UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto De Ciências Biológicas

Programa Interunidades De Pós-Graduação Em Bioinformática

Mário Henrique De Assis Silva

PADRONIZAÇÃO E AUTOMAÇÃO DE DADOS DA COLEÇÃO ACAROLÓGICA DO CENTRO DE COLEÇÕES TAXONÔMICAS DA UFMG: UM ENFOQUE NA GESTÃO DE METADADOS E MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIES

Belo Horizonte

2025

Mário Henrique De Assis Silva

PADRONIZAÇÃO E AUTOMAÇÃO DE DADOS DA COLEÇÃO ACAROLÓGICA DO CENTRO DE COLEÇÕES TAXONÔMICAS DA UFMG: UM ENFOQUE NA GESTÃO DE METADADOS E MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIES

Dissertação apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Mestre em Bioinformática.

Orientador: Aristóteles Góes Neto

Coorientador: Almir Rogério Pepato

Belo Horizonte

2025

043 Assis-Silva, Mário Henrique de.
Padronização e automação de dados da coleção acarológica do centro de coleções taxonômicas da UFMG: um enfoque na gestão de metadados e modelagem de distribuição de espécies [manuscrito] / Mário Henrique de Assis Silva. – 2025.
91 f. : il. ; 29,5 cm.

Orientador: Aristóteles Góes Neto. Coorientador: Almir Rogério Pepato.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Biodiversidade. 3. Ecologia. 4. Aprendizado de Máquina. 5. Padrões de Referência. I. Neto, Aristóteles Góes. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 577.1

Ficha Catalográfica elaborada pelo bibliotecário Rafael Gonçalves Dias (CRB: 6 - 3277)

UNIVERSIDADE FEDERAL DE MINAS GERAIS

ATA

INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

DEFESA DE DISSERTAÇÃO

Mário Henrique de Assis Silva

Às quatorze horas do dia **24 de junho de 2025**, reuniu-se, por videoconferência através do aplicativo Zoom, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Padronização e Automação de Dados da Coleção Acarológica do Centro de Coleções Taxonômicas da UFMG: Um Enfoque na Gestão de Metadados e Modelagem de Distribuição de Espécies"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Aristóteles Góes Neto**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dr. Aristóteles Góes Neto	Universidade Federal de Minas Gerais	Aprovado
Dr. Almir Rogério Pepato	Universidade Federal de Minas Gerais	Aprovado
Dra. Joicymara Santos Xavier	Instituto Tecnológico de Aeronáutica	Aprovado
Dr. Philip Russo da Silva	Universidade Federal de Minas Gerais	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 24 de junho de 2025.

Documento assinado eletronicamente por **Philip Russo da Silva**, **Usuário Externo**, em 25/06/2025, às 09:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Joicymara Santos Xavier**, **Usuário Externo**, em 25/06/2025, às 11:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Professor do Magistério Superior**, em 25/06/2025, às 15:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Almir Rogerio Pepato**, **Professor do Magistério Superior**, em 25/06/2025, às 16:40, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4312374** e o código CRC **DD94B09B**.

Referência: Processo nº 23072.238268/2025-31 SEI nº 4312374

Agradecimentos

Primeiramente, agradeço à minha mãe, Glacimar, por ser meu maior exemplo e incentivo ao longo de toda a minha vida. Seu apoio incondicional e constante encorajamento aos estudos foram fundamentais para que eu chegasse até aqui. Sem ela, certamente eu não teria trilhado esse caminho.

Estendo meus agradecimentos a toda a minha família — minhas avós e avôs, minhas tias e tios, meus primos e primas — pelo carinho, suporte e palavras de encorajamento nos momentos mais difíceis.

Sou profundamente grato aos Ramos, que foram essenciais para que eu não desistisse do mestrado ainda no primeiro semestre. O acolhimento, amizade e apoio que recebi foram determinantes para a continuidade dessa jornada.

