WESLEY LOURENCO BARBOSA

AVALIAÇÃO DO IMPACTO DA QUALIDADE DE DADOS EM MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES

SÃO PAULO

2023

WESLEY LOURENCO BARBOSA

AVALIAÇÃO DO IMPACTO DA QUALIDADE DE DADOS EM MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências.

Área de concentração:

Engenharia de Computação

Orientador(a):

Profa. Dra. Solange Nice Alves de Souza

SÃO PAULO

2023

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 08 de Maio de 2023

Assinatura do autor: Wesley Lourenco Barbosa

Assinatura do orientador:

Catalogação-na-publicação

Barbosa, Wesley Lourenco
Avaliação do Impacto da Qualidade de Dados em Modelos de Distribuição de Espécies / W. L. Barbosa -- versão corr. -- São Paulo, 2023.
121 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Qualidade de dados 2.Modelos de distribuição de espécies 3.Simulação 4.Maxent 5.Espécies virtuais I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

RESUMO

Os modelos de distribuição de espécies se tornaram uma ferramenta importante em ecologia, biogeografia, sustentabilidade e, mais recentemente, em gestão de conservação. No entanto, problemas de qualidade presentes nos dados utilizados na modelagem de distribuição de espécies podem resultar em modelos imprecisos e que não refletem o real padrão de distribuição das espécies. Como consequência, estratégias de conservação baseadas em modelos de distribuição gerados por dados enviesados, podem resultar em desperdício de recursos financeiros ou perda importante de biodiversidade. Assim, o objetivo deste trabalho é investigar como problemas de qualidade de dados afetam os resultados dos modelos de distribuição de espécies. A metodologia do trabalho emprega uma estratégia de simulação que consiste na criação de duas bases de dados, uma base de controle e outra de erros. A base de controle é constituída por dados ambientais e dados simulados de presença e ausência de uma espécie virtual. A base de erros é imputada com problemas de qualidade e utilizada para a amostragem de diferentes gradientes de erros para teste. Os resultados da revisão de escopo indicaram que erros de localização, erros de identificação e viés geográfico são os mais comuns em dados de ocorrência de espécies. Os algoritmos de Maximum Entropy Modeling (Maxent), Random Forest (RF) e Generalized Linear Model (GLM), Neural Network (NN) e Extreme Gradient Boosting (XGBoost) foram utilizados e avaliados quanto a robustez e capacidade de generalização mesmo para amostras de treinamento com erros de qualidade de dados. O XGBoost gerou modelos bastante robustos a diversos tipos e intensidades de erros. O GLM gerou os modelos mais sensíveis aos problemas de qualidade. O tipo de erro de viés geográfico foi o que teve maior efeito sobre os resultados dos modelos, enquanto os erros de localização, embora muito discutidos na literatura científica, só geraram impacto expressivo quando a amostra estava contaminada por erros de alta intensidade. A métrica AUC (Area Under the Curve), comumente utilizada para validar modelos de aprendizado de máquina para tarefas de classificação, mostrou-se pouco susceptível à presença de erros nos dados de treinamento, por outro lado, as métricas Kappa, MCC (Matthews Correlation Coefficient), TSS (True Statistics Skill) estão entre as mais sensíveis a problemas de qualidade. Este trabalho empregou a simulação de espécies virtuais, geradas a partir do comportamento

identificado de 6 espécies de nicho ecológico amplo e restrito obtidas no repositório de dados de biodiversidade GBIF, para avaliar o impacto de diferentes gradientes de três tipos de erros de qualidade de dados em modelos de distribuição de espécie. Os resultados trazem um aprofundamento importante no entendimento dos impactos dos erros nos dados de ocorrência de espécies, e contribuem para avanço da área de estudos de qualidade de dados em estudos de biodiversidade e conservação.

Palavras-chave: Qualidade de dados. Modelos de distribuição de espécies. Simulação. Maxent. Espécies virtuais.

ABSTRACT

Species distribution models have become an important tool in ecology, biogeography, sustainability and, more recently, in conservation management. However, quality problems present in the data used in species distribution modeling can result in inaccurate models that do not recognize the actual pattern of species distribution. Therefore, conservation strategies based on distribution models generated by biased data can result in wasted financial resources or important loss of biodiversity. Thus, the objective of this work is to investigate how data quality problems affect the results of species distribution models. The methodology employs a simulation strategy that consists of creating two databases, a control database and an error database. The control base consists of environmental data and simulated data of the presence and absence of a virtual species. The error database is imputed with quality problems and used for sampling different error gradients for testing. The results of the scoping review indicated that location errors, identification errors and geographic bias are the most common in species occurrence data. The Maximum Entropy Modeling (Maxent), Random Forest (RF) and Generalized Linear Model (GLM), Neural Network (NN) and Extreme Gradient Boosting (XGBoost) algorithms were used and evaluated for robustness and generalization ability even for training samples with data quality errors. XGBoost generated models that are quite robust to different types and intensities of errors. The GLM generated the models most sensitive to data quality issues. The type of geographic bias error had the greatest effect on the results of the models, while location errors, although much discussed in the scientific literature, only generated a significant impact when the sample was contaminated by high-intensity errors. The AUC (Area Under the Curve) metric, commonly used to validate machine learning models for classification tasks, proved to be little susceptible to the presence of errors in the training data, on the other hand, Kappa, MCC (Matthews Correlation Coefficient) and TSS (True Statistics Skill), are among the most sensitive to data quality problems. This work used the simulation of virtual species, generated from the identified behavior of 6 species of broad and narrow ecological niche obtained from the biodiversity data repository GBIF, to evaluate the impact of different gradients of three types of data quality errors in species distribution models. The results bring an important deepening in the understanding of the impacts of errors in species occurrence data, which

contributes to advancing the area of data quality studies in biodiversity and conservation studies.

Keywords: Data quality. Species distribution models. Simulation. Maxent. Virtual species.

LISTA DE FIGURAS

Figura 1: Algoritmos utilizados em MDEs.....40

Figura 2: Fluxo do processo de criação de MDEs.....43

Figura 3: Representação do erro de localização.....50

Figura 4: Erro de identificação da espécie.....51

Figura 5: Erro de viés geográfico.....52

Figura 6: Mapa de com pontos de ocorrência das espécies.....55

Figura 7: Matriz de correlação de variáveis com nível de significância de 0.01 - andorinha.....60

Figura 8: Importância de variáveis DALEX/RFE – andorinha.....62

Figura 9: Importância de variáveis Variable importance/Boruta - andorinha.....63

Figura 10: Ranking de variáveis capivara/andorinha.....64

Figura 11: Ranking de variáveis canada/pardo.....65

Figura 12: Ranking de variáveis euro/pireneus.....66

Figura 13: Projeção das ocorrências das bases de controle.....69

Figura 14: Boxplot das métricas AUC e F1 para modelos de controle e de erro.....71

Figura 15: Gráfico de densidade com análise de normalidade e igualdade das variâncias.....72

Figura 16: *Boxplot* da métrica F1 para cada tipo de problema de qualidade em comparação aos resultados dos modelos de controle.....74

Figura 17: Teste de Wilcoxon emparelhado para F1 de controle e cada tipo de erro.....75

Figura 18: *Boxplot* da métrica F1 estratificado por tipo de erro e por espécie.....76

Figura 19: Média de F1 estratificada por tipo de erro e por espécie 77

Figura 20: Tamanho de efeito dos resultados dos modelos com erros em relação aos modelos de controle.....79

Figura 21: Métrica F1 estratificada por intensidade de erro 80

Figura 22: Tamanho de efeito por intensidade dos erros de localização estratificado por espécie 81

Figura 23: Degradação da métrica F1 estratificado pela intensidade e proporção de presença de erro de localização 82

Figura 24: Tamanho de efeito por intensidade dos erros de viés geográfico estratificado por espécie 86

Figura 25: Tamanho de efeito por intensidade dos erros de identificação estratificado por espécie 88

Figura 26: Taxa de variação das métricas de validação em função da intensidade de erro.....90

Figura 27: Taxa de variação da métrica composta por algoritmo em função da intensidade de erro – espécies de nicho ecológico amplo.....95

Figura 28: Taxa de variação da métrica composta por algoritmo em função da intensidade de erro – espécies de nicho ecológico restrito.....96

LISTA DE QUADROS E TABELAS

Tabela 1: Definição das dimensões de qualidade. 21

Tabela 2: Variáveis ambientais que podem ser utilizadas para espécies terrestres. 26

Tabela 3: Erros Identificados. 38

Tabela 4: Consolidado da extração dos dados dos artigos 39

Tabela 5: Espécies de interesse e suas características 55

Tabela 6: Lista de variáveis selecionadas 67

Tabela 7: Número de ocorrências na base de controle gerada para cada espécie virtual 68

Tabela 8: Valor-p para os testes de Wilcoxon de erros de localização emparelhados 83

Tabela 9: Valor-p para os testes de Wilcoxon de erros de viés geográfico emparelhados 87

Tabela 10: Teste de Wilcoxon emparelhado por métrica e tipo de erro 93

LISTA DE ABREVIAÇÕES

ANN – *Artificial Neural Network*

CS – *Citizen Science*

CTA – Classification Tree Analysis

FDA – *Flexible Discriminant Analysis*

GAM – *Generalized Additive Model*

GARP – *Genetic Algorithm for Rule-set Production*

GBM – *Generalized Boosting Model*

GLM – *Generalized Linear Model*

GLMM – *Generalized Linear Mixed Model*

GPP – *Gaussian Process Prior*

IPP – *Inhomogeneous Poisson Process*

LDA – *Linear Discriminant Analysis*

MARS – *Multivariate Adaptive Regression Splines*

MCC – *Matthews Correlation Coefficient*

MDE – Modelo de Distribuição de Espécies

MDEs – Modelos/Modelagem de Distribuição de Espécies

RF – *Random Forest*

RSL – Revisão Sistemática da Literatura

SPEDInstabR – *Species Distribution Instability*

SPP – *Spatial Point Process*

SVM – *Support Vector Machine*

TSS – *True Statistics Skill*

XGBoost – *Extreme Gradient boosting*

WRF – *Weighted Random Forest*

SUMÁRIO

1 INTRODUÇÃO 12

1.1 Justificativa 14

1.2 Objetivo 15

1.3 Metodologia 15

2 CONTEXTUALIZAÇÃO TEÓRICA 17

2.1 Gerenciamento de Dados de Pesquisa 17

2.2 Qualidade de Dados 18

2.3 Dimensões de Qualidade de Dados 20

2.4 Modelos de Distribuição de Espécies 21

2.5 Tipos de Dados 23

2.5.1 Dados de Ocorrência de Espécies 23

2.5.2 Dados Ambientais 25

2.6 Problemas de Qualidade em Dados de Ocorrência de Espécies 27

2.6.1 Erros de Localização 28

2.6.2 Erros de Identificação 29

2.6.3 Viés Geográfico 30

2.7 Métricas de Validação 30

2.8 Técnicas de Modelagem 32

2.8.1 *Maximum Entropy Modeling* (Maxent) 32

2.8.2 *Random Forests* 33

2.8.3 Modelos Lineares Generalizados 34

3 RESULTADOS DA REVISÃO DE ESCOPO DA LITERATURA 36

3.1 Metodologia da Revisão de Escopo da Literatura 36

3.1.1. Planejamento 36

3.1.2. Condução 37
3.1.3 Comunicação 38
3.2 Resultados da Revisão 38
3.3 Conclusão da REL 41
4 PROCESSO DE CRIAÇÃO DE MDES PARA EXPERIMENTO 43
4.1 Dados de Ocorrência 44
4.2 Dados Ambientais 46
4.3 Matriz de Dados e Base de Controle 46
4.4 Base de Erros 47
4.4.1 Erro de localização 47
4.4.2 Erro de identificação 48
4.4.3 Viés Geográfico 49
4.5 Amostragem 49
4.6 Modelagem e Validação 53
4.7 Caracterização dos dados de ocorrência 54
4.8 Caracterização dos dados ambientais 56
4.9 Construção da base de controle 57
4.9.1 Análise de correlação dos atributos 59
4.9.2 Sorteio de amostras 60
4.9.3 Ranqueamento de atributos 61
4.9.4 Seleção final de atributos 66
4.9.5 Modelagem de algoritmos 67
4.9.6 Geração da base de controle 67
5 ANÁLISE DOS RESULTADOS 70
6 CONCLUSÃO 99
7 REFERÊNCIAS 102

1 INTRODUÇÃO

Modelos de Distribuição de Espécies (MDEs) são ferramentas estatísticas utilizadas para a geração de predições probabilísticas da presença de entidades biológicas no espaço geográfico (ELITH et al., 2006; GUISAN; ZIMMERMANN, 2000). De acordo com (BROTONS, 2014), ecologistas e biólogos têm, tradicionalmente, construído MDEs utilizando bases de dados que residem principalmente em suas estações de trabalho ou em redes locais, coletados por eles mesmos ou por pequenos grupos de pesquisa. Porém, essa estratégia de análise está se tornando um desafio à medida que os cientistas precisam aperfeiçoar seus modelos, empregando novas técnicas e adicionando novos componentes às suas análises, para aprimorar o embasamento para decisões de conservação sob diferentes cenários de mudanças climáticas. Os requisitos de dados para este tipo de modelagem consistem na presença conhecida e, quando disponível, na ausência de avistamento ou localizações das espécies, bem como os valores de variáveis ambientais ou climáticas que definem a adequabilidade do habitat dessas espécies nesses locais. Os dados climáticos são geralmente extraídos de dados históricos, medições de estações meteorológicas e sensoriamento remoto. Os dados de biodiversidade, por sua vez, são obtidos por expedições de campo realizadas por especialistas, ou, ainda, reunidas a partir de dados fornecidos por voluntários em projetos de *Citizen Science* (*CS*).

As relações entre o organismo e as variáveis ambientais são os aspectos considerados pelos MDEs. No entanto, para estudar e compreender sistemas ecológicos em uma escala maior, mais dados precisam ser coletados com resoluções mais precisas em amplos limites espaciais e temporais (BARBOSA et al., 2019). No entanto, custeios para instalação de estações meteorológicas e a disponibilidade de especialistas para coletar as quantidades necessárias de dados de biodiversidade não são suficientes para a obtenção desses dados nas extensões e periodicidades requeridas. Nesse cenário, projetos de *CS* surgiram como uma maneira eficiente de reunir esses dados, envolvendo um número maior de pessoas e compilando suas observações ecológicas, propiciando o crescimento mais rápido dos dados de distribuição de espécies que são obtidos por voluntários que participam desses projetos (KELLING et al., 2015b; STUART et al., 2010). Como resultado dessas iniciativas, tem-se o aumento da quantidade de dados em resolução espacial e

temporal, da velocidade com que os dados são gerados, e da heterogeneidade das fontes a partir dos quais esses dados podem ser coletados. No entanto, a falta de consideração de aspectos importantes que podem interferir na adequabilidade dos dados para a modelagem de distribuição de espécies pode resultar em modelos cujo valor prático é limitado. A riqueza de informações que esses dados podem prover e que poderia ajudar a melhorar os MDEs, pode ficar perdida se aspectos sobre a qualidade dos dados utilizados não forem levados em conta.

Em alguns campos de pesquisa, questões relacionadas a problemas de qualidade de dados foram extensamente exploradas. Na área de saúde, (CHEN et al., 2014; NDABARORA; CHIPPS; UYS, 2014) apontam que a má qualidade dos dados pode resultar na alocação inadequada do financiamento do sistema de saúde e na falha na vigilância da saúde pública, ou ainda, em casos extremos, colocar em risco a segurança dos pacientes (MAURICIO PINTO-VALVERDE et al., 2013). Em segurança pública, dados imprecisos sobre criminalidade interferem no processo de alocação de recursos financeiros e humanos, além de ter potencial de prejudicar o tempo de resposta para eventos atípicos (BENNETT, 2018). Em biodiversidade, decisões sobre conservação biológica, ou em agricultura, baseadas em modelos não representativos da realidade não só podem ocasionar prejuízos financeiros substanciais, como também podem gerar efeitos irreversíveis a longo prazo.

Uma das principais fontes de incertezas em modelagens de distribuição de espécies está relacionada à qualidade dos dados utilizados (DORMANN et al., 2008). Em estudos sobre distribuição de espécies, diversas dimensões de qualidade podem ser comprometidas devido à problemas que podem ocorrer desde a etapa de coleta, até a agregação de fontes heterogêneas de dados para utilização nos modelos. Na literatura, alguns trabalhos buscaram avaliar a presença de problemas de qualidade em dados bioclimáticos. (DORMANN et al., 2008; TROIA; MCMANAMAY, 2016) analisaram observações de algumas espécies em determinados pontos no tempo e no espaço, e identificaram problemas de completude e temporalidade em dados primários de biodiversidade em muitas regiões e grupos taxonômicos. (SYFERT; SMITH; COOMES, 2013a; TROIA; MCMANAMAY, 2016) abordaram os problemas de qualidade causados pelo enviesamento na amostragem dos dados. (BARBOSA et al., 2019) realizaram um levantamento dos principais problemas de qualidade

identificados no processo de coleta de dados bioclimáticos. O trabalho de (HOWARD et al., 2014) trata dos problemas e limitações de MDEs criados a partir de dados escassos. Outro problema identificado são as incompatibilidades geográficas, que ocorrem quando uma espécie é registrada em uma localidade muito distinta da sua zona de ocorrência padrão (DEVICTOR et al., 2010). O comprometimento das dimensões de confiabilidade, completude e precisão são levantados por (LUKYANENKO; PARSONS; WIERSMA, 2016), que aborda problemas e desafios de qualidade dos dados em projetos de *CS*. Portanto, a qualidade de dados em estudos de biodiversidade é uma preocupação da comunidade científica. Em MDEs, a qualidade de dados é um aspecto importante devido à complexidade e custo para a obtenção dos dados, e a importância dos resultados dos modelos gerados que são utilizados para a tomada de decisão na definição de políticas de conservação e na agricultura.

1.1 Justificativa

A proposta deste trabalho é examinar como a qualidade de dados (QD) afeta os MDEs. Embora algumas iniciativas, como as de (ARENAS-CASTRO et al., 2022; AUBRY; RALEY; MCKELVEY, 2017a; STOCK; MICHELI, 2016; VEIGA; CARTOLANO; SARAIVA, 2014), tenham buscado discutir problemas de qualidade nos dados utilizados em MDEs, a quantificação e caracterização dos impactos desses problemas no processo de modelagem ainda não foi adequadamente explorada na literatura. Realizar a avaliação de QD tornou-se uma questão crítica no contexto dos estudos biogeográficos, especialmente porque, em geral, não são fornecidas informações suficientes sobre a qualidade dos dados, dificultando a compreensão de como possíveis problemas de qualidade podem afetar a utilidade de um subconjunto de dados para propósitos específicos de MDEs.

Os resultados e informações obtidos neste trabalho têm o potencial de serem utilizados para aumentar a conscientização sobre a importância de uma investigação detalhada para verificar a conformação dos dados utilizados no processo de modelagem de distribuição de espécies. Compreender como o elemento de qualidade interfere na adequação dos modelos auxilia os especialistas a definirem metodologias de coleta de dados mais robustas. Os procedimentos de amostragem e as técnicas

de análise de dados também podem ser aperfeiçoados a partir dessa compreensão. Além disso, os resultados deste trabalho podem resultar em orientações importantes sobre abordagens para o tratamento de dados de biodiversidade de acordo com a dimensão de qualidade comprometida.

1.2 Objetivo

O objetivo deste trabalho é avaliar os impactos que problemas de QD causam em MDEs. Os problemas de QD avaliados nos MDEs foram os que mais são discutidos na literatura.

1.3 Metodologia

A partir de dados reais de ocorrências de espécies de nicho ecológico amplo e restrito, foram simuladas espécies virtuais considerando fatores ambientais que caracterizam a existência de cada espécie nas regiões em que são encontradas. Essas simulações foram necessárias para a obtenção de número de dados suficientes para a aplicação de algoritmos para modelar as distribuições de cada espécie. Posteriormente, diferentes intensidades de erros (problemas de qualidade de dados) foram introduzidas nas bases das espécies virtuais simuladas, para verificar o impacto nos modelos de distribuição de cada espécie para os principais problemas de qualidade dos dados discutidos na literatura para MDEs.

Como parte inicial da metodologia seguida fez-se uma revisão de escopo da literatura (REL), para identificação dos problemas de qualidade comumente observados em dados de biodiversidade, dos principais algoritmos empregados na modelagem de distribuição de espécies e como os erros interferem nos modelos. Na sequência selecionou-se bases de dados que poderiam fornecer dados ambientais e de ocorrência de espécies com resolução adequada para a simulação. Os dados obtidos foram tratados e utilizados para a criação de uma base de dados de controle, sem problema de qualidade. Em seguida, uma base de erros foi gerada a partir da base de controle, imputando diferentes intensidades de erros simulados. Finalmente,

amostras foram extraídas das bases de controle e de erro para gerar os MDEs e verificar o impacto dos erros nos resultados dos modelos.

2 CONTEXTUALIZAÇÃO TEÓRICA

2.1 Gerenciamento de Dados de Pesquisa

O gerenciamento de dados de pesquisa é um termo que abrange atividades relacionadas ao armazenamento, organização, documentação e disseminação de dados científicos (BORGHI et al., 2018). Esse gerenciamento preocupa-se em garantir a integridade do processo de pesquisa (COLLINS; TABAK, 2014). Segundo (BERTIN; VISOLI; DRICKER, 2017), nos últimos anos, diversas agências internacionais de fomento à pesquisa têm passado a requerer, como critério para a concessão de recursos, que os pesquisadores se comprometam a seguir um método formal de gerenciamento de dados de pesquisa, de modo a garantir a preservação em longo prazo e maior facilidade de compartilhamento.

Iniciativas que visam prover infraestrutura de armazenamento de dados de pesquisa foram estabelecidas em vários países no mundo. Na Holanda tem-se o 3TU.Datacentrum, com foco em ciência e tecnologia; na Inglaterra tem-se o Data Archive que armazena dados sobre estudos de ciências sociais e humanidades; o americano BioLINCC (*Biological Specimen and Data Repository Information Coordinating Center*) possui um extenso acervo de dados biológicos; na área médica, o Dryad, dos Estados Unidos, é referência em curadoria de dados; para dados climáticos, a iniciativa internacional *WorldClim* merece destaque; na área de biodiversidade, o dinamarquês GBIF (*Global Biodiversity Information Facility*) é o provedor de dados mais proeminente. No Brasil, alguns repositórios de dados também foram implementados: da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, surgiu o Banco de Dados de Exploração e Produção (BDEP); o Centro de Estudos Integrados da Biodiversidade Amazônica implementou o Repositório de Dados de Levantamentos Biológicos; por fim, o PortalBio, que disponibiliza bases de dados de biodiversidade recebidas pelo Instituto Chico Mendes de Conservação da Biodiversidade.

De acordo com (BARONE; WILLIAMS; MICKLOS, 2017; COX et al., 2017; WILMS et al., 2016), pesquisadores frequentemente reconhecem que lhes faltam as habilidades e a experiência necessárias para gerenciar e compartilhar com eficácia os

dados de pesquisa gerados. O fluxo do gerenciamento de dados de pesquisa tende a ser contínuo, interativo, e vai sendo incorporado ao longo do projeto de pesquisa, conforme demonstrado por (COX; AM; TAM, 2018; WISSIK; DURCO, 2016). O planejamento adequado do gerenciamento dos dados torna o processo de pesquisa mais eficiente, facilita a colaboração e ajuda a evitar a perda de dados.

O gerenciamento efetivo de dados requer uma reflexão cuidadosa sobre cada estágio do processo de uso dos dados, incluindo descrição e documentação do processo, conteúdo e tipo dos dados; persistência e armazenamento dos dados em um local a partir do qual possa ser acessado, ou compartilhado, e preservados utilizando um formato adequado para reutilização a longo prazo (GOBEN; RASZEWSKI, 2015). Além disso, métodos de validação e garantia de qualidade devem ser estabelecidos, e a definição de diretrizes deve orientar a implementação de políticas para a disponibilização de dados e resultados reprodutíveis.

Idealmente, a qualidade dos dados de pesquisa deveria ser tratada durante todo o ciclo de vida dos dados, pois em cada estágio há implicações sobre como a qualidade dos dados impacta na utilidade e valor gerados pelos dados disponibilizados pelos pesquisadores. No contexto de gerenciamento de dados científicos, a QD é importante por dois aspectos principais. Primeiramente, os dados disponibilizados permitem que outros pesquisadores validem os métodos e resultados experimentais. Em segundo lugar, permite a reutilização dos dados em outros contextos, incluindo outros campos de estudo, com diferentes objetivos de pesquisa.

Assim, os dados devem estar corretos e completos principalmente para casos em que não se tem acesso aos detalhes do processo de obtenção e manipulação dos dados, de forma que os dados possam ser reutilizados. O acesso a dados de qualidade permite que análises secundárias possam impulsionar o progresso científico.

2.2 Qualidade de Dados

Segundo (CAI; ZHU, 2015), questões relacionadas a qualidade começaram a ser estudadas na década de 50. No início o foco dos pesquisadores estava voltado

para a questão de qualidade de produtos de manufatura. Nesse período, a noção de qualidade estava relacionada ao grau de conformidade de um conjunto de características com os requisitos do produto (CROSBY, 1979). Nas décadas seguintes, com o desenvolvimento da tecnologia da informação, a pesquisa voltou-se para o estudo da QD.

Os estatísticos foram os primeiros a investigar alguns dos problemas relacionados à qualidade dos dados, ao propor uma teoria matemática para considerar em conjuntos de dados estatísticos duplicados, no final dos anos 60 (FELLEGI; SUNTER, 1969). No início dos anos 80, de acordo com os trabalhos de (GARVIN, 1988; GERWIN, 1981; SON; PARK, 1987), foi a área de gestão que se preocupou em detectar problemas QD para o controle de sistemas de manufatura. Somente no início dos anos 90 que a comunidade científica passou a tratar esse tema com mais seriedade.

Diversas definições sobre QD foram apresentadas (REDMAN, 2001; REDMAN; BY-GODFREY; BLANTON, 1996; WAND; WANG, 1996). Uma das definições mais bem aceitas foi apresentada por (WAND; WANG, 1996), na qual os autores estabelecem que a qualidade dos dados pode ser intuitivamente caracterizada como adequação ao uso. Nos últimos 20 anos, o campo de estudos de QD evoluiu. Pesquisadores acadêmicos como Richard Wang, Carlo Batini e Thomas Redman lançaram luz sobre correlação entre informação, processos organizacionais e qualidade dos dados. O grupo de pesquisa do MIT, liderado pelo Professor Richard Y. Wang, elaborou a metodologia *Total Data Quality Management* (TDQM) (WANG, 1998) e realizou diversas pesquisas na área de QD. Esse grupo foi o responsável por apresentar a definição formal para QD, assim como a categorização das dimensões de qualidade (CAI; ZHU, 2015). No campo de gestão, iniciativas como o *Data Management Body of Knowledge* (DMBOK) (DAMA INTERNATIONAL, 2017) e *Data Management Maturity Model* (DMMM) (CMMI INSTITUTE, 2014) estabeleceram diretrizes formais para a implementação do gerenciamento de ativos de dados. Em ambas as referências a qualidade dos dados assume um papel crucial que precisa ser implementada para a melhoria efetiva dos ativos de dados de uma organização. QD também é definida como “conformidade com a realidade” e “satisfaz os requisitos do

usuário”. Nesta dissertação segue-se a definição de que um dado tem qualidade se ele está adequado ao uso (“*fitness for use*”).

2.3 Dimensões de Qualidade de Dados

Para (FIRMANI et al., 2016), qualidade de dados é um conceito multifacetado composto por diferentes dimensões. A literatura fornece uma classificação abrangente das dimensões da qualidade dos dados (DQD). No entanto, não há um consenso quanto ao conjunto nem quanto ao significado da maioria das dimensões devido à natureza contextual da qualidade (REDMAN; BY-GODFREY; BLANTON, 1996; WAND; WANG, 1996). Como o objetivo deste trabalho não é explorar todas as definições, serão apresentadas as definições que melhor se aplicam ao contexto de dados ambientais e de biodiversidade. (CAI; ZHU, 2015) realizaram um levantamento extensivo das classificações de dimensões de qualidade. Ao analisar essas classificações, os autores definiram um conjunto básico de categorias de DQD e que são apresentadas na Tabela 1. As categorias mostradas são consideradas características inerentes e indispensáveis da qualidade dos dados (SCANNAPIECO; CATARCI, 2002) e são definidas como:

- Disponibilidade é o grau de conveniência para os usuários obterem dados e informações. Essa categoria engloba duas dimensões: acessibilidade e temporalidade.
- Usabilidade constitui a noção de insuspeição dos dados, de forma que eles atendam às necessidades dos usuários, apresentando descrição, documentação e metadados. A dimensão dessa categoria é a credibilidade.
- Confiabilidade está relacionada à certeza e a segurança de que se pode confiar nos dados, isso consiste em acurácia, consistência, completude e integridade.

Tabela 1: Definição das dimensões de qualidade.

CATEGORIA	DIMENSÃO	DEFINIÇÃO
Disponibilidade	**Acessibilidade**	Diz respeito a até que ponto as informações estão disponíveis ou podem ser recuperadas com facilidade e rapidez (WANG; STRONG, 1996).
	Temporalidade	Uma medida do grau em que os dados são atuais e estão disponíveis para uso e no período de tempo em que são esperados (MCGILVRAY, 2008).
Usabilidade	**Credibilidade**	Relacionado a até que ponto as informações são consideradas verdadeiras e credíveis (WANG; STRONG, 1996).
Confiabilidade	**Acurácia**	O grau no qual os dados corretamente descrevem o objeto ou evento do mundo real (CUPOLI; EARLEY; HENDERSON, 2014).
	Completude	O grau em que os dados contêm os atributos requeridos e um número suficiente de registros e o grau em que os atributos são preenchidos de acordo com as expectativas do consumidor dos dados (SEBASTIAN-COLEMAN, 2013).
	Consistência	Refere-se a se ao relacionamento lógico entre os dados correlacionados (CAI; ZHU, 2015).
	Integridade	O grau em que os dados estão em conformidade com as regras de relacionamento de dados (conforme definido no modelo de dados) (SEBASTIAN-COLEMAN, 2013).

Fonte**:** próprio autor (2023).

2.4 Modelos de Distribuição de Espécies

(GHINS, 2011) definem um modelo como sendo uma representação que reflete parcialmente as propriedades da realidade. Os modelos são, portanto, simplificações, devido tanto à necessidade de reduzir a complexidade do objeto real quanto ao próprio desconhecimento do pesquisador quanto a muitas de suas propriedades. Assim, os Modelos de Distribuição de Espécies (MDEs) são representações da adequação de um espaço geográfico para a presença ou ausência de espécies com base nas variáveis ambientais utilizadas para gerar essas representações. Para (MATEO; FELICÍSIMO; MUÑOZ, 2011), essa adequação é a relação matemática ou estatística entre a distribuição real conhecida da espécie e um conjunto de variáveis independentes que são usadas como preditoras. Essas variáveis podem consistir em informações climáticas da região, ou ainda, em variáveis abióticas do meio como topografia, composição do solo e cobertura vegetal. De acordo com (GUISAN;

ZIMMERMANN, 2000), é a combinação dessas variáveis que define as condições favoráveis para a presença da espécie.

A construção de MDEs é essencialmente um processo de aprendizado supervisionado de classificação, no qual a variável dependente é, em geral, dicotômica, ou seja, só aceita valores de presença ou ausência, e as variáveis independentes que podem ser tanto quantitativas (e.g. precipitação, temperatura) quanto nominais (e.g. litologia, vegetação). Os modelos de classificação estabelecem um valor numérico de adequação da presença das espécies de acordo com os valores das variáveis ambientais. Esse valor pode ser discreto, de presença ou não presença, ou também um valor probabilístico de possibilidade de ocorrência.

As primeiras tentativas de quantificar a relação da distribuição das espécies em função das informações ambientais foram feitas há mais de 60 anos, com a formalização do conceito de nicho ecológico por (HUTCHINSON, 1957), como uma série de variáveis ambientais independentes que definem um "hiperespaço dimensional", no qual a espécie pode sobreviver e se reproduzir. No entanto, somente em meados da década de 1980 é que houve o desenvolvimento de métodos formais mais confiáveis. O primeiro modelo moderno de MDE foi desenvolvido em 1986, na Austrália, sob a liderança de Henry Nix. Esse modelo, gerado pelo algoritmo chamado de Bioclim, foi o primeiro pacote de modelagem de distribuição de espécies que relacionou dados de ocorrência de espécies espacialmente explícitas com mapas de variáveis ambientais (BOOTH et al., 2014b). De acordo com (BOOTH et al., 2014a), o Bioclim é um pacote cujo diferencial, na época, era a capacidade de gerar mapas com estimativas de distribuições de espécies e a incorporação de ferramentas exploratórias de análise de dados, permitindo que os usuários identificassem variáveis importantes e valores extremos. Bioclim utiliza apenas dados de ocorrência para definir um espaço ambiental multidimensional, no qual uma espécie pode ocorrer. Segundo (BCCVL, 2016), esse espaço ambiental é construído como uma caixa delimitadora em torno dos valores mínimo e máximo das variáveis ambientais para todas as ocorrências, resultando em um envelope retilíneo multidimensional. Esse algoritmo ainda é amplamente usado porque é fácil de entender, e é bastante utilizado, principalmente para fins didáticos e comparativos (DUAN et al., 2014).

