correlation metric for different algorithms, partition type, and absence amount.

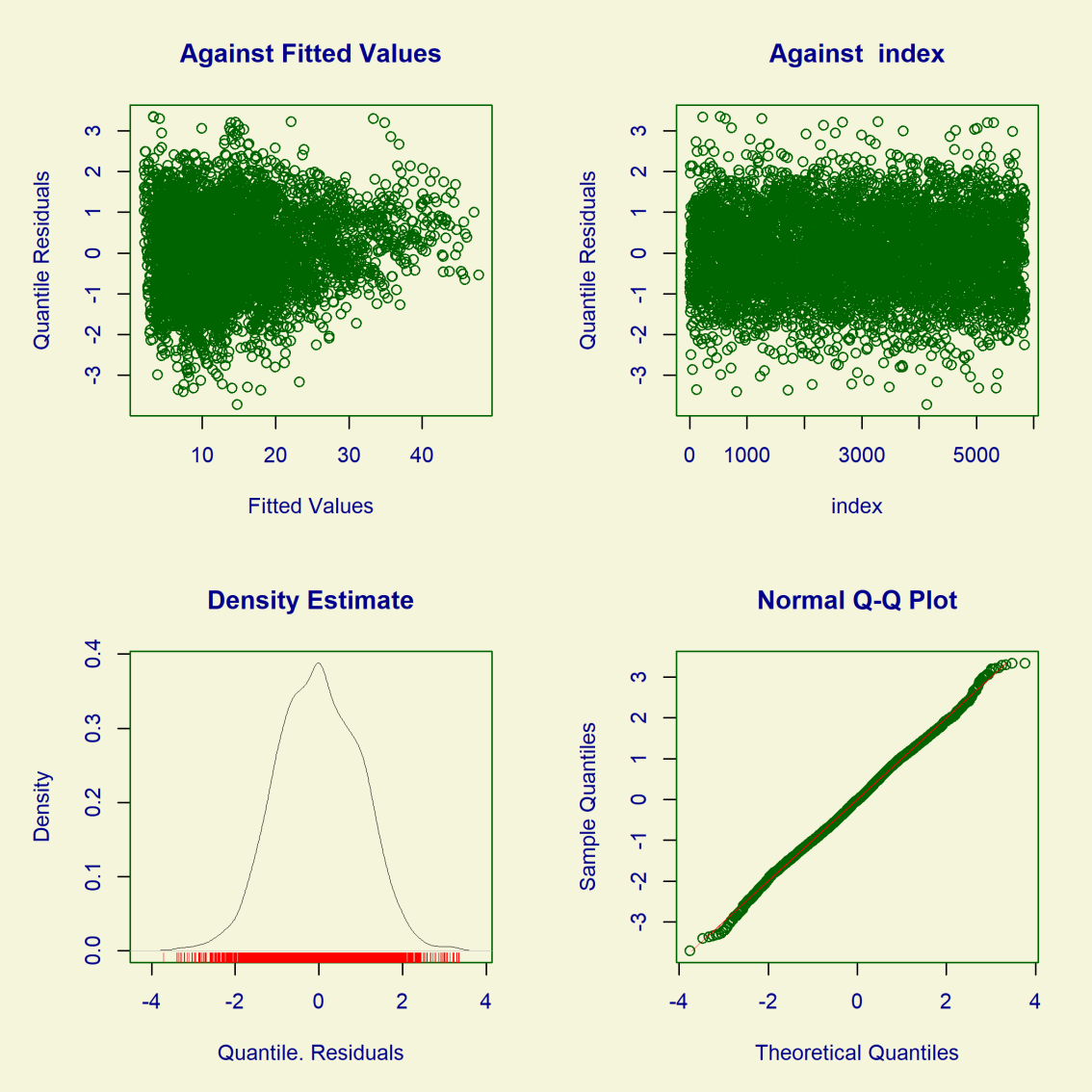


Figure S3. Residuals of GAMLSS model used for evaluating the performance ADM based Mean Absolute Error metric for different algorithms, partition type, and absence amount.