Aos meus amigos, agradeço pela presença constante, pelas conversas, pelo apoio emocional, pelas distrações necessárias nos momentos de tensão, pelas quintas-feiras e, principalmente, por me lembrarem que eu não estava sozinho. Cada um, à sua maneira, contribuiu para que esse caminho fosse mais leve e possível.

Agradeço também a todos os meus professores, desde a primeira professora no maternal, passando pelos anos do Ensino Fundamental e Médio, pelos professores particulares, de cursos e pré-vestibulares, até os docentes da graduação e pós-graduação. Citar nomes seria injusto, pois inevitavelmente esqueceria alguém; cada um deles teve um papel importante na minha formação e ajudou a moldar quem sou hoje.

Mas a vida escolar não é feita apenas de professores. Por isso, registro minha gratidão também a todos os outros profissionais que fizeram parte desse percurso: porteiros, disciplinários, bibliotecárias, tios e tias do serviço geral, coordenadores e pedagogos — todos contribuíram para a minha formação de maneira direta ou indireta.

Um agradecimento especial ao meu co-orientador, Almir, que, aceitou receber um aluno da Física em um laboratório de acarologia, me permitindo explorar uma área completamente nova e fascinante, muito além da minha formação original.

Também agradeço ao meu orientador, Aristóteles, que acreditou no meu potencial e me acolheu no Programa de Pós-Graduação em Bioinformática, viabilizando o desenvolvimento deste trabalho.

Por fim, agradeço às agências de fomento, em especial à CAPES, cujo apoio financeiro tornou possível a realização deste mestrado.

Resumo

A digitalização e padronização de dados biológicos são fundamentais para a gestão eficiente de coleções científicas, permitindo sua integração a repositórios globais e ampliando seu potencial para pesquisas ecológicas e biogeográficas. Esta dissertação buscou a padronização e automação dos dados da Coleção Acarológica UFMG-AC, visando solucionar problemas de fragmentação e inconsistência dos registros. O estudo implementou metodologias computacionais para a conversão e organização dos dados, garantindo conformidade com o padrão Darwin Core (DwC). Foram desenvolvidos scripts e pipelines para a estruturação de metadados, correção de inconsistências taxonômicas e geoespaciais, e integração dos registros a plataformas internacionais como o Global Biodiversity Information Facility (GBIF) e o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr). Paralelamente, a modelagem de distribuição de espécies (Species Distribution Modeling – SDM) foi aplicada para inferir padrões biogeográficos e prever áreas potenciais de ocorrência das espécies catalogadas. A biblioteca EcoDistrib foi utilizada para otimizar a seleção de variáveis ambientais e a execução de modelos baseados em aprendizado de máquina, enquanto técnicas de validação cruzada e métricas estatísticas asseguraram a robustez dos resultados. A análise comparativa dos dados antes e após a padronização evidenciou um aumento significativo na qualidade, consistência e acessibilidade das informações, com a correção de redundâncias taxonômicas e a melhoria da precisão espacial dos registros. A integração dos dados aprimorados às plataformas GBIF e SiBBr conferiu maior visibilidade à coleção, facilitando seu uso em estudos biogeográficos e de conservação. Os modelos de distribuição revelaram padrões espaciais relevantes, identificando lacunas nos registros e sugerindo novas áreas de ocorrência para futuras coletas. A pesquisa demonstra que a combinação de técnicas computacionais com abordagens tradicionais de curadoria pode aprimorar significativamente a gestão de coleções biológicas, destacando a importância da automação e padronização para a preservação, compartilhamento e aplicabilidade científica dos dados de biodiversidade. Os resultados reforçam a relevância da digitalização e do uso de ferramentas computacionais na modernização de acervos científicos, contribuindo para o avanço da biogeografia, ecologia e conservação da biodiversidade.

Palavras-chave: Digitalização de coleções biológicas; Modelagem de distribuição de espécies; Darwin Core; Aprendizado de máquinas; Integração de dados.