Com a facilitação do acesso a bases climáticas e de ocorrência de espécies, foram desenvolvidas técnicas de modelagem mais complexas e flexíveis que buscam superar limitações que os métodos antigos apresentam como: colinearidade entre variáveis independentes, vieses de amostragem e inclusão de variáveis nominais (FELICÍSIMO; GÓMEZ-MUÑOZ, 2004; PHILLIPS et al., 2009b).

2.5 Tipos de Dados

Os MDEs, amplamente utilizados em biologia e ecologia de conservação, quantificam a relação entre os dados coletados sobre as espécies e as características ambientais e espaciais dos locais (CRUZ-CÁRDENAS et al., 2014; ELITH; LEATHWICK, 2009). Assim, os dados utilizados consistem em ocorrências de espécies e variáveis ambientais que, na maioria dos casos, são representadas por dados climáticos (como temperatura e precipitação), mas outras variáveis como tipo de solo ou cobertura do solo também podem ser empregadas (BOTELLA et al., 2018). Neste trabalho, a referência ao termo dados ou variáveis climáticas equivale-se a dados meteorológicos por ser o termo empregado pelas bases de dados utilizadas na realização dos experimentos (FICK; HIJMANS, 2017; NOCE; CAPORASO; SANTINI, 2019).

A implementação dos MDEs requer a coleta de dados que constituem a base experimental a partir do qual os modelos são construídos. Os dados utilizados em MDE são de dois tipos:

- Dados relativos à ocorrência das espécies em uma determinada região;
- Dados sobre a região que apresentam as variáveis ambientais.

2.5.1 Dados de Ocorrência de Espécies

Os dados de ocorrência podem ser relativos à presença, abundância (como a densidade de animais ou o número total de indivíduos), ou ausência das espécies. Esse tipo de dado pode ser obtido por meio de observações diretas dos animais, ou através de índices indiretos de presença das espécies estimados pela observação de vestígios, excrementos, ninhos, etc. (KALAN et al., 2016). Estas observações podem

ser obtidas em iniciativas desenvolvidas e realizadas diretamente por pesquisadores em pesquisas de campo, ou ainda por meio de programas de CS. No contexto de MDE, é essencial que esses dados sejam espacialmente explícitos, ou seja, eles precisam se referir a entidades geográficas georreferenciadas e bem definidas, normalmente áreas ou pontos da região de interesse.

Registros de abundância de espécies são difíceis de serem obtidos, pois requerem esforços que demandam intensa exploração do espaço geográfico para mapear o número máximo de indivíduos únicos de uma região, ou ainda, a implementação de técnicas para aquisição de dados que são caras e não acessíveis para a maioria dos projetos. Os registros de ausência também são difíceis de obter, pois “a ausência de presença não é igual à presença de ausência” (BARBET-MASSIN et al., 2012), ou seja, a não identificação de uma espécie em uma região não indica necessariamente que ela não possa ocorrer nessa região.

Os primeiros algoritmos para MDE, como o Bioclim, utilizavam somente dados de presença no processo de modelagem. No entanto, abordagens estatísticas clássicas como *Generalized Linear Model* (GLM) e *Generalized Additive Model* (GAM), necessitam de dados de presença e ausência. Para (HIJMANS; ELITH, 2017), quando o pesquisador tem acesso a dados de presença e ausência, ele deve empregar técnicas que tirem vantagem de toda essa informação e não optar por métodos de presença somente por comodidade na modelagem. Embora muitos algoritmos necessitem de dados de ausência para funcionar adequadamente, eles possuem uma flexibilidade quanto a forma como essa ausência é apresentada. Como na maioria dos casos os dados reais de ausência de espécies não estão disponíveis, é possível emulá-los simulando pseudo-ausências ou incluindo dados do background da região de interesse.

Dados do background estabelecem o domínio ambiental do estudo, enquanto os dados de presença devem estabelecer em que condições uma espécie tem maior probabilidade de estar presente do que em média (HIJMANS; ELITH, 2019; PHILLIPS et al., 2009a). No caso das pseudo-ausências, os pesquisadores estabelecem onde as ausências podem ocorrer, adotando duas possíveis estratégias: amostrando toda a região, exceto em locais com dados de presença, ou podem amostrar locais que não são adequados para a espécie.

A quantidade de dados de presença também pode impactar os modelos de distribuição de espécies. Em alguns casos, o número de registros disponíveis para modelar as distribuições de espécies pode ser limitado, como é o caso de espécies raras ou em risco de extinção. Na literatura, não existe unanimidade sobre a quantidade ideal mínima de registros espacialmente únicos.

Assim, conclui-se que os tipos dos dados de ocorrência disponíveis determinam os métodos estatísticos que podem ser usados para analisá-los.

Muitas bases de dados dedicadas passaram a disponibilizar dados de ocorrência de espécies nos últimos anos. O *Global Biodiversity Information Facility* (GBIF) é atualmente o maior portal com acervo de registros de coleta de espécimes de todos os continentes. Na América do Sul, o portal Link é uma base importante com mais de 4 milhões de registros georreferenciados. A digitalização dos acervos de herbários e museus de história natural também constitui uma importante fonte geradora de dados de ocorrência das espécies.

2.5.2 Dados Ambientais

Os dados ambientais podem ser medidos em campo simultaneamente com os de ocorrência das espécies. Por exemplo, a localização do animal é detectada e as características da área ao redor do local são descritas. Esse modo de proceder, embora custoso, permite detectar as variáveis ambientais com alta precisão, mas apenas em áreas de dimensões limitadas. Por isso, a abordagem mais comum utiliza dados relativos ao ambiente de áreas próximas ao local de ocorrência.

Para (GUISAN; ZIMMERMANN, 2000; VELEZ-LIENDO; STRUBBE; MATTHYSEN, 2013), as variáveis ambientais a serem consideradas em um MDE dependem fortemente das espécies em estudo porque fatores ambientais têm um efeito direto sobre a distribuição das espécies. Essas relações entre os organismos e o ambiente são uma das causas dos padrões de distribuição espacial. A Tabela 2 agrupa em 5 classes principais as variáveis usadas para modelos de distribuição de espécies terrestres.

Tabela 2: Variáveis ambientais que podem ser utilizadas para espécies terrestres.

CLASSE DE VARIÁVEIS	VARIÁVEIS
Morfológicas	• Altitude • Inclinação • Litologia • Geomorfologia • Conformação rochosa • Rede hidrográfica • Corpos de água
Vegetação	• Cobertura vegetal • Tipo de vegetação • Fitossociologia • Estrutura Florestal • Idade da vegetação • Altura das plantas • Densidade de plantas
Tróficas	• Distribuição ou densidade de presas, predadores e competidores • Locais de nidificação, abrigo, e reprodução
Climáticas	• Temperatura média, máxima e mínima • Precipitação sazonal • Velocidade e direção dos ventos
Antrópicas	• Uso da terra • Rede rodoviária • Ocupação humana • Urbanização, distribuição de aterros e fontes de poluição • Distribuição de atividades pecuárias • Distribuição de atividades agrícolas • Risco de incêndios

Fonte**:** próprio autor (2023).

Para desenvolver MDEs, é necessário ter matrizes de dados que associem cada observação da espécie a ser modelada com os valores das variáveis ambientais necessárias para descrever as condições do local de ocorrência. Por variável ambiental, entende-se qualquer fator abiótico ou biótico que caracteriza o habitat das espécies que estão sendo investigadas. (MATEO; FELICÍSIMO; MUÑOZ, 2011) estabelece que a inclusão das variáveis ambientais em MDE deve atender a três condições:

- **Elucidativo**: as variáveis escolhidas precisam ter uma relação explicativa com a distribuição da espécie objeto, seja como fator potencializador ou limitante.
- **Variabilidade significativa**: as variáveis precisam apresentar variação dentro da região de estudo para distintivamente caracterizar as células do espaço geográfico.
- **Não Correlação**: as variáveis escolhidas não podem estar excessivamente correlacionadas entre si. Para (DE MARCO; NÓBREGA, 2018; MUÑOZ; FELICÍSIMO, 2004), esse é um aspecto importante especialmente se uma interpretação biológica dos resultados é buscada.

Apesar do grande número de variáveis que podem ser utilizadas em um MDE, como apresentado na Tabela 2, em geral os modelos utilizam uma quantidade reduzida de variáveis, devido principalmente a dificuldade de acesso à maioria delas. As variáveis ambientais comumente utilizadas são as variáveis climáticas, normalmente geradas a partir da interpolação de dados de estações meteorológicas, que são mais facilmente obtidas em repositórios online (SUGGITT et al., 2017).

2.6 Problemas de Qualidade em Dados de Ocorrência de Espécies

O comprometimento da qualidade pode ocorrer em diversas etapas do ciclo de vida dos dados de ocorrência de espécies, desde o momento da coleta até a aquisição dos dados para amostragem. Operadores humanos podem inserir erros e vieses nos dados coletados, que podem propagar pelo processo de análise resultando em modelos imprecisos e que não refletem o real padrão de distribuição das espécies.

No contexto de dados de biodiversidade, dados já coletados com problemas são difíceis de serem tratados porque tendem a ser fortemente dependentes do contexto, situação na qual técnicas simples de limpeza não são capazes de melhorar a qualidade dos dados. Dados de baixa qualidade limitam o modo como podem ser analisados e reutilizados, reduzindo a utilidade dos mesmos. Por isso, é necessário entender como a ocorrência desses erros impactam o processo de criação de MDEs.

Neste trabalho, três tipos de dados, cujo comprometimento pode impactar os resultados dos MDEs, são explorados:

- **Localização**: dados relacionados à localização geográfica do ponto de ocorrência de uma espécie.
- **Taxonômico**: nomenclatura e hierarquia taxonômica do organismo observado.
- **Geoespacial**: pontos georreferenciados dos locais de ocorrência da espécie.

Esses tipos de dados foram selecionados porque identificou-se na literatura (capítulo 3) que em estudos de distribuição de espécies, erros nesses tipos de dados são comuns e tem potencial de comprometer a utilidade dos modelos criados.

2.6.1 Erros de Localização

O erro de localização ocorre quando o registro do local de avistamento da espécie é diferente do local de ocorrência real. Na literatura, os autores descrevem a ocorrência desse tipo de erro em diferentes cenários. (CUGLER et al., 2013) descreve esse erro em metadados de observação de espécies em que os campos de geolocalização não são preenchidos ou então são preenchidos com dados muito vagos como o município ou estado de ocorrência. Esse é um problema predominante em dados históricos e pode acarretar erros que variam de algumas dezenas de metros a centenas de quilômetros. De acordo com (GÁBOR et al., 2022), erros de localização não são exclusivos de bases que consolidam dados antigos, esse tipo de erro é inerente a dados georreferenciados que utilizam sistemas globais de navegação por satélite. Para estudos de distribuição de espécies, é importante que os pontos de ocorrência possuam alta resolução geográfica, sendo georreferenciados por um sistema de coordenadas preciso. Descrições generalistas sobre o local de coleta da informação podem inviabilizar a utilização desses dados para MDEs que requerem fina resolução espacial. Outro cenário em que a ocorrência desse tipo de erro é comum é em dados coletados por CS, em que voluntários frequentemente registram o local de ocorrência como a estrada mais próxima, cidade ou algum outro ponto de referência (HEFLEY; BROST; HOOTEN, 2017). De acordo com (MALDONADO et al., 2015), dados gerados pela digitalização de acervos de herbários e museus também pode apresentar registros de ocorrências com localizações incertas porque muitas vezes os dados geográficos exatos não estão disponíveis. Em casos em que operadores humanos registram dados no sistema de coleta, (SIMÕES; PETERSON,

2018) afirmam que os erros podem ocorrer pela comutação entre os dados de longitude e latitude, ou confusão dos algarismos numéricos.

2.6.2 Erros de Identificação

O segundo erro mais comum identificado na literatura são os erros de identificação de espécies. Esse tipo de erro está relacionado à classificação taxonômica da espécie e ocorre quando o observador registra outra espécie no lugar da espécie real observada. De acordo com (CLARE et al., 2019; KOSMALA et al., 2016; RATNIEKS et al., 2016), esse erro é muito comum em dados gerados por projetos de CS. (KOSMALA et al., 2016; RATNIEKS et al., 2016; YU; WONG; KELLING, 2014), indicam que erros na identificação da espécie normalmente são causados pela falta de experiência dos observadores, que podem ter dificuldades para identificar espécies de animais a distância, e de diferenciar espécimes semelhantes (BIRD et al., 2014).

Nos últimos anos, armadilhas fotográficas têm sido utilizadas para estudar uma ampla gama de espécies permitindo a exploração de diferentes questões de pesquisa a partir dos dados capturados. A identificação de espécies utilizando armadilhas fotográficas pode ser feita por um observador remoto (CHOO et al., 2020; MCKIBBEN; FREY, 2021)ou de forma automática por meio de software (SCHNEIDER et al., 2020). O aperfeiçoamento de técnicas de visão computacional para processamento e classificação de imagens como *deep learning,* permitiu adotar estratégias de obtenção de dados de ocorrência e abundância que reduzam os custos de aquisição (SCHNEIDER et al., 2019). No entanto, seja por observador humano ou por software, problemas como o erro na identificação de espécies ainda persistem. (PERES et al., 2021)analisaram métodos de identificação de espécies de 118 planos de gestão de biodiversidade e constataram que 60% deles empregaram métodos inadequados para identificação de espécies de cervídeos. (ZETT; STRATFORD; WEISE, 2022) verificaram o impacto da experiência do observador e a variação fenotípica da espécie na taxa de erros de identificação. (MIAO et al., 2021)destacaram os desafios da identificação automática das espécies utilizando algoritmos de classificação de imagens em virtude da limitação de volume de dados para

treinamento, principalmente relacionados a espécies raras ou com risco de extinção, as mais susceptíveis a decisões de conservação enviesadas.

2.6.3 Viés Geográfico

No contexto deste trabalho, os erros que afetam os tipos de dados geoespaciais são aqueles relacionados aos vieses de coleta. Segundo (SMITH; NOBLE, 2014), os vieses impactam a validade e confiabilidade dos resultados dos estudos. O viés geográfico no processo de coleta é recorrente em dados coletados em projetos de CS (BIRD et al., 2014; KOSMALA et al., 2016; LEWANDOWSKI; SPECHT, 2015a). Para (DORAZIO, 2014), os vieses geográficos na coleta dos dados ocorrem porque a facilidade de acesso a algumas áreas faz com que o registro de ocorrências seja muito maior do que em áreas mais remotas. Assim, registros de ocorrência tendem a ocorrer em localidades próximas a estradas, rodovias e trilhas.

O viés geográfico ocorre quando a intensidade de amostragem é inconsistente no espaço, tornando problemático o cálculo de estimativas precisas para a distribuição e densidade das espécies. De acordo com (FEI; YU, 2016a), a amostragem incompleta e o viés do espaço geográfico podem ter um efeito importante nas estratégias de conservação, pois a escolha das áreas de conservação prioritárias pode ser afetada.

2.7 Métricas de Validação

A validação é uma etapa fundamental de todo o processo de construção de modelos computacionais, e é necessária para quantificar e garantir a credibilidade deles (B.H.THACKER et al., 2004; ROBERTS et al., 2017). Se o modelo não for considerado adequado para representar a distribuição da espécie em questão, pode ser necessário refazer as etapas anteriores do processo de modelagem para tentar entender se as premissas sobre o uso do habitat devem ser ajustadas, se os dados utilizados devem ser reamostrados, ou se o tipo de modelo empregado precisa ser mudado.

Para a avaliação das previsões no caso de presença ou ausência, diversas estatísticas são utilizadas.

- **Acurácia** – fração de predições corretas totais indicadas pelo modelo.
- **Sensibilidade** (*recall*) – probabilidade condicional de classificar corretamente uma presença.
- **Especificidade** – probabilidade condicional de classificar corretamente uma ausência.
- **Precisão** – indica o percentual de acertos de presenças em relação a todas as presenças indicadas pelo modelo.
- *Receiver Operating Characteristic* **(ROC)** – reflete o quanto o modelo é capaz de distinguir entre as classes.
- *Area Under the Curve* **(AUC)** – é o valor da área sob a curva do gráfico ROC e indica a capacidade discriminativa dos modelos.
- *True Statistics Skill* **(TSS)** – é a média da taxa de sucesso da previsão para pontos de ocorrência e ausência.
- **Kappa de Cohen** – mede até que ponto a concordância entre o observado e o previsto é superior ao esperado pelo acaso.
- **F1 Score** – média harmônica que combina as medidas de precisão e sensitividade em uma única métrica.

A maioria das métricas de validação dos MDEs dependem de uma matriz de confusão binária. Uma matriz de confusão é uma tabela de contingência 2 × 2 que captura as seguintes informações:

I. O número de presenças corretamente previstas (“sensibilidade”)

II. O número de ausências falsamente previstas como presentes (“falsos positivos”)

III. O número de presenças falsamente previstas como ausentes (“falsos negativos”)

IV. O número de ausências previstas corretamente como ausentes (“especificidade”)

Para a utilização da matriz confusão, as saídas dos MDEs precisam ser convertidas para valores discretos definindo a presença ou a ausência das espécies. A sensibilidade e especificidade da matriz confusão são utilizadas para a plotagem da curva ROC, no qual a ordenada corresponde à sensibilidade e a abscissa representa a especificidade. A estatística derivada é a AUC, cujo valor varia entre 0 e 1. O valor 1 indica que todos os casos foram classificados corretamente e o valor 0,5 indica que o modelo não é diferente da classificação aleatória; valores abaixo de 0,5 indicam que o modelo é ruim, pois classifica erroneamente mais casos do que o acaso.

Os dados de validação devem ser diferentes daqueles usados para o desenvolvimento do modelo. No caso de um único conjunto de dados de presença, é apropriado dividi-lo em duas partes e usar parte para a criação do modelo e parte para a validação. A divisão da amostra pode ser aleatória ou seguir um critério de escolha (por exemplo, no caso de dados obtidos ao longo de vários anos, os dados dos primeiros anos podem ser usados para desenvolver o modelo e os dos anos seguintes para testá-lo). Segundo (MATEO; FELICÍSIMO; MUÑOZ, 2011), os métodos de modelagem geralmente oferecem resultados muito diferentes, então a seleção do método que será usado na interpretação é muito importante. Na literatura, grande parte dos estudos implementam mais de uma métrica de validação para um mesmo modelo. A validação dos resultados é importante para avaliar o quão generalizável é o modelo implementado.

2.8 Técnicas de Modelagem

MDEs podem ser gerados, em princípio, com qualquer classificador estatístico apropriado para o tipo de variável modelada (presença/ausência ou densidade). (ELITH et al., 2006) apresenta uma revisão completa dos principais algoritmos utilizados pela literatura. Alguns tradicionais são detalhados a seguir.

2.8.1 *Maximum Entropy Modeling* (Maxent)

De acordo com a literatura, Maxent consiste em uma das melhores abordagens para modelos de distribuição de espécies existentes (AGUIRRE-GUTIÉRREZ et al.,

2013; FOURCADE et al., 2014; GOMES et al., 2018a). Maxent emprega aprendizado automático para estimar a distribuição de uma espécie, ou habitat, com base na entropia máxima, sujeita a um conjunto de restrições baseadas no conhecimento das condições ambientais em locais de ocorrência conhecidos. Este método avalia a adequação de cada ponto de ocorrência de acordo com as variáveis ambientais disponíveis (MOUSAZADE et al., 2019; PHILLIPS; ANDERSON; SCHAPIRE, 2006). Através do uso de dados de presença de espécies e variáveis ambientais (contínuas ou categóricas) para a área de estudo, é gerada uma estimativa da probabilidade de presença de uma espécie que varia entre 0 e 1.

Entropia é um conceito fundamental na teoria da informação e é definida por (JUKNA, 2011) como a medida da quantidade de informação perdida quando o valor de uma variável aleatória não é conhecido. A entropia é geralmente considerada como a expressão do distúrbio de um sistema físico, ou como uma medida da falta de informação sobre as características de um sistema físico: quanto maior a informação, menor é a entropia (CHAN, 2015). Em Maxent, a entropia medida dentro de uma célula correspondente a um registro de presença da espécie investigada deve ser baixa, enquanto a entropia medida dentro de uma célula de presença incerta deve ser alta.

Uma característica importante da abordagem usada pelo Maxent é a sua eficácia mesmo com amostras pequenas, como aquelas geralmente disponíveis para estudos de espécies raras e arquivos históricos de faunas ou herbários. Além disso, com o Maxent é possível trabalhar somente com dados de presença de espécies. Assim, pode-se utilizar como fontes de dados as coleções de história natural, evitando os altos custos de amostragem das espécies ao longo de sua extensão de ocorrência (GOMES et al., 2018a).

2.8.2 *Random Forests*

Random Forests (RF) é um modelo de classificação, ou regressão, baseado em árvores (BREIMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2001). O uso de agregação via *bootstrap* seleciona diversas subamostras dos dados, e um algoritmo de ensacamento gera um grande número de árvores de regressão não correlacionadas.

No aprendizado de máquina, uma árvore de decisão (CART – *Classification And Regression Tree*) é um algoritmo que gera um modelo preditivo (classificação ou regressão), em que cada nó interno representa uma variável, conexões para um nó filho representam um valor possível para essa variável e uma folha apresenta o valor previsto para a variável de resposta.

Um algoritmo de RF consiste em muitas árvores de decisão, cuja resposta é dada pela combinação das previsões das árvores individuais (BREIMAN, 2001; LINDHOLM et al., 2019). As árvores são criadas selecionando-se um subconjunto de dados de calibração e, em seguida, combinadas seguindo um procedimento de ensacamento. Ao contrário das abordagens de regressão clássicas, em que a relação entre a variável de resposta e os preditores é inicialmente especificada (por exemplo, linear, quadrática, etc.), no caso de RF, não existe uma hipótese a priori sobre a forma de relação. As altas capacidades preditivas desta técnica, destacadas em alguns estudos comparativos (PRASAD; IVERSON; LIAW, 2006), bem como para outras técnicas similares derivadas da aprendizagem de máquina (ELITH; GRAHAM, 2009), estão contribuindo para uma rápida difusão de suas aplicações no campo ecológico (MI et al., 2017; VINCENZI et al., 2011).

2.8.3 Modelos Lineares Generalizados

O *Generalized Linear Model* (GLM) é uma extensão da abordagem clássica dos algoritmos lineares de regressão (MCCULLAGH; NELDER, 1989; NELDER; WEDDERBURN, 1972). Os modelos lineares generalizados foram os métodos de regressão mais comuns usados para prever a distribuição espacial das espécies (ELITH et al., 2006).

Enquanto a suposição do modelo linear é que a variável alvo é normalmente distribuída, no GLM a variável alvo pode ser uma distribuição da classe de famílias exponenciais (por exemplo, Normal, Binomial, Poisson, Gama, Gaussiana e Inversa) (CORDEIRO; DEMÉTRIO, 2008). No contexto de MDEs, a regressão logística é adequada para dados de ocorrência e ausência de espécies e a regressão de Poisson é adequada para dados de contagem de espécies. A função de link logit para a regressão logística e a função log para a regressão de Poisson permitem que a

variável dependente (indicando presença ou abundância de espécies) seja linearmente relacionada com as variáveis explicativas. As variáveis explicativas dos GLMs podem conter termos de interação e termos polinomiais, sendo preferíveis para relações não-lineares simples entre espécies e variáveis de ambiente.

As principais vantagens da adoção de modelos lineares generalizados em relação aos modelos lineares consistem:

- Na capacidade de lidar com uma ampla gama de distribuições;
- Na relação entre a variável resposta e preditores lineares, através da função de link, que, além de garantir a linearidade, representa um método eficiente de forçar predições dentro do intervalo de valores possíveis da variável resposta (por exemplo, entre 0 e 1 para regressões logísticas);
- Na capacidade de gerenciar grandes dispersões nas distribuições.

3 RESULTADOS DA REVISÃO DE ESCOPO DA LITERATURA

Para a realização deste trabalho, uma revisão de escopo da literatura (REL) foi realizada com o intuito de avaliar estudos primários da literatura e identificar os aspectos teóricos e metodológicos relacionados à QD em MDEs. O objetivo deste capítulo é apresentar resumidamente a abordagem metodológica adotada e os principais resultados obtidos na REL, principalmente em relação aos problemas de qualidade identificados e principais algoritmos utilizados.

3.1 Metodologia da Revisão de Escopo da Literatura

O protocolo de revisão adotado na revisão foi adaptado a partir das etapas estabelecidas por (KITCHENHAM, 2004; KITCHENHAM; CHARTERS, 2007). O processo é composto de três fases: planejamento, condução e relato da revisão. Uma breve explicação das subtarefas de cada uma das fases é apresentada a seguir.

3.1.1. Planejamento

A) Especificação das questões de pesquisa: define os tópicos direcionadores dos objetivos da revisão.
    a. RQ1: Quais são os principais problemas de qualidade identificados em dados utilizados na modelagem de distribuição de espécies?
    b. RQ2: Quais são os algoritmos mais utilizados na modelagem de distribuição de espécies?
    c. RQ3: Como os problemas de qualidade afetam a análise de dados climáticos e de biodiversidade?
B) Seleção das bases de busca: define os repositórios mais pertinentes à área de interesse da pesquisa.
    a. As Bases de dados empregadas para a seleção dos artigos foram: Web of Science (WoS), Scopus, IEEE e Science Direct. A escolha dessas bases foi feita baseada em critérios de qualidade (MARTÍN-MARTÍN et al., 2018; ŠUBELJ et al., 2015), pertinência ao campo de estudos de computação e completude dos metadados retornados por essas bases.

C) Definição dos critérios de inclusão e exclusão: são parâmetros dicotômicos para julgar um artigo.
    a. Artigos em inglês, publicados de janeiro de 2013 a abril de 2019, cujo texto completo esteja disponível em revistas ou anais de congressos, e que tenham metadados disponíveis em formato BibTeX.

3.1.2. Condução

A) Identificação dos estudos relevantes: são os resultados das buscas com os critérios de inclusão e exclusão aplicados.
    a. Ao todo foram retornados 1492 artigos únicos das quatro bases de dados utilizadas.
B) Avaliação de pertinência: amostragem de um subconjunto de artigos e classificação dessa amostra em 5 graus de pertinência.
    a. Dos 1492 artigos retornados, 100 deles foram sorteados para a criação de um arquivo de avaliação. O sorteio desses artigos foi feito a partir de uma amostragem aleatoriamente estratificada sem reposição. Essa estratégia foi adotada para garantir que cada questão de pesquisa tivesse um número representativo de artigos dentro da amostra.
C) Análise de similaridade: busca de artigos com padrão similar aos artigos bem classificados.
    a. Foi empregada a técnica de similaridade de cossenos. A similaridade do cosseno é uma métrica de mineração de textos empregada para mensurar o quanto informações entre dois vetores textuais estão correlacionadas (KITASUKA; ARITSUGI; RAHUTOMO, 2012; ZAHROTUN, 2016). A partir de uma matriz de similaridade é possível agrupar informações textuais com padrões semelhantes.
    b. Dos 1492 artigos obtidos inicialmente, foram selecionados 180 artigos com o padrão de interesse mais similar aos artigos com alto grau de pertinência na classificação da amostra de avaliação.
D) Refinamento da literatura: seleção manual dos artigos similares retornados pela etapa anterior.

a. Selecionou-se os trabalhos com maior potencial de responder as perguntas de pesquisa definidas. Ao final, 59 artigos foram selecionados para a análise completa.

E) Análise dos artigos selecionados: leitura completa da lista final de artigos.

3.1.3 Comunicação

A) Extração e síntese dos dados: identificação e destaque dos pontos de interesse do artigo.
B) Avaliação dos resultados: análise crítica dos resultados.
C) Sumarização e apresentação dos resultados: criação de gráficos, tabelas e mapas para representação visual dos resultados.

3.2 Resultados da Revisão

Um dos pontos mais importantes dessa revisão foi verificar quais os problemas de qualidade afetam dados climáticos ou de ocorrência de espécies. A Tabela 03 apresenta os erros identificados na literatura com a REL.

Tabela 3: Erros Identificados.

	PROBLEMA	DESCRIÇÃO
1	Localização imprecisa	Dados com registros de localização incorretos ou imprecisos.
2	Identificação Incorreta	Detecções incorretas de uma espécie que é confundida com outra similar.
3	Viés geográfico	Tendência de registro de observações de espécies de ocorrerem em regiões de mais fácil acesso.
4	Habilidade do observador	Habilidade ou experiência de um observador não profissional na detecção de ocorrência de espécies.
5	Incompleto ou faltante	Registro com atributos de dados ou metadados parcialmente preenchidos ou não preenchidos.
6	Viés de detecção	Tendência do observador de registrar determinadas espécies em detrimento de outras, causando super e subamostragem de ocorrências.
7	Viés temporal	Tendência de capturas sazonais de ocorrências de espécies.
8	Viés de amostragem	Ocorre quando os dados não estão aleatoriamente distribuídos por todo o nicho ecológico ou estão desbalanceados.
9	Resolução espacial	Registros de dados cuja amplitude e granularidade de ocorrência impedem análises de microrregiões.
10	Registros Duplicados	Múltiplo registros que representam uma mesma entidade do mundo real.
11	Temporalidade	Dados em intervalo de tempo restrito e insuficiente para a análise de mudança de distribuição a longo prazo.

12	Limites esperados excedidos	Valores de medição de informações climáticas inconsistentes com a realidade local.
13	Erro posicional	Registro de localização incorreto em função da posição do observador com relação ao espécime.
14	Incerteza de região	Conjuntos de dados climáticos espaciais estimados em que algumas regiões são consistentemente superestimadas enquanto outras são subestimadas.
15	Região de ocorrência	Identificação de espécies fora do hábitat esperado.

Fonte**:** próprio autor (2023).

A Tabela 04 consolida os resultados dos problemas de qualidade com os tipos de dados que podem ser afetados e a quantidade de artigos que mencionaram o respectivo tipo de problema.