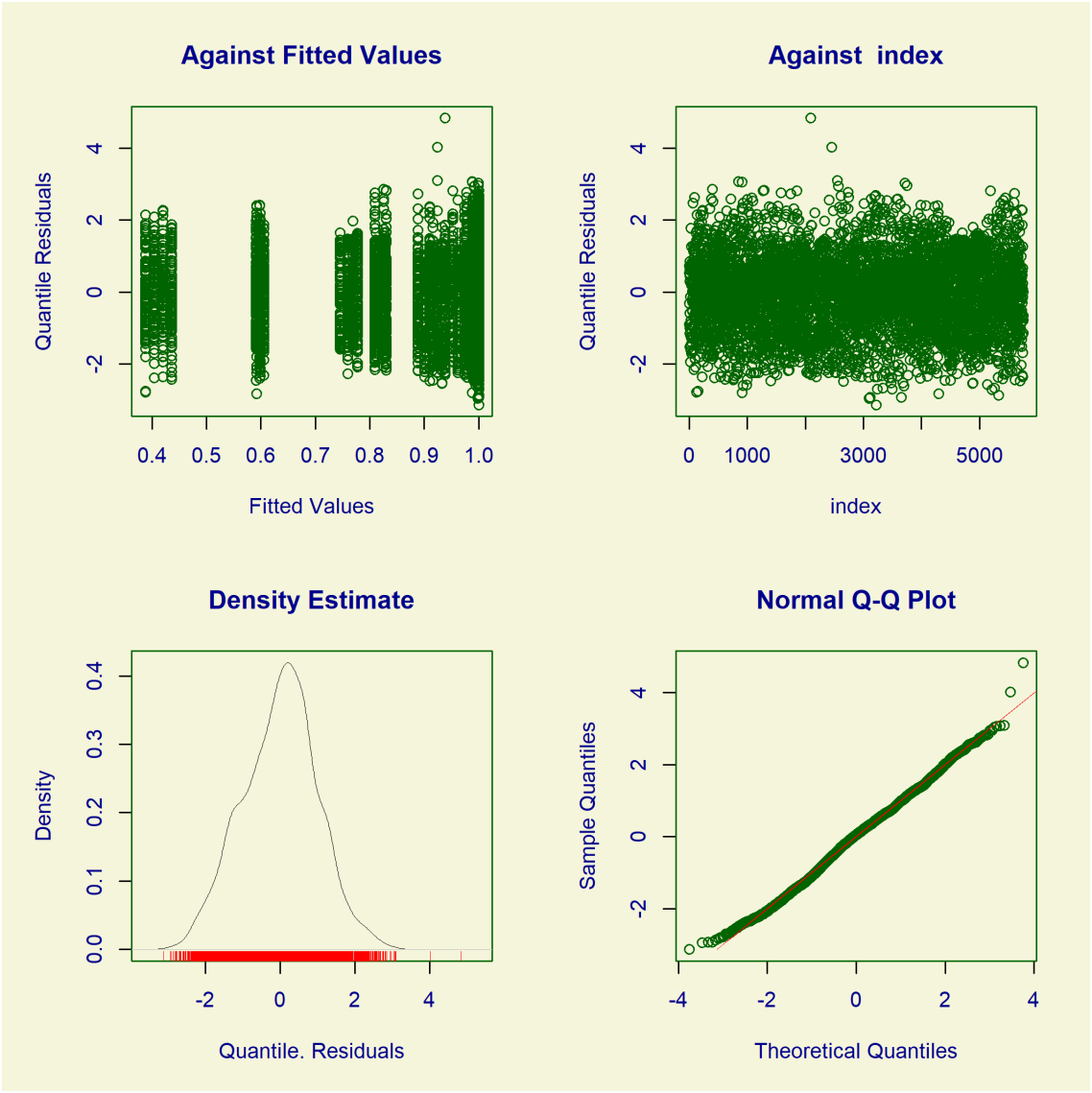


Figure S4. Residuals of GAMLSS model used for evaluating the performance ADM based Dispersion metric for different algorithms, partition type, and absence amount.

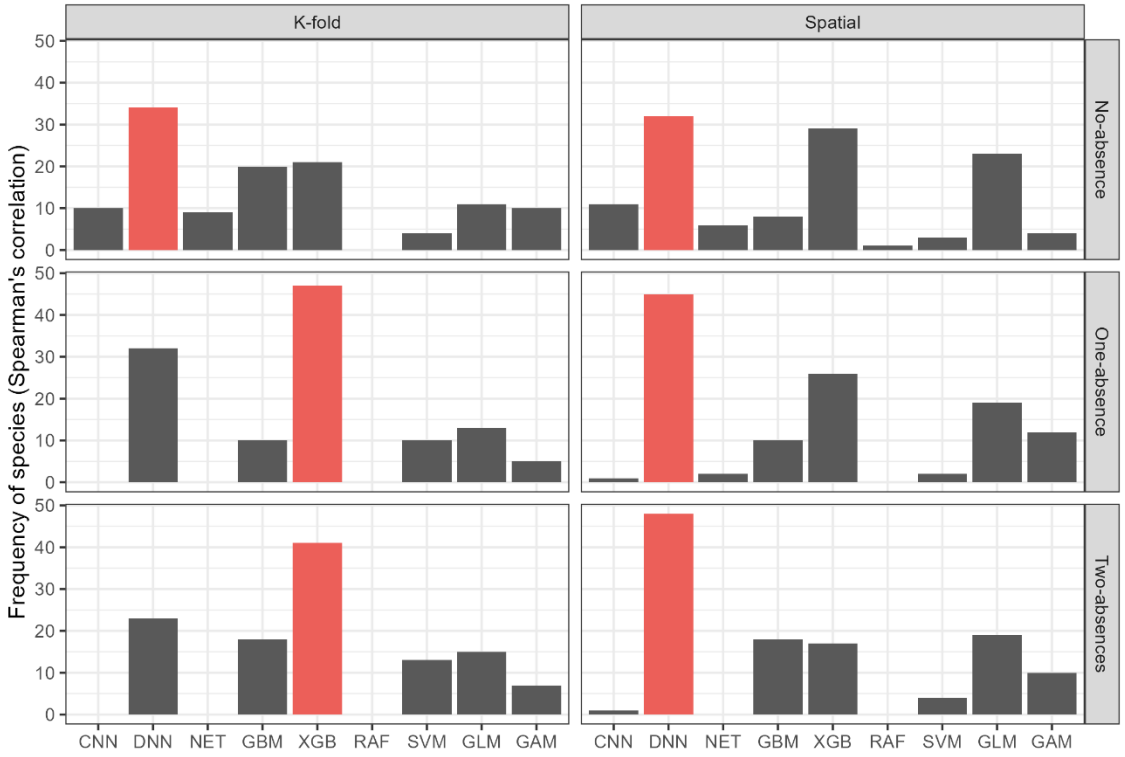


Figure S5. Frequency of species with the highest discrimination (i.e., the highest Spearman's correlation) different algorithms, partition type (columns), and absence amount (rows). Red-colored bars represent the best algorithm.

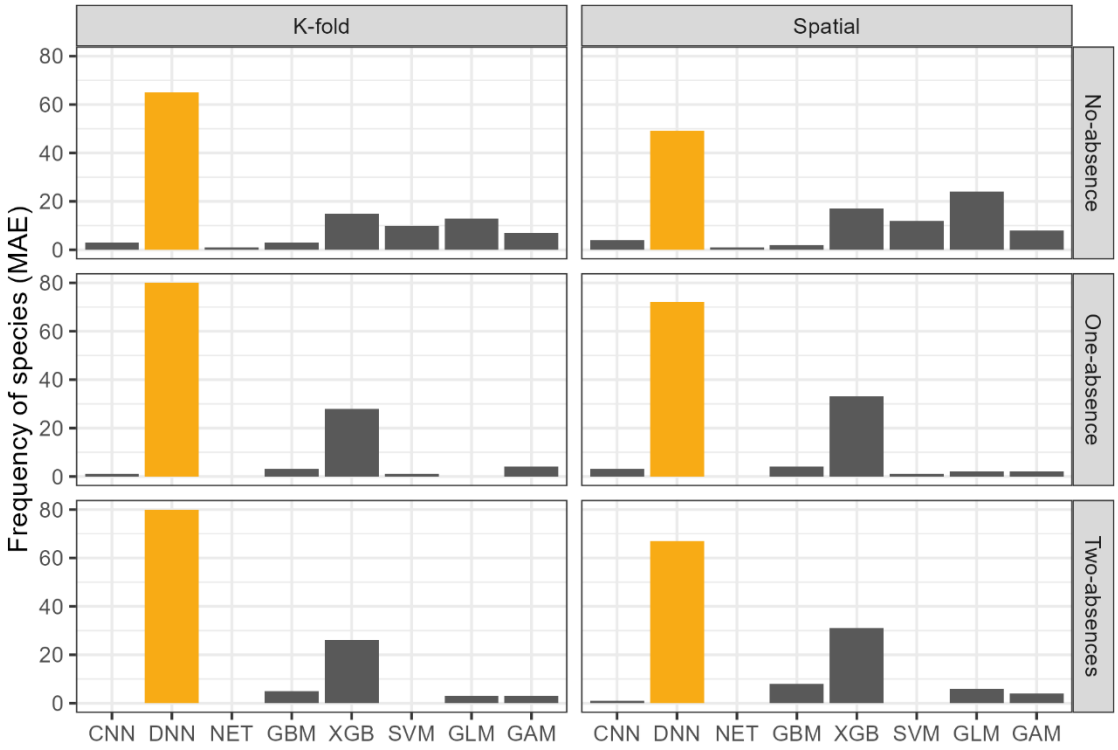


Figure S6. Frequency of species with the highest accuracy (i.e., the lowest MAE) different algorithms, partition type (columns), and absence amount (rows). Yellow-colored bars represent the best algorithm.