Abstract

The digitization and standardization of biological data are essential for the efficient management of scientific collections, enabling their integration into global repositories and expanding their potential for ecological and biogeographical research. This dissertation sought to standardize and automate data from the UFMG-AC Acarological Collection to address challenges concerning data fragmentation and inconsistency. The study implemented computational methodologies for data conversion and organization, ensuring compliance with the Darwin Core (DwC) standard. Scripts and pipelines were developed to structure metadata, correct taxonomic and geospatial inconsistencies, and integrate records into international platforms such as the Global Biodiversity Information Facility (GBIF) and the Brazilian Biodiversity Information System (SiBBr). Concurrently, Species Distribution Modeling (SDM) was applied to infer biogeographical patterns and predict potential occurrence areas for cataloged species. The EcoDistrib library was used to optimize environmental variable selection and machine learning-based model execution, while cross-validation techniques and statistical metrics ensured result robustness. Comparative analysis of data before and after standardization revealed significant improvements in quality, consistency, and accessibility, including taxonomic redundancy correction and enhanced spatial accuracy. The integration of enhanced data into GBIF and SiBBr increased the collection's visibility, facilitating its use in biogeographical and conservation studies. Distribution models identified relevant spatial patterns, highlighting gaps in records and suggesting new areas for future sampling. The research demonstrates that combining computational techniques with traditional curation approaches can significantly improve biological collection management, emphasizing the importance of automation and standardization for preserving, sharing, and applying biodiversity data. The findings reinforce the relevance of digitization and computational tools in modernizing scientific collections, advancing biogeography, ecology, and biodiversity conservation.

Keywords: Digitization of biological collections; Species distribution modeling; Darwin Core; Machine Learning; Data integration.

Lista de figuras

Lista de tabelas

Lista de abreviações

ABCD: Acesso a Dados de Coleções Biológicas

ANN: Artificial Neural Network

CCT: Centro de Coleções Taxonômicas

DwC: Darwin Core

GAM: Generalized Additive Models

GBIF: Global Biodiversity Information Facility

GLM: Generalized Linear Models

IABIN: Inter-American Biodiversity Information Network

INPA: Instituto Nacional de Pesquisa Amazônica

JBRJ: Jardim Botânico do Rio de Janeiro

MARS: Multivariate Adaptive Regression Splines

MNRJ: Museu Nacional do Rio de Janeiro

MPEG: Museu Paraense Emílio Goeldi

MZUSP: Museu de Zoologia da Universidade de São Paulo

OBIS: Ocean Biodiversity Information System

PCA: Principal Component Analysis

RF: Random Forest

SDM: Species Distribution Models

SiBBr: Sistema de informação sobre a Biodiversidade Brasileira

SVM: Support Vector Machine

TDWG: Bioinformatic Information Standards

UFMG-AC: Coleção de Ácaros do Centro de Coleções Taxonômicas

Sumário

1. Introdução

1.1. Coleções

1.1.1. Importância das coleções

O hábito de colecionar remonta a diferentes períodos históricos, sendo uma prática que atravessa culturas e contextos. As coleções biológicas, em particular, ganharam destaque na Europa a partir do século XV, com o surgimento dos Gabinetes de Curiosidades na Europa (Lima & Faleiro, 2020). Esses gabinetes reuniam uma ampla variedade de materiais, incluindo itens artísticos, biológicos, geológicos, astronômicos e arqueológicos. Frequentemente denominadas de "história natural", essas iniciativas não se limitavam a uma única área do conhecimento, mas abrangiam diversas disciplinas.

Entre as coleções mais antigas que conhecemos está a de Ulisses Aldrovandi (1522-1605), que incluía espécimes de herpetologia e foi documentada no século XVII (Bauer, Ceregato & Delfino, 2013). Outra coleção histórica é a da família Linck (Bauer & Wahlgren, 2013). Essas coleções pioneiras serviram como base para o desenvolvimento de museus de história natural e da pesquisa científica moderna.