Tabela 4: Consolidado da extração dos dados dos artigos

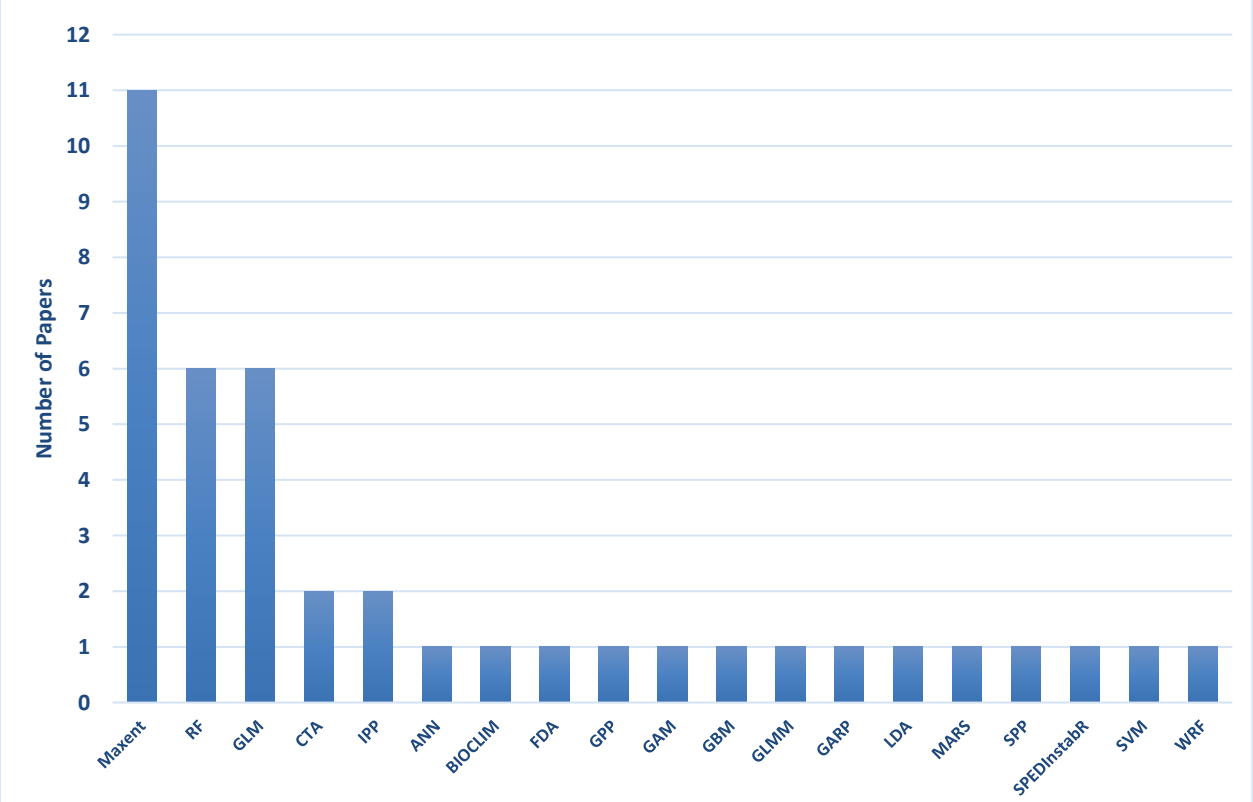
	PROBLEMA	**TIPO DE DADOS AFETADO**	**QTD DE ARTIGOS**	**ARTIGOS**
1	Localização imprecisa	Biodiversidade CS Climáticos	9	(AUBRY; RALEY; MCKELVEY, 2017b; CUGLER et al., 2013; GOMES et al., 2018c; HEFLEY et al., 2014; HEFLEY; BROST; HOOTEN, 2017; MALDONADO et al., 2015; RADOVIĆ et al., 2018; SIMÕES; PETERSON, 2018; VELÁSQUEZ-TIBATÁ; GRAHAM; MUNCH, 2016)
2	Identificação Incorreta	Biodiversidade CS	8	(AUBRY; RALEY; MCKELVEY, 2017b; BIRD et al., 2014; CLARE et al., 2019; CUGLER et al., 2013; DORAZIO, 2014; GOMES et al., 2018c; KOSMALA et al., 2016; MALDONADO et al., 2015)
3	Viés geográfico	Biodiversidade CS Climáticos	7	(BIRD et al., 2014; DORAZIO, 2014; FEI; YU, 2016b; HEFLEY; BROST; HOOTEN, 2017; KELLING et al., 2015a; PARK; DAVIS, 2017; ROBINSON; RUIZ-GUTIERREZ; FINK, 2018)
4	Habilidade do observador	CS	6	(KELLING et al., 2015a; KOSMALA et al., 2016; LEWANDOWSKI; SPECHT, 2015b; LIN et al., 2015; RATNIEKS et al., 2016; YU; WONG; KELLING, 2014)
5	Incompleto ou faltante	Biodiversidade CS Climáticos	4	(CUGLER et al., 2013; FEI; YU, 2016b; GOMES et al., 2018c; SERRA-DIAZ et al., 2017)
6	Viés de detecção	CS	3	(LEWANDOWSKI; SPECHT, 2015b; PARK; DAVIS, 2017; ROBINSON; RUIZ-GUTIERREZ; FINK, 2018)
7	Viés temporal	CS	3	(KELLING et al., 2015a; KOSMALA et al., 2016; PARK; DAVIS, 2017)
8	Viés de amostragem	Biodiversidade	3	(AUBRY; RALEY; MCKELVEY, 2017b; RADOVIĆ et al., 2018; ROBINSON; RUIZ-GUTIERREZ; FINK, 2018)
9	Resolução espacial	Biodiversidade Climáticos	3	(BEDIA; HERRERA; GUTIÉRREZ, 2013; GRAHAM; HAINES-YOUNG; FIELD, 2015; PARK; DAVIS, 2017)

10	Registros Duplicados	Biodiversidade CS Climáticos	2	(CUGLER et al., 2013; SERRA-DIAZ et al., 2017)
11	Temporalidade	Biodiversidade CS Climáticos	2	(BIRD et al., 2014; LEWANDOWSKI; SPECHT, 2015b)
12	Limites esperados excedidos	Climáticos	1	(CUGLER et al., 2013)
13	Erro posicional	Biodiversidade	1	(ZHANG et al., 2018)
14	Incerteza de região	Climáticos	1	(STOKLOSA et al., 2015)
15	Região de ocorrência	Biodiversidade	1	(SERRA-DIAZ et al., 2017)

Fonte**:** próprio autor (2023).

O processo modelagem de distribuição de espécies é apoiado por algoritmos que analisam os padrões de ocorrência de espécies em função das condições ambientais da região de interesse. Para o levantamento das principais técnicas utilizadas para modelagem de distribuição de espécies, 21 artigos foram analisados. A Figura 01 apresenta os resultados consolidados.

Figura 1: Algoritmos utilizados em MDEs



Ao longo do tempo, novos algoritmos foram sendo incorporados pela área de ecologia para trabalhar com dados de diferentes tipos e características para descrever

padrões de distribuição de espécies (BOOTH et al., 2014b; BUSBY, 1991; GUISAN; EDWARDS; HASTIE, [s.d.]; PHILLIPS; ANDERSON; SCHAPIRE, 2006). Em MDEs os algoritmos utilizados são, em geral, extensões de técnicas estatísticas mais tradicionais, como métodos lineares, e algoritmos de *machine learning* supervisionados (MLS).

O Maxent foi o algoritmo mais utilizado pelos trabalhos analisados, ocorrendo em 11 (52.38%) artigos, seguidos por RF e GLM, empregados em 6 (28.57%) trabalhos. Maxent é o algoritmo mais popular em MDEs, o que pode ser explicado pelos seguintes motivos:

- possibilidade de trabalhar com dados somente de presenças;
- em geral apresenta melhor acurácia preditiva do que outros modelos, sendo utilizado como parâmetro de comparação para outros algoritmos implementados (GUISANDE et al., 2017)
- robustez ao problema de qualidade de geolocalização imprecisa que pode afetar dados de ocorrência de espécies (GRAHAM et al., 2008).

RF é um algoritmo de MLS conhecido por ter um excelente desempenho em previsões ecológicas (MI et al., 2017). E o GLM, um modelo linear, é usado para estimar uma resposta ecológica como uma combinação linear de variáveis preditoras independentes.

3.3 Conclusão da REL

A REL permitiu realizar o levantamento dos problemas de qualidade que são preocupação da comunidade científica. Assim, a localização imprecisa, a identificação incorreta e o viés geográfico são os principais erros presentes em dados de ocorrência de espécies e que tem potencial de impactar negativamente os resultados dos MDEs. Do ponto de vista de criação dos modelos, diferentes algoritmos são empregados. Algoritmos de aprendizado de máquina como o Maxent e RF estão entre as principais técnicas utilizadas. No entanto, métodos lineares tradicionais, como o GLM, ainda encontram muitas possibilidades de aplicação. Outros algoritmos como redes neurais e máquinas de vetores de suporte, que são empregados em outras áreas de aplicação, em MDEs acabam não sendo amplamente utilizados, em virtude da

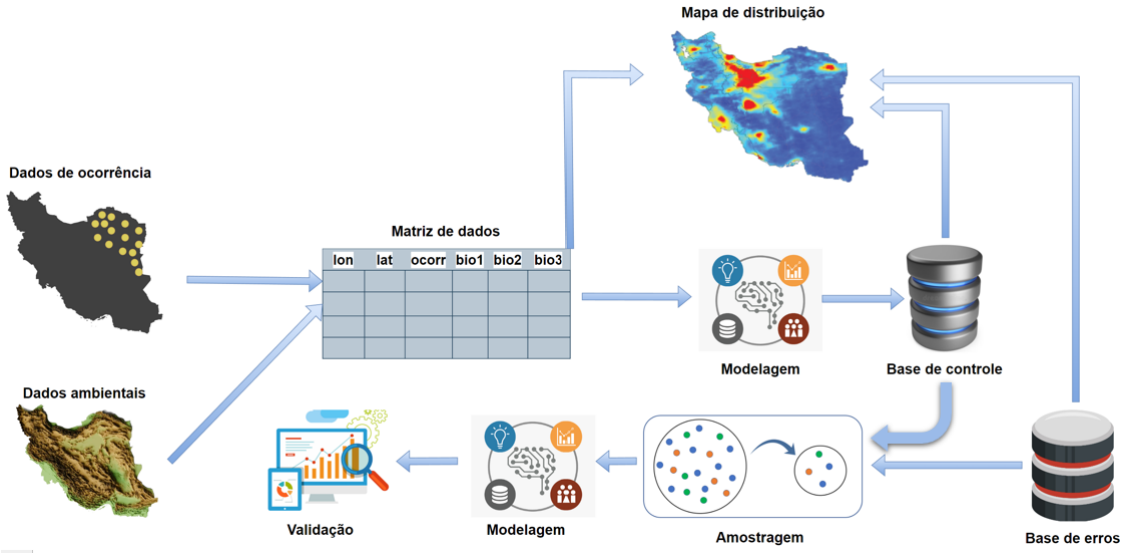
complexidade dos modelos e por ainda não superarem a capacidade preditiva dos métodos tradicionais (CARVALHO et al., 2017; REISS et al., 2011).

4 PROCESSO DE CRIAÇÃO DE MDEs PARA EXPERIMENTO

Neste capítulo, descreve-se a abordagem experimental que foi empregada para verificar como diferentes níveis de problemas de qualidade dos dados afetam os MDEs.

No processo padrão de criação dos MDEs os dados de ocorrência de espécies são geograficamente vinculados a informações ambientais. Essa vinculação alimenta algoritmos de modelagem e as regras e os coeficientes computados são utilizados para gerar mapas de distribuição contínua que podem mostrar, por exemplo, a probabilidade relativa de ocorrência de uma determinada espécie na região de estudo. O fluxo apresentado na Figura 4 apresenta uma adaptação do processo padrão de criação de MDEs para o contexto deste trabalho. Neste fluxo, adicionou-se as etapas de geração das bases de controle e bases de erros que são amostradas para a obtenção dos dados de entrada dos modelos para análise dos impactos da qualidade. Os detalhes das etapas do fluxo apresentado são descritos nos tópicos a seguir.

Figura 2: Fluxo do processo de criação de MDEs



Fonte: adaptado de (FRANKLIN, 2009)

4.1 Dados de Ocorrência

Para o experimento, dados de ocorrência de espécies foram criados baseados na geração de uma espécie virtual. De acordo com (GRIMMETT; WHITSED; HORTA, 2021), essa estratégia tem sido utilizada em ecologia para testar vários aspectos do processo de modelagem de distribuição de espécies, como testagem de novos métodos de modelagem (QIAO; SOBERÓN; PETERSON, 2015)e avaliação dos efeitos das características dos dados (LIU; NEWELL; WHITE, 2019). Normalmente, as distribuições de espécies virtuais são simuladas a partir de variáveis ambientais que formam funções de resposta cuja combinação definem o nicho ecológico da espécie simulada. Mas ainda há algumas lacunas nessa estratégia. (GRIMMETT; WHITSED; HORTA, 2021) apontaram a não consideração de processos endógenos como dispersão e dinâmica populacional na geração de espécies virtuais. (MEYNARD; LEROY; KAPLAN, 2019) concluíram que ainda não há substituto à altura de dados reais de presença e ausência de qualidade. Porém, segundo (BARBET-MASSIN et al., 2012), os dados de presença são abundantes, mas os dados de ausência são difíceis de obter e muitas vezes não são confiáveis devido ao esforço ou recursos insuficientes para a coleta extensiva dos dados.

Portanto, as espécies do experimento deste trabalho foram criadas a partir de ocorrências de espécies reais, sendo os pontos de ocorrências e ausências gerados a partir do comportamento identificado dessas espécies reais. Essa estratégia permitiu obter dados de presença e ausência que podem ser testados em algoritmos que exigem esses tipos de dados como o GLM. Além disso, com uma espécie virtual gerada a partir de espécies reais, foi possível garantir que a quantidade de pontos de ocorrência não fosse um impeditivo para a utilização de algoritmos específicos que necessitam de mais pontos de dados para convergir a função de mapeamento da distribuição, como as redes neurais (ALWOSHEEL; VAN CRANENBURGH; CHORUS, 2018; BENKENDORF; HAWKINS, 2020); e nem que tivessem influência negativa na capacidade preditiva dos modelos, que já foi consensuado na literatura (BEAN; STAFFORD; BRASHARES, 2012; LIU; NEWELL; WHITE, 2019; PERRY; DICKSON, 2018; SHIROYAMA; WANG; YOSHIMURA, 2020; WISZ et al., 2008). Assim, seriam necessários menos 10 mil pontos de dados gerados, para simular tanto

os pontos de ocorrência como os de ausências, como indicado por (BARBET-MASSIN et al., 2012) e (FOURCADE et al., 2014).

O procedimento para a criação das espécies virtuais segue as seguintes etapas:

A. Escolha das espécies de interesse: dados reais de ocorrência obtidos em um repositório de dados.
B. Aquisição de dados ambientais: obtenção dos dados climáticos e topográficos da região de ocorrência das espécies definidas.
C. Criação do modelo de distribuição: utilizando os dados de ocorrência juntamente com os dados ambientais, os MDEs são criados e ajustados empregando os algoritmos RF, GLM e Maxent. Estes foram escolhidos com base no resultado da REL.
D. Geração de espécie virtual: os modelos criados na etapa anterior são utilizados para gerar pontos de ocorrência e ausência baseados na combinação das probabilidades de ocorrência dos algoritmos utilizados.

As distribuições de probabilidade resultantes da etapa de criação do modelo de distribuição são transformadas em dados de ausência-presença, que representam a “verdadeira” distribuição da espécie virtual. De acordo com (FERNANDES; SCHERRER; GUISAN, 2019a), essa abordagem metodológica garante que a espécie virtual apresente respostas realistas com relação ao padrão de distribuição na região de estudo. Assim, a partir do comportamento identificado em uma espécie real, gera-se um n número de novas ocorrências projetadas sobre um espaço geográfico mais amplo. No cenário de simulação deste trabalho, este espaço geográfico será toda a extensão do continente ou área continental predominante da espécie real de interesse, garantindo que a espécie virtual represente a totalidade da distribuição possível da espécie de interesse.

A escolha das espécies de interesse dá-se também pela amplitude do nicho ecológico da espécie. Segundo (GÁBOR et al., 2020), há uma ligação entre erros de localização e a largura do nicho ecológico com o desempenho de MDEs. (CONNOR et al., 2018; TESSAROLO et al., 2021) também identificaram que diversas características ecológicas e tipos de distribuições geográficas das espécies como a

amplitude de nicho, podem afetar os resultados de MDEs. Por isso, espécies com nichos ecológicos amplos e restritos de diferentes países foram selecionadas.

4.2 Dados Ambientais

O repositório de dados do WorldClim (www.worldclim.org) foi utilizado como uma das fontes para os dados ambientais. Essa base fornece uma interpolação de um conjunto de 19 variáveis climáticas em quatro níveis de granularidade com uma resolução espacial de até 1 km^2 (FICK; HIJMANS, 2017). As variáveis de temperatura e precipitação e suas variações representam condições climáticas responsáveis pela maior parte da variação espacial na estimativa da probabilidade de ocorrência de uma espécie terrestre (BOUCHER-LALONDE; MORIN; CURRIE, 2012). Ademais, dada a disponibilidade de acesso via repositório do WorldClim, a variável topográfica de elevação também foi incluída na simulação. Além do repositório do WorldClim, o conjunto de dados globais de indicadores bioclimáticos CMCC-BioClimInd também foi utilizado. Este novo conjunto de dados complementa a disponibilidade de informações bioclimáticas com 35 indicadores com resolução de 30 segundos, tendo aplicabilidade em estudos ecológicos e ambientais em larga escala (NOCE; CAPORASO; SANTINI, 2020).

4.3 Matriz de Dados e Base de Controle

Nesta etapa, é realizada o georreferenciamento dos pontos de ocorrência originais das espécies selecionadas com as variáveis ambientais locais. Com a construção dessa matriz, três diferentes modelos de classificação são criados. Pontos de cortes são definidos para os modelos de cada um dos algoritmos para indicar o limite a partir do qual considera-se uma ocorrência. Para os resultados de ocorrências, aplica-se a predição de cada modelo para todo o conjunto de dados ambientais disponível. Considera-se como pontos de ocorrência válidos, as coordenadas para as quais os três modelos indicaram ocorrências, para todos os pontos que não obedecem a essa condição, considera-se como ponto de ausência.

A unificação dos pontos de presença-ausência com os dados ambientais resulta em uma base de dados de controle, que será utilizada como referência da distribuição real da espécie virtual. Essa base de controle contém o que se classifica como base de distribuição, simulando uma base levantada a partir do monitoramento/escaneamento intensivo/extensivo e completo de todo o espaço geográfico possível.

4.4 Base de Erros

A construção da base de erros é a etapa fundamental do experimento. Nessa base são inseridos diferentes gradientes de problemas de qualidade para posterior avaliação dos impactos. Três tipos de erros são considerados neste trabalho: erros de localização, erros de identificação e viés geográfico, conforme resultado da REL.

A base de erros é construída a partir da base de controle. Cada um dos pontos de ocorrência e ausência são tratados dentro das especificidades requeridas para cada tipo de erro que se pretende simular, como detalhado a seguir.

4.4.1 Erro de localização

A partir da posição geográfica do ponto de ocorrência, sorteia-se aleatoriamente se a inserção do erro será efetuada na coordenada da latitude, longitude ou em ambas. Em seguida, mais um sorteio é realizado para definir o sentido da modificação. Se o erro tiver que ser inserido na coordenada da latitude em sentido positivo, indica uma movimentação do ponto de ocorrência para o Norte, caso contrário, o ponto deve ser movido em sentido negativo para o Sul. No caso da longitude, se o sentido de modificação for positivo, move-se o ponto de ocorrência ao longo do meridiano para o Leste, caso contrário, o ponto deve ser movido à Oeste. Se a inserção do erro for efetuada simultaneamente nas duas coordenadas, é sorteado aleatoriamente uma direção de modificação para cada uma delas individualmente. Latitude positiva e longitude positiva move-se o ponto à Nordeste, latitude positiva e longitude negativa move-se o ponto à Noroeste, latitude negativa e longitude positiva move-se o ponto à

Sudeste, e, finalmente, latitude negativa e longitude negativa move-se o ponto à Sudoeste.

4.4.2 Erro de identificação

Para simular os erros de identificação duas abordagens podem ser empregadas. A primeira adota elementos da abordagem de ecologista virtual proposta por (ZURELL et al., 2010), na qual, “o modelo ecológico virtual representa uma espécie e/ou ecossistema virtual e inclui processos-chave do sistema ecológico relevantes para a questão em estudo”. Assim, o modelo ecológico virtual pode compreender uma única ou múltiplas espécies, com o nicho ecológico em escala ampla ou restrita, e ser orientado por fatores abióticos. Essa estratégia possibilitou a criação de diversos pacotes de software dedicados à simulação de espécies virtuais, como (DUAN et al., 2015; LEROY et al., 2016; QIAO et al., 2016). Nessa estratégia, é possível gerar espécies virtuais por meio de várias relações espécie-ambiente. Uma das abordagens disponíveis em (LEROY, 2018), por exemplo, permite gerar uma espécie virtual a partir da definição das funções de resposta a cada uma das variáveis ambientais. Essas respostas são combinadas para calcular a adequação ambiental da espécie virtual. É possível combinar funções aditivas ou multiplicativas de resposta para calcular a adequação ambiental e simular uma nova espécie que tenha uma amplitude de ocorrência mais ampla ou mais restrita que a espécie de interesse.

No entanto, para a simulação pretendida neste trabalho, essa abordagem é inadequada porque pode-se rapidamente atingir um problema inextricável de requisitos ambientais irreais devido ao número de variáveis climáticas inicialmente consideradas (55 no total). Como exemplificado por (LEROY, 2018), gerar uma espécie simulada que depende de 5 variáveis de temperatura, 3 variáveis de precipitação e 2 variáveis de uso da terra torna quase impossível saber se suas funções de resposta são realistas em relação às condições ambientais. Portanto, a percepção da degradação do modelo pela presença ou intensidade desse tipo de erro, ficaria comprometida, dada a grande quantidade de variáveis influenciando simultaneamente os resultados.

Assim, para erros de identificação, a estratégia escolhida foi a abordagem de (FERNANDES; SCHERRER; GUISAN, 2019b), na qual ao invés de se trabalhar sobre as funções de resposta para cada uma das variáveis ambientais no nicho ecológico da espécie, pode-se atuar somente sobre os pontos de ocorrência, parametrizando uma espécie cujo nicho de ocorrência seja mais amplo ou restrito que a da espécie de interesse. Dessa forma, o erro pode ser adicionado (i) apenas às presenças, criando falsos negativos, alterando presenças para ausências, ou (ii) apenas para ausências, criando falsos positivos, alterando ausências para presenças. Isso resulta na amplificação ou restrição do nicho ecológico que impacta na mesma proporção em todas as variáveis climáticas simultaneamente.

4.4.3 Viés Geográfico

A simulação do tipo de erro com viés geográfico requer a seleção de uma região geográfica restrita a partir da qual a amostra de treinamento é extraída. Este erro é realizado selecionando aleatoriamente um ponto de ocorrência e simulando um polígono de distribuição cuja amplitude é uma fração variável dos limites de alcance de distribuição da espécie virtual.

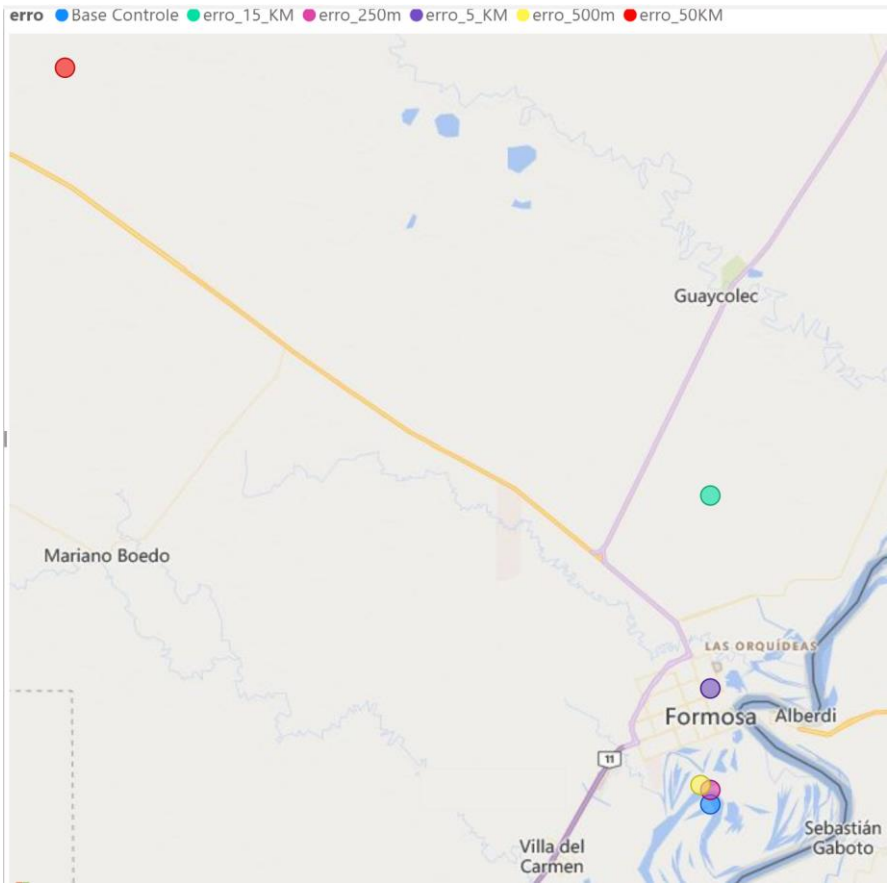
4.5 Amostragem

Na etapa de amostragem são construídas bases de treinamento (extraídas da base de controle e da base de erros) para os algoritmos de modelagem. Cada amostra reflete um cenário diferente com relação ao tipo, intensidade e proporção dos erros. As possibilidades de construções de cenários de simulações são diferentes em virtude da natureza do erro.

Para erros da localização, a partir de um ponto de ocorrência da base de controle, é possível variar a intensidade do erro. No experimento realizado, as intensidades de erro utilizadas foram: 250 metros, 500 metros, 1 quilômetro, 5 quilômetros, 10 quilômetros, 15 quilômetros, 50 quilômetros e 100 quilômetros. A Figura 03 apresenta em azul o ponto de ocorrência de referência na base de controle, os pontos rosa e amarelo representam os erros de 250 e 500 metros, enquanto os

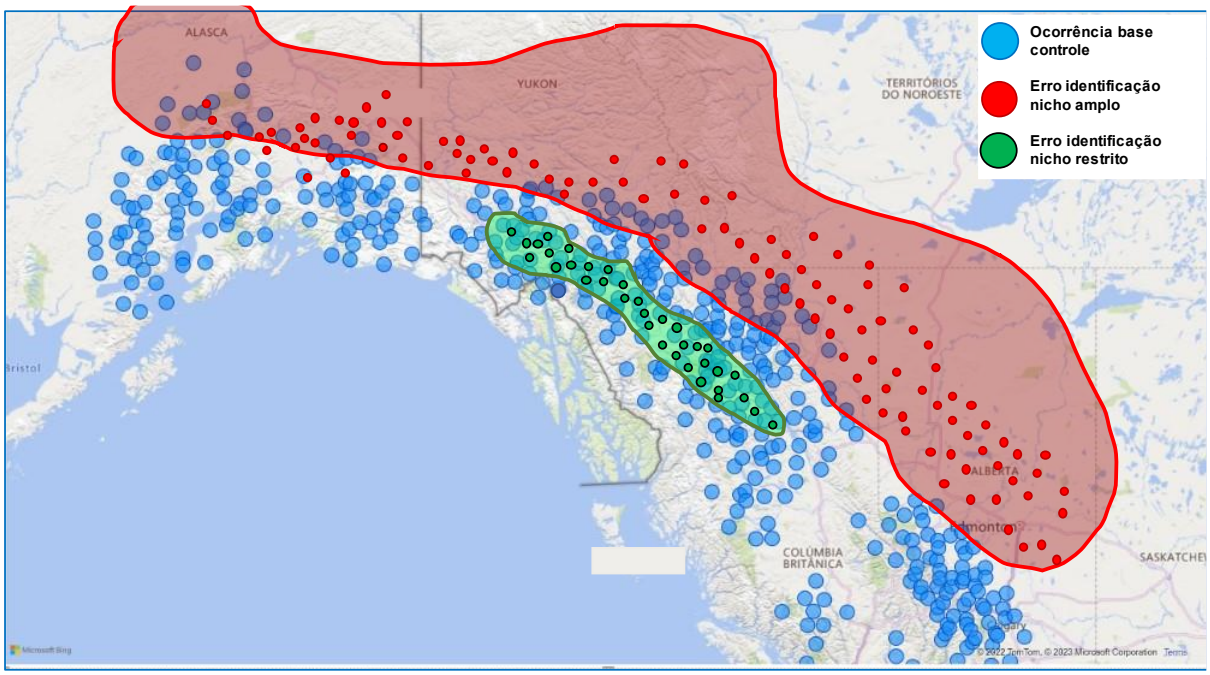
pontos lilás, verde e vermelho represente as intensidades de erro de 5,15 e 50 quilômetros, respectivamente.

Figura 3: Representação do erro de localização



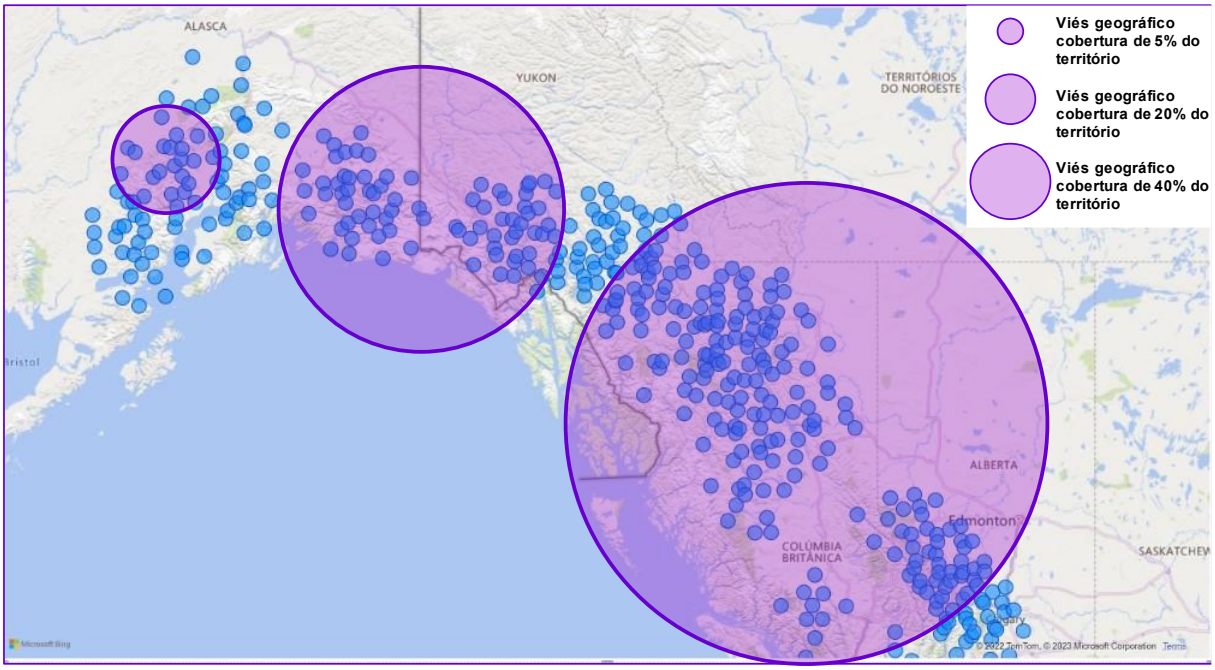
Para erro de identificação da espécie, considera-se apenas dois cenários, um com “espécie contaminante” com nicho ecológico mais amplo e outra com nicho mais restrito como ilustrado na Figura 04.

Figura 4: Erro de identificação da espécie



Para o erro de viés geográfico, as amostras foram construídas de forma a apresentar o erro baseado na cobertura geográfica. As amostras foram integralmente extraídas de polígonos que representam uma fração da amplitude geográfica máxima possível da espécie de interesse. Assim, polígonos que representam 2%, 5%, 10%, 15%, 20%, 30% e 40% da extensão máxima do território de distribuição foram utilizados para obter as amostras. A Figura 05 exemplifica os erros considerando os polígonos de 5%, 20% e 40%.

Figura 5: Erro de viés geográfico



Para os erros de localização, com as 8 intensidades de erro (250 e 500 metros, e 5, 10, 15, 50 e 100 km), e os erros de identificação com dois cenários apenas, nicho amplo e restrito, foram geradas amostras com gradientes de erros de 2%, 5%, 10%, 15%, 20%, 30% e 40%, o complemento da amostra foi obtido da base de controle da espécie de interesse. Dessa forma, um percentual da amostra é extraído da base de controle e outro da base de erro, representando a contaminação da amostra pelos problemas de qualidade. Para cada combinação possível de análise (cenário de erro), 30 amostras de 5000 registros cada foram construídas. Por exemplo, para o erro de localização de 250 metros de uma espécie de interesse com uma proporção de erro de 20%, cada uma das 30 amostras contém 1000 registros da base de erro de localização de 250 metros e 4000 registros da base de controle da espécie de interesse em questão. As 30 amostras são necessárias para que haja representação diversa em cada cenário de erro cujas amostras são obtidas amostrando aleatoriamente as bases de erros e controle.

As 8 intensidades de erros de localização foram amostradas em 7 proporções de erro, totalizando 56 cenários de teste. As amostras do erro de identificação, que contemplam nichos amplos e restritos também são geradas com gradientes de 2%, 5%, 10%, 15%, 20%, 30% e 40%, assim, 14 cenários de testes foram gerados para este tipo de erro. Para erro de viés geográfico não há gradientes de erros dado que

cada amostra é obtida integralmente de dentro do polígono que representa um percentual de cobertura do espaço geográfico, desta forma, mais 7 cenários de teste foram gerados, representando percentuais de 2%, 5%, 10%, 15%, 20%, 30% e 40% de cobertura do espaço geográfico total. Isso totaliza 77 cenários de teste e 2.310 amostras geradas, 30 de cada cenário, para cada espécie de interesse.

Além das amostras de treinamento construídas a partir da composição das bases de erro e de controle, mais 30 amostras de dados foram obtidas exclusivamente da base de controle para treinar modelos de controle, o que permite comparar os desempenhos e resultados dos modelos gerados com dados ideais e aqueles treinados com dados “contaminados”.