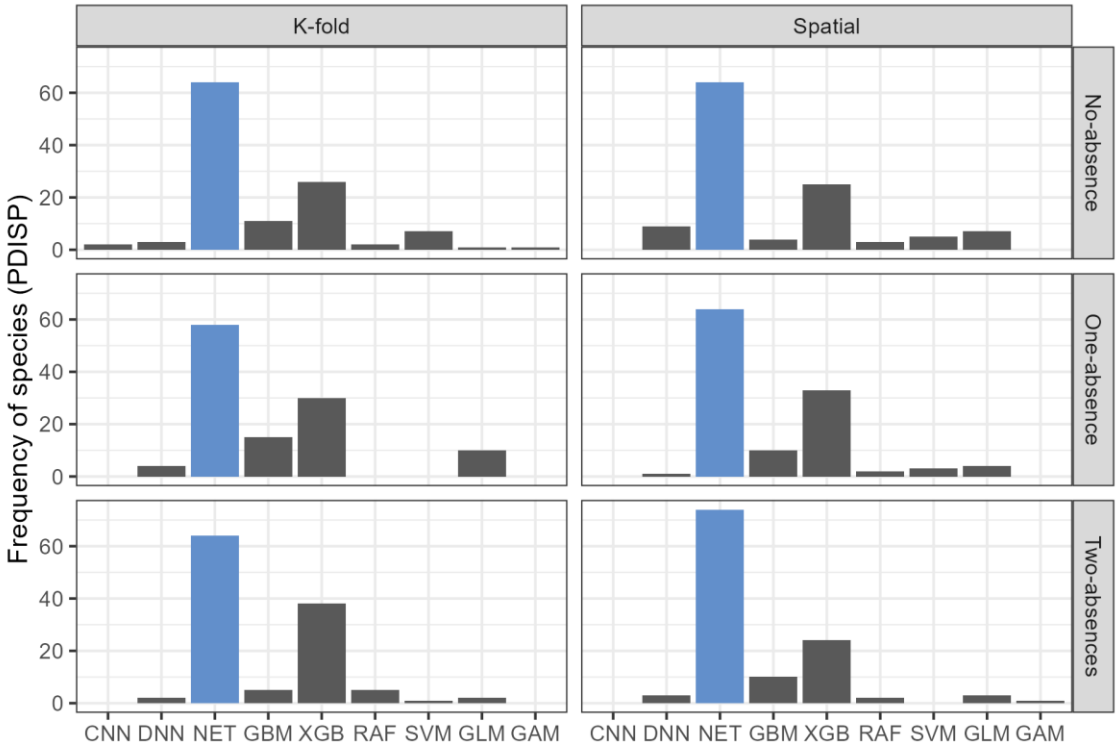


Figure S7. Frequency of species with the highest precision (i.e., PDISP near to 1) different algorithms, partition type (columns), and absence amount (rows). Blue-colored bars represent the best algorithm.

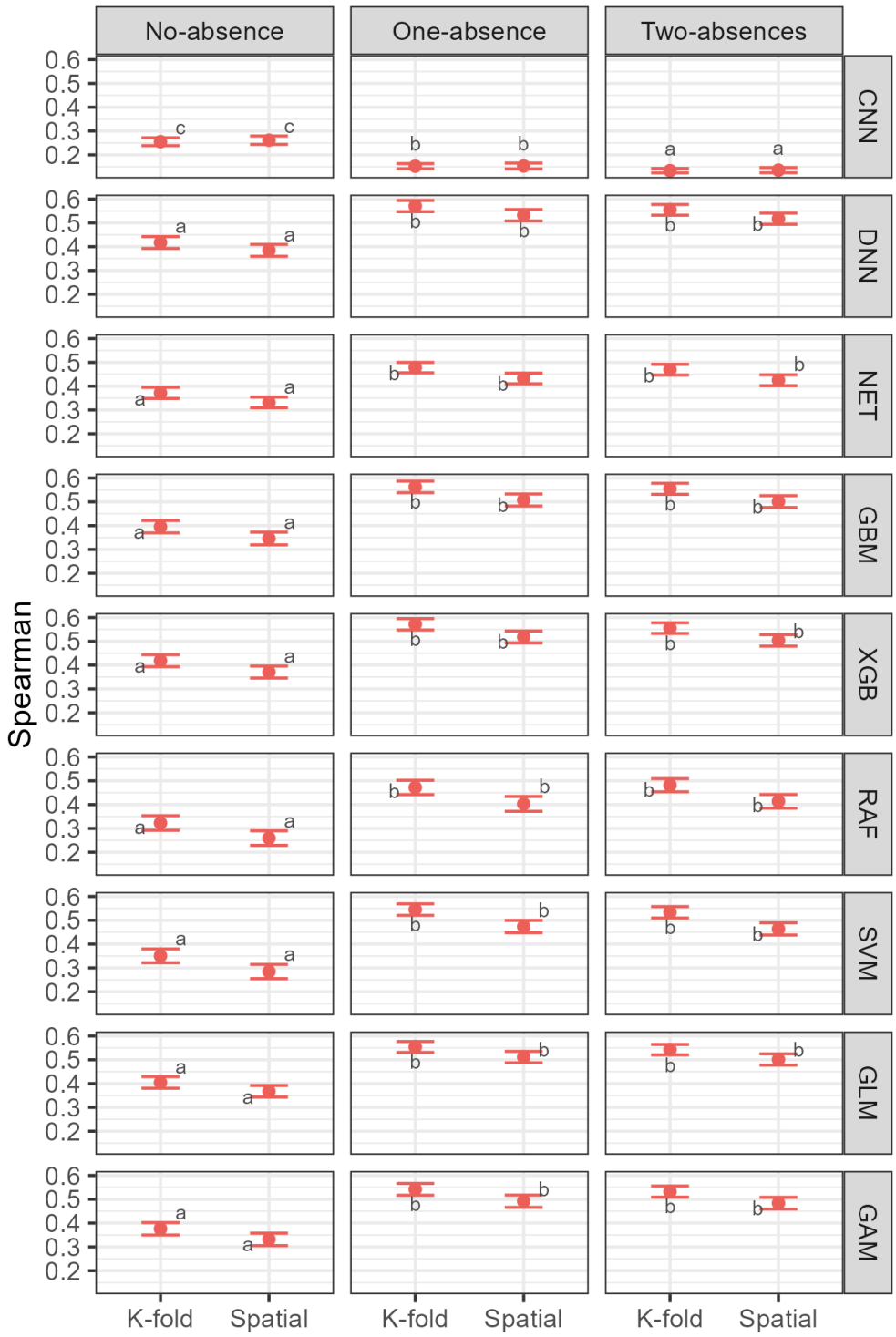


Figure S8. Post-hoc analysis of model discrimination based on Spearman metric for different partitions, partition type (columns), and algorithms (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