No Brasil, o desenvolvimento das coleções biológicas seguiu um caminho diferente, influenciado por fatores históricos e políticos. Um marco importante foi a fundação do Museu Nacional do Rio de Janeiro (MNRJ) em 1818, por D. João VI, após a vinda da família real portuguesa para o Brasil. O MNRJ é a mais antiga instituição científica do país, consolidando-se como um pilar do avanço científico, acadêmico e cultural brasileiro. Seu acervo inclui nove coleções centenárias (Basílio et al., 2024), que preservam espécimes de inestimável valor histórico e científico, além de desempenhar um papel crucial na formação acadêmica e na pesquisa científica (MNRJ, 2025).

Outro exemplo relevante é o Museu Paraense Emílio Goeldi (MPEG), fundado em 1871. Com mais de 150 anos de história, o MPEG é o segundo museu mais antigo de história natural do Brasil e se destaca pela dedicação à preservação e ao estudo da biodiversidade amazônica. Desde sua fundação, tem sido um pilar fundamental na pesquisa científica sobre a fauna e flora da Amazônia, especialmente em um período em que naturalistas já exploravam a região, enviando espécimes para instituições no Brasil e no exterior (Santos, Aviz & Albuquerque, 2019).

Além do MNRJ e do MPEG, outras instituições brasileiras têm desempenhado papéis importantes na conservação e estudo da biodiversidade. O Jardim Botânico do Rio de Janeiro (JBRJ), fundado em 1808, abriga uma rica coleção de plantas nativas e

exóticas, destacando-se pela pesquisa e conservação da flora brasileira (JBRJ, 2025). O Museu de Zoologia da Universidade de São Paulo (MZUSP), criado em 1890, possui um acervo vasto de espécimes zoológicos que contribuem significativamente para o estudo da biodiversidade brasileira e global (MZUSP, 2025). O Instituto Butantan, estabelecido em 1901, é amplamente reconhecido por suas coleções de serpentes e aracnídeos, além de suas pesquisas em biomedicina (BUTANTAN,2025).

Mais recentemente, o Centro de Coleções Taxonômicas da Universidade Federal de Minas Gerais (CCT-UFMG), institucionalizado em 2015, consolidou-se como uma referência de armazenamento, catalogação e disponibilização de material biológico. Vinculado ao Instituto de Ciências Biológicas (ICB) da UFMG, o CCT gerencia 26 coleções, incluindo exemplares zoológicos, botânicos e micológicos, microbiológicos, bem como subamostras de tecidos, DNA e células em cultivo. Além de promover a pesquisa, o ensino e a inovação científica, o CCT também é ativo na formação de recursos humanos e na divulgação do conhecimento (CCT-UFMG, 2025).

As coleções zoológicas desempenham um papel fundamental na preservação e no estudo da biodiversidade, atuando como repositórios indispensáveis para o avanço científico e tecnológico. Além de registrarem a diversidade biológica, essas coleções constituem fontes inesgotáveis de dados para pesquisas científicas, tecnológicas e relacionadas à saúde pública (Zaher & Young, 2003; De Vivo, Silveira & Nascimento, 2014). No Brasil, essas coleções integram museus de história natural e instituições de ensino e pesquisa, sendo fundamentais para a preservação de espécimes ao longo do tempo, juntamente com dados biogeográficos, taxonômicos e de procedência (Zaher & Young, 2003). Sua função não se limita à conservação: as coleções também são centros para estudos moleculares, armazenando amostras de material genético, como DNA, tecidos e ossos, além de contribuir para a saúde pública ao identificar espécies vetores de doenças (De Vivo, Silveira & Nascimento, 2014).