4.6 Modelagem e Validação

Para a etapa de criação dos modelos, cinco algoritmos foram empregados. Os três mais tradicionais em MDEs, GLM, RF e Maxent, como apontado pela REL, foram incluídos. O quarto algoritmo foi o de redes neurais. Embora este não seja amplamente utilizado em MDEs, com o sucesso desse algoritmo em outras áreas como no setor financeiro (RYLL; SEIDENS, [s.d.]), previsão do tempo (SUBHAJINI, 2018) e saúde (SHAHID; RAPPON; BERTA, 2019), resolveu-se investigar seus resultados, quando dados com problemas de qualidade são empregados. E como novas técnicas estão continuamente sendo introduzidas no campo de estudos de distribuição de espécies, o quinto algoritmo selecionado foi o XGBoost (*Extreme Gradient Boosting*) proposto em 2016 por (CHEN; GUESTRIN, 2016). Esse é um algoritmo relativamente novo que é uma evolução do algoritmo de árvores de decisão impulsionadas. Nos últimos dois anos, começou a ser adotado em MDEs, como nos trabalhos de (ALDOSSARI; HUSMEIER; MATTHIOPOULOS, 2022; LEROY, 2022; VALAVI et al., 2022), e vem apresentando bom desempenho, superando em alguns casos modelos criados por algoritmos bem estabelecidos como o Maxent (CAI et al., 2022; EFFROSYNIDIS et al., 2020; FENG et al., 2021; ZHAO et al., 2022). Para a implementação dos modelos GLM, RF, redes neurais e Maxent, utilizou-se os pacotes dismo e biomod da linguagem R. Os modelos de XGBoost foram implementados com o pacote Caret.

Para a análise dos resultados, em virtude do grande número de cenários de teste, fez-se a consolidação das médias dos resultados de cada métrica de validação da seção 2.7. Além das métricas da seção 2.7, adicionamos também duas métricas de validação propostas por (WUNDERLICH et al., 2019) para o contexto de MDEs, *Odds Ratio Skill Score* (ORSS) e *Symmetric Extremal Dependence Index* (SEDI). Os autores propuseram essas métricas em substituição à métrica de TSS porque, de acordo com eles, se o número total de registros previstos indicados na matriz de confusão exceder 30.000 pontos, a métrica TSS converge para a métrica de sensibilidade, sendo fortemente enviesado para ausências "verdadeiras", comprometendo a avaliação dos modelos.

Para cada amostra de cada cenário de teste, os 5 algoritmos foram utilizados para gerar os modelos, e para cada modelo foi aplicado a totalidade da base de controle da respectiva espécie de interesse para indicar a capacidade de generalização dos modelos gerados a partir de uma base de treinamento com a qualidade de dados comprometida. Para cada previsão, as métricas de validação foram calculadas e posteriormente consolidadas em resultados médios para análise e comparação com os resultados obtidos a partir dos modelos de controle. Construção da base de controle para cada espécie virtual

4.7 Caracterização dos dados de ocorrência

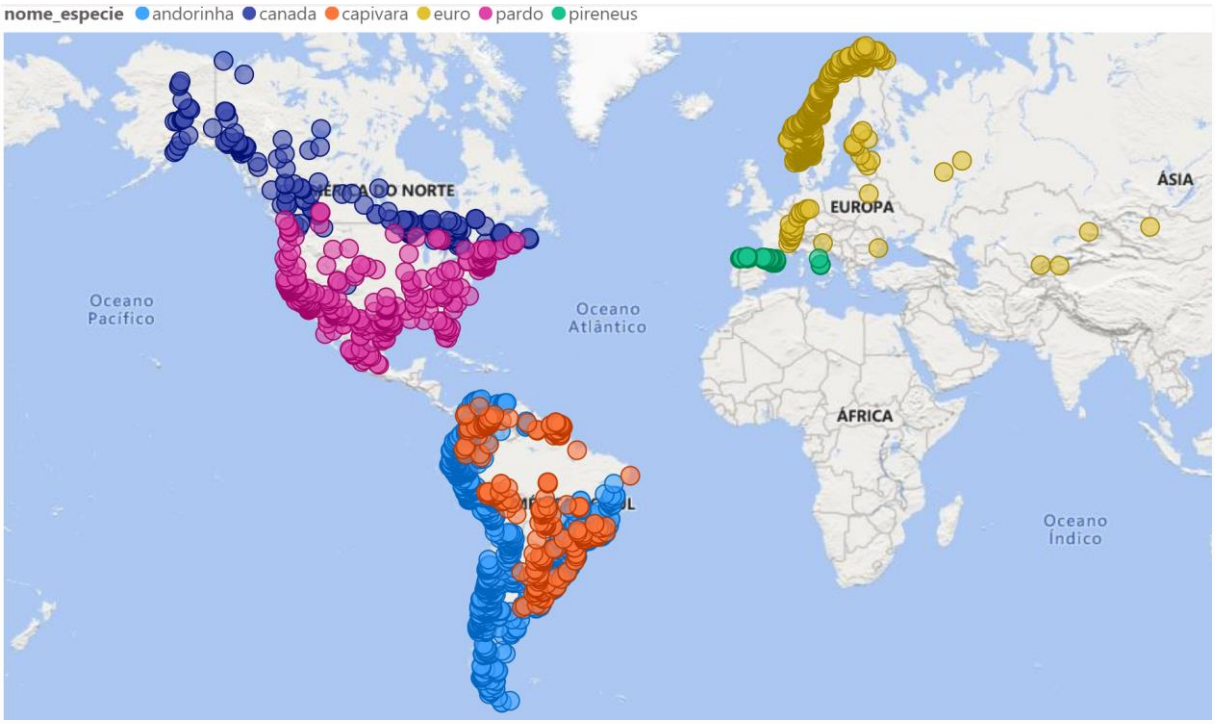
Os dados de ocorrência foram coletados no repositório de dados do Gbif. Para a seleção das espécies de interesse, primeiramente definiu-se três regiões de estudo com características climáticas distintas. As regiões foram: América do Sul, América do Norte e Eurásia. Essa diferenciação de regiões foi feita para ponderar o aspecto climático regional com os problemas de qualidade. Outro ponto de análise considerado foi a amplitude do nicho ecológico da espécie de interesse. Por isso, para cada região de estudo foram coletados dados de uma espécie generalista de nicho ecológico amplo e de uma espécie específica de nicho restrito. A cobertura temporal dos dados de ocorrência foi de janeiro de 2010 a agosto de 2020. A Tabela 05 detalha a espécie alvo de cada nicho ecológico em cada uma das regiões do estudo.

Tabela 5: Espécies de interesse e suas características

Referência	Região	Área/País predominante	Espécie	Nome científico	Nicho ecológico	Tamanho da amostra original
Andorinha	América do Sul	América do Sul	Andorinha-azul-e-branca	Notiochelidon cyanoleuca	Amplo	33.435
Capivara	América do Sul	Brasil	Capivara	Hydrochoerus hydrochaeris	Restrito	1.038
Pardo	América do Norte	Estados Unidos	Lince-pardo	Lynx rufus	Amplo	6.460
Canada	América do Norte	Canadá	Lince-do-canadá	Lynx canadensis	Restrito	223
Euro	Eurásia	Europa/Sibéria	Lince-euroasiático	Lynx lynx	Amplo	15.536
Pirineus	Eurásia	Montanhas Cantábricas	Camurça-dos-pirenéus	Rupicapra pyrenaica	Restrito	2.592

A Figura 06, apresenta um mapa com os pontos de ocorrência originais projetados sobre o espaço geográfico.

Figura 6: Mapa de com pontos de ocorrência das espécies



4.8 Caracterização dos dados ambientais

20 variáveis ambientais, com cobertura histórica de 1970 e 2000, foram capturadas no repositório do WorldClim com resolução de 30 segundos. A base de dados do CMCC-BioClimInd (NOCE; CAPORASO; SANTINI, 2020) disponibiliza 35 indicadores bioclimáticos, no entanto, capturou-se somente 16 variáveis complementares às variáveis já disponibilizadas pelo WorldClim. Essas variáveis complementares cobrem o período histórico de 1960 a 1999. Diversas variáveis ambientais do CMCC-BioClimInd são fortemente correlacionadas às variáveis do WorldClim, por isso optou-se por capturar somente o complemento àquelas já obtidas. A matriz de cada uma das variáveis climáticas, na resolução de 30 segundos, que equivale a aproximadamente 1 km^2, apresenta dimensão de 21.600 x 43.200, totalizando 933.120.000 pontos de dados. O detalhamento e definição de cada uma das variáveis utilizadas estão em (FICK; HIJMANS, 2017; NOCE; CAPORASO; SANTINI, 2019). Para a geração da base de controle as variáveis obtidas foram as seguintes:

- BIO1 = Temperatura Média Anual
- BIO2 = Range Diurno Médio
- BIO3 = Isotermalidade
- BIO4 = Sazonalidade da Temperatura
- BIO5 = Temperatura máxima do mês mais quente
- BIO6 = Temperatura mínima do mês mais frio
- BIO7 = Faixa anual de temperatura
- BIO8 = Temperatura média do trimestre mais úmido
- BIO9 = Temperatura média do trimestre mais seco
- BIO10 = Temperatura Média do trimestre mais quente
- BIO11 = Temperatura média do trimestre mais frio
- BIO12 = Precipitação anual
- BIO13 = Precipitação do mês mais úmido
- BIO14 = Precipitação do mês mais seco
- BIO15 = Sazonalidade de precipitação
- BIO16 = Precipitação do trimestre mais úmido

- BIO17 = Precipitação do trimestre mais seco
- BIO18 = Precipitação do quarto mais quente
- BIO19 = Precipitação do trimestre mais frio
- ELEV = Elevação
- BIO20 = Quociente de Ellenberg
- BIO21 = Temperatura positiva anual
- BIO22 = Soma da temperatura anual
- BIO23 = Índice ombrotérmico
- BIO24 = Precipitação positiva anual
- BIO25 = Índice de frio Kira modificado
- BIO26 = Índice de calor Kira modificado
- BIO27 = Índice de continentalidade simplificado
- BIO28 = Temperatura média do mês mais quente
- BIO29 = Temperatura média do mês mais frio
- BIO30 = Temperatura média do mês mais seco
- BIO31 = Temperatura média do mês mais chuvoso
- BIO32 = Índice de termicidade modificado
- BIO33 = Índice ombrotérmico do verão e do mês anterior
- BIO34 = Evapotranspiração Potencial Hargreaves
- BIO35 = Evapotranspiração Potencial Thornthwaite

4.9 Construção da base de controle

A matriz de dados de cada uma das espécies de interesse (Tabela 05) foi construída considerando as informações climáticas dos pontos geográficos de ocorrência das espécies. Esse processo inicia-se com a seleção dos dados de latitude e longitude de ocorrência das espécies, a partir disso busca-se, dentro do *raster* (forma de representação matricial de células de um sistema de informação geográfica) (HIJMANS et al., 2022) de cada variável bioclimática as informações ambientais daquela posição geográfica. A obtenção dessas informações ambientais é realizada por meio de interpolação bilinear. Além dos pontos de ocorrência das espécies,

obtém-se também todas as coordenadas e informações ambientais disponíveis dentro do *raster* de cada variável ambiental. Consolida-se então um *dataframe*, que são objetos de dados genéricos da linguagem R, usados para armazenar dados tabulares. A consolidação desse *dataframe* dá-se pela conjunção das coordenadas geográficas dos pontos de ocorrência com seus respectivos dados ambientais, além de todas as coordenadas e informações ambientais restantes do *raster*. Nessa etapa de consolidação, as variáveis BIO23, BIO24, BIO25 e BIO26 foram excluídas porque a extração das informações resultou em muitos valores vazios.

Com a matriz de dados consolidada, inicia-se o pré-tratamento dos dados para garantir que a base de controle seja gerada obedecendo as condições de (MATEO; FELICÍSIMO; MUÑOZ, 2011). Nesta etapa faz-se a seleção de atributos. O melhor subconjunto de atributos é o que permite, ao mesmo tempo, reduzir a dimensionalidade dos atributos e contribuir para a construção de modelos mais simples e compreensíveis, melhorando o desempenho de mineração de dados (LI et al., 2017). Além disso, essa etapa é importante para proteger os modelos do sobreajuste que pode ocorrer quando os modelos são treinados com muitos atributos (TUV; BORISOV; TORKKOLA, 2006).

Para a seleção de atributos, empregou-se algumas técnicas tradicionalmente utilizadas:

- DALEX: técnica que oferece um procedimento independente de um modelo previamente treinado para calcular a importância do atributo (LAW BIECEK, 2018) (LAW BIECEK, 2018).
- RFE: é um tipo de seleção de atributos, cujo objetivo principal é reduzir a dimensão dos dados escolhendo um subgrupo de variáveis com maior capacidade de diferenciação (CHEN et al., 2018).
- *Variable Importance*: técnica utilizada para caracterizar o efeito geral dos preditores sobre os modelos (LOH; ZHOU, 2021).
- Boruta: algoritmo de classificação e seleção de atributos baseado no algoritmo de RF que ajuda a escolher os atributos mais importantes a partir da avaliação da significância estatística de cada variável (KURSA, 2020; KURSA; JANKOWSKI; RUDNICKI, 2010).

A iniciativa de utilizar quatro técnicas distintas foi para encontrar, por meio de combinação, dos 36 disponíveis, os melhores atributos para cada espécie de interesse. Assim, cada técnica resultou em um ranking próprio de atributos, e os que foram mais bem ranqueados na combinação das 4 técnicas foram selecionados. Como o objetivo final era obter uma compreensão ecológica sobre a dinâmica de distribuição de cada espécie de interesse, avançou-se para a seleção de atributos a fim de encontrar o menor conjunto possível que mantivesse excelente capacidade de explicação da distribuição da espécie de interesse.

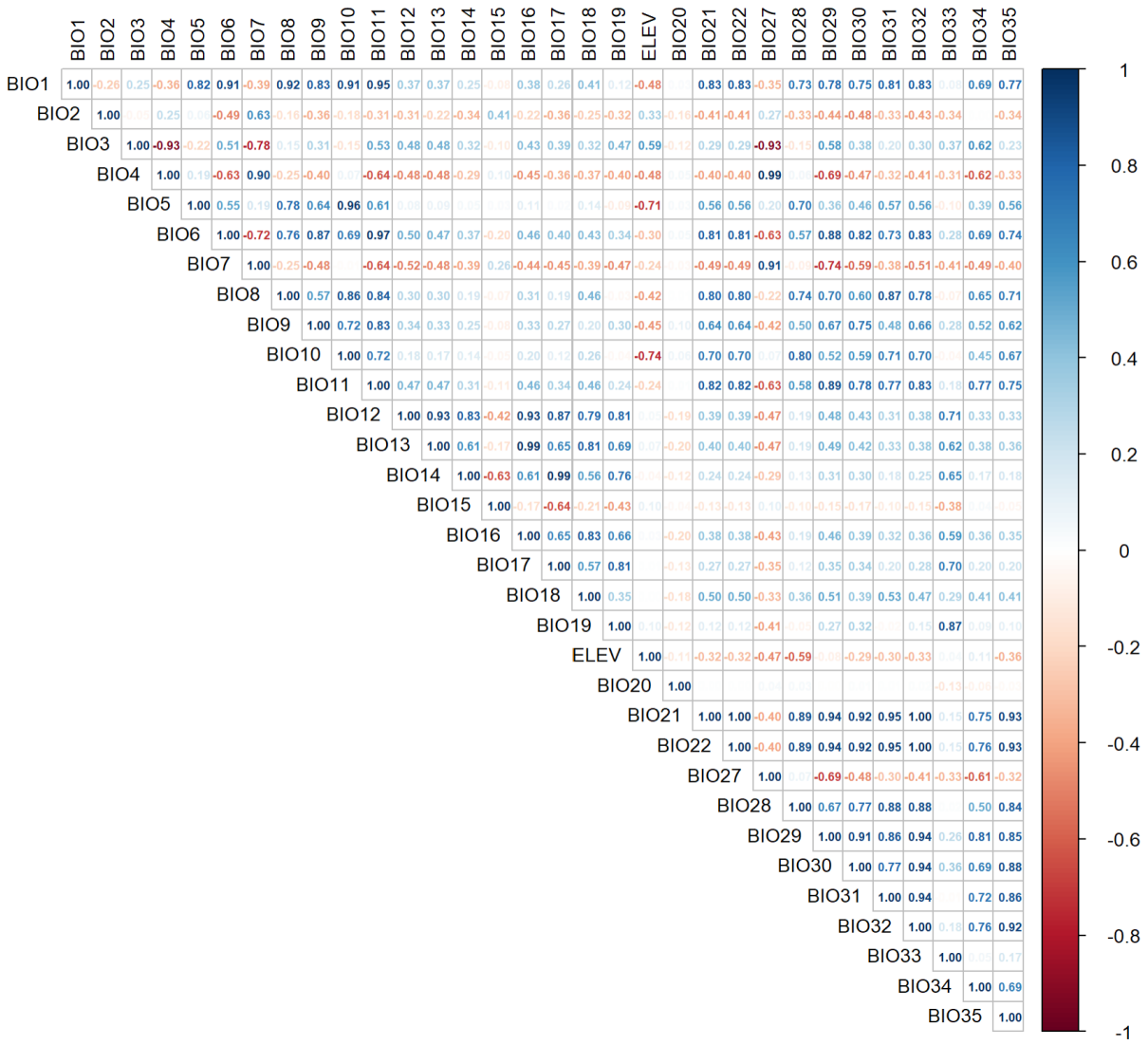
As etapas de pré-processamento foram:

1. Análise de correlação dos atributos
2. Sorteio de amostras
3. Ranqueamento de atributos
4. Seleção final de atributos
5. Treinamento dos modelos
6. Geração da base de controle

4.9.1 Análise de correlação dos atributos

Na primeira etapa do processamento foi realizada uma análise de correlação dos atributos. Essa análise foi realizada a partir de uma matriz de correlação que foi utilizada para examinar a relação linear entre as variáveis ambientais. O coeficiente de correlação pode variar em valor de -1 a +1, e quanto maior o valor absoluto do coeficiente, mais forte é a relação entre as variáveis. Para este trabalho, definiu-se o limiar de 0,75, assim, atributos com coeficiente de correlação absoluto acima deste valor foram descartados. A análise para exclusão foi iniciada a partir das primeiras variáveis, de acordo com a ordem apresentada na seção 4.8. Dessa forma, manteve-se uma variável em análise e eliminou-se todas as variáveis sequentes que estivessem fortemente correlacionadas a esta. A Figura 07 apresenta a matriz de correlação para a espécie de interesse andorinha.

Figura 7: Matriz de correlação de variáveis com nível de significância de 0.01 - andorinha



4.9.2 Sorteio de amostras

Para cada espécie de interesse foram sorteadas 10 amostras de acordo com os seguintes critérios: (i) cada amostra deve conter em torno de 5.000 registros, se o número de ocorrências da espécie de interesse for maior que 2.500; caso contrário, a amostra deve conter aproximadamente o dobro do número de ocorrências disponível; (ii) as amostras devem estar balanceadas, ou seja, o número de ocorrências e não ocorrências devem ser similares. As amostras são sorteadas a partir dos pontos de ocorrências e ausências. O objetivo de construir 10 amostras de 5.000 registros e não uma única amostra contemplando a todos os pontos da amostra original foi para se

aproximar a simulação ao cenário real de MDEs em que o número de ocorrências normalmente disponível é limitado (BRACKEN et al., 2022; LUAN et al., 2021; SHIREY et al., 2019).

4.9.3 Ranqueamento de atributos

Com as 10 amostras de cada espécie de interesse, aplicou-se cada uma das 4 técnicas de seleção de atributos, ou seja, cada técnica foi executada 10 vezes para cada espécie de interesse. Em cada técnica, construiu-se uma tabela com os resultados de cada amostra. Também se criou gráficos consolidados para cada técnica. As Figuras de 08 e 09 apresentam os resultados das amostras da espécie de interesse andorinha aplicadas a cada técnica.

Figura 8: Importância de variáveis DALEX/RFE – andorinha

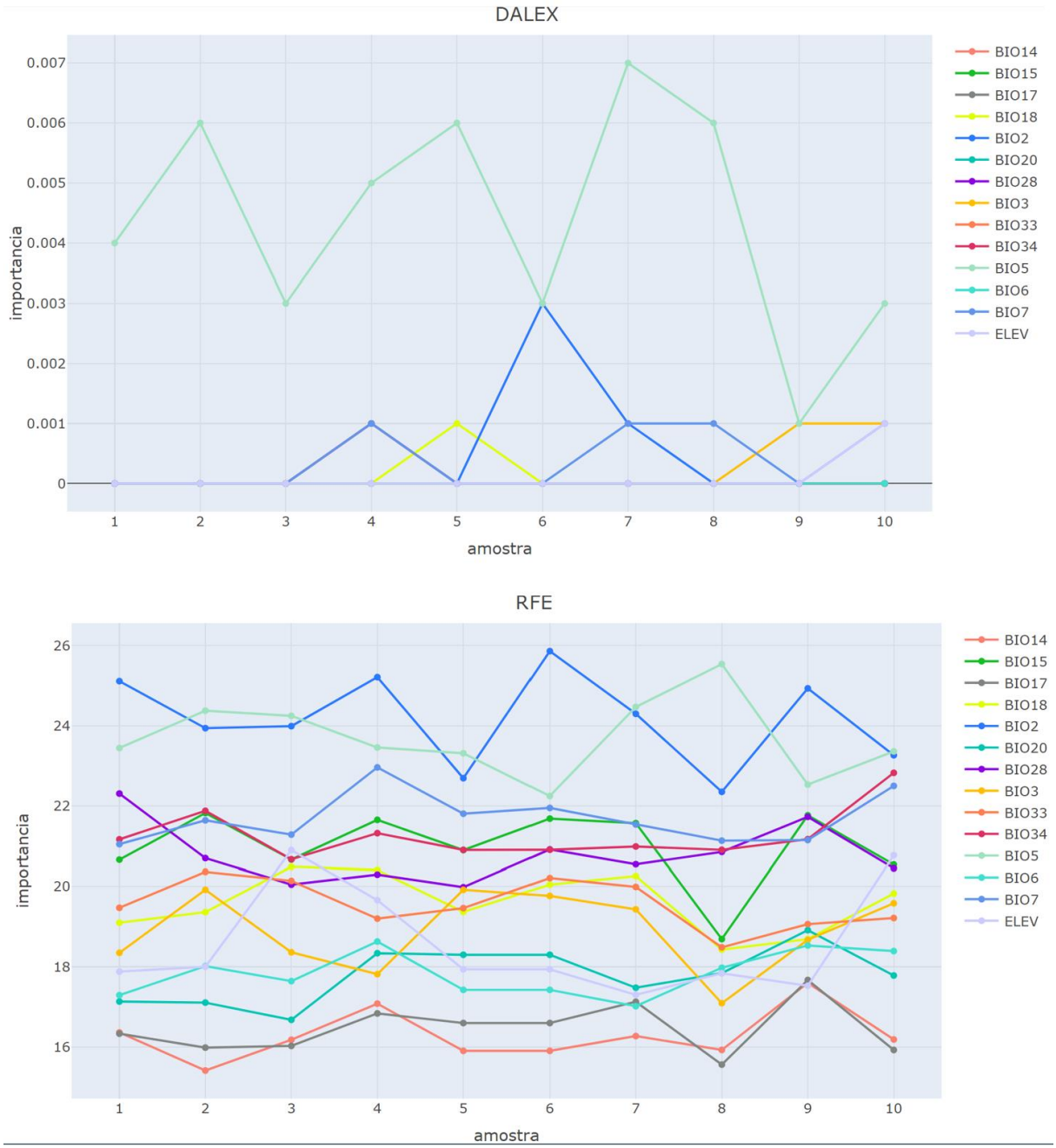
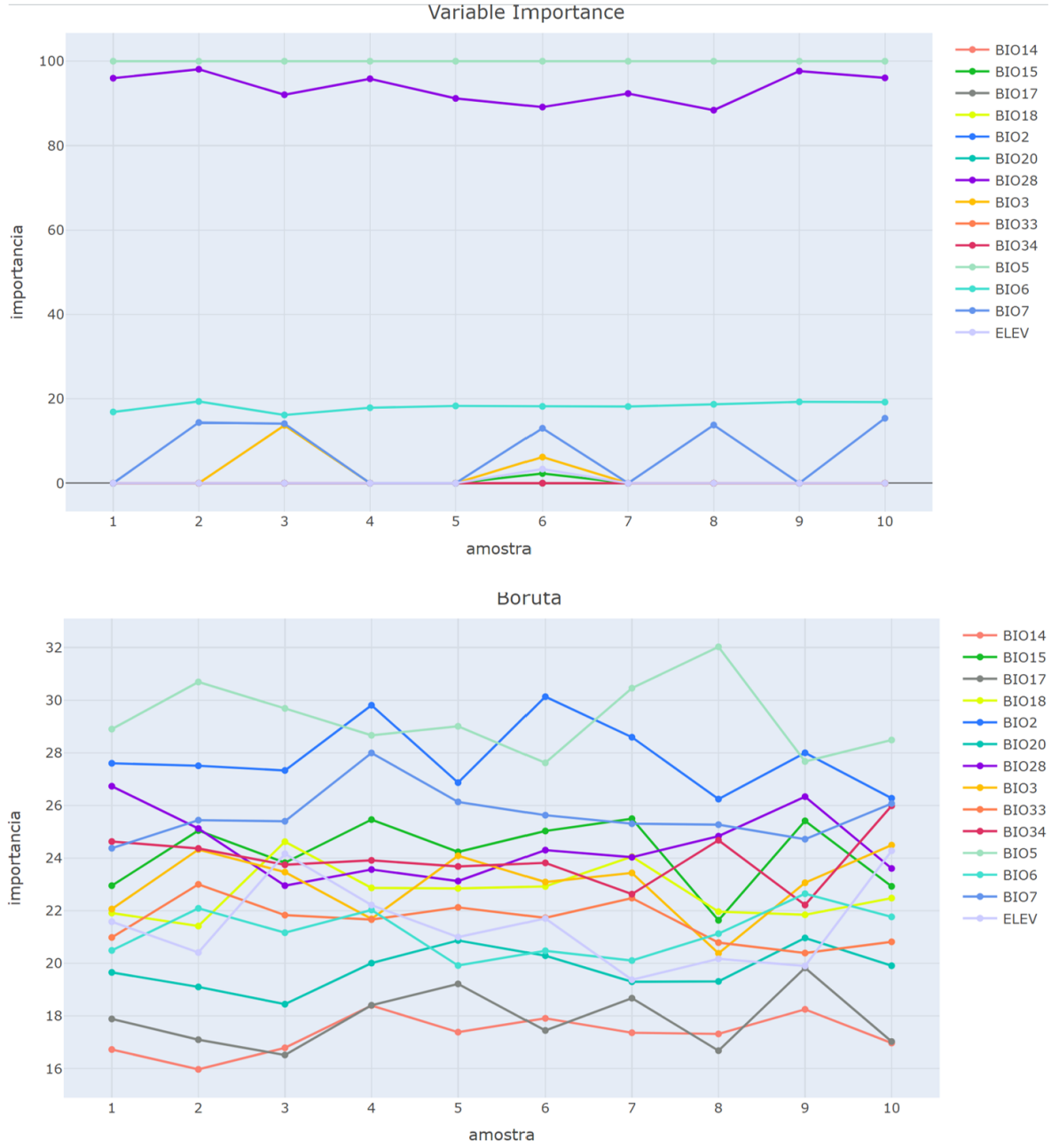


Figura 9: Importância de variáveis Variable importance/Boruta - andorinha



Finalmente, os resultados das médias de cada técnica foram normalizados para o intervalo de 0 a 1 e o somatório das quatro técnicas foram consolidados em um gráfico com o ranking final para cada espécie de interesse. Os resultados finais de cada espécie de interesse são apresentados nas Figuras 10, 11 e 12.