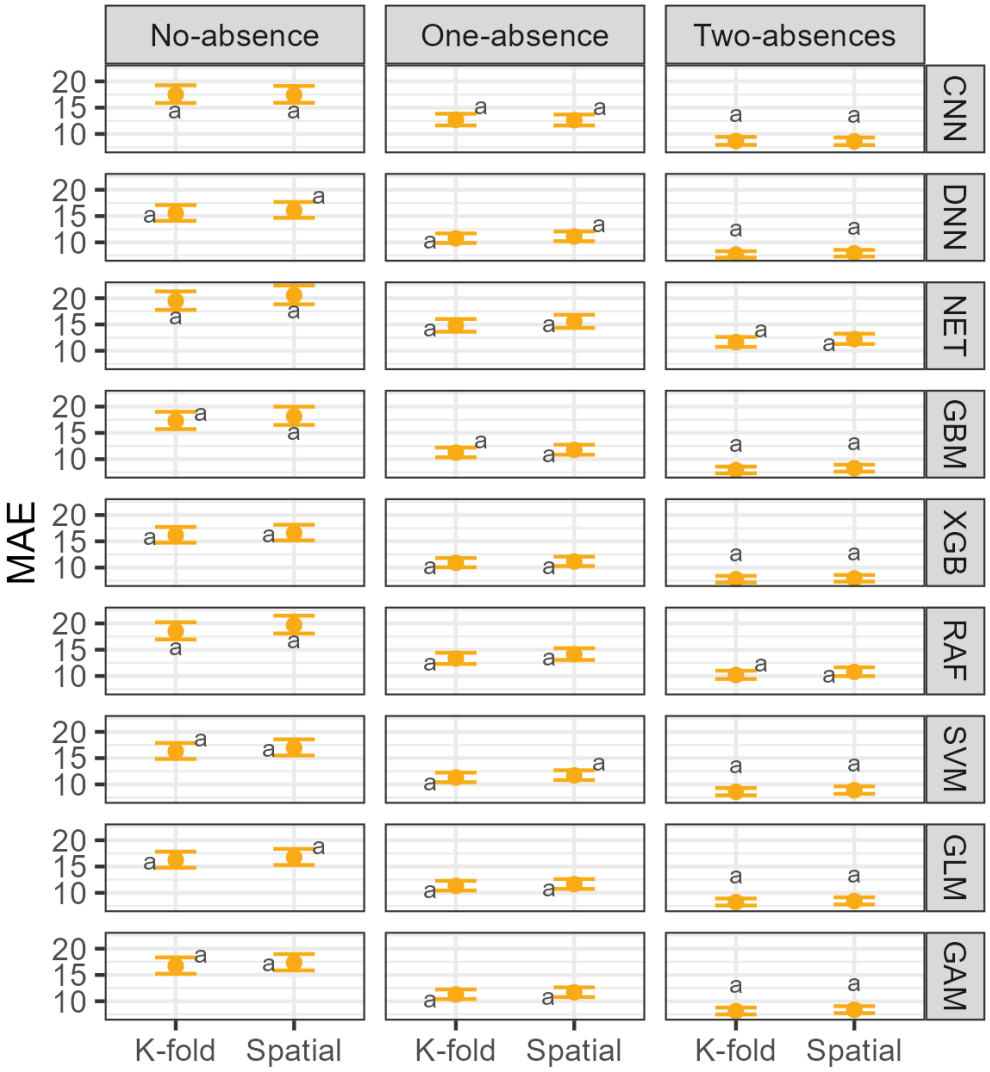


Figure S9. Post-hoc analysis of model discrimination based on MAE metric for different partitions, partition type (columns), and algorithms (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

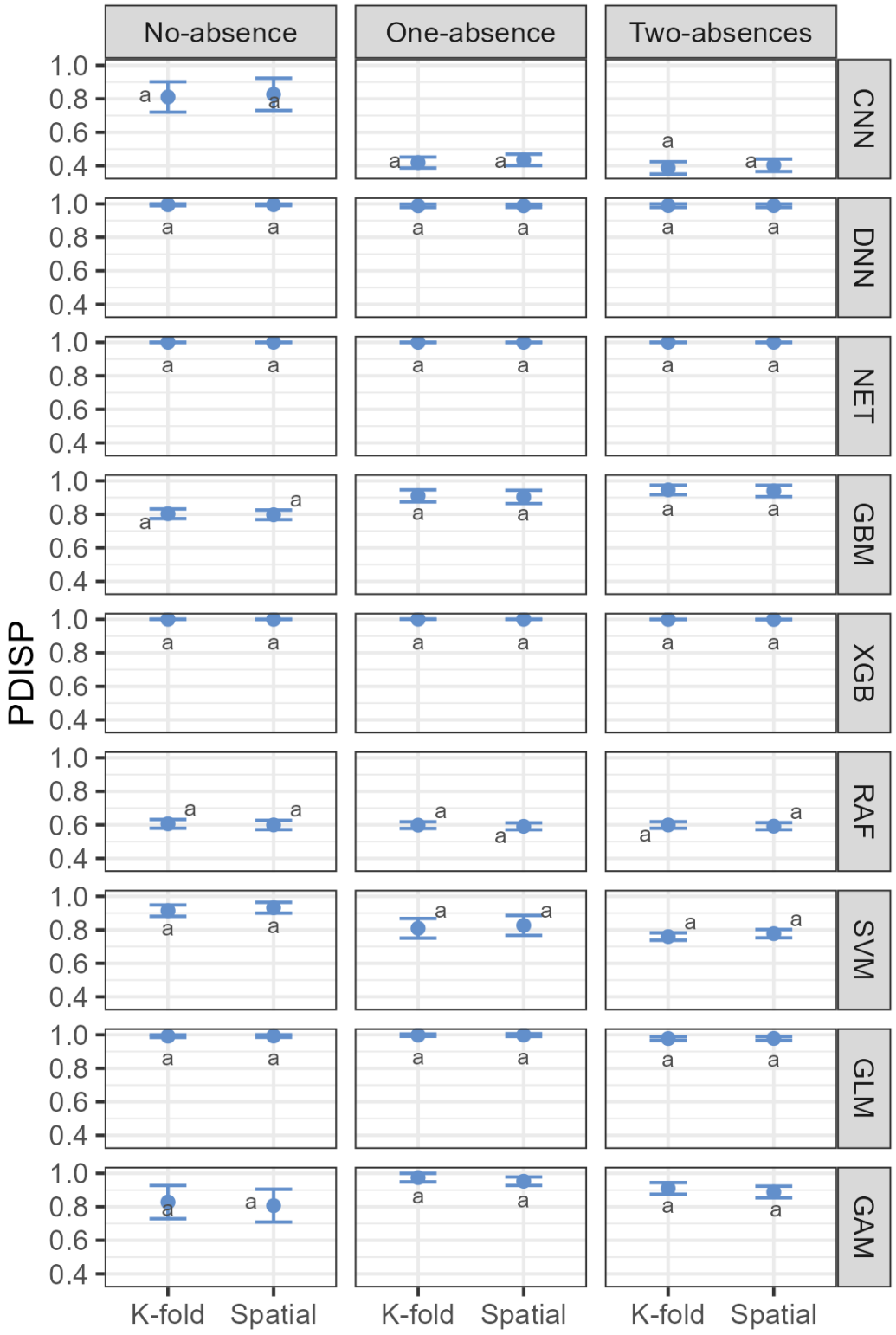


Figure S10. Post-hoc analysis of model discrimination based on PDISP metric for different partitions, partition type (columns), and algorithms (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