O desenvolvimento e a ampliação de coleções científicas no Brasil são reflexos de políticas públicas e investimentos direcionados ao fortalecimento da pesquisa e da educação superior no país. Essas coleções desempenham um papel essencial na preservação da biodiversidade e na formação de taxonomistas especializados, além de serem fundamentais para o avanço do conhecimento em áreas como zoologia, botânica e microbiologia. O impacto dessas iniciativas é evidente ao se observar o crescimento das coleções nas últimas décadas, como descrito por Basílio e colaboradores (2024):

> "De acordo com os resultados, observa-se que houve um aumento do número de coleções no Brasil nos últimos 30 anos, sendo que 33,5% (114) dos acervos zoológicos possui entre 11 e 30 anos de existência. Esses números são coincidentes com os observados nas coleções botânicas e microbiológicas. Da mesma forma que para aquelas coleções, acredita-se que o número de coleções zoológicas nas últimas décadas tenha sido alavancado por iniciativas públicas que geraram subsídios financeiros e investimentos às instituições de ensino superior. Dentre as iniciativas, merecem destaque o programa de Reestruturação e Expansão das Universidades Federais (REUNI), com objetivo de ampliar o acesso e a permanência no ensino superior; e o Programa de Capacitação em Taxonomia (PROTAX), que visou o desenvolvimento e capacitação de taxonomistas nas três áreas, botânica, microbiologia e zoologia."

As coleções taxonômicas contribuem substancialmente para o progresso científico, apoiando estudos em sistemática, ecologia e conservação. Elas fornecem dados essenciais sobre morfologia, fisiologia e características do ciclo de vida, possibilitam pesquisas sobre distúrbios nos habitats e servem como registros históricos das mudanças ambientais globais (Suarez & Tsutsui, 2004; Wen et al., 2015; Castillo-Figueroa, 2018). Avanços recentes em bioinformática e infraestrutura cibernética aumentaram a acessibilidade e a utilidade dessas coleções, promovendo a colaboração e abrindo novas oportunidades de pesquisa (Wen et al., 2015).

1.1.2. Problemas das coleções

Apesar de desempenharem funções importantes na área de conservação, filogenia e outras áreas, as coleções enfrentam uma série de problemas que comprometem seu pleno potencial. Entre os principais desafios estão a falta de recursos financeiros, humanos e espaciais, além de infraestrutura inadequada. Adicionalmente, os curadores dessas coleções frequentemente enfrentam responsabilidades concorrentes, precisando equilibrar atividades de ensino, pesquisa e gestão das coleções (Snow, 2005; De Vivo, Silveira, & Nascimento, 2014).

No Brasil e na América Latina, regiões reconhecidas por sua biodiversidade, essas dificuldades são agravadas por cortes orçamentários, insuficiência de pessoal e entraves burocráticos, como a obtenção de permissões para coleta e transporte de espécimes. Esses fatores combinados têm um impacto negativo no avanço da taxonomia e nos estudos sobre biodiversidade (Rafael, Aguiar & Amorim, 2009; Glienke et al., 2024).

Além disso, eventos catastróficos têm revelado a vulnerabilidade dessas coleções. Um exemplo marcante foi o incêndio do MNRJ, que resultou na perda irreparável de

material tipo e táxons únicos (Zamudio, et al., 2018), destacando a necessidade de medidas preventivas e de digitalização para proteger os dados dos espécimes (Kellner, 2024). Tais desastres representam uma perda científica incalculável e reforçam a importância crítica dos esforços de digitalização, tanto para a preservação quanto para a acessibilidade futura das informações (Miller et al., 2020).

Outro problema significativo é a distribuição global desigual das coleções científicas. Enquanto algumas regiões concentram vastos acervos, outras enfrentam limitações em sua infraestrutura e legislação, dificultando a criação de novas coleções. Burocracias e políticas inadequadas também agravam essas disparidades (Kellner, 2024).

Um desafio adicional, particularmente relevante para a taxonomia, é o chamado "*shelf-life*", ou tempo de prateleira, que se refere ao intervalo entre a coleta de um espécime e sua descrição formal como uma nova espécie. Estudos apontam tempos médios de prateleira variados: 21 anos considerando todos os reinos (Fontaine et al., 2012), 16 anos para árvores neotropicais (Luján et al., 2024) e 19 anos para angiospermas do Cerrado (Cavallin et al., 2016). Diversos fatores contribuem para esses atrasos, incluindo características biológicas dos organismos, vieses sociais e geopolíticos e práticas de coleta (Fontaine et al., 2012; Luján, et al., 2024).