Figura 10: Ranking de variáveis capivara/andorinha

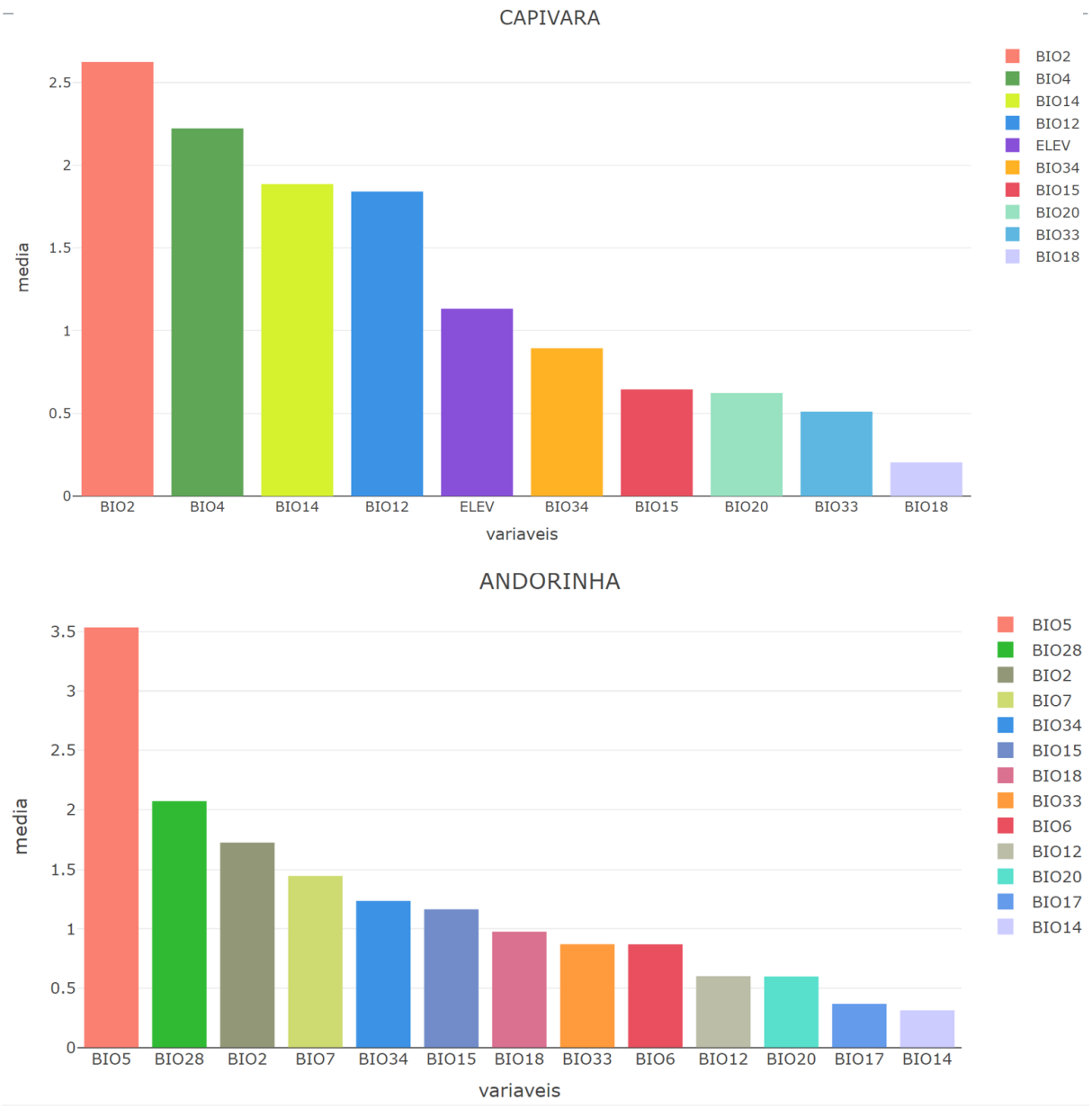


Figura 11: Ranking de variáveis canada/pardo

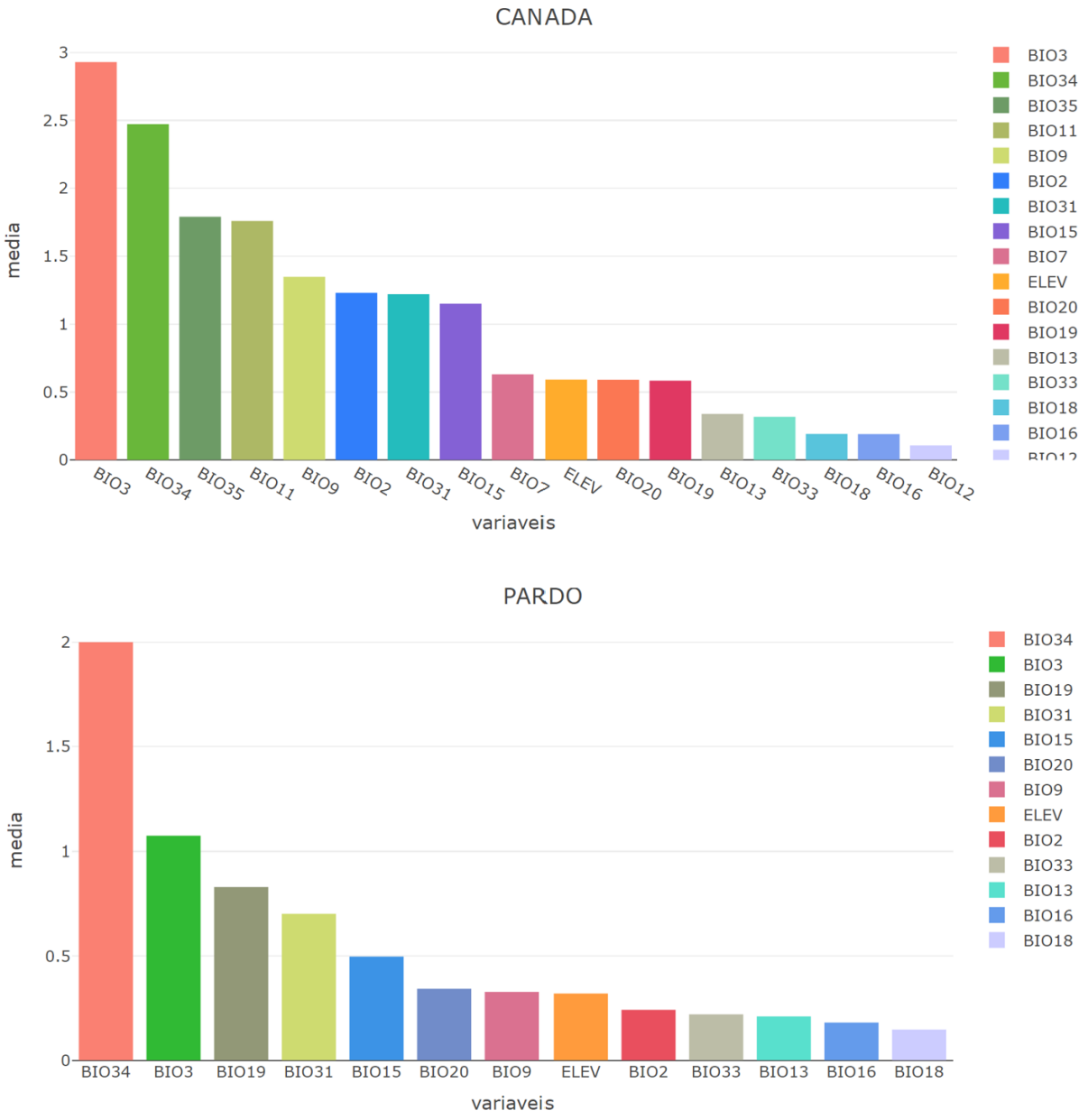
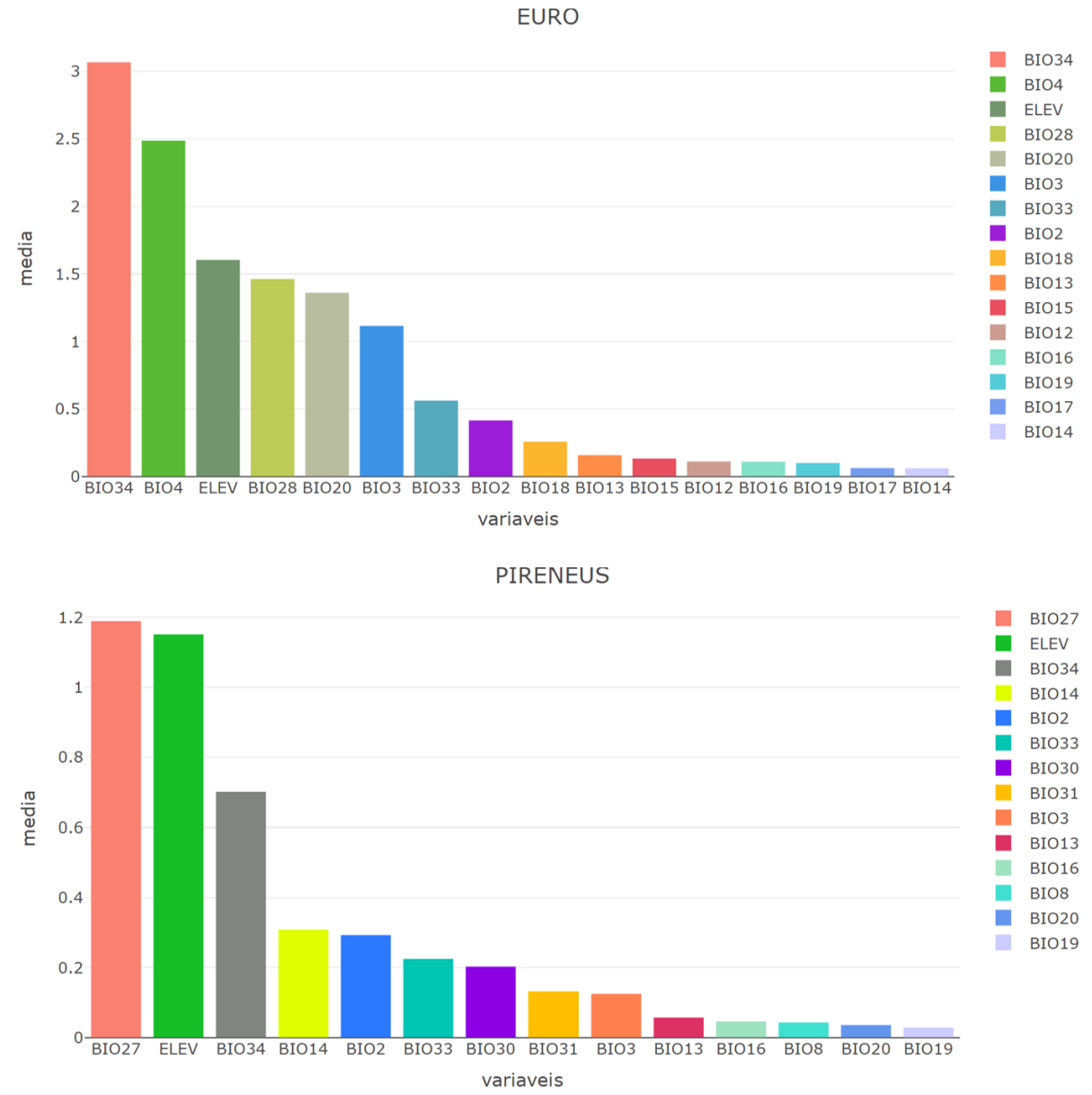


Figura 12: Ranking de variáveis euro/pireneus



4.9.4 Seleção final de atributos

Para a seleção final dos atributos, ponderou-se o tempo computacional de treinamento dos modelos, a melhoria no desempenho dos algoritmos e o risco de sobreajuste, optando-se por manter somente os 8 que melhor explicaram a distribuição das espécies de interesse. Ademais, em MDEs, em virtude da dificuldade de obtenção de dados ambientais na diversidade, localidade e resolução necessárias, costuma-se trabalhar com um pequeno número de atributos, como (GOMES et al., 2018b) que trabalharam com 6 preditores e (GÁBOR et al., 2019) que utilizaram

somente 3. Assim, as variáveis finais selecionadas para cada espécie de interesse são apresentadas na Tabela 06.

Tabela 6: Lista de variáveis selecionadas

Variáveis selecionadas Capivara	Variáveis selecionadas Andorinha	Variáveis selecionadas Canada	Variáveis selecionadas Pardo	Variáveis selecionadas Euro	Variáveis selecionadas Pirineus
1. bio2	1. bio5	1. bio3	1. bio34	1. bio34	1. bio27
2. bio4	2. bio28	2. bio34	2. bio3	2. bio4	2. elev
3. bio14	3. bio2	3. bio35	3. bio19	3. elev	3. bio34
4. bio12	4. bio7	4. bio11	4. bio31	4. bio28	4. bio14
5. elev	5. bio34	5. bio9	5. bio15	5. bio20	5. bio2
6. bio34	6. bio15	6. bio2	6. bio20	6. bio3	6. bio33
7. bio15	7. bio18	7. bio31	7. bio9	7. bio33	7. bio30
8. bio20	8. bio33	8. bio15	8. elev	8. bio2	8. bio31

4.9.5 Modelagem de algoritmos

Na construção dos modelos, a partir da base integral da matriz de dados uma amostra balanceada foi extraída e dividida em um grupo de treinamento e outro de validação na proporção de 70-30 (70% de treinamento e 30% de validação). Neste sorteio, para manter a proporção dos pontos de ocorrência e não ocorrência, as classes de ocorrência foram separadas para o sorteio da amostra, ou seja, da classe de ocorrência sorteou-se 70% para treinamento e 30% para validação, e a classe de não ocorrência seguiu o mesmo procedimento. Em seguida, a amostra de treinamento foi utilizada para treinar o algoritmo de um modelo linear generalizado para classificação binária, um algoritmo de RF e um algoritmo Maxent.

4.9.6 Geração da base de controle

Na geração da base de controle, a totalidade da matriz de dados de cada espécies de interesse foi aplicada a cada um dos modelos, deste modo cada registro da matriz de dados foi classificado como ocorrência ou ausência. Finalmente, considera-se como ponto de ocorrência válido aquele registro para o qual os três

algoritmos o indicaram como ocorrência. Assim, a Tabela 07 apresenta o resultado das bases de controle.

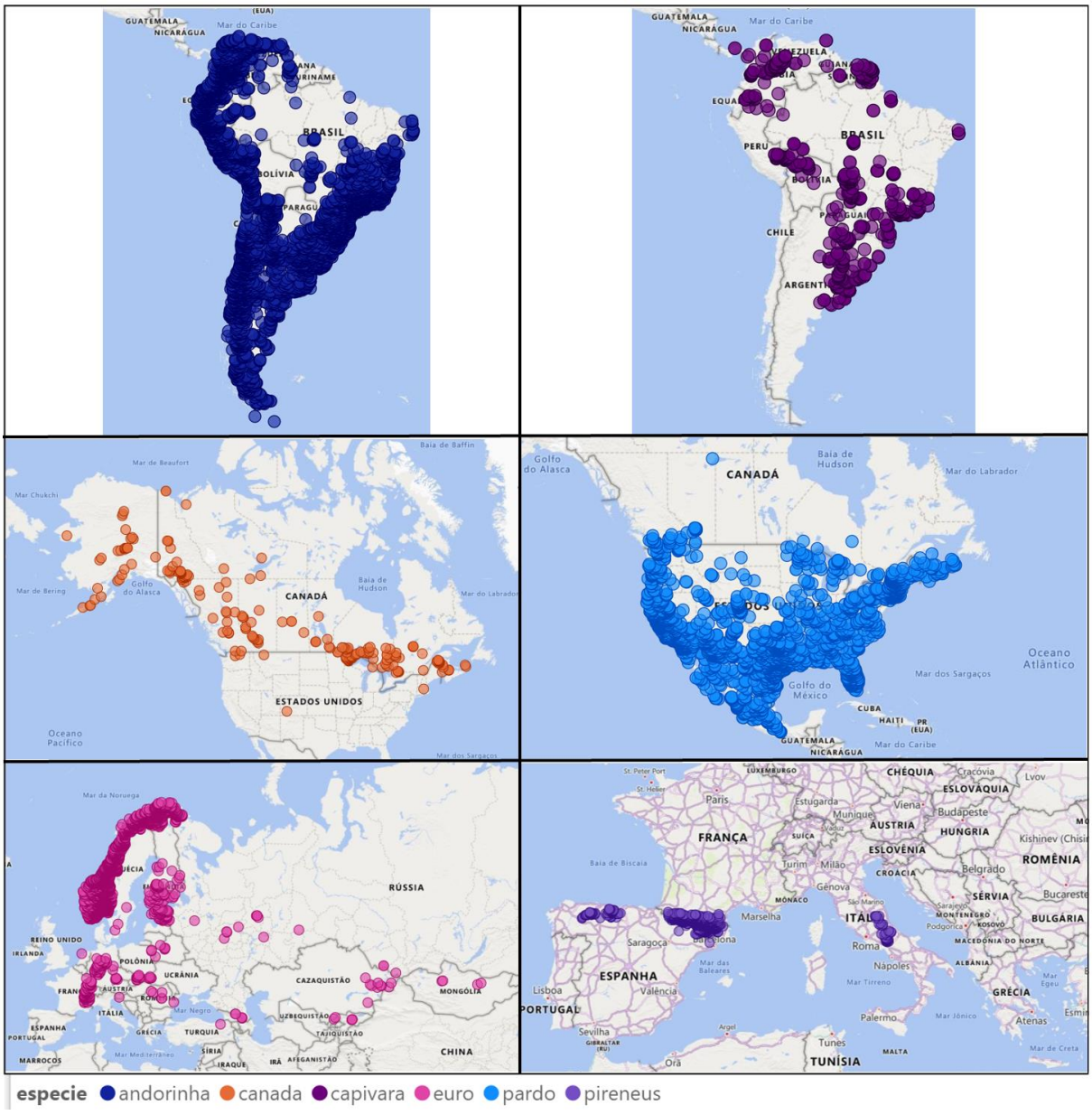
Tabela 7: Número de ocorrências na base de controle gerada para cada espécie virtual

Referência	Tamanho da amostra original	Número de ocorrências na base de controle	Aumento percentual
Andorinha	33.435	486.507	1.455%
Capivara	1.038	183.071	17.637%
Pardo	6.460	21.865	338%
Canada	223	113.220	50.771%
Euro	15.536	127.495	821%
Pirineus	2.592	31.985	1.234%

(BARBET-MASSIN et al., 2012) e (FOURCADE et al., 2014) indicam que pelo menos 10 mil pontos de dados devem ser gerados, para simular tanto os pontos de ocorrência como os de pseudo-ausências. Mais pontos de dados podem ser criados porque alguns algoritmos necessitam de mais dados para convergir a função de mapeamento da distribuição. Mas conforme apresentado na Tabela 07, o número de ocorrências superou em pelo menos 219% esse número mínimo recomendado pelos autores.

A Figura 13 apresenta a projeção de uma amostra da base de controle de cada espécie de interesse sobre o espaço geográfico predominante da espécie.

Figura 13: Projeção das ocorrências das bases de controle

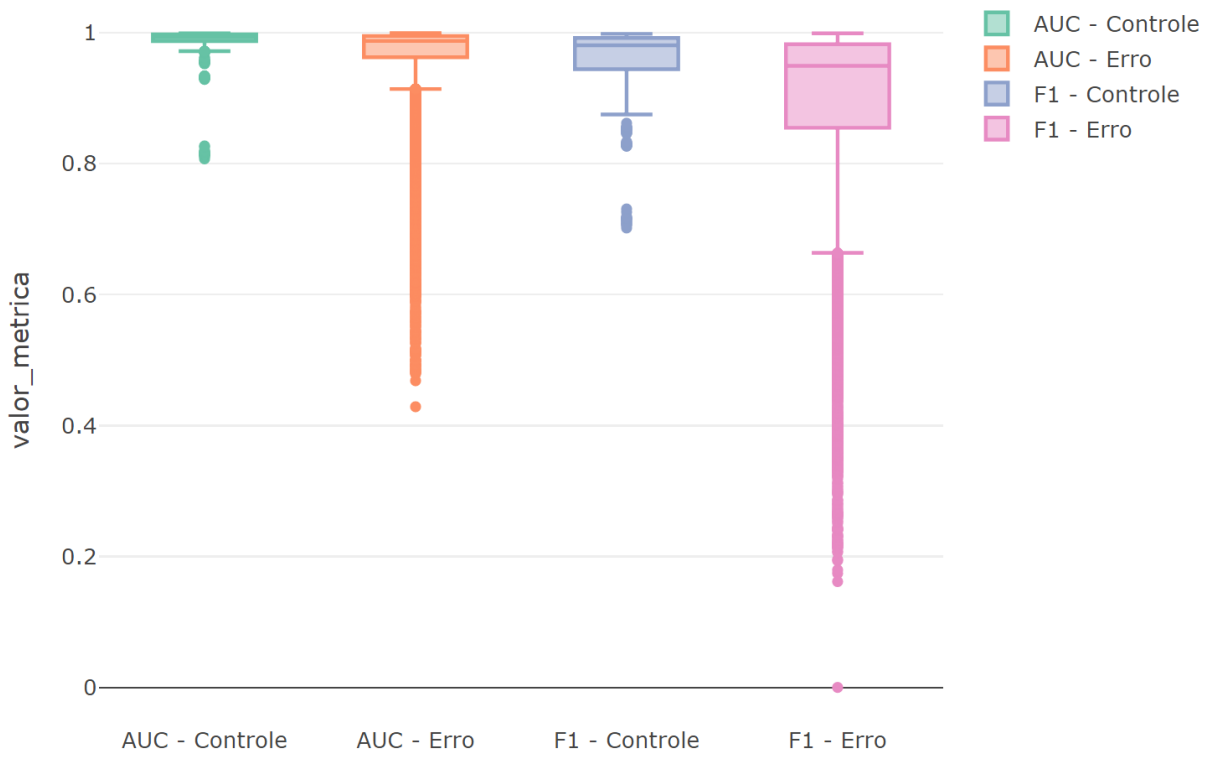


5 ANÁLISE DOS RESULTADOS

Os problemas de QD em MDEs avaliados foram: erros de localização, identificação e viés geográfico conforme já discutido. Para avaliar o impacto causado nos MDEs por tais problemas de QD, primeiro considerou-se os modelos de controle, gerados com dados sem “contaminação”, avaliando se, em média, realmente geram resultados mais consistentes se comparado aos modelos gerados a partir de dados com problemas de qualidade. Para essa comparação, utilizou-se duas métricas de validação: AUC e F1 Score. Embora todas as métricas descritas da seção 2.7 possam ser empregadas na avaliação dos resultados dos modelos, optou-se por AUC porque é uma métrica amplamente utilizada em MDEs (JIMÉNEZ; SOBERÓN, 2020), enquanto F1 Score é uma média harmônica que congrega as métricas de precisão e sensibilidade. A acurácia, embora seja uma métrica importante por descrever o número de previsões corretas sobre todas as previsões, em um cenário que o número de classes negativas (ausência) é significativamente maior que de positivas (ocorrências), a acurácia pode mascarar deficiências dos modelos em prever a classe de interesse.

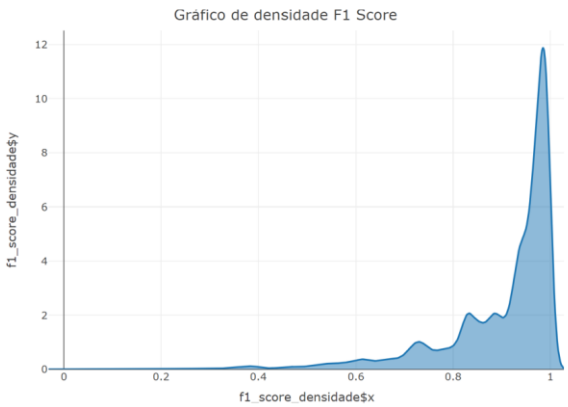
A Figura 14 apresenta um *boxplot*, que é uma representação gráfica de quartis e permite visualizar a simetria e distorção dos dados. A figura representa o mais amplo nível de abstração dos resultados, permitindo a comparação dos resultados para os modelos de controle e os modelos afetados por problemas de QD.

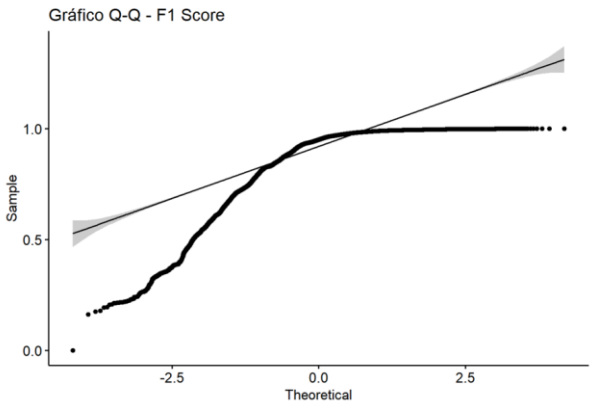
Figura 14: Boxplot das métricas AUC e F1 para modelos de controle e de erro



A Figura 14 mostra que os modelos criados a partir de dados com problemas de QD apresentam resultados com muitos *outliers*, enquanto os modelos de controle geram métricas mais consistentes. A AUC apresenta uma variabilidade menor nos resultados se comparado ao F1 Score. Para verificar se existe diferença estatisticamente significativa entre as duas amostras de resultados, controle e erro, aplicou-se o teste não paramétrico de Mann-Whitney. O teste de Mann-Whitney é indicado para comparação de dois grupos não pareados para se verificar se pertencem ou não à mesma população (NACHAR, 2008). O intuito era utilizar o teste t não pareado, no entanto as amostras não atendiam ao pressuposto de normalidade da distribuição, nem o de igualdade das variâncias como mostrado na Figura 15.

Figura 15: Gráfico de densidade com análise de normalidade e igualdade das variâncias





H_0: a população é normalmente distribuída
H_1: a população não é normalmente distribuída
Teste de Shapiro: *p-value* < 0,001

H_0: $\sigma^2_1 = \sigma^2_2$, as variâncias são iguais
H_1: $\sigma^2_1 \neq \sigma^2_2$, as variâncias não são iguais
Teste F: *p-value* < 0,001

Na Figura 15, o gráfico de densidade do F1 score indica uma assimetria negativa e o gráfico Q-Q (permite comparar duas distribuições de probabilidade e descobrir o tipo de distribuição de uma variável) apresentou um comportamento não normal da distribuição. A análise gráfica dá forte indicativos de que as distribuições não são gaussianas. O resultado da execução do teste de Shapiro-Wilk, teste que avalia se uma distribuição é semelhante a uma distribuição normal, (MOHD RAZALI; BEE WAH, 2011), permitiu rejeitar a hipótese de que a amostra possui distribuição normal. O teste F, usado para avaliar se as variâncias de duas populações são iguais (NGUYEN et al., 2019), também permitiu rejeitar a hipótese nula e concluir que as variâncias não são iguais. Consequentemente, o teste t não é adequado para as comparações pretendidas e aplicou-se o teste de Mann-Whitney, cujos resultados são apresentados a seguir.

H_0: F1 score (controle) ≥ F1 (erros)
H_1: F1 score (controle) < F1 (erros)
Teste de Mann-Whitney: *p-value* > 0,99

H_0: AUC (controle) ≥ AUC (erros)
H_1: AUC (controle) < AUC (erros)
Teste de Mann-Whitney: *p-value* > 0,501

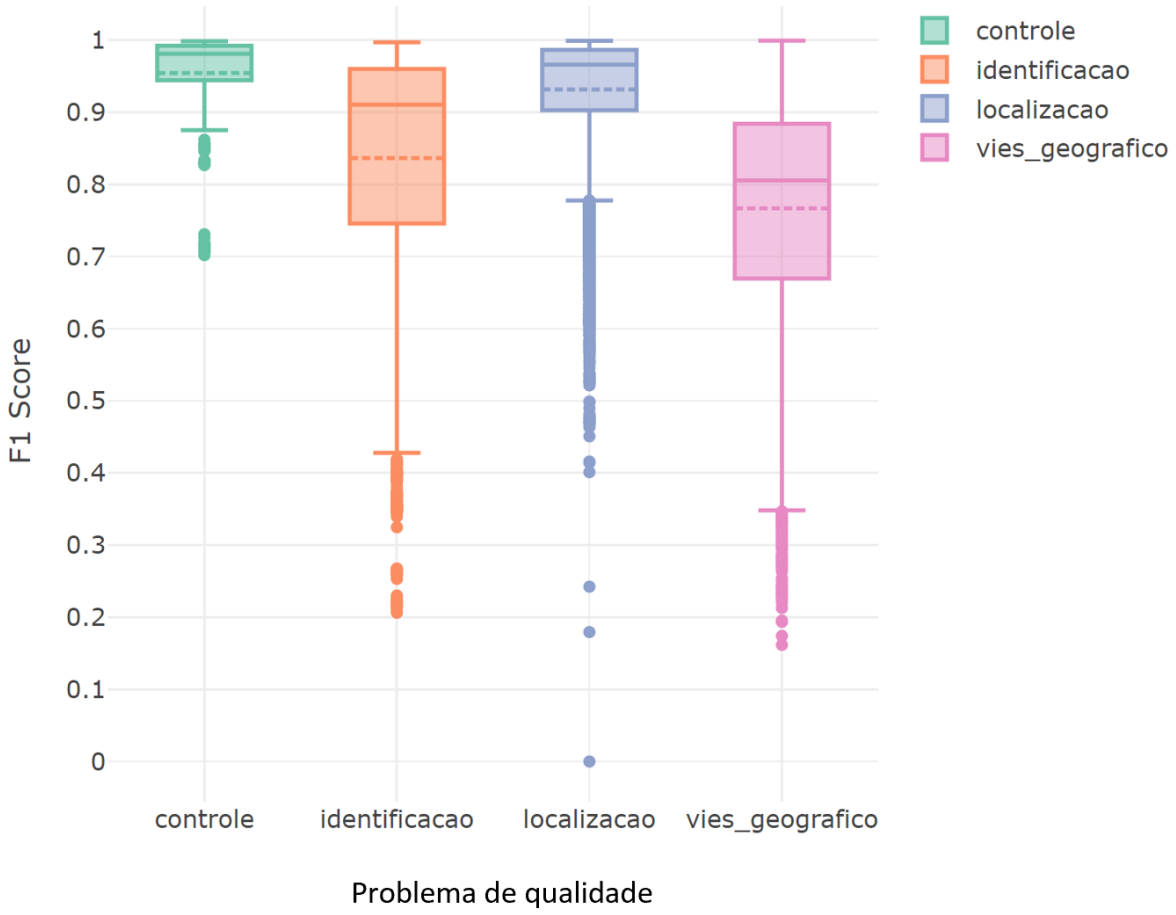
O teste de hipótese indicou que não há evidências suficientes para rejeitar a hipótese nula, ou seja, há indicativo de que os modelos de controle performam melhor

do que os modelos com erro. Os testes para F1 score fornecem um indicativo ainda mais forte. Este era o resultado esperado, provando que os erros nas amostras de treinamento degradam o desempenho dos modelos.

As análises seguintes consideram apenas F1 score porque os resultados e testes realizados para essa métrica podem ser replicados utilizando o AUC, além disso, o F1 score é mais adequado para análise de desempenho de conjuntos de dados desbalanceados, que é o caso das bases de controle que contém muito mais pontos de ausências do que presenças. Em aprendizado de máquina, deve-se estabelecer as previsões mais críticas para o contexto de análise. Nos MDEs, se um modelo indicar um local com ocorrência de uma espécie como sendo ausência, pode enviesar decisões de conservação importantes, tendo um impacto mais crítico do que indicar um local real de ausência como sendo presença. Por isso, a detecção alvo dos MDEs são as ocorrências das espécies, consequentemente, métricas que tratam da assertividade das previsões positivas são as mais adequadas. E o F1 Score consolida a sensibilidade, métrica que avalia a capacidade do modelo de detectar com sucesso resultados classificados como presença, e a precisão, que mede a quantidade de pontos classificados como presença quantos são realmente presença (seção 2.7).

De acordo com o resultado anterior, problemas de QD presentes nos dados afetam MDEs. Assim, avalia-se qual tipo de erro (problema de qualidade), no geral, provoca os piores resultados nos modelos gerados. A Figura 16 apresenta os resultados de F1 para cada um dos tipos de problemas de QD em comparação aos modelos de controle.

Figura 16: *Boxplot* da métrica F1 para cada tipo de problema de qualidade em comparação aos resultados dos modelos de controle



Os modelos de controle geraram a métrica F1 com média de 0,954, os modelos contaminados com erros de identificação tiveram média de 0,836, enquanto erros de localização obtiveram média de 0,931 e os modelos criados com dados contaminados por viés geográfico atingiram média de 0,767. Os resultados mostram que os modelos gerados com dados contaminados por erros de localização, apesar de apresentarem mais outliers, obtiveram resultados menos dispersos comparado aos outros erros. O valor da média é equivalente à 97,6% daquele obtido pelos modelos de controle. Isso indica que, embora seja um erro recorrente na literatura, o impacto do erro de localização nos modelos parece não ser tão intenso. Por outro lado, erros de identificação e viés geográfico tem um impacto maior, obtive-se médias equivalentes à 87,6% e 80,4% da média dos modelos de controle.

Para verificar se os resultados possuem resultados estatisticamente diferentes realizou-se um teste de hipótese para comparar médias de mais de um grupo. A Análise de Variância Simples (ANOVA), que é um teste paramétrico, seria a estratégia

inicial para a realização do teste, no entanto, a natureza dos dados gerados só obedece à condição de independência, falhando nos pressupostos de distribuição normal e homocedacidade (homogeneidade de variâncias). Assim, adotou-se o teste de Kruskal-Wallis, que é uma alternativa não paramétrica ao teste ANOVA quando os pressupostos deste não são atendidos, e que estende o teste de Mann-Whitney para amostras com mais de dois grupos (OSTERTAGOVÁ; OSTERTAG; KOVÁČ, 2014).

Para verificar se existe alguma diferença significativa entre as médias geradas pelos modelos de controle e os modelos gerados pelas 3 condições de erro, realizou-se o teste Kruskal-Wallis com as seguintes hipóteses:

H_0: As amostras têm médias iguais.
H_1: As amostras têm médias que não são iguais.
Teste de Kruskal-Wallis: *p-value* < 2.2e-16

Como o valor de p é menor que o nível de significância 0,05, pode-se concluir que existem diferenças significativas entre os resultados das médias dos modelos dos grupos de erro e de controle. Mas o teste realizado permite concluir somente que há uma diferença significativa entre os grupos de modelos. Para uma comparação entre cada um dos grupos de modelos, realizou-se o teste de Wilcoxon não pareado. Este teste é utilizado para realizar a comparação emparelhada de populações (KELLEY; DONNELLY, 2009). O resultado é apresentado na Figura 17.