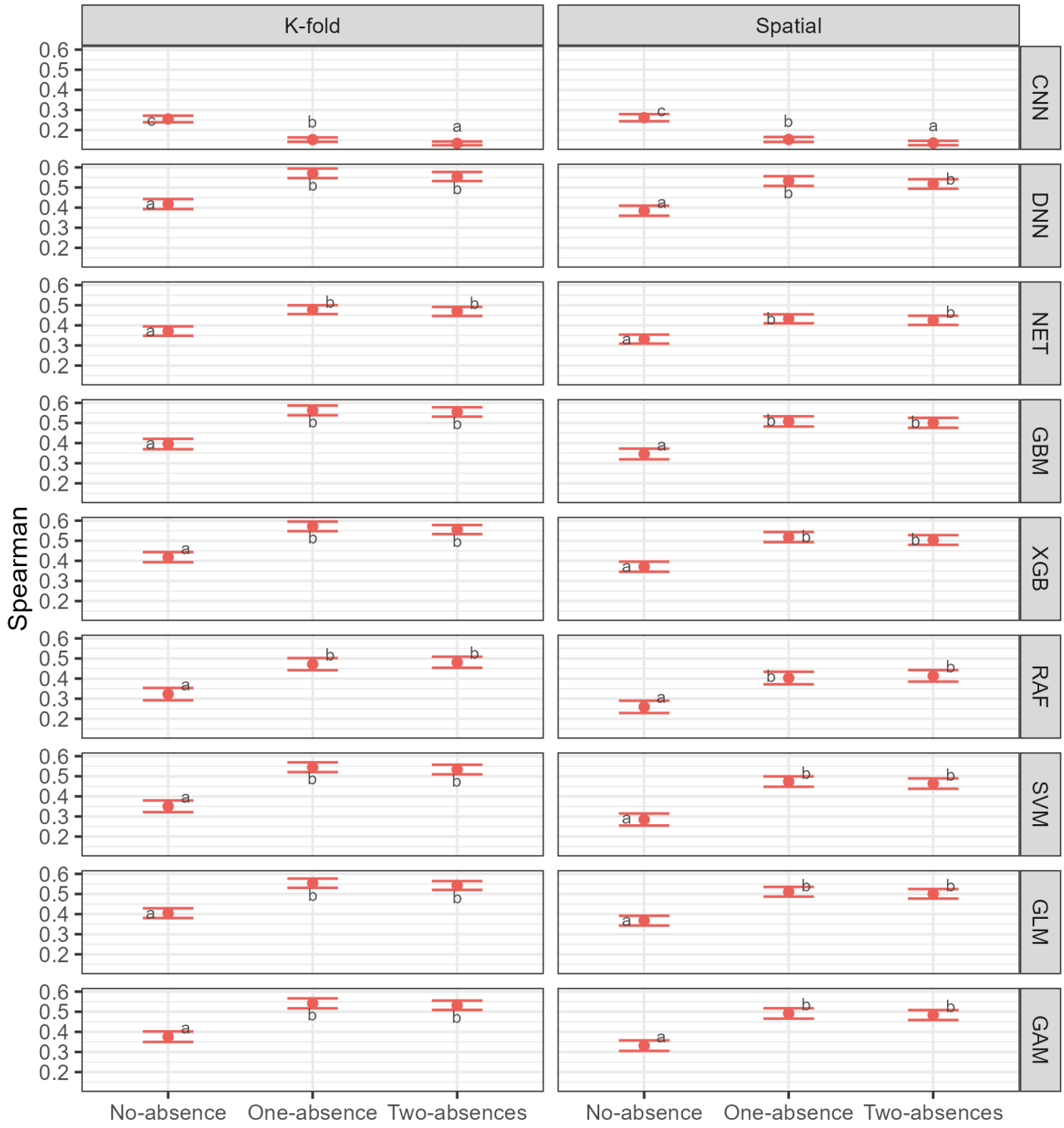


Figure S11. Post-hoc analysis of model discrimination based on Spearman metric for different absence amounts, partition type (columns), and algorithms (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