Colaborações entre coletores e autores, bem como revisões taxonômicas, têm sido associadas a tempos de prateleira mais curtos (Guedes et al., 2020). No entanto, desafios como instabilidade política e conflitos dificultam ainda mais o progresso. Essas questões ressaltam a necessidade de campanhas de coleta seguras e da promoção de colaborações internacionais para superar barreiras e acelerar o avanço taxonômico (Luján et al., 2024).

1.1.3. A coleção UFMG-AC

A Coleção Acarológica do Centro de Coleções Taxonômicas da Universidade Federal de Minas Gerais (CCT-UFMG), ou é UFMG-AC, possui mais de 16 mil itens tombados, incluindo indivíduos montados em lâminas e material preservado em álcool (Dados até outubro de 2023). Formalmente criada em 2012, inclui registros de coleta desde 1984, com cerca de 7,5% dos itens coletados antes de 2010. O tempo médio de espera para identificação é de 9 anos, e a descrição de novas espécies leva cerca de 2 anos.

A coleção abrange espécimes de diversas regiões do mundo, como Rússia (219), Espanha (200), Austrália (117), Nova Zelândia (94), Estados Unidos (72), Chile (62), Alemanha (53), Azerbaijão (39), Irã (28), Bolívia (12), Panamá (12), Peru (6), Myanmar (5), Cuba (2), Guiana Francesa (2), Honduras (2), Tajiquistão (2) e Equador (1). Além de ácaros marinhos, coletados em erupções vulcânicas de grandes profundidades no Oceano

Índico e na Antártida. No Brasil, todas as unidades federativas estão representadas, exceto Amapá, Distrito Federal, Goiás, Roraima e Tocantins.

Atualmente, são registradas 141 famílias de ácaros, com destaque para aquelas com mais de 200 indivíduos: Halacaridae (1865), Erythraeidae (1084), Spinturnicidae (487), Erythracaridae (445), Trhypochthoniidae (358), Macronyssidae (343), Laelapidae (330), Microtrombidiidae (277), Leeuwenhoekiidae (241), Smarididae (227), Rhagidiidae (210), Trombiculidae (200). A coleção também possui material tipo, consistindo em 20 holótipos, 299 parátipos e 1 neótipo, além de 3711 indivíduos que são material testemunho de trabalhos que envolveram extração e sequenciamento de DNA para trabalhos filogenéticos ou filogeográficos, os quais já produziram mais de 1500 sequências armazenadas em bases públicas de dados, como o GenBank e o BOLD.

Parte significativa dos exemplares é oriunda das atividades de coleta do Laboratório de Acarologia do Departamento de Zoologia da UFMG. Os ácaros oriundos das atividades de coletas de outros laboratórios ou de empresas passam pela triagem e eventualmente preparação em lâminas neste mesmo laboratório. O fluxo de trabalho envolve as seguintes etapas:

1- Coleta, adequada aos diferentes ambientes, incluindo inclusive a captura seguida de soltura ou coleta de hospedeiros. As amostras têm sua origem em diferentes ambientes, incluindo marinhos da região entre-marés até grandes profundidades, cavernícolas, de água doce e serapilheira, associados ou não a outros organismos. Para garantir a preservação do DNA, as amostras coletadas são armazenadas em álcool antes de serem transportadas para o laboratório.

2 - No laboratório, a triagem das amostras é conduzida para selecionar e organizar o material a ser analisado, separando os ácaros de outros grupos de animais, do substrato e identificados até onde é possível, sob o aumento permitido pelo estereomicroscópio.

3- Parte dos ácaros, a depender do interesse pelo grupo ou disponibilidade de reagentes, tem seu DNA extraído e mantido em coleção de extratos de DNA a -80°C para a posterior obtenção de sequências. Os extratos são numerados sequencialmente, até a escrita desta dissertação haviam mais de 4000 extratos mantidos no freezer do laboratório.