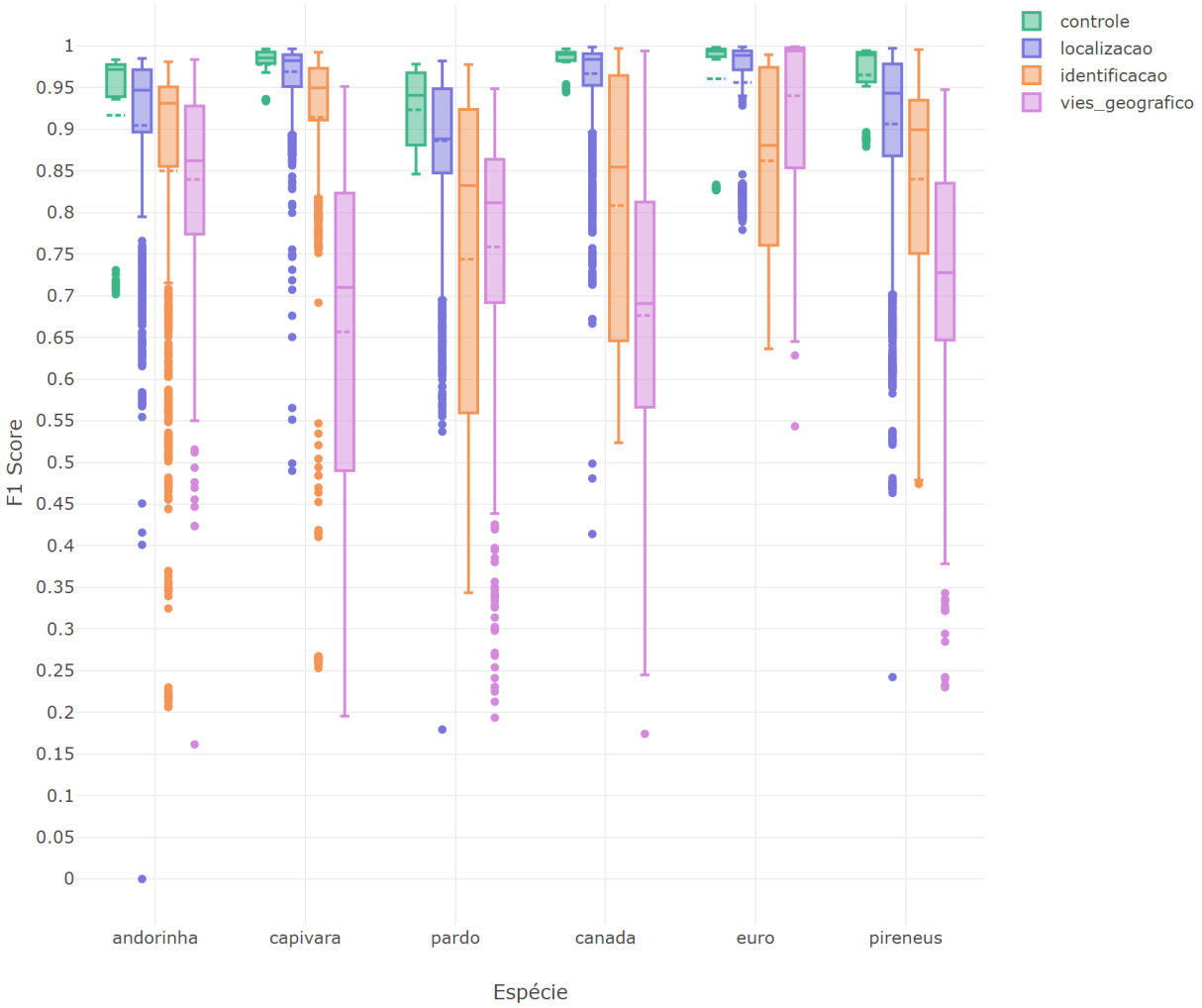
Figura 17: Teste de Wilcoxon emparelhado para F1 de controle e cada tipo de erro

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  df$F1 and df$erro

                controle localizacao identificacao
localizacao     4e-16    -           -
identificacao   <2e-16   <2e-16      -
vies_geografico <2e-16   <2e-16      <2e-16