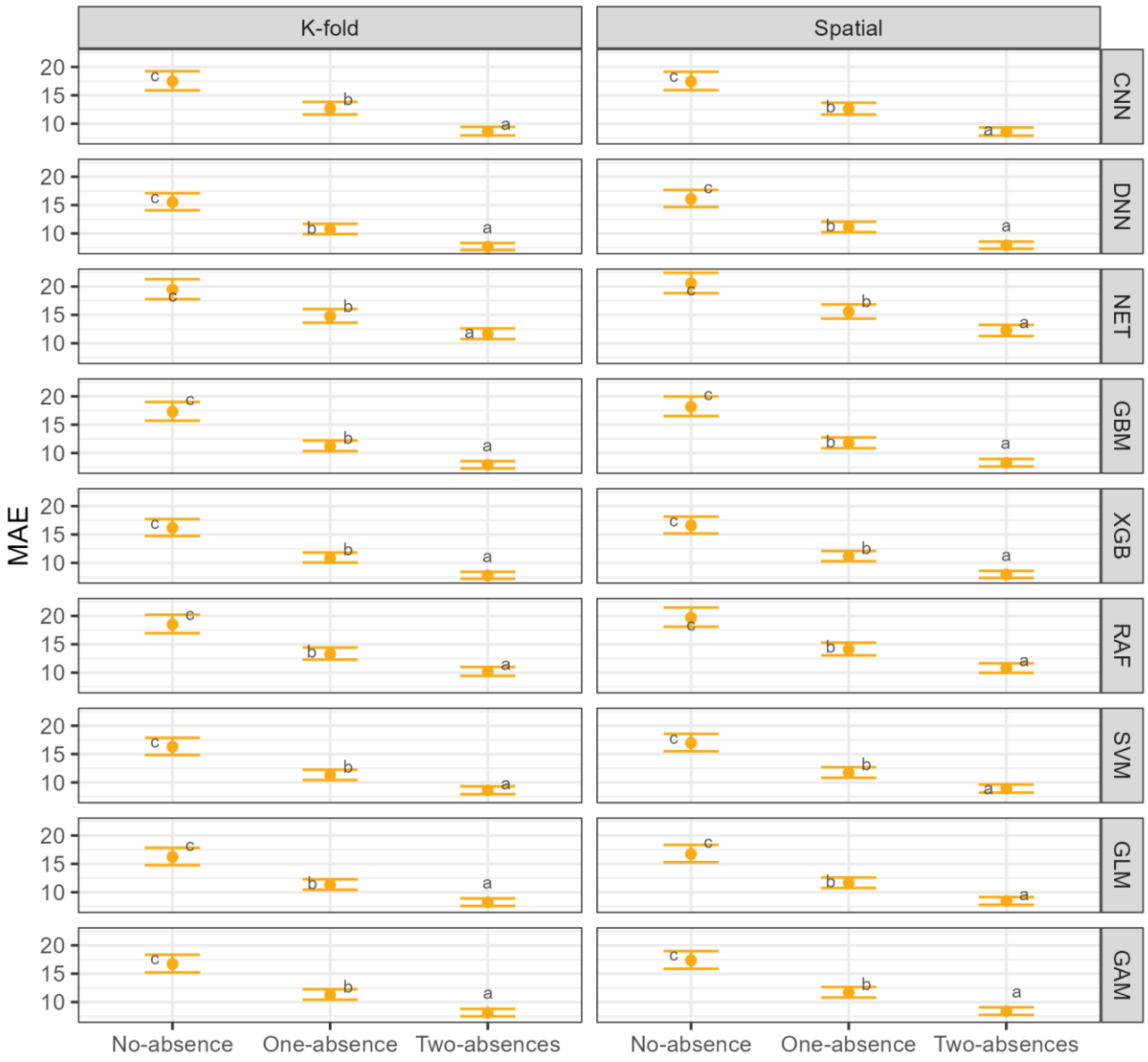


Figure S12. Post-hoc analysis of model discrimination based on MAE metric for different absence amounts, partition type (columns), and algorithms (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

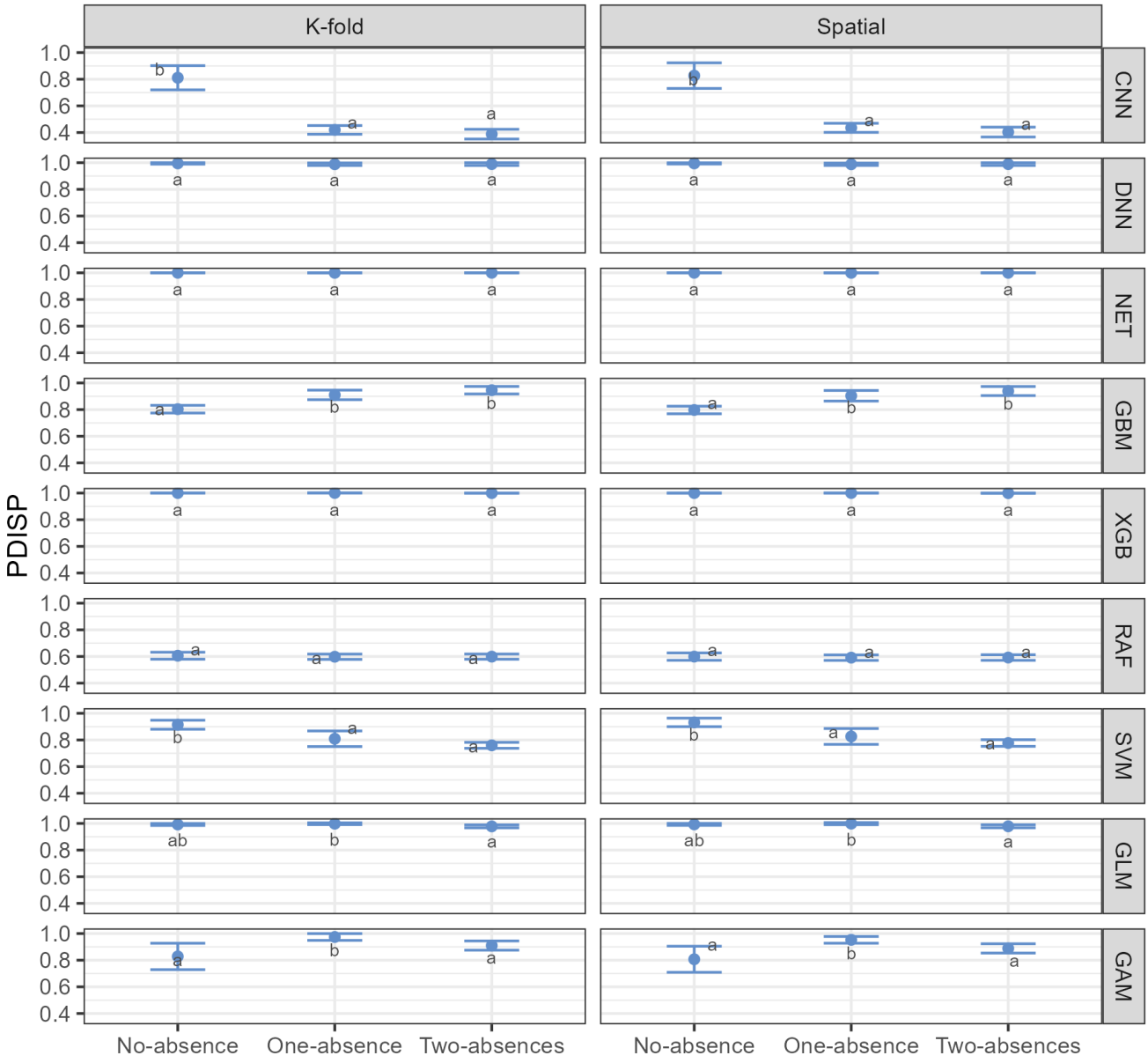


Figure S13. Post-hoc analysis of model discrimination based on Spearman metric for different absence amounts, partition type (columns), and algorithms (rows). Different letters within each panel indicate statistical significance differences according to the HDS Tukey test ($p < 0.05$).

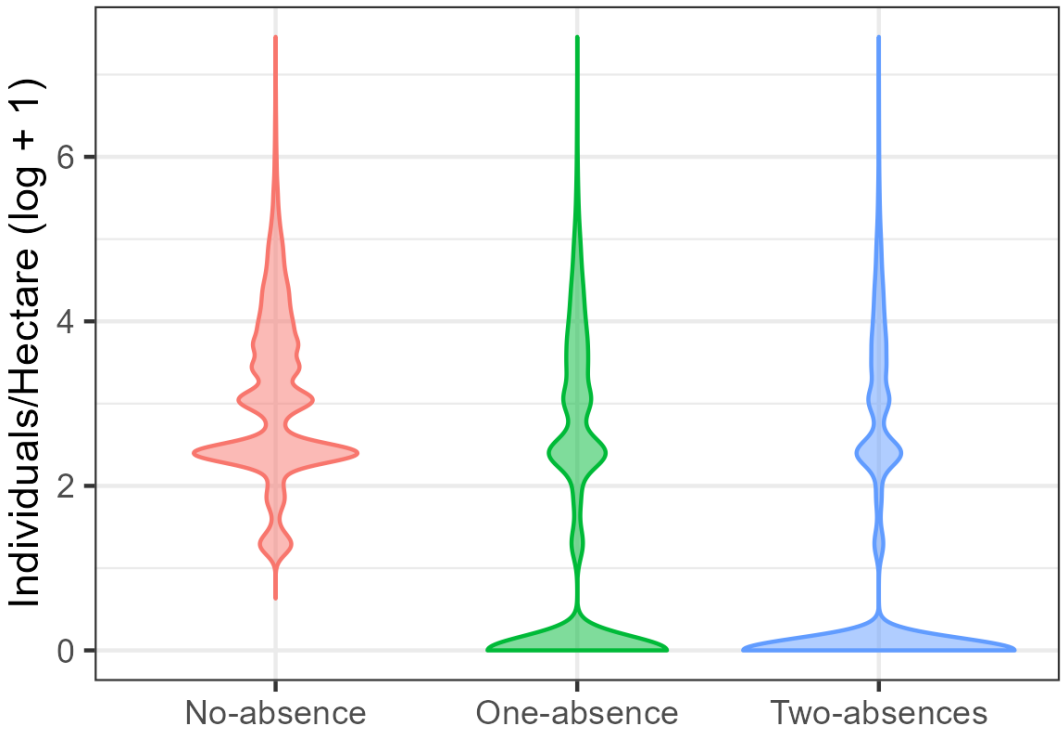


Figure S14. Density distribution of abundance (Individuals/Hectare), transformed with log+1, for the three absences amount treatment tested, No-absence, One-absence, Two-absences. Densities include entire data from the 117 species.

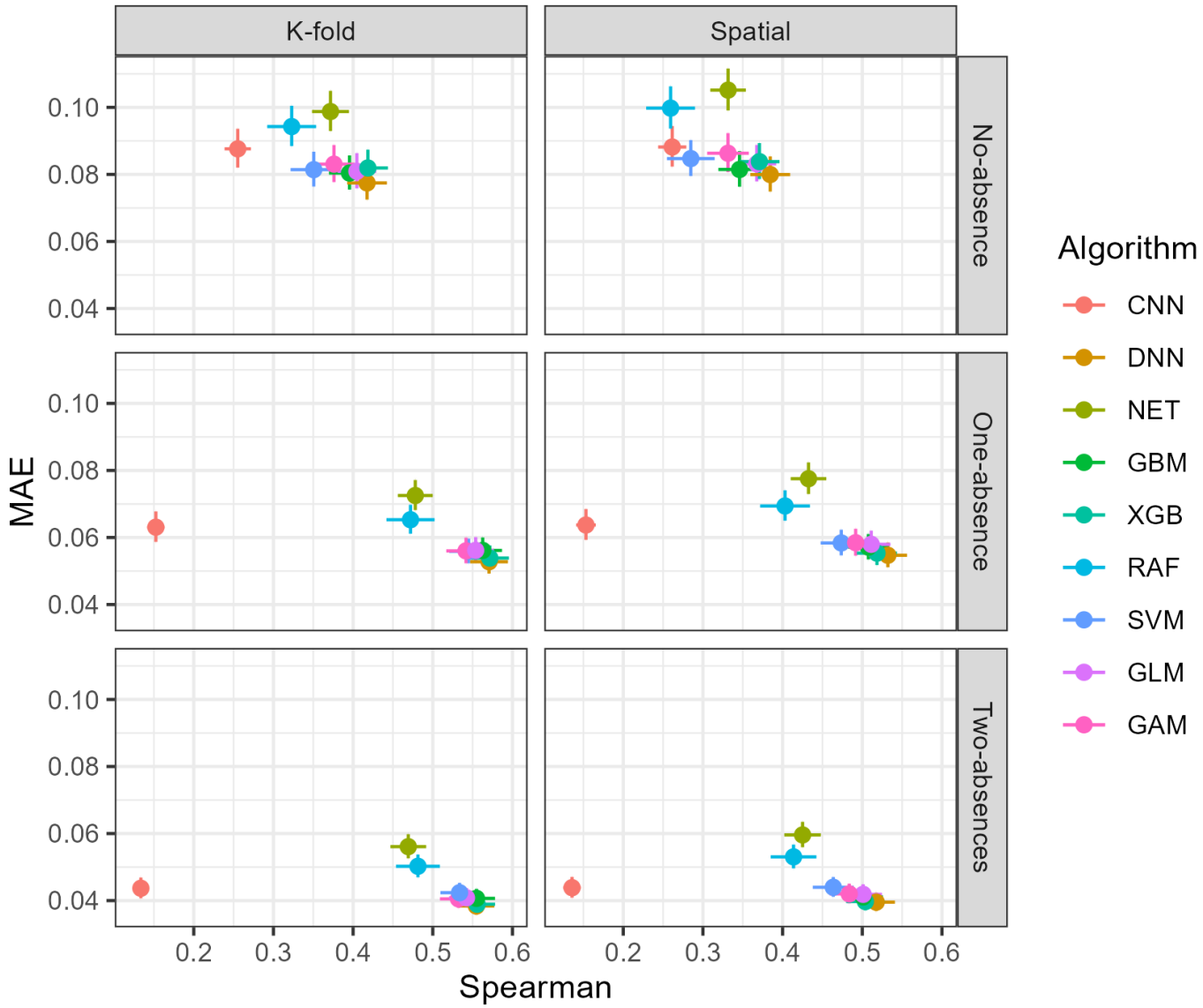


Figure S15. Algorithms' performance under different absence amount and data partitioning treatments. Spearman correlation (x-axis) and MAE (y-axis) are shown with uncertainty bars. Higher Spearman and lower MAE indicate better performance.