P value adjustment method: BH
```

A comparação emparelhada indica que entre todos os grupos de modelos há diferenças estatisticamente significativas. Então, embora a diferença entre as médias dos modelos de controle e dos modelos com erro de localização seja pequena, ela é estatisticamente significativa. Estratificando ainda mais os resultados, é possível

qualificar o impacto de cada tipo de erro nas espécies de interesse de acordo com a amplitude de seus nichos ecológicos, como mostrado na Figura 18.

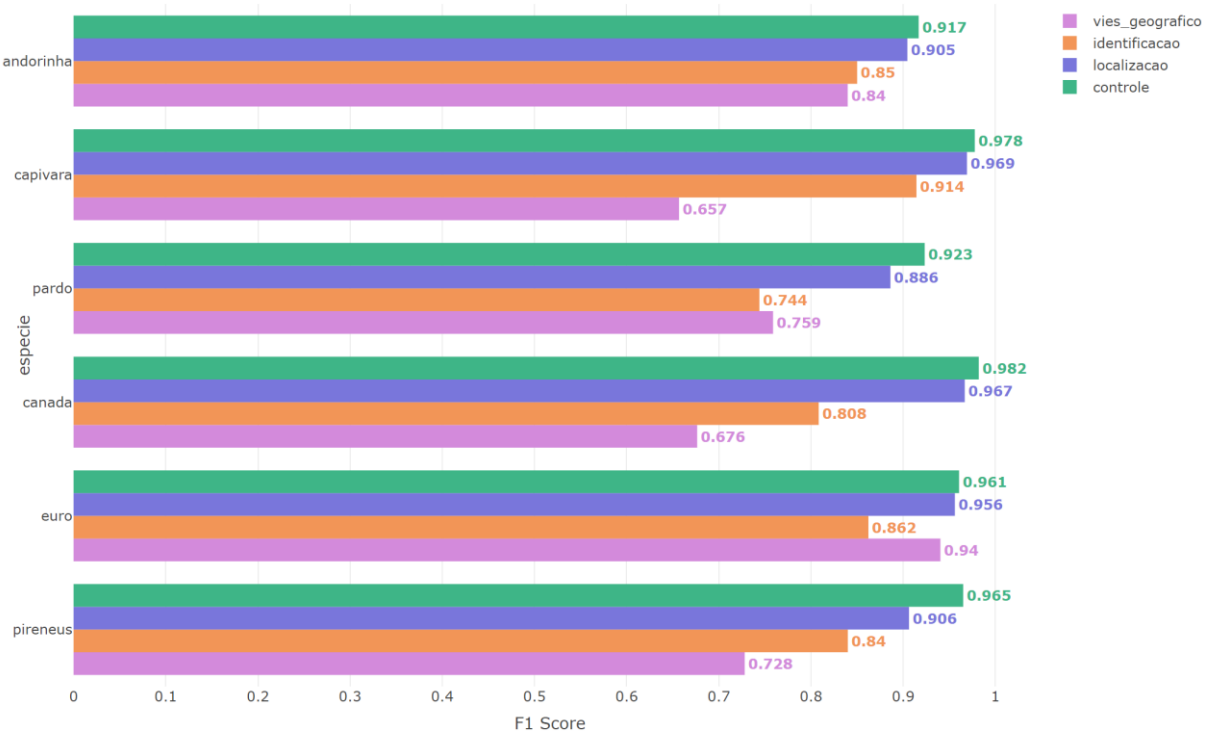
Figura 18: *Boxplot* da métrica F1 estratificado por tipo de erro e por espécie



Na Figura 18 é possível observar o comportamento de cada tipo de erro para cada espécie de interesse. Nas 6 espécies do experimento, os erros de localização são os que geram menor variação de desempenho dos modelos se comparado aos respectivos modelos de controle. Em 4 casos, com as espécies de andorinha, capivara, canada e pireneus, o viés geográfico é o tipo de erro que mais impacta os resultados dos MDEs. Nos casos da espécie pardo e euro, o erro de identificação é o gerou as menores médias para a métrica F1. Essas duas espécies têm em comum o nicho ecológico amplo e são predominantes de regiões frias. Para a espécie andorinha, de nicho amplo e predominante em região tropical, a diferença média entre os dois tipos de erro de maior impacto foi pequena, com média de 0,85 para os erros de identificação e 0,84 para o erro de viés geográfico.

A partir do gráfico da Figura 18 também é possível observar uma diferença dos efeitos dos problemas de QD entre espécies de diferentes amplitudes de nicho e predominantes de uma mesma região. Há um indicativo de que espécies de nicho restrito são mais impactadas por problemas de qualidade nos dados do que espécies de nicho amplo. A Figura 19 mostra os resultados médios da métrica F1 de cada espécie de interesse para cada tipo de erro. Essa Figura evidencia ainda mais a degradação que os erros induzem nos modelos das espécies de nicho restrito.

Figura 19: Média de F1 estratificada por tipo de erro e por espécie



Para confirmar o indicativo visual, realizou-se testes de Mann-Whitney para responder à questão:

- modelos gerados com dados contaminados com problemas de qualidade afetam mais o desempenho de MDEs de espécies de nicho restrito do que os de espécies de nicho amplo?

Os testes foram implementados para cada par de espécie de cada região de predominância (andorinha/capivara, pardo/canada/ euro/pireneus) com as seguintes hipóteses:

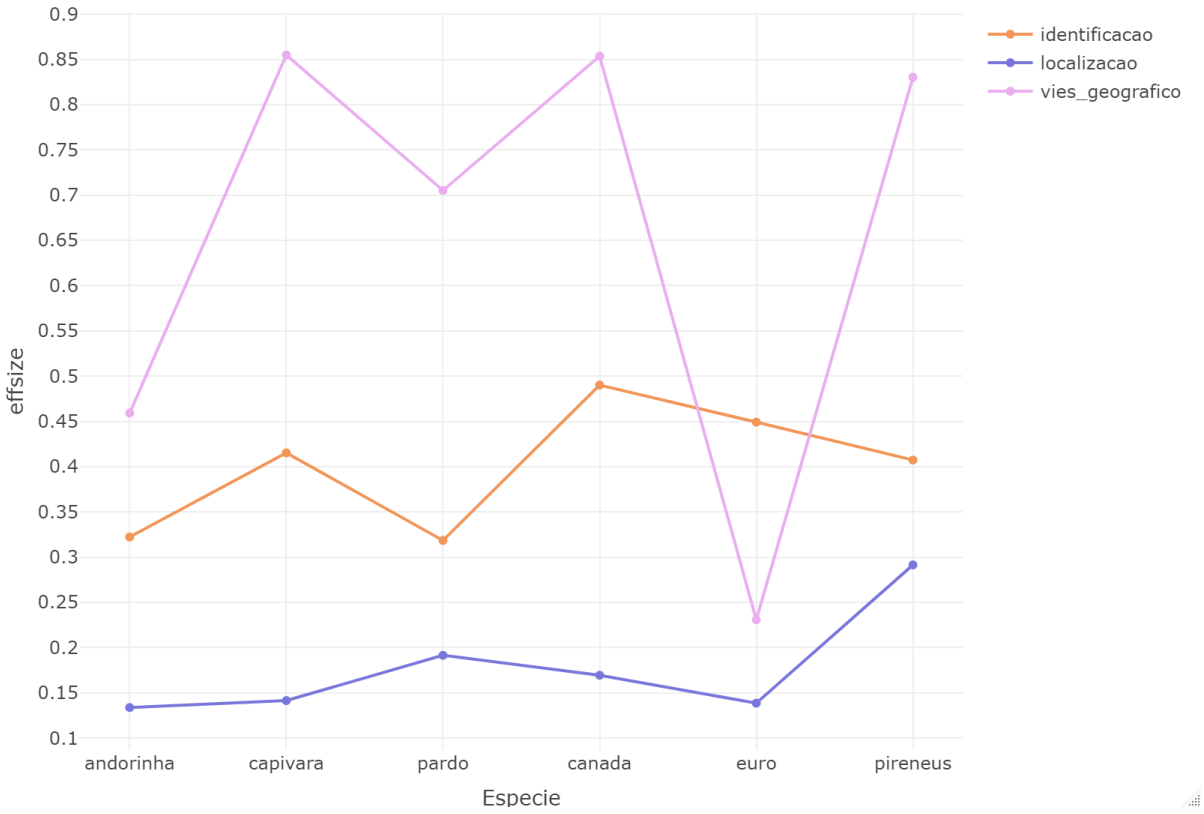
H_0: Média diferença controle-erro Espécie nicho restrito >= Média diferença controle-erro Espécie nicho amplo.
H_1: Média diferença controle-erro Espécie nicho restrito < Média diferença controle-erro Espécie nicho amplo.

Teste de Mann-Whitney (andorinha - capivara): *p-value* = 0,9
Teste de Mann-Whitney (pardo - canada): *p-value* = 0,69
Teste de Mann-Whitney (euro - pireneus): *p-value* > 0,99

O valor-p de todos os testes resultaram em valores superiores ao nível de significância de 0,05, o que permite aceitar a hipótese nula e concluir há indícios estatisticamente significativos de que, em média, a diferença entre o desempenho dos modelos de controle e modelos com problemas de QD é maior para as espécies de nicho ecológico restrito. Esse é um resultado consistente com a resposta ecológica de espécies de nicho restrito, que tem menor tolerância à variabilidade de variáveis ambientais como temperatura e precipitação. Erros de localização podem indicar a ocorrência em um local com padrão climático discrepante do esperado para a ocorrência da espécie. Erro de identificação pode contaminar a amostra com espécies cujo espalhamento sobre o espaço geográfico pode ser o suficiente para ampliar artificialmente o padrão de distribuição da espécie de interesse. Enquanto o viés geográfico e o erro de identificação com espécie contaminante de nicho restrito, impactam os resultados por restringirem a condição de variabilidade significativa das variáveis ambientais dentro das células do espaço geográfico da região de estudo.

A análise a partir do valor-p, no entanto, não se destina a fornecer informações sobre a força real da relação entre os resultados dos modelos de controle e de erros, e não permite determinar o efeito dos erros sobre o controle. Para realizar esse tipo de análise, utilizou-se as medidas de tamanho de efeito que permitem avaliar a força da relação entre os cenários investigados e que, na prática, permitem avaliar a magnitude e a importância do resultado obtido (TOMCZAK; TOMCZAK, 2014). Os resultados das medidas de tamanho de efeito dos resultados dos erros em relação aos controles são apresentados na Figura 20.

Figura 20: Tamanho de efeito dos resultados dos modelos com erros em relação aos modelos de controle



Os resultados de tamanho de efeito da Figura 20 corroboram as análises e testes estatísticos de que os modelos das espécies de nicho ecológico restrito são mais susceptíveis aos problemas de QD. O erro de viés geográfico tem impacto significativamente maior em 5 das 6 espécies de interesse. Para a espécie de interesse pardo, embora a média do resultado da métrica F1 para o erro de identificação tenha sido ligeiramente superior à do erro de viés geográfico, o gráfico de tamanho de efeito dos resultados mostrou que o viés geográfico tem um efeito significativamente maior nos modelos do que o erro de identificação. Para as espécies de nicho ecológico restrito, capivara, canada e pireneus, a degradação nos resultados com o viés geográfico é maior do que suas contrapartes de nicho ecológico amplo. O tamanho de efeito desse tipo de erro nas espécies de nicho restrito foi sempre acima de 0,80, caracterizando uma magnitude de relação forte. Já os erros de identificação possuem impacto de magnitude moderada, com tamanho de efeito entre 0,32 e 0,49, e os erros de localização possuem magnitude fraca, com efeito máximo de 0,29.

Após a análise do impacto de cada tipo de erro nos resultados dos MDEs, avaliou-se também o efeito das intensidades de erros nos resultados. A Figura 21 apresenta os *boxplots* da métrica F1 para as diferentes intensidades de cada tipo de erro.

Figura 21: Métrica F1 estratificada por intensidade de erro

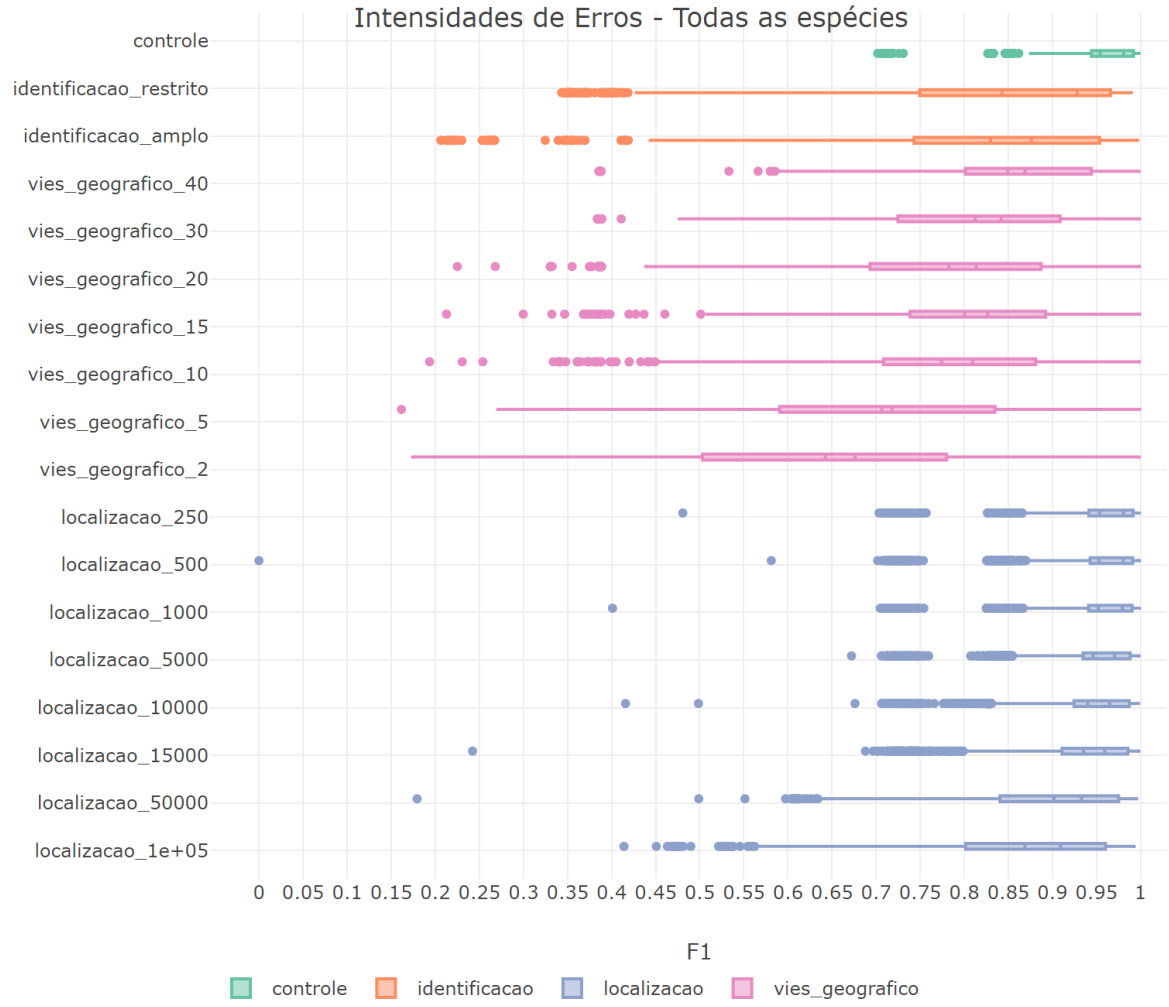
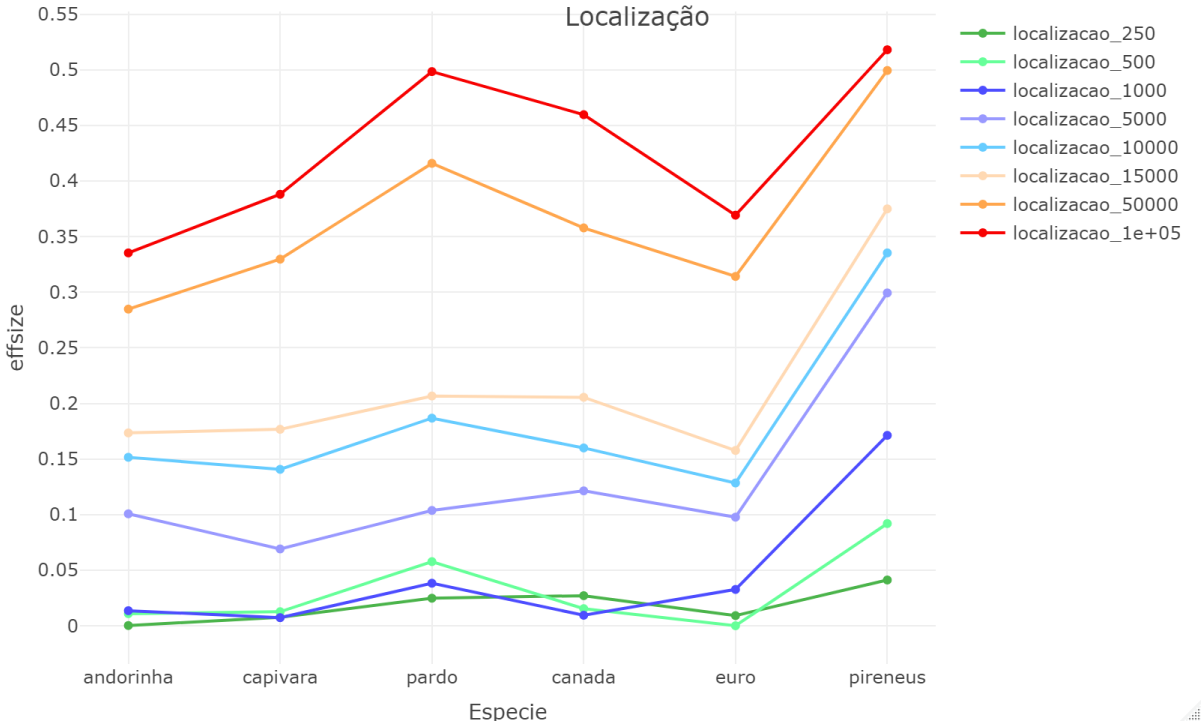


Figura 21 mostra que as duas intensidades de erros de identificação, no geral, geram efeito similar no desempenho dos MDEs, com a contaminação por espécie de nicho mais amplo causando uma degradação um pouco mais expressiva. Já nos modelos com erros de viés geográfico, a degradação é maior quanto maior for a intensidade dele. Para erros de viés geográfico, erro 2 representa 2% do território total possível, e apresenta maior intensidade do que o erro 40, que representa 40% do território, pois constitui uma restrição maior sobre o espaço geográfico. Já os erros de localização de intensidade de 250 metros a 5 quilômetros parecem ter efeito muito

semelhante sobre os resultados, não apresentando grande variação entre sim e nem mesmo em comparação aos resultados dos modelos de controle, a diferença das médias é sutil. O impacto parece ser perceptível somente nas maiores intensidades de erro. Essa análise geral pode camuflar diferenças importantes entre as espécies com relação às suas susceptibilidades aos erros. Para isso, os resultados das medidas de tamanho de efeito das intensidades dos erros de localização em relação aos resultados dos modelos controles são apresentados na Figura 22. Esta Figura permite caracterizar melhor os efeitos das intensidades dos erros de localização sobre cada espécie de interesse.

Figura 22: Tamanho de efeito por intensidade dos erros de localização estratificado por espécie

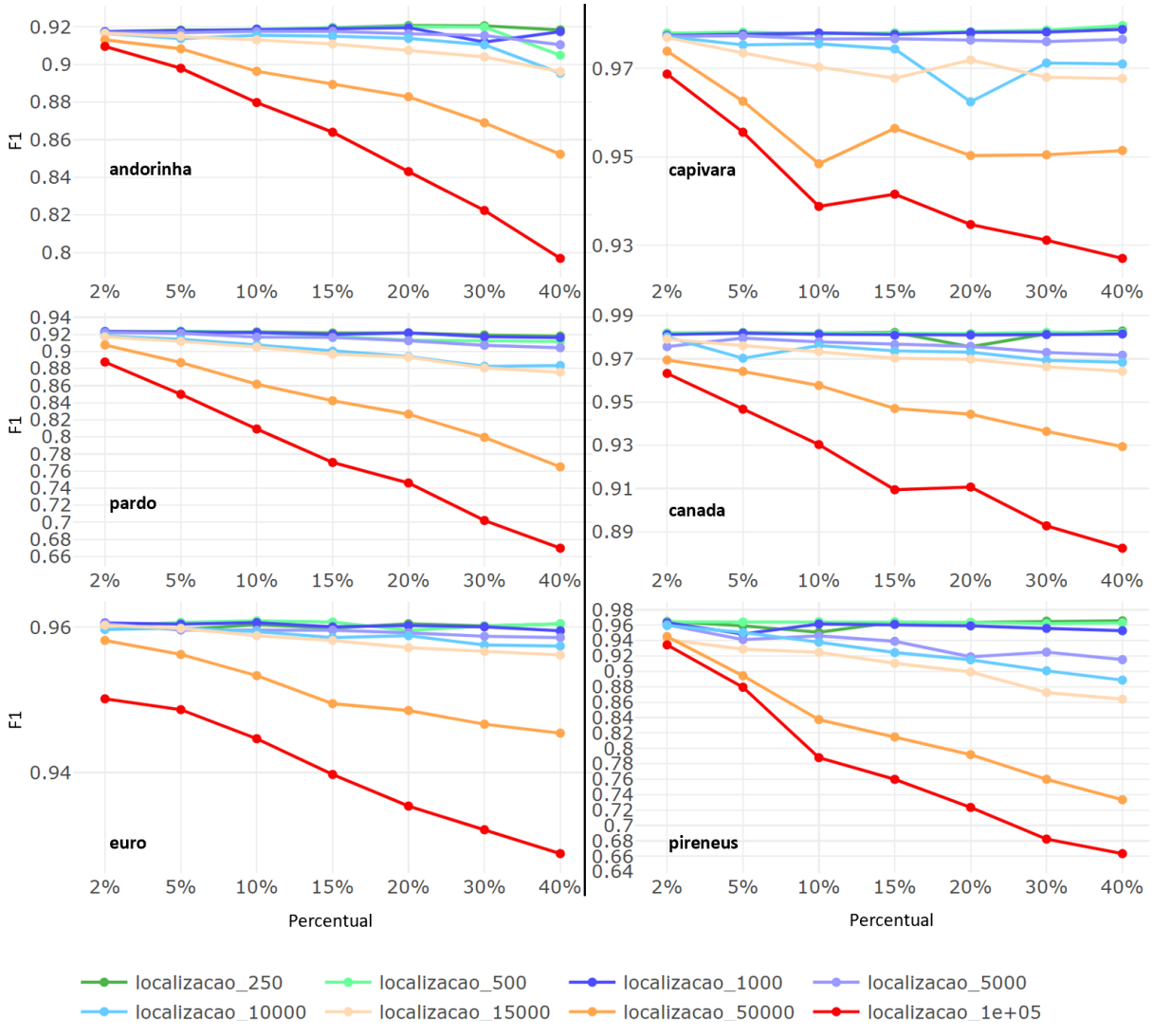


Na Figura 22 verifica-se que a partir de 5 quilômetros, os efeitos estão correlacionados às intensidades do erro, ou seja, aumentando o erro há um aumento também no efeito, ainda que esse aumento não seja linear. Comparando à amplitude de nicho das espécies de interesse, não há um padrão discernível, não sendo possível tirar uma conclusão direta dos dados. Entre andorinha (nicho amplo) e capivara (nicho restrito), os modelos da espécie de nicho restrito são mais afetados por erros de maior intensidade (50 quilômetros e 100 quilômetros). Para erros de média (5 quilômetros, 10 quilômetros, 15 quilômetros) e baixa intensidade (250 metros, 500 metros, 1

quilômetro), não há diferença clara entre os nichos. Para as espécies pardo (nicho amplo) e canada (nicho restrito), o efeito da intensidade do erro de localização tende a ser menor para a espécie de nicho restrito. Já para as espécies de interesse do continente europeu, a espécie de nicho restrito pireneus é muito mais afetada pelos erros de localização do que sua contraparte de nicho ecológico amplo, euro. Enquanto para as outras espécies não há diferenciação clara do efeito da intensidade de erros de 250 metros a 1000 metros nos resultados dos modelos, para a espécie pireneus, essa diferenciação ocorre desde o menor gradiente de erro até o maior.

A Figura 23 apresenta a abstração das intensidades dos erros em relação ao percentual do erro presente nas amostras utilizadas para criar os MDEs e a degradação para a métrica F1.

Figura 23: Degradação da métrica F1 estratificado pela intensidade e proporção de presença de erro de localização



Assim como os resultados do tamanho de efeito da Figura 22, não é possível identificar variação significativa na degradação do desempenho médio dos modelos para erros de baixa e média intensidade em função da variação percentual de presença do erro nas amostras. Para verificar a força da diferenciação dessas intensidades de erros, realizou-se o teste de Wilcoxon para calcular comparações emparelhada entre grupos, com as seguintes hipóteses:

H_0: Não há diferença nas médias H_1: Há diferença nas médias

Os resultados dos testes são apresentados na Tabela 08. As células destacadas são as que tiveram valor-p superior ao nível de significância de 0.05.

Tabela 8: Valor-p para os testes de Wilcoxon de erros de localização emparelhados

Espécie	Erro	localizacao_500	localizacao_1000	localizacao_5000	localizacao_10000	localizacao_15000	localizacao_50000	localizacao_1e+05
andorinha	localizacao_500	0.617						
	localizacao_1000	0.5353	0.8847					
	localizacao_5000	0	0	0				
	localizacao_10000	0	0	0	0.0002			
	localizacao_15000	0	0	0	0	0.0067		
	localizacao_50000	0	0	0	0	0	0	
	localizacao_1e+05	0	0	0	0	0	0	0
capivara	localizacao_500	0.7747						
	localizacao_1000	0.4958	0.3291					
	localizacao_5000	0.0002	0.0001	0.0012				
	localizacao_10000	0	0	0	0			
	localizacao_15000	0	0	0	0	0.0019		
	localizacao_50000	0	0	0	0	0	0	
	localizacao_1e+05	0	0	0	0	0	0	0
pardo	localizacao_500	0.1566						
	localizacao_1000	0.552	0.3544					
	localizacao_5000	0.0001	0.0117	0.0004				
	localizacao_10000	0	0	0	0			

	localizacao_15000	0	0	0	0	0.1573		
	localizacao 50000	0	0	0	0	0	0	
	localizacao_1e+05	0	0	0	0	0	0	0

(continua)

(continuação)

Espécie	Erro	localizacao_500	localizacao_1000	localizacao_5000	localizacao_10000	localizacao_15000	localizacao_50000	localizacao_1e+05
canada	localizacao_500	0.5725						
	localizacao_1000	0.3674	0.6767					
	localizacao_5000	0	0	0				
	localizacao_10000	0	0	0	0.012			
	localizacao_15000	0	0	0	0	0.0018		
	localizacao_50000	0	0	0	0	0	0	
	localizacao_1e+05	0	0	0	0	0	0	0
euro	localizacao_500	0.6524						
	localizacao_1000	0.2946	0.1324					
	localizacao_5000	0	0	0.0003				
	localizacao_10000	0	0	0	0.0699			
	localizacao_15000	0	0	0	0.0001	0.0265		
	localizacao_50000	0	0	0	0	0	0	
	localizacao_1e+05	0	0	0	0	0	0	0
pireneus	localizacao_500	0.0048						
	localizacao_1000	0	0					
	localizacao_5000	0	0	0				
	localizacao_10000	0	0	0	0			
	localizacao_15000	0	0	0	0	0		
	localizacao_50000	0	0	0	0	0	0	
	localizacao_1e+05	0	0	0	0	0	0	0

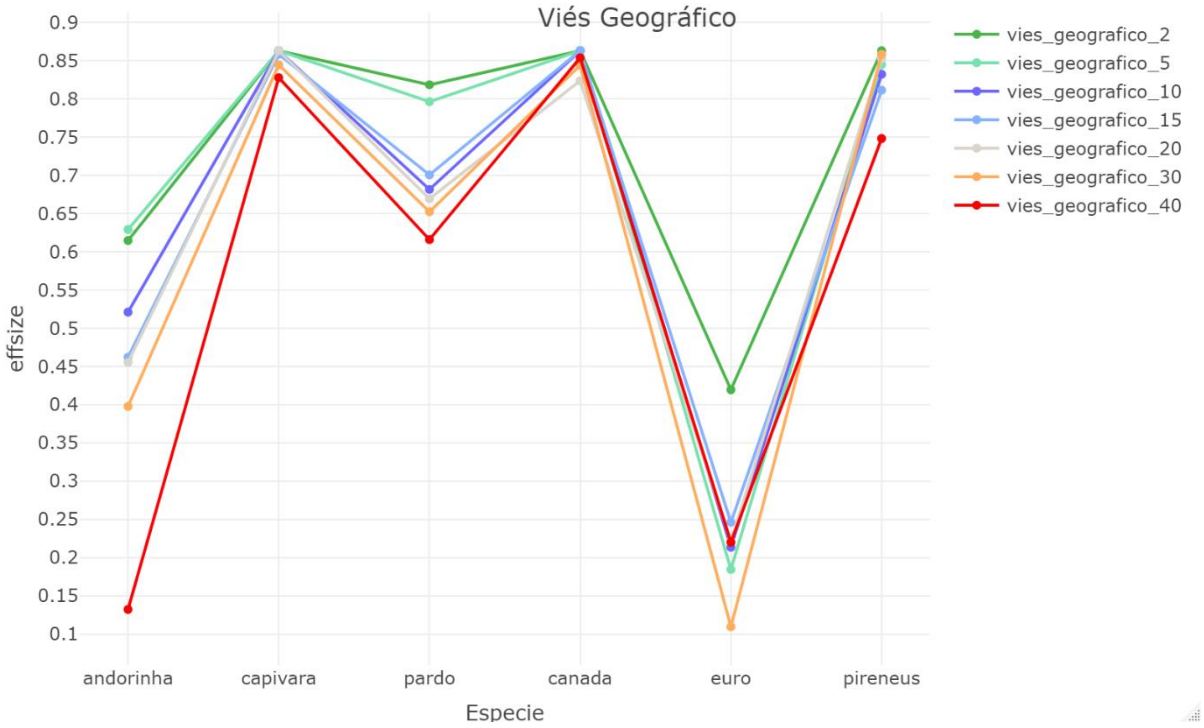
Os resultados do valor-p apresentados na Tabela 08 nas comparações das intensidades de erro de 250 metros, 500 metros e 1 quilômetro, para 5 das espécies de interesse, são maiores que o nível de significância 0,05, o que implica na aceitação

da hipótese nula, de que não há diferenças estatisticamente significativas nas médias das baixas intensidades para todas as espécies, com exceção de pireneus.

Os erros com alta intensidade, conforme Figura 23, tendem a aumentar o impacto sobre o resultado à medida que o percentual da sua presença na amostra aumenta. Nesse caso, somente a espécie capivara quebrou o padrão nos erros de alta intensidade entre 10% e 15%. A variação em pontos percentuais (pp), do cenário menos crítico com erros de 250 metros na proporção de 2% para o mais crítico na proporção de 40% para o erro de 100 quilômetros, a espécie andorinha teve variação de 12,05 pp, capivara 5,08 pp, pardo 25,36 pp, canada 9,95 pp, euro 3,15 pp e pireneus 30,13 pp. Para os erros de localização, a partir dos resultados não é possível concluir se a amplitude do nicho ecológico é fator predominante para sensibilidade ao efeito deste tipo de erro. Para esse tipo de erro, características biológicas e o perfil de distribuição geográfica de cada espécie podem indicar a sua susceptibilidade a diferentes gradientes dos erros. Por exemplo, para os pireneus, espécie com maior variação em pp, o aumento da intensidade de erro impacta diretamente os resultados. A característica de distribuição dessa espécie pode explicar esse resultado. Embora neste trabalho só sejam apresentadas duas estratificações de nicho ecológico, amplo e restrito, os pireneus são uma espécie de nicho muito restrito, ocorrendo apenas em regiões montanhosas com características ambientais específicas. Até por isso, a elevação é uma variável importante para explicar o padrão de distribuição desta espécie. Assim, para esse perfil de nicho, os erros de localização podem colocar a espécie em local não tão propício, e quanto mais distante do ponto real, maior o impacto. Já a espécie euro, que é o Lince Euroasiático, caracteriza-se por ser um dos predadores com maior nicho ecológico do mundo (SUÁREZ, 2022), sendo assim, mesmo um erro de 100 quilômetros nos pontos de ocorrência pode fazer com que esses pontos caiam sobre um espaço geográfico que ainda mantém condições propícias para a ocorrência dessa espécie.

Para os erros de viés geográfico, os resultados das medidas de tamanho de efeito em relação aos resultados dos modelos controles são apresentados na Figura 24.

Figura 24: Tamanho de efeito por intensidade dos erros de viés geográfico estratificado por espécie



A Figura 24 detalha mais os resultados da Figura 21 para as diferentes intensidades de erros de viés geográfico. Ao contrário do que ocorre com os erros de localização, erros de viés geográfico apresentam impacto mais consistente na análise por nicho ecológico das espécies. Espécies de nicho restrito aparentam sofrer fortemente com a presença desse tipo de erro, e, independente da intensidade do erro, o impacto é de magnitude forte, mas sem grande variabilidade. Já as espécies de nicho amplo, andorinha, pardo e euro, há uma variação mais notável na magnitude da maior intensidade para a menor. O teste de Wilcoxon foi empregado novamente para calcular comparações emparelhada entre as intensidades do erro. A hipótese nula declara que as médias são iguais. Os resultados do valor-p são apresentados na Tabela 09. As células destacadas são as que tiveram valor-p superior ao nível de significância de 0.05.

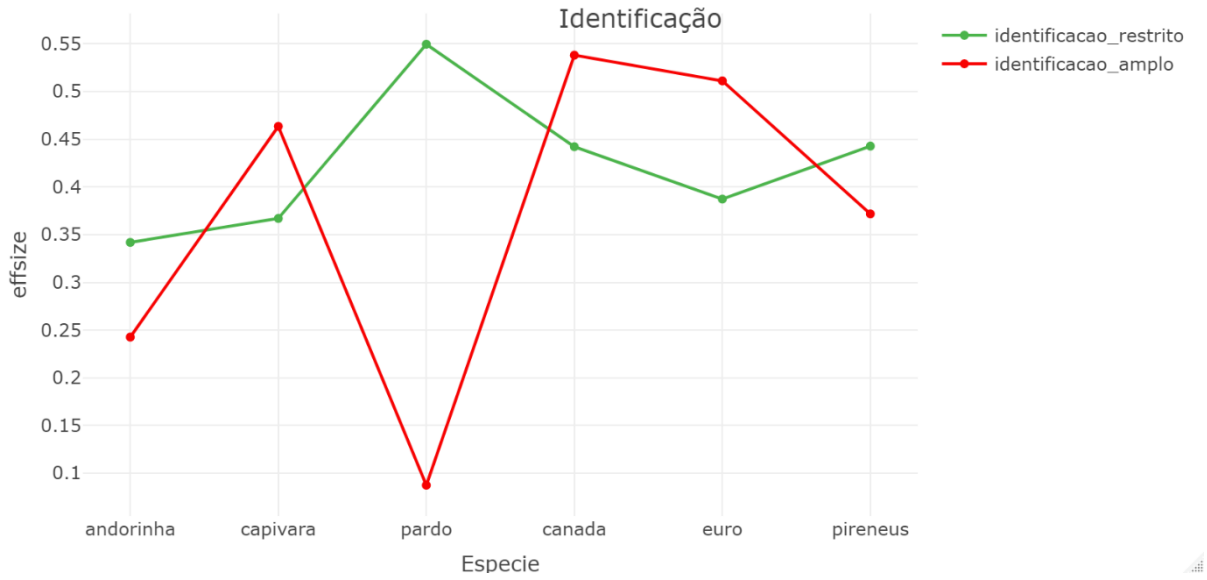
Tabela 9: Valor-p para os testes de Wilcoxon de erros de viés geográfico emparelhados

Especie	Erro	vies_geografico_40	vies_geografico_30	vies_geografico_20	vies_geografico_15	vies_geografico_10	vies_geografico_5
andorinha	vies_geografico_30	0					
	vies_geografico_20	0	0.0075				
	vies_geografico_15	0	0.0107	0.8392			
	vies_geografico_10	0	0	0	0		
	vies_geografico_5	0	0	0	0	0	
	vies_geografico_2	0	0	0	0	0	0.3647
capivara	vies_geografico_30	0.5203					
	vies_geografico_20	0	0				
	vies_geografico_15	0	0	0.955			
	vies_geografico_10	0	0	0.6938	0.5355		
	vies_geografico_5	0	0	0.0005	0.0001	0.0005	
	vies_geografico_2	0	0	0.0019	0.0001	0.0005	0.8351
pardo	vies_geografico_30	0.542					
	vies_geografico_20	0.557	0.9491				
	vies_geografico_15	0.0629	0.21	0.2439			
	vies_geografico_10	0.0213	0.0729	0.1703	0.6163		
	vies_geografico_5	0	0	0	0	0.0025	
	vies_geografico_2	0	0	0	0	0.0008	0.76
canada	vies_geografico_30	0.0817					
	vies_geografico_20	0	0.0023				
	vies_geografico_15	0.0001	0.0678	0.1255			
	vies_geografico_10	0.0001	0.0135	0.9072	0.2691		
	vies_geografico_5	0	0	0.0986	0.0013	0.1255	
	vies_geografico_2	0	0	0	0	0	0
euro	vies_geografico_30	0.7666					
	vies_geografico_20	0.997	0.7666				
	vies_geografico_15	0.7666	0.476	0.7666			
	vies_geografico_10	0.997	0.7666	0.997	0.7666		
	vies_geografico_5	0.997	0.7666	0.997	0.7666	0.997	
	vies_geografico_2	0	0	0	0	0	0
pireneus	vies_geografico_30	0					
	vies_geografico_20	0.0153	0.0013				
	vies_geografico_15	0.7275	0	0			
	vies_geografico_10	0.6471	0	0	0.2299		
	vies_geografico_5	0	0.2299	0.0004	0	0	
	vies_geografico_2	0	0	0	0	0	0

O teste de Wilcoxon indicou que os modelos da espécie euro, que possui nicho ecológico amplo, sofrem impacto moderado para o erro de mais alta intensidade. Para todos os outros gradientes, os modelos desta espécie não sofrem efeito variável em função da intensidade do erro, corroborado pelos valores-p obtidos, que permitem aceitar a hipótese de que as médias são iguais. Na caracterização por espécies, os modelos da espécie andorinha são os mais consistentes no que tange a variação de degradação em função do aumento da intensidade do erro. Para as outras, a variação só é mais clara para casos de intensidades de erros mais fortes com 2%.

Para os dois gradientes de erros de identificação, os resultados das medidas de tamanho de efeito para cada espécie de interesse são apresentados na Figura 25.

Figura 25: Tamanho de efeito por intensidade dos erros de identificação estratificado por espécie



A separação dos gradientes de erro é importante para entender melhor os resultados da Figura 21. Esta indica que, no geral, os erros de identificação com espécie contaminante de nicho amplo e restrito apresentam resultados similares. No entanto, em uma análise mais detalhada, apresentada na Figura 25, verifica-se que há distinções claras entre as intensidades de erros de identificação nas diferentes espécies de interesse. A magnitude da relação do erro de identificação de intensidade restrita é sempre moderada, enquanto a intensidade ampla varia entre fraca e moderada. Os modelos da espécie pardo apresentam diferença expressiva entre as duas intensidades. Mas não é possível identificar nenhum padrão de variação ou

susceptibilidade aos erros em função da amplitude de nicho ou região predominantes de ocorrência.

Embora não tenha sido possível destacar padrões de susceptibilidade a cada tipo de erro de identificação, foi possível quantificar a magnitude possível desse impacto. Como não é o objetivo deste trabalho realizar uma análise do impacto do erro sobre a projeção geográfica do padrão de distribuição das espécies, destaca-se que há uma oportunidade em aberto de se qualificar e quantificar se os erros de identificação provocam uma distorção do padrão de distribuição na direção da espécie contaminante, causando uma contração ou expansão da distribuição esperada da espécie de interesse.

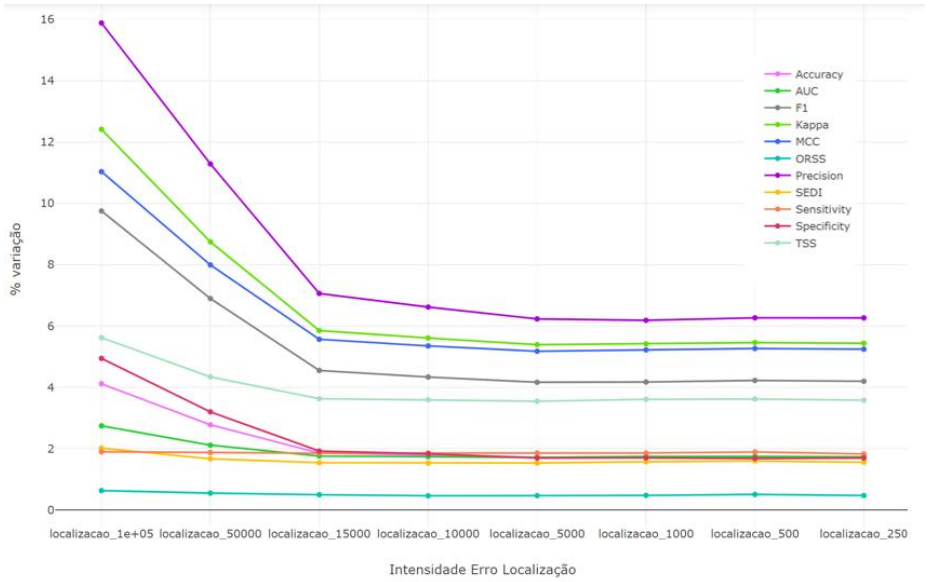
Avaliou-se também a resposta das métricas de validação ao aumento das intensidades dos erros de localização, identificação e viés geográfico. O intuito foi caracterizar as métricas para identificar aquelas mais sensíveis à contaminação dos dados utilizados na construção dos modelos. Métricas com baixa variabilidade comprometem a análise e decisão sobre a adequabilidade do MDE, pois mascaram deficiências dos modelos e indicam uma capacidade de generalização incorreta que pode gerar respostas enviesadas. Por outro lado, as métricas que se deterioram mais em função da intensidade de erro, podem contribuir para que se possa rever o procedimento de modelagem, desde a obtenção, até o tratamento, pré-processamento dos dados e seleção de algoritmos.

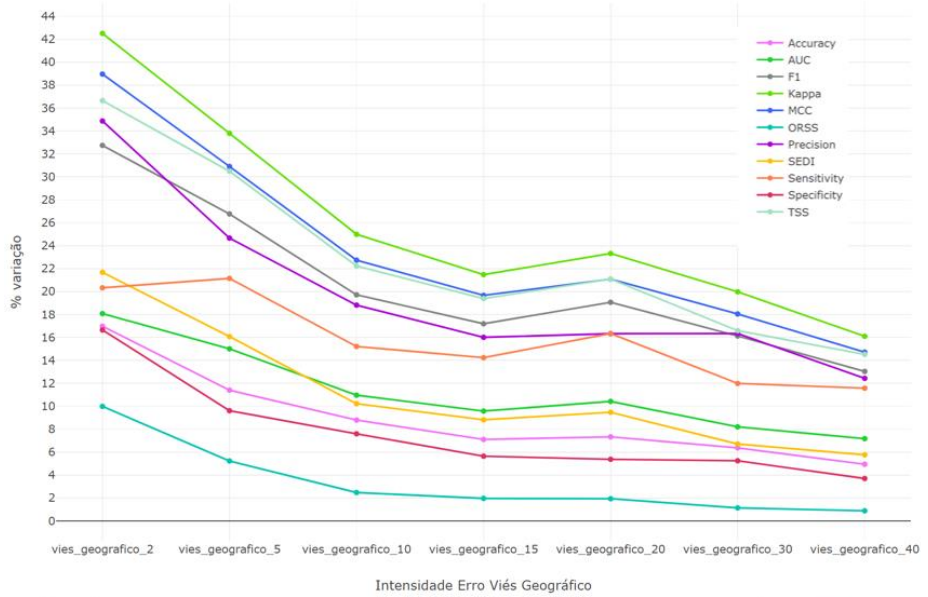
Para realizar essa análise, computou-se a taxa de variação do valor da métrica de cada modelo, de cada amostra e de cada intensidade de erro, em relação ao resultado médio da respectiva métrica para os modelos de controle. Para cada métrica, foi computado do valor médio obtido pelos modelos de controle. Em seguida, calculou-se a taxa de variação (%) absoluta de acordo com a seguinte equação:

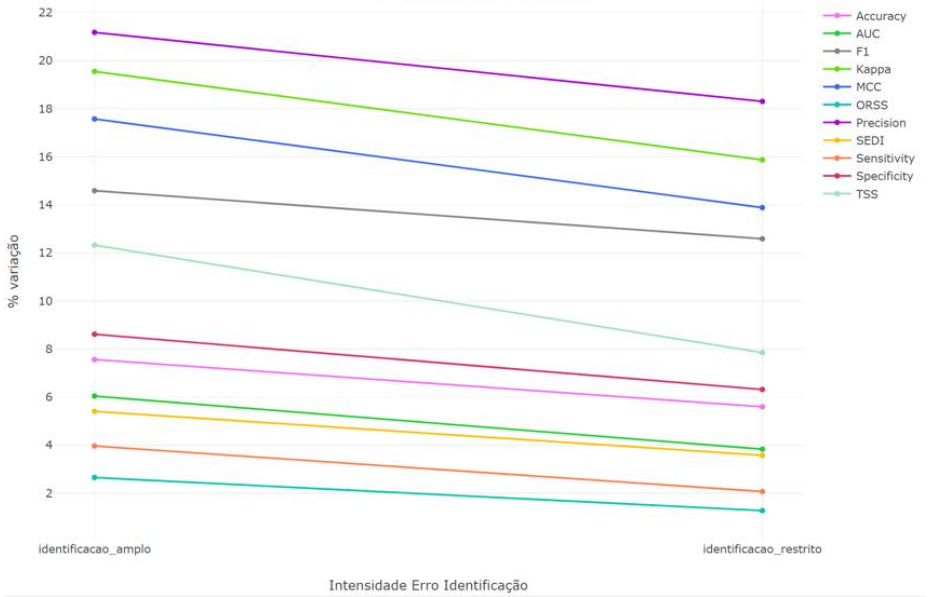
$$\Delta metrica_x = \frac{abs(media\_metrica_{x_{modelos\_controle}} - metrica_{x_{modelo\_erro}})}{media\_metrica_{x_{modelos\_controle}}} \times 100 \ (\%) \quad (1)$$

Os resultados das taxas de variação das métricas são apresentados na Figura 26.

Figura 26: Taxa de variação das métricas de validação em função da intensidade de erro







Para os erros de localização, as 5 métricas que apresentaram maior variabilidade entre as intensidades de erros foram Precisão, Kappa, MCC, F1 e TSS. Os gradientes dos erros de viés geográfico são mais percebidos pelas métricas Kappa, MCC, TSS, F1 e Precisão. Os dois tipos de erros de identificação impactam mais as métricas Precisão, Kappa, MCC, F1 e TSS. Portanto, para os três tipos de erros simulados, as métricas que apresentam maior variabilidade percentual em comparação às métricas dos modelos de controle são as mesmas.

Na literatura, AUC é a métrica mais utilizada para avaliar a capacidade preditiva de modelos de aprendizado de máquina. Em MDEs, essa métrica também está presente na maioria dos artigos que tratam da avaliação comparativa (i) dos resultados dos modelos (AGUIRRE-GUTIÉRREZ et al., 2013; BUCKLAND; SMITH; THOMAS, 2022; BUCKLIN et al., 2015), (ii) dos estudos sobre vieses e ruídos (INMAN et al., 2021; SOULTAN; SAFI, 2017), (iii) da complexidade de modelos (BRUN et al., 2020; SYFERT; SMITH; COOMES, 2013b) e (iv) incertezas (BRODIE et al., 2022; QAZI; SAQIB; ZAMAN-UL-HAQ, 2022). No entanto, no que tange à possibilidade de detectar problema de QD a partir da degradação da métrica, AUC acaba superestimando os resultados dos modelos. Para erros de localização, AUC variou somente 1.015 pp, entre a maior e menor intensidade de erro. Como comparação, a métrica que mais variou nesse tipo de erro foi a de Precisão 9.6 pp entre os gradientes extremos. Mas é importante destacar que assim como a caracterização dos erros de localização apontou, não há diferenças estatisticamente significativas entre os erros de baixa intensidade. Nas métricas de validação, a variação entre elas foi muito sutil para erros de intensidade fraca e moderada. Os erros de viés geográfico são o que mais causam efeito sobre as métricas de validação. Adicionalmente, dos três tipos de erros analisados, o do viés geográfico é o que provoca maior variação entre as intensidades. No entanto, mais uma vez a métrica AUC não varia na mesma intensidade do que as métricas de Kappa, MCC, TSS, F1 e Precisão. Para os erros de identificação, a variação das métricas para dois gradientes de erros seguiu um padrão similar, com os erros de identificação com espécie contaminantes de nicho ecológico amplo tendo maior impacto nos resultados. É interessante destacar que as duas métricas propostas por (WUNDERLICH et al., 2019), ORSS e SEDI, como substitutos à métrica TSS apresentaram resultados pouco consistentes. Segundo os autores, a ORSS seria uma alternativa melhor que a TSS para modelos em que os

dados de ausência estivessem disponíveis, como é o caso da simulação realizada. ORSS, apresentou resultados com os menores percentuais de variação em relação aos modelos de controle em todos os cenários avaliados, apresentado a menor sensibilidade aos problemas de qualidade entre todas as métricas comparadas.

Para fazer uma análise da capacidade de cada métrica distinguir os diferentes gradientes de erros, realizou-se o teste de Wilcoxon para calcular comparações emparelhada da métrica em relação às intensidades do erro. Os resultados são apresentados na Tabela 10 e codificados de modo que a cor verde indica um valor-p menor que a significância estatística de 0,05, e vermelho representa o valor-p maior que 0,05. O teste estatístico permite a comparação sob a hipótese nula, de que as médias das duas amostras são iguais. O teste compara grupos de dois em dois, por exemplo, para a métrica de acurácia, compara o erro de localização de 100 quilômetros com o de 50 quilômetros, 15 quilômetros, 10 quilômetros até o de 1 quilômetro. Depois inicia-se pelo erro de 50 quilômetros compara com o de 15 quilômetros, depois de 10 quilômetros e assim sucessivamente até comparar todas as intensidades dos erros de localização entre si. As comparações são feitas somente entre os tipos de erro, localização com localização, identificação com identificação e viés geográfico com viés geográfico. Uma célula é preenchida com verde se, para uma coluna, a métrica da respectiva intensidade de erro não é estatisticamente igual a mesma métrica de qualquer outra intensidade. Uma célula é preenchida com vermelho se, para uma coluna, a métrica da respectiva intensidade de erro é estatisticamente igual a mesma métrica de qualquer outra intensidade, indicando que não ouve variação significativa.

Tabela 10: Teste de Wilcoxon emparelhado por métrica e tipo de erro

Métricas / Intensidade de erro	Acurácia	AUC	F1	Kappa	MCC	ORSS	Precisão	SEDI	Sensibilidade	Especificidade	TSS
Localização 100KM											
Localização 50KM											
Localização 15KM											
Localização 10KM											
Localização 5KM											
Localização 1KM											
Viés Geográfico 2%											
Viés Geográfico 5%											
Viés Geográfico 10%											
Viés Geográfico 15%											
Viés Geográfico 20%											
Viés Geográfico 30%											
Viés Geográfico 40%											
Identificação Amplo											
Identificação Restrito											

A Tabela 10 mostra que para erros de localização de baixa e média intensidade, nenhuma métrica é capaz de distinguir entre os gradientes de erros. Para erros de localização de alta intensidade, 8 das 11 métricas distinguem entre essas intensidades, ou seja, a média da amostra de uma métrica para a intensidade de erro de localização de 100 quilômetros é estatisticamente diferente da média da amostra para a intensidade de erro de localização de 50 quilômetros. Portanto, os erros de localização são difíceis de serem percebidos a partir da avaliação da degradação da

métrica de validação. Isso só é possível para grandes intensidades de erros. Por outro lado, as duas intensidades de erros de identificação são distinguíveis pelos resultados de qualquer uma das métricas. Os erros de vieses geográficos, que são o tipo de erro que causa maior impacto nos modelos, podem ser distinguidos por quase todas as métricas para gradientes extremos, como os de cobertura de 2% e 5% do território. Os gradientes de erros intermediários, de 10%, 15%, 20% e 30%, destacam a alta sensibilidade das métricas Kappa, MCC e TSS para a variação da intensidade do erro.

Nas análises seguinte, os algoritmos utilizados na construção dos MDEs são avaliados para detectar quais deles mantém a capacidade de generalização, mesmo na presença de erros de qualidade na amostra de dados de treinamento. Essa avaliação consiste em identificar os algoritmos que são capazes de gerar modelos que obtém métricas consistentes em relação às métricas dos modelos de controle. Na etapa de caracterização das métricas de validação o objetivo foi identificar aquelas que mais variavam em função da intensidade de erro. Na avaliação dos algoritmos, o objetivo é determinar aqueles que minimizam a degradação dos resultados mesmo na presença dos problemas de QD. Para isso, as 5 métricas mais susceptíveis aos erros, Kappa, MCC, TSS, F1 e Precisão, foram consolidadas em uma única para que a sensibilidade aos problemas de QD fosse mantida, e a capacidade de generalização do algoritmo pudesse ser investigada. A taxa de variação foi utilizada para plotar os resultados da Figura 27 com as espécies de nicho ecológico amplo e da Figura 28 com as espécies de nicho ecológico restrito. À esquerda nessas figuras estão registrados às taxas de variação da métrica composta dos modelos gerados por cada algoritmo em função das intensidades dos erros. À direita está a agregação das taxas de variação dos algoritmos em função do tipo de erro.

Figura 27: Taxa de variação da métrica composta por algoritmo em função da intensidade de erro – espécies de nicho ecológico amplo

Figura 28: Taxa de variação da métrica composta por algoritmo em função da intensidade de erro – espécies de nicho ecológico restrito

Os resultados mostram que dos 5 algoritmos utilizados, o GLM é o que apresenta pior desempenho no geral. Em erros do tipo identificação e localização, o GLM gera modelos que apresentam um desempenho significativamente pior que os modelos gerados por outros algoritmos para 5 das 6 espécies simuladas. Somente para a espécie pardo, o algoritmo Maxent tem pior desempenho que o GLM. Para erros do tipo identificação, o Maxent gera o segundo pior resultado, perdendo somente para o GLM, o que é surpreendente visto que o Maxent é tratado como referência quando se trata de algoritmos para MDEs. Para erros de identificação, o algoritmo

XGBoost se destaca por ter gerado os melhores resultados para todas as espécies de nicho ecológico restrito e duas das espécies de interesse de nicho amplo. Xgboost, RF e Redes Neurais apresentam os melhores desempenhos para cenários com erros de identificação. Com exceção da espécie pardo, os algoritmos Xgboost, RF e redes neurais geram modelos com variação dos resultados para o erro de identificação com espécies contaminante restrita abaixo dos 7%. Para verificar se esse resultado é significativo, realizou-se um teste de hipótese de Mann-Whitney para verificar se os resultados obtidos por esses três modelos têm média igual às métricas computadas para os modelos de controle. Formulou-se as seguintes hipóteses:

H_0: Média métricas modelos de controle = Média métricas modelos de erro	Teste de Mann-Whitney (erro identificação restrito): *p-value* = 0,9
H_1: Média métricas modelos de controle ≠ Média métricas modelos de erro	Teste de Mann-Whitney (erro identificação amplo): *p-value* = 2e-07

Os resultados dos testes de hipótese permitiram concluir que, para o erro de identificação com espécies contaminantes de nicho restrito, com o valor-p superior ao nível de significância de 0,05, pode-se aceitar a hipótese de que as médias das métricas consolidadas dos modelos gerados com dados contaminados por problemas de qualidade são estatisticamente iguais às médias obtidas pelos modelos de controle. Por outro lado, para o erro de identificação com espécies contaminantes de nicho amplo, o valor-p é inferior ao nível de significância, o que resulta na rejeição da hipótese nula, permitindo concluir que nesse caso as médias das métricas são diferentes.

Para os erros de localização, o Maxent consegue resultados melhores, sendo o algoritmo que gera os modelos com melhores resultados para espécies de nicho amplo andorinha e euro, e as de nicho restrito capivara e canada. As redes neurais não têm sido muito aplicadas em MDEs, mas no cenário de simulação, com abundância de dados para treinamento, o algoritmo gerou modelos que apresentaram resultado consistentes, sendo bastante robusto para erros de identificação e localização. O RF, algoritmo tradicional e reconhecido pela capacidade preditiva e robustez, nos resultados obtidos está quase sempre bem próximo do algoritmo que gerou os modelos com menor variabilidade. Para erros de identificação de espécies de nicho restrito, o XGBoost apresentou o melhor desempenho nos dois gradientes

de erros, mostrando ser a técnica mais robusta para essa característica de nicho. Para as espécies de nicho amplo também apresenta baixa volatilidade nos resultados. Para erro de viés geográfico, não foi possível observar um padrão de desempenho consistente dos algoritmos, até mesmo o GLM teve o melhor resultado em um dos cenários analisados.

6 CONCLUSÃO

Este trabalho buscou caracterizar e avaliar os impactos que problemas de QD causam em MDEs. A partir da simulação de espécies virtuais, foi possível realizar a análise dos efeitos dos problemas de QD dos três erros mais comuns que afetam dados ecológicos, localização imprecisa do avistamento, identificação incorreta de espécie e viés geográfico no registro de observações de espécies. A influência dos erros nos resultados foi analisada para contrastar os impactos de cada tipo de erro sobre uma métrica de validação. As análises de impacto foram feitas também sob as perspectivas de diferenciação de amplitude de nichos ecológicos das espécies simuladas, e a distinção das intensidades e proporções dos erros nas amostras de treinamento dos modelos. Três algoritmos tradicionais em MDEs, GLM, Maxent e RF, um algoritmo ainda não tão estabelecido e aplicado no contexto ecológico, XGboost, e um algoritmo de redes neurais de aplicação limitada em MDEs, foram utilizados na criação dos modelos. Os modelos criados foram comparados para identificar aqueles que conseguem manter a capacidade de generalização mesmo sob diferentes tipos, intensidades e proporções de erros de QD. As métricas de validação de modelos de classificação também foram discutidas sob o prisma da sensibilidade aos erros de QD e variabilidade.

Na maioria das espécies virtuais simuladas, o erro de viés geográfico foi o que mais impactou os resultados dos modelos, com magnitude de moderada a forte no tamanho de efeito desse tipo de erro sobre os modelos de controle. Somente a espécie de interesse euro foi mais impactada por erros de identificação. Os erros de localização, que na literatura científica são os mais comumente discutidos, apresentaram fraca capacidade de deterioração dos resultados dos modelos. Mesmo para as maiores intensidades de erros de 50 e 100 quilômetros, os erros de localização não tiveram tanto efeito nos modelos quanto as menores intensidades do erro de viés geográfico. Nos erros de localização, considerando as altas intensidades (50 quilômetros e 100 quilômetros), a maior diferença ocorreu entre espécies de amplitude de nicho ecológico muito discrepantes. Para a espécie euro de nicho ecológico muito amplo, altas intensidades de erros geraram impacto 10 vezes menor do que a espécie pireneus, espécie específica de nicho restrito. Os erros de identificação têm impacto moderado sobre os resultados e não foi possível discernir

um padrão de efeito nem sob a dimensão de amplitude de nicho ecológico, e nem sobre a intensidade do erro com espécie contaminante de nicho restrito e amplo. Para esse tipo de erro os resultados variaram muito de uma espécie de interesse para outra e de uma intensidade para outra.