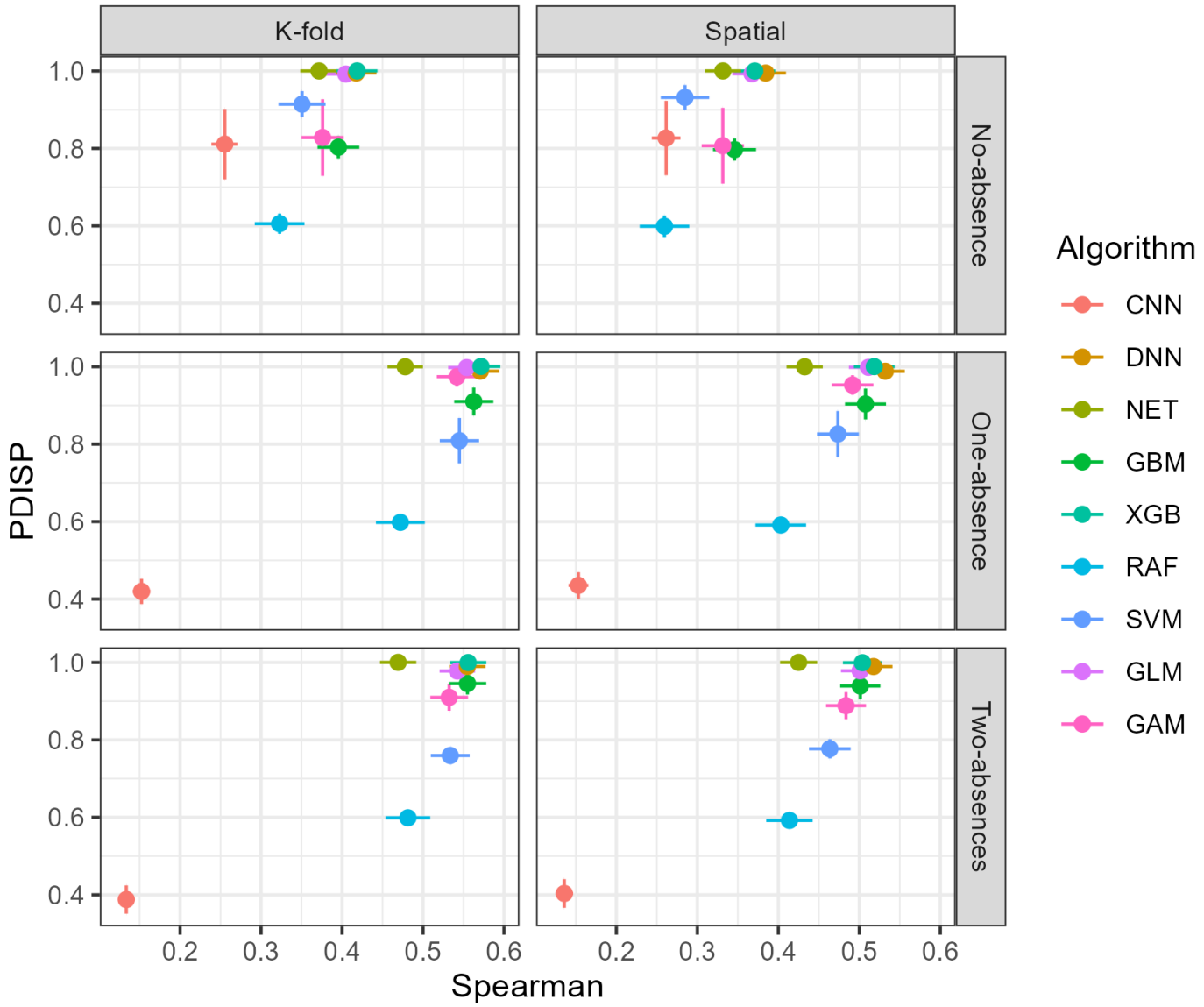


Figure S16. Algorithms' performance under different absence amount and data partitioning treatments. Spearman correlation (x-axis) and PDISP (y-axis) are shown with uncertainty bars. Higher Spearman and PDISPT closer to 1 indicate better performance.

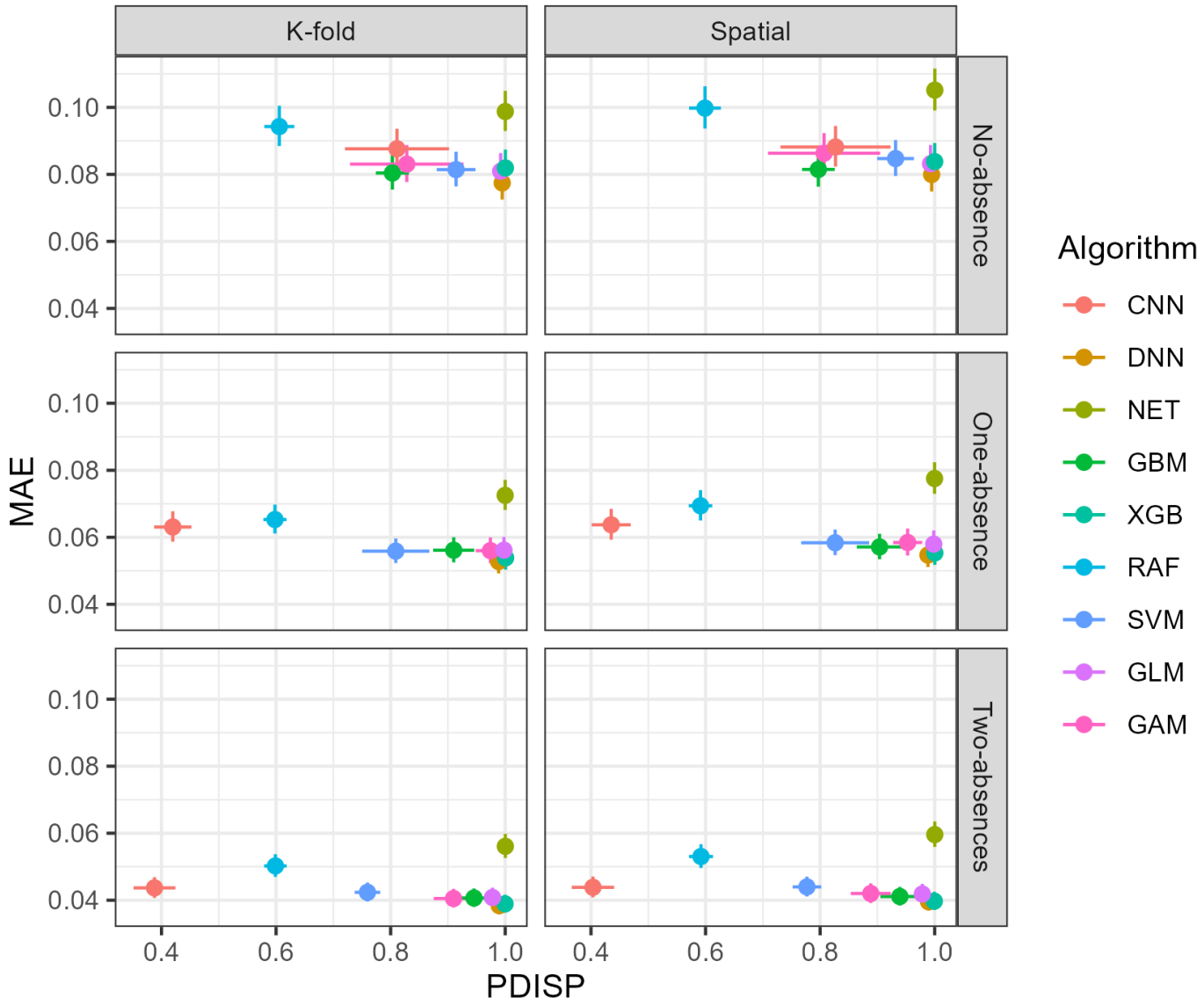


Figure S17. Algorithms' performance under different absence amount and data partitioning treatments. PDISP (x-axis) and MAE (y-axis) are shown with uncertainty bars. PDISP closer to 1 and lower MAE indicate better performance.