Sobre as métricas de validação, Precisão, Kappa, MCC, F1 e TSS, foram as mais adequadas em relação à sensibilidade à presença dos erros e à variação da intensidade dos erros. Sobre os algoritmos para construção dos modelos, o XGBoost, apresentou bons resultados com capacidade de generalização mesmo em cenários com forte presença de problemas de QD. O algoritmo de RF, comprovou o porquê é uma das técnicas mais bem estabelecidas em aplicações gerais e em MDEs, obtendo resultados satisfatórios. As redes neurais, que é um algoritmo normalmente desconsiderado em aplicações de MDEs, manteve resultados próximos aos XGBoost e RF, com bons resultados mesmo para cenários de erros. Já o Maxent, algoritmo tradicional em aplicações ecológicas mostrou ser o mais adequado somente para algumas configurações de erros e para algumas espécies, não mantendo um desempenho homogêneo entre os cenários analisados.

Este trabalho contribui para o corpo de conhecimento de qualidade de dados em estudos de distribuição de espécies ao utilizar espécies virtuais baseadas no comportamento de espécies reais, de diferentes regiões e amplitudes de nicho ecológico, para comparar e quantificar o impacto de três tipos de erros de QD nos MDEs, caracterizando os diferentes gradientes de erros. O estudo comparativo da sensibilidade das métricas de validação para os três erros de qualidade trouxe resultados importantes para a seleção oportuna de métricas que podem evidenciar deficiências nos dados ou no processo de modelagem. Além disso, mostrou também o potencial que o algoritmo de XGBoost, relativamente recente em aplicações de MDEs, tem para construir modelos robustos, com capacidade de generalização mesmo com problemas QD.

Porém, mais estudos são necessários para explorar alguns aspectos não contemplados neste trabalho como investigar o impacto do problema de qualidade em uma estratificação mais granular da amplitude no nicho ecológico. As análises realizadas trataram de nichos amplos e restritos, mas dentro destas categorias é possível detalhar ainda mais para lançar luz sobre os efeitos dos problemas de QD

nos modelos de espécies específicas de nicho muito restrito e micro-habitats. Outra oportunidade está em explicar se o erro de viés geográfico continua tendo forte efeito nos MDEs para estudos em áreas mais restritas, dado que a redução do espaço geográfico total de interesse acarretaria o aumento proporcional do percentual de área coberta pelos dados. Mais uma possibilidade de exploração é em relação à resolução dos dados utilizados para modelagem, investigando se em níveis de resolução menores a sensibilidade aos erros se mantém.

Dados de ocorrência de espécies são recursos imprescindíveis para a compreensão do padrão de distribuição geográfica e ocupação de habitats das espécies. O conhecimento potencial que pode ser obtido a partir desses dados constitui base fundamental para a tomada de decisão sobre estratégias de mitigação e atenuação de impactos da atividade humana na biodiversidade. No entanto, problemas de qualidade comprometem esses dados e interferem na capacidade explicativa dos MDEs, e consequentemente na efetividade de estratégias de preservação. Assim, é importante um aprofundamento dos estudos científicos com o intuito de prover um entendimento mais amplo de como os impacto de problemas de DQ afetam os MDEs para que propostas de mitigação e aperfeiçoamento das técnicas envolvidas possam ser elaboradas.

7 REFERÊNCIAS

AGUIRRE-GUTIÉRREZ, J. et al. Fit-for-purpose: species distribution model performance depends on evaluation criteria - Dutch Hoverflies as a case study. **PloS one**, v. 8, n. 5, p. e63708, 2013.

ALDOSSARI, S.; HUSMEIER, D.; MATTHIOPOULOS, J. Transferable species distribution modelling: Comparative performance of Generalised Functional Response models. **Ecological Informatics**, v. 71, p. 101803, 1 nov. 2022.

ALWOSHEEL, A. ; VAN CRANENBURGH, S. ;; CHORUS, C. G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. 2018.

ARENAS-CASTRO, S. et al. Effects of input data sources on species distribution model predictions across species with different distributional ranges. **Journal of Biogeography**, v. 49, n. 7, p. 1299–1312, 1 jul. 2022.

AUBRY, K. B.; RALEY, C. M.; MCKELVEY, K. S. The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. **PLoS ONE**, v. 12, n. 6, p. 1–17, 2017a.

AUBRY, K. B.; RALEY, C. M.; MCKELVEY, K. S. The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. **PLOS ONE**, v. 12, n. 6, p. e0179152, 22 jun. 2017b.

BARBET-MASSIN, M. et al. Selecting pseudo-absences for species distribution models: how, where and how many? **Methods in Ecology and Evolution**, v. 3, n. 2, p. 327–338, 1 abr. 2012.

BARBOSA, W. L. et al. **Data Quality Problems Identified in the Bioclimatic Data Collection Process - A Survey**. 2019 14th Iberian Conference on Information Systems and Technologies (CISTI). **Anais**...IEEE, jun. 2019.

BARONE, L.; WILLIAMS, J.; MICKLOS, D. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. **PLOS Computational Biology**, v. 13, n. 10, p. e1005755, 19 out. 2017.

BCCVL. **Bioclim: BCCVL**. Disponível em: <https://support.bccvl.org.au/support/solutions/articles/6000083201-bioclim>. Acesso em: 21 ago. 2019.

BEAN, W. T.; STAFFORD, R.; BRASHARES, J. S. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. **Ecography**, p. 250–258, mar. 2012.

BEDIA, J.; HERRERA, S.; GUTIÉRREZ, J. M. Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections. **Global and Planetary Change**, v. 107, p. 1–12, 2013.

BENKENDORF, D. J.; HAWKINS, C. P. Effects of sample size and network depth on a deep learning approach to species distribution modeling. **Ecological Informatics**, v. 60, p. 101137, 1 nov. 2020.

BENNETT, D. S. **Police Response Times to Calls for Service: Fragmentation, Community Characteristics, and Efficiency**. [s.l: s.n.].

BERTIN, P. R. B.; VISOLI, M. C.; DRICKER, D. P. A GESTÃO DE DADOS DE PESQUISA NO CONTEXTO DA E-SCIENCE: BENEFÍCIOS, DESAFIOS E OPORTUNIDADES PARA ORGANIZAÇÕES DE P&D. **Ponto de Acesso**, v. 11, n. 2, p. 15, 2017.

B.H.THACKER et al. **Concepts of Model Verification and Validation**. Los Alamos, NM: [s.n.].

BIRD, T. J. et al. Statistical solutions for error and bias in global citizen science datasets. **Biological Conservation**, v. 173, p. 144–154, 2014.

BOOTH, T. H. et al. bioclim : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. **Diversity and Distributions**, v. 20, n. 1, p. 1–9, jan. 2014a.

BOOTH, T. H. et al. bioclim : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. **Diversity and Distributions**, v. 20, n. 1, p. 1–9, 1 jan. 2014b.

BORGHI, J. et al. Support Your Data: A Research Data Management Guide for Researchers. **Research Ideas and Outcomes**, v. 4, p. 112, 2018.

BOTELLA, C. et al. A deep learning approach to Species Distribution Modelling. **Multimedia Tools and Applications**, p. 978, 2018.

BOUCHER-LALONDE, V.; MORIN, A.; CURRIE, D. J. How are tree species distributed in climatic space? A simple and general pattern. **Global Ecology and Biogeography**, v. 21, n. 12, p. 1157–1166, 1 dez. 2012.

BRACKEN, J. T. et al. Maximizing species distribution model performance when using historical occurrences and variables of varying persistency. **Ecosphere**, v. 13, n. 3, p. e3951, 1 mar. 2022.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BRODIE, S. et al. Recommendations for quantifying and reducing uncertainty in climate projections of species distributions. **Global Change Biology**, v. 28, n. 22, p. 6586–6601, 1 nov. 2022.

BROTONS, L. Species Distribution Models and Impact Factor Growth in Environmental Journals: Methodological Fashion or the Attraction of Global Change Science. **PLoS ONE**, v. 9, n. 11, p. e111996, 11 nov. 2014.

BRUN, P. et al. Model complexity affects species distribution projections under climate change. **Journal of Biogeography**, v. 47, n. 1, p. 130–142, 1 jan. 2020.

BUCKLAND, C. E.; SMITH, A. J. A. C.; THOMAS, D. S. G. A comparison in species distribution model performance of succulents using key species and subsets of environmental predictors. **Ecology and Evolution**, v. 12, n. 6, p. e8981, 1 jun. 2022.

BUCKLIN, D. N. et al. Comparing species distribution models constructed with different subsets of environmental predictors. **Diversity and Distributions**, v. 21, n. 1, p. 23–35, 1 jan. 2015.

BUSBY, JR. BIOCLIM - a bioclimate analysis and prediction system. **Plant Protection Quarterly**, v. 61, p. 8–9, 9 abr. 1991.

CAI, L. et al. Machine learning improves global models of plant diversity. **bioRxiv**, p. 2022.04.08.487610, 9 abr. 2022.

CAI, L.; ZHU, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. **Data Science Journal**, v. 14, n. 0, p. 2, 2015.

CARVALHO, M. C. et al. Modeling Ecological Niche of Tree Species in Brazilian Tropical Area. **Cerne**, v. 23, n. 2, p. 229–240, 2017.

CHAN, D. **Environmental science and information application technology : proceedings of the 2014 5th international conference on environmental science and information application technology**. 1. ed. Hong Kong: CRC Press, 2015.

CHEN, H. et al. A Review of Data Quality Assessment Methods for Public Health Information Systems. **International Journal of Environmental Research and Public Health**, v. 11, n. 5, p. 5170–5207, 14 maio 2014.

CHEN, Q. et al. Decision Variants for the Automatic Determination of Optimal Feature Subset in RF-RFE. **Genes**, v. 9, n. 6, 15 jun. 2018.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, v. 13-17- August-2016, p. 785–794, 9 mar. 2016.

CHOO, Y. R. et al. Best practices for reporting individual identification using camera trap photographs. **Global Ecology and Conservation**, v. 24, p. e01294, 1 dez. 2020.

CLARE, J. D. J. et al. Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved? **Ecological Applications**, v. 29, n. 2, 2019.

CMMI INSTITUTE. Data Management Maturity Model Introduction. 12 dez. 2014.

COLLINS, F. S.; TABAK, L. A. Policy: NIH plans to enhance reproducibility. **Nature**, v. 505, n. 7485, p. 612–613, 27 jan. 2014.

CONNOR, T. et al. Effects of grain size and niche breadth on species distribution modeling. **Ecography**, v. 41, n. 8, p. 1270–1282, 1 ago. 2018.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados e Extensões**. Piracicaba: [s.n.].

COX, A. M. et al. Developments in research data management in academic libraries: Towards an understanding of research data service maturity. **Journal of the Association for Information Science and Technology**, v. 68, n. 9, p. 2182–2200, set. 2017.

COX; AM; TAM. A critical analysis of lifecycle models of the research process and research data management. 2018.

CROSBY, P. B. **Quality is free : the art of making quality certain**. [s.l.] McGraw-Hill, 1979.

CRUZ-CÁRDENAS, G. et al. Potential species distribution modeling and the use of principal component analysis as predictor variables. **Revista Mexicana de Biodiversidad**, v. 85, n. 1, p. 189–199, 2014.

CUGLER, D. C. et al. A geographical approach for metadata quality improvement in biological observation databases. **Proceedings - IEEE 9th International Conference on e-Science, e-Science 2013**, p. 212–220, 2013.

CUPOLI, P.; EARLEY, S.; HENDERSON, D. **DAMA-DMBOK2 Framework Production Editor**. [s.l: s.n.].

DAMA INTERNATIONAL. DAMA-DMBOK Enterprise version. p. 644, 5 jul. 2017.

DE MARCO, P.; NÓBREGA, C. C. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. **PLOS ONE**, v. 13, n. 9, p. e0202403, 11 set. 2018.

DEVICTOR, V. et al. Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. **Ecology Letters**, p. no-no, 10 jun. 2010.

DORAZIO, R. M. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. **Global Ecology and Biogeography**, v. 23, n. 12, p. 1472–1484, 2014.

DORMANN, C. F. et al. COMPONENTS OF UNCERTAINTY IN SPECIES DISTRIBUTION ANALYSIS: A CASE STUDY OF THE GREAT GREY SHRIKE. **Ecology**, v. 89, n. 12, p. 3371–3386, dez. 2008.

DUAN, R. Y. et al. SDMvspecies: a software for creating virtual species for species distribution modelling. **Ecography**, v. 38, n. 1, p. 108–110, 1 jan. 2015.

DUAN, R.-Y. et al. The Predictive Performance and Stability of Six Species Distribution Models. **PLoS ONE**, v. 9, n. 11, p. e112764, 10 nov. 2014.

EFFROSYNIDIS, D. et al. Species Distribution Modelling via Feature Engineering and Machine Learning for Pelagic Fishes in the Mediterranean Sea. **Applied Sciences 2020, Vol. 10, Page 8900**, v. 10, n. 24, p. 8900, 13 dez. 2020.

ELITH, J. et al. **Novel methods improve prediction of species' distributions from occurrence data**. [s.l: s.n.].

ELITH, J.; GRAHAM, C. H. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. **Ecography**, v. 32, n. 1, p. 66–77, 3 abr. 2009.

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, 2009.

FEI, S.; YU, F. Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. **Landscape Ecology**, v. 31, n. 1, p. 31–42, 2016a.

FEI, S.; YU, F. Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. **Landscape Ecology**, v. 31, n. 1, p. 31–42, 2016b.

FELICÍSIMO, Á.; GÓMEZ-MUÑOZ, A. GIS and predictive modelling: A comparison of methods for forest management and decision-making. **GIS for Environmental Decision-Making**, n. January 2004, p. 117–129, 2004.

FELLEGI, I. P.; SUNTER, A. B. A Theory for Record Linkage. **Journal of the American Statistical Association**, v. 64, n. 328, p. 1183–1210, dez. 1969.

FENG, L. et al. Predicting Suitable Habitats of Melia Azedarach L. Using Data Mining. 2021.

FERNANDES, R. F.; SCHERRER, D.; GUISAN, A. Effects of simulated observation errors on the performance of species distribution models. **Diversity and Distributions**, v. 25, n. 3, p. 400–413, 2019a.

FERNANDES, R. F.; SCHERRER, D.; GUISAN, A. Effects of simulated observation errors on the performance of species distribution models. **Diversity and Distributions**, v. 25, n. 3, p. 400–413, 2019b.

FICK, S. E.; HIJMANS, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology**, v. 37, n. 12, p. 4302–4315, 1 out. 2017.

FIRMANI, D. et al. On the Meaningfulness of “Big Data Quality” (Invited Paper). **Data Science and Engineering**, v. 1, n. 1, p. 6–20, 2016.

FOURCADE, Y. et al. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. **PLoS ONE**, v. 9, n. 5, p. e97122, 12 maio 2014.

FRANKLIN, J. **Mapping species distributions**. Cambridge: Cambridge University Press, 2009.

GÁBOR, L. et al. How do species and data characteristics affect species distribution models and when to use environmental filtering? **https://doi.org/10.1080/13658816.2019.1615070**, v. 34, n. 8, p. 1567–1584, 2 ago. 2019.

GÁBOR, L. et al. The effect of positional error on fine scale species distribution models increases for specialist species. **Ecography**, v. 43, n. 2, p. 256–269, 1 fev. 2020.

GÁBOR, L. et al. Positional errors in species distribution modelling are not overcome by the coarser grains of analysis. **Methods in Ecology and Evolution**, v. 13, n. 10, p. 2289–2302, 1 out. 2022.

GARVIN, D. A. **Managing quality : the strategic and competitive edge**. [s.l.] Free Press, 1988.

GERWIN, D. Control and evaluation in the innovation process: The case of flexible manufacturing systems. **IEEE Transactions on Engineering Management**, v. EM-28, n. 3, p. 62–70, 1981.

GHINS, M. Models, truth and realism: assessing Bas van Fraassen's views on scientific representation. **Manuscrito**, v. 34, n. 1, p. 207–232, jun. 2011.

GOBEN, A.; RASZEWSKI, R. The data life cycle applied to our own data. **Journal of the Medical Library Association : JMLA**, v. 103, n. 1, p. 40–4, jan. 2015.

GOMES, V. H. F. et al. Species Distribution Modelling: Contrasting presence-only models with plot abundance data. **Scientific Reports**, v. 8, n. 1, p. 1–12, 2018a.

GOMES, V. H. F. et al. Species Distribution Modelling: Contrasting presence-only models with plot abundance data. **Scientific Reports 2018 8:1**, v. 8, n. 1, p. 1–12, 17 jan. 2018b.

GOMES, V. H. F. et al. Species Distribution Modelling: Contrasting presence-only models with plot abundance data. **Scientific Reports**, v. 8, n. 1, p. 1003, 17 dez. 2018c.

GRAHAM, C. H. et al. The influence of spatial errors in species occurrence data used in distribution models. **Journal of Applied Ecology**, v. 45, n. 1, p. 239–247, fev. 2008.

GRAHAM, L. J.; HAINES-YOUNG, R. H.; FIELD, R. Using citizen science data for conservation planning: Methods for quality control and downscaling for use in stochastic patch occupancy modelling. **Biological Conservation**, v. 192, p. 65–73, 2015.

GRIMMETT, L.; WHITSED, R.; HORTA, A. Creating virtual species to test species distribution models: the importance of landscape structure, dispersal and population processes. **Ecography**, v. 44, n. 5, p. 753–765, 1 maio 2021.

GUISAN, A.; EDWARDS, T. C.; HASTIE, T. **Generalized linear and generalized additive models in studies of species distributions: setting the scene**. [s.l: s.n.].

GUISAN, A.; ZIMMERMANN, N. E. **Predictive habitat distribution models in ecologyEcological Modelling**. [s.l: s.n.].

GUISANDE, C. et al. SPEDInstabR: An algorithm based on a fluctuation index for selecting predictors in species distribution modeling. **Ecological Informatics**, v. 37, p. 18–23, 2017.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. [s.l: s.n.].

HEFLEY, T. J. et al. Correction of location errors for presence-only species distribution models. **Methods in Ecology and Evolution**, v. 5, n. 3, p. 207–214, 2014.

HEFLEY, T. J.; BROST, B. M.; HOOTEN, M. B. Bias correction of bounded location errors in presence-only data. **Methods in Ecology and Evolution**, v. 8, n. 11, p. 1566–1573, 2017.

HIJMANS, R. J. et al. Package “raster” Geographic Data Analysis and Modeling. **Cran R-Project**, p. 1–249, 16 set. 2022.

HIJMANS, R. J.; ELITH, J. Species distribution modeling with R Introduction. **R Manual**, p. 71, 2017.

HIJMANS, R. J.; ELITH, J. **Spatial Distribution Models - R Spatial**. [s.l: s.n.].

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, 2014.

HUTCHINSON, G. E. Concluding Remarks. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 22, p. 415–427, 1 jan. 1957.

INMAN, R. et al. Comparing sample bias correction methods for species distribution modeling using virtual species. **Ecosphere**, v. 12, n. 3, p. e03422, 1 mar. 2021.

JIMÉNEZ, L.; SOBERÓN, J. Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data. **Methods in Ecology and Evolution**, v. 11, n. 12, p. 1571–1586, 1 dez. 2020.

JUKNA, S. **Extremal combinatorics : with applications in computer science**. [s.l.] Springer, 2011.

KALAN, A. K. et al. Passive acoustic monitoring reveals group ranging and territory use: a case study of wild chimpanzees (Pan troglodytes). **Frontiers in zoology**, v. 13, p. 34, 2016.

KELLEY, W. MICHAEL.; DONNELLY, R. A. **The humongous book of statistics problems : translated for people who don't speak math!!** [s.l.] Alpha, 2009.

KELLING, S. et al. Taking a "Big Data" approach to data quality in a citizen science project. **Ambio**, v. 44, p. 11, 2015a.

KELLING, S. et al. Taking a "Big Data" approach to data quality in a citizen science project. **Ambio**, v. 44 Suppl 4, n. Suppl 4, p. 601–11, nov. 2015b.

KITASUKA, T.; ARITSUGI, M.; RAHUTOMO, F. **Semantic Cosine Similarity**. [s.l: s.n.]. Disponível em: <https://www.researchgate.net/publication/262525676>.

KITCHENHAM, B. Procedures for Performing Systematic Literature Reviews. **Keele University & Durham University, UK**, 2004.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. **EBSE Technical Report**, v. 2, 2007.

KOSMALA, M. et al. Assessing data quality in citizen science. **Frontiers in Ecology and the Environment**, v. 14, n. 10, p. 551–560, 1 dez. 2016.

KURSA, M. B. Package "Boruta" Title Wrapper Algorithm for All Relevant Feature Selection. p. 1–17, 21 maio 2020.

KURSA, M. B.; JANKOWSKI, A.; RUDNICKI, W. R. Boruta - A system for feature selection. **Fundamenta Informaticae**, v. 101, n. 4, p. 271–285, 2010.

LAW BIECEK, P. DALEX: Explainers for Complex Predictive Models in R. **Journal of Machine Learning Research**, v. 19, p. 1–5, 2018.

LEROY, B. et al. virtualspecies, an R package to generate virtual species distributions. **Ecography**, v. 39, n. 6, p. 599–607, 1 jun. 2016.

LEROY, B. **The virtualspecies R package: a complete tutorial**. Disponível em: <http://borisleroy.com/files/virtualspecies-tutorial.html>. Acesso em: 17 out. 2022.

LEROY, B. Choosing presence-only species distribution models. **Journal of Biogeography**, 2022.

LEWANDOWSKI, E.; SPECHT, H. Influence of volunteer and project characteristics on data quality of biological surveys. **Conservation Biology**, v. 29, n. 3, p. 713–723, 1 jun. 2015a.

LEWANDOWSKI, E.; SPECHT, H. Influence of volunteer and project characteristics on data quality of biological surveys. **Conservation Biology**, v. 29, n. 3, p. 713–723, 1 jun. 2015b.

LI, J. et al. Feature selection: A data perspective. **ACM Computing Surveys**, v. 50, n. 6, 1 dez. 2017.

LIN, Y. P. et al. Uncertainty analysis of crowd-sourced and professionally collected field data used in species distribution models of Taiwanese moths. **Biological Conservation**, v. 181, p. 102–110, 2015.

LINDHOLM, A. et al. **Supervised Machine Learning: Statistical Machine Learning course**. Uppsala: [s.n.].

LIU, C.; NEWELL, G.; WHITE, M. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. **Ecography**, v. 42, n. 3, p. 535–548, 1 mar. 2019.

LOH, W.-Y.; ZHOU, P. Variable Importance Scores. 13 fev. 2021.

LUAN, J. et al. Matching Data Types to the Objectives of Species Distribution Modeling: An Evaluation With Marine Fish Species. **Frontiers in Marine Science**, v. 8, p. 1544, 22 out. 2021.

LUKYANENKO, R.; PARSONS, J.; WIERSMA, Y. F. Emerging problems of data quality in citizen science. **Conservation Biology**, v. 30, n. 3, p. 447–449, jun. 2016.

MALDONADO, C. et al. Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? **Global Ecology and Biogeography**, v. 24, n. 8, p. 973–984, 2015.

MARTÍN-MARTÍN, A. et al. Scopus: a systematic comparison of citations in 252 subject categories. **Journal of Informetrics**, v. 12, n. 4, p. 1160–1177, 2018.

MATEO, R. G.; FELICÍSIMO, Á. M.; MUÑOZ, J. Modelos de distribución de especies: Una revisión sintética. **Revista chilena de historia natural**, v. 84, n. 2, p. 217–240, jun. 2011.

MAURICIO PINTO-VALVERDE, J. et al. **HDQM2: Healthcare Data Quality Maturity Model WMU ScholarWorks Citation**. [s.l: s.n.].

MCCULLAGH, P. (PETER); NELDER, J. A. **Generalized linear models**. [s.l.] Chapman and Hall, 1989.

MCGILVRAY, DANETTE. **Executing data quality projects : ten steps to quality data and trusted information**. [s.l.] Morgan Kaufmann/Elsevier, 2008.

MCKIBBEN, F. E.; FREY, J. K. Linking camera-trap data to taxonomy: Identifying photographs of morphologically similar chipmunks. **Ecology and Evolution**, v. 11, n. 14, p. 9741–9764, 1 jul. 2021.

MEYNARD, C.; LEROY, B.; KAPLAN, D. M. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? 2019.

MI, C. et al. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. **PeerJ**, v. 5, p. e2849, 12 jan. 2017.

MIAO, Z. et al. Iterative Human and Automated Identification of Wildlife Images. 5 maio 2021.

MOHD RAZALI, N.; BEE WAH, Y. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. **Journal of Statistical Modeling and Analytics**, v. 2, n. 1, p. 13–14, 2011.

MOUSAZADE, M. et al. Maxent Data Mining Technique and Its Comparison with a Bivariate Statistical Model for Predicting the Potential Distribution of Astragalus Fasciculifolius Boiss. in Fars, Iran. **Sustainability**, v. 11, n. 12, p. 3452, 2019.

MUÑOZ, J.; FELICÍSIMO, Á. M. Comparison of statistical methods commonly used in predictive modelling. **Journal of Vegetation Science**, v. 15, n. 2, p. 285–292, 2004.

NACHAR, N. The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. **Tutorials in Quantitative Methods for Psychology**, v. 4, n. 1, p. 13–20, 1 mar. 2008.

NDABARORA, E.; CHIPPS, J. A.; UYS, L. Systematic review of health data quality management and best practices at community and district levels in LMIC. **Information Development**, v. 30, n. 2, p. 103–120, 27 maio 2014.

NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized Linear ModelsSource: Journal of the Royal Statistical Society. Series A (General)**. [s.l: s.n.].

NGUYEN, D. et al. Empirical comparison of tests for one-factor ANOVA under heterogeneity and non-normality: A Monte Carlo study. **Journal of Modern Applied Statistical Journal of Modern Applied Statistical Methods Methods**, v. 18, n. 2, p. 2906, 2019.

NOCE, S.; CAPORASO, L.; SANTINI, M. **CMCC-BioClimInd. A new globaldataset of bioclimatic indicators**. Disponível em: <https://doi.pangaea.de/10.1594/PANGAEA.904278>. Acesso em: 10 jan. 2023.

NOCE, S.; CAPORASO, L.; SANTINI, M. A new global dataset of bioclimatic indicators. **Scientific Data 2020 7:1**, v. 7, n. 1, p. 1–12, 16 nov. 2020.

OSTERTAGOVÁ, E.; OSTERTAG, O.; KOVÁČ, J. Methodology and application of the Kruskal-Wallis test. **Applied Mechanics and Materials**, v. 611, p. 115–120, 2014.

PARK, D. S.; DAVIS, C. C. Implications and alternatives of assigning climate data to geographical centroids. **Journal of Biogeography**, v. 44, n. 10, p. 2188–2198, 2017.

PERES, P. H. DE F. et al. Implications of unreliable species identification methods for Neotropical deer conservation planning. **Perspectives in Ecology and Conservation**, v. 19, n. 4, p. 435–442, 1 out. 2021.

PERRY, G. L. W.; DICKSON, M. E. Using Machine Learning to Predict Geomorphic Disturbance: The Effects of Sample Size, Sample Prevalence, and Sampling Strategy. **Journal of Geophysical Research: Earth Surface**, v. 123, n. 11, p. 2954–2970, 2018.

PHILLIPS, S. J. et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. **Ecological Applications**, v. 19, n. 1, p. 181–197, jan. 2009a.

PHILLIPS, S. J. et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. **Ecological Applications**, v. 19, n. 1, p. 181–197, 1 jan. 2009b.

PHILLIPS, S. J.; ANDERSON, R. P.; SCHAPIRE, R. E. Maximum entropy modeling of species geographic distributions. **Ecological Modelling**, v. 190, p. 231–259, 2006.

PRASAD, A. M.; IVERSON, L. R.; LIAW, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. **Ecosystems**, v. 9, n. 2, p. 181–199, 15 mar. 2006.

QAZI, A. W.; SAQIB, Z.; ZAMAN-UL-HAQ, M. Trends in species distribution modelling in context of rare and endemic plants: a systematic review. **Ecological Processes**, v. 11, n. 1, p. 1–11, 1 dez. 2022.

QIAO, H. et al. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. **Ecography**, v. 39, n. 8, p. 805–813, 1 ago. 2016.

QIAO, H.; SOBERÓN, J.; PETERSON, A. T. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. **Methods in Ecology and Evolution**, v. 6, n. 10, p. 1126–1136, 1 out. 2015.

RADOVIĆ, A. et al. Impact of biased sampling effort and spatial uncertainty of locations on models of plant invasion patterns in Croatia. **Biological Invasions**, v. 20, n. 12, p. 3527–3544, 2018.

RATNIEKS, F. L. W. et al. Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. **Methods in Ecology and Evolution**, v. 7, n. 10, p. 1226–1235, 2016.

REDMAN, T. C. **Data quality : the field guide**. [s.l.] Digital Press, 2001.

REDMAN, T. C.; BY-GODFREY, T. C. /FOREWORD; BLANTON, A. **Data quality for the information age**. [s.l.] Artech House, 1996.

REISS, H. et al. Species distribution modelling of marine benthos: A North Sea case study. **Marine Ecology Progress Series**, v. 442, p. 71–86, 2011.

ROBERTS, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. **Ecography**, v. 40, n. 8, p. 913–929, 1 ago. 2017.

ROBINSON, O.; RUIZ-GUTIERREZ, V.; FINK, D. Correcting for bias in distribution modelling for rare species using citizen science data. **Diversity and Distributions**, v. 24, n. 4, p. 460–472, 1 abr. 2018.

RYLL, L.; SEIDENS, S. **Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey**. [s.l: s.n.].

SCANNAPIECO, M.; CATARCI, T. Data Quality under the Computer Science perspective. **Computer Engineering**, v. 2, n. 2, p. 1–12, 2002.

SCHNEIDER, S. et al. Past, present and future approaches using computer vision for animal re-identification from camera trap data. **Methods in Ecology and Evolution**, v. 10, n. 4, p. 461–470, 1 abr. 2019.

SCHNEIDER, S. et al. Three critical factors affecting automated image species recognition performance for camera traps. **Ecology and Evolution**, v. 10, n. 7, p. 3503–3517, 1 abr. 2020.

SEBASTIAN-COLEMAN, LAURA. **Measuring data quality for ongoing improvement : a data quality assessment framework**. [s.l.] Elsevier Science, 2013.

SERRA-DIAZ, J. M. et al. Big data of tree species distributions: how big and how good? **Forest Ecosystems**, v. 4, n. 1, 2017.

SHAHID, N.; RAPPON, T.; BERTA, W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. **PLOS ONE**, v. 14, n. 2, p. e0212356, 19 fev. 2019.

SHIREY, V. et al. Current GBIF occurrence data demonstrates both promise and limitations for potential red listing of spiders. **Biodiversity Data Journal 7: e47369**, v. 7, p. e47369-, 2019.

SHIROYAMA, R.; WANG, M.; YOSHIMURA, C. Effect of sample size on habitat suitability estimation using random forests: a case of bluegill, Lepomis macrochirus. **Annales de Limnologie - International Journal of Limnology**, v. 56, p. 13, 2020.

SIMÕES, M. V. P.; PETERSON, A. T. Utility and limitations of climate-matching approaches in detecting different types of spatial errors in biodiversity data. **Insect Conservation and Diversity**, v. 11, n. 5, p. 407–414, 2018.

SMITH, J.; NOBLE, H. Bias in research. **Evidence Based Nursing**, v. 17, n. 4, p. 100–101, 2014.

SON, Y. K.; PARK, C. S. Economic measure of productivity, quality and flexibility in advanced manufacturing systems. **Journal of Manufacturing Systems**, v. 6, n. 3, p. 193–207, 1 jan. 1987.

SOULTAN, A.; SAFI, K. The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. **PLOS ONE**, v. 12, n. 11, p. e0187906, 1 nov. 2017.

STOCK, A.; MICHELI, F. Effects of model assumptions and data quality on spatial cumulative human impact assessments. **Global Ecology and Biogeography**, v. 25, n. 11, p. 1321–1332, 1 nov. 2016.

STOKLOSA, J. et al. A climate of uncertainty: Accounting for error in climate variables for species distribution models. **Methods in Ecology and Evolution**, v. 6, n. 4, p. 412–423, 2015.

STUART, S. N. et al. Ecology. The barometer of life. **Science (New York, N.Y.)**, v. 328, n. 5975, p. 177, 9 abr. 2010.

SUÁREZ, E. F. **Espécies de lince pelo planeta**. Disponível em: <https://meusanimais.com.br/especies-de-lince-pelo-planeta/>. Acesso em: 23 jan. 2023.

ŠUBELJ, L. et al. Quantifying the Consistency of Scientific Databases. 2015.

SUBHAJINI, A. C. APPLICATION OF NEURAL NETWORKS IN WEATHER FORECASTING. **Climate Change and Conservation Research**, v. 4, n. 1, p. 8–18, 2018.

SUGGITT, A. J. et al. Conducting robust ecological analyses with climate data. **Oikos**, v. 126, n. 11, p. 1533–1541, 1 nov. 2017.

SYFERT, M. M.; SMITH, M. J.; COOMES, D. A. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. **PloS one**, v. 8, n. 2, p. e55158, 2013a.

SYFERT, M. M.; SMITH, M. J.; COOMES, D. A. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. **PloS one**, v. 8, n. 2, 14 fev. 2013b.

TESSAROLO, G. et al. High uncertainty in the effects of data characteristics on the performance of species distribution models. **Ecological Indicators**, v. 121, p. 107147, 1 fev. 2021.

TOMCZAK, M.; TOMCZAK, E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. **TRENDS in Sport Sciences**, v. 1, n. 21, p. 19–25, 2014.

TROIA, M. J.; MCMANAMAY, R. A. Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. **Ecology and Evolution**, v. 6, n. 14, p. 4654–4669, 1 jul. 2016.

TUV, E.; BORISOV, A.; TORKKOLA, K. Feature selection using ensemble based ranking against artificial contrasts. **IEEE International Conference on Neural Networks - Conference Proceedings**, p. 2181–2185, 2006.

VALAVI, R. et al. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. **Ecological Monographs**, v. 92, n. 1, p. e01486, 1 fev. 2022.

VEIGA, A. K.; CARTOLANO, E. A.; SARAIVA, A. M. Data quality control in biodiversity informatics: The case of species occurrence data. **IEEE Latin America Transactions**, v. 12, n. 4, p. 683–693, 2014.

VELÁSQUEZ-TIBATÁ, J.; GRAHAM, C. H.; MUNCH, S. B. Using measurement error models to account for georeferencing error in species distribution models. **Ecography**, v. 39, n. 3, p. 305–316, 1 mar. 2016.

VELEZ-LIENDO, X.; STRUBBE, D.; MATTHYSEN, E. **Effects of variable selection on modelling habitat and potential distribution of the Andean bear in BoliviaUrsus**. [s.l: s.n.].

VINCENZI, S. et al. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. **Ecological Modelling**, v. 222, p. 1471–1478, 2011.

WAND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological foundations. **Communications of the ACM**, v. 39, n. 11, p. 86–95, 1 nov. 1996.

WANG, R. Y. A Product Perspective on Total Data Quality Management. **COMMUNICATIONS OF THE ACM**, v. 41, n. 2, 1998.

WANG, R. Y.; STRONG, D. M. **Beyond Accuracy: What Data Quality Means to Data ConsumersSource: Journal of Management Information Systems**. [s.l: s.n.].

WILMS, K. et al. How to Improve Research Data Management. Em: [s.l: s.n.]. p. 434–442.

WISSIK, T.; DURCO, M. **Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions**. [s.l: s.n.].

WISZ, M. S. et al. Effects of sample size on the performance of species distribution models. **Diversity and Distributions**, v. 14, n. 5, p. 763–773, set. 2008.

WUNDERLICH, R. F. et al. Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models. **Nature Conservation 35: 97-116**, v. 35, p. 97–116, 20 jun. 2019.

YU, J.; WONG, W.-K.; KELLING, S. Clustering species accumulation curves to identify skill levels of citizen scientists participating in the eBird project. **Twenty-sixth IAAI Conference**, p. 3017–3023, 2014.

ZAHROTUN, L. Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method. **Computer Engineering and Applications**, v. 5, n. 1, 2016.

ZETT, T.; STRATFORD, K. J.; WEISE, F. J. Inter-observer variance and agreement of wildlife information extracted from camera trap images. **Biodiversity and Conservation**, v. 31, n. 12, p. 3019–3037, 1 out. 2022.

ZHANG, G. et al. A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. **Transactions in GIS**, v. 22, n. 1, p. 202–216, 2018.

ZHAO, Y. et al. A comparative mapping of plant species diversity using ensemble learning algorithms combined with high accuracy surface modeling. **Environmental Science and Pollution Research**, v. 29, n. 12, p. 17878–17891, 1 mar. 2022.

ZURELL, D. et al. The virtual ecologist approach: Simulating data and observers. **Oikos**, v. 119, n. 4, p. 622–635, mar. 2010.