4- Os ácaros muito grandes ou muito esclerotizados são mantidos em via úmida (álcool 70-100%). Os demais são montados em lâminas, no caso do Laboratório de Acarologia em meio de Hoyers ou de Gelatina-Glicerina.

5- Os ácaros são identificados e eventualmente descritos, muitas vezes utilizando uma abordagem integrativa, que combina dados morfológicos e moleculares. As lâminas

recebem etiquetas e os metadados são incluídos na planilha da coleção, que a intervalos de tempos regulares são submetidos a bases públicas (SiBBr e GBIF).

Durante todo o processo, são coletados e registrados dados detalhados sobre cada etapa, desde a coleta inicial até a extração molecular e a identificação morfológica. Esses registros são fundamentais para garantir a integridade dos dados e permitir futuras análises comparativas, contribuindo para a ampliação do conhecimento acarológico e sua aplicação em estudos científicos.

1.1.4. Dados de ocorrências

Dados de ocorrência de espécies referem-se a observações de presença ou abundância de espécies em locais específicos, desempenhando um papel crucial na pesquisa em biodiversidade. Provenientes de fontes como museus e herbários, esses dados estão se tornando cada vez mais acessíveis por meio de bancos de dados online, contribuindo para diversas aplicações em ecologia e conservação (Ball-Damerow et al., 2019). Eles são utilizados para estimar a riqueza de espécies, desenvolver inventários e descrever novas espécies. No entanto, persistem desafios relacionados ao controle de qualidade dos dados e à abordagem de vieses intrínsecos em conjuntos de dados (Ball-Damerow et al., 2019).

Intrinsecamente ligados às coleções científicas, os dados de ocorrência funcionam como registros indispensáveis associados aos espécimes armazenados. Eles incluem informações detalhadas sobre a localidade, data e condições de coleta, além da classificação taxonômica e outros atributos relevantes. Dessa forma, não apenas documentam a diversidade biológica, mas também servem como base para estudos fundamentais em áreas como distribuição de espécies, biogeografia e mudanças ambientais (Speed, 2018; Guralnick & Hill, 2009). As coleções científicas, como as zoológicas, frequentemente contêm maior riqueza de espécies do que os dados observacionais e fornecem registros verificáveis, essenciais para rastrear a origem das observações relatadas (Marinoni et al., 2024).

As coleções zoológicas, como discutido anteriormente, não apenas preservam os espécimes físicos, mas também geram e armazenam um grande volume de informações associadas. Esses registros são essenciais para análises que orientam o manejo e a conservação da biodiversidade, bem como para a formulação de políticas ambientais e o planejamento de estratégias de mitigação de impactos climáticos (Arengo et al., 2017; Benham & Bowie, 2023). Além de fornecerem dados temporais e espaciais únicos, essas coleções permitem o acompanhamento das mudanças na diversidade biológica global ao

longo do tempo, sendo cruciais para o entendimento de padrões biogeográficos e impactos ambientais (Monfils et al., 2017). A digitalização de espécimes tem expandido significativamente o acesso às informações de biodiversidade, contribuindo para a resolução de questões globais, regionais e locais (Hilton et al., 2021).

Apesar de sua relevância, os dados de ocorrência frequentemente enfrentam desafios relacionados à sua padronização e qualidade. Problemas como coordenadas geográficas incorretas, nomenclatura taxonômica desatualizada e a duplicação de registros são problemas comuns que podem comprometer os resultados das análises (Ball-Damerow et al., 2019). Iniciativas como as do Global Biodiversity Information Facility (GBIF) têm trabalhado para melhorar a integração e a acessibilidade dos dados, potencializando sua aplicação científica (Heberling et al., 2021).

A integração de fontes de dados não-convencionais, como informações de "táxons associados" extraídas de registros digitalizados, pode ampliar os registros de ocorrência conhecidos e melhorar a compreensão da distribuição dos táxons (Pearson, 2018). Contudo, a heterogeneidade entre coleções e a falta de uniformidade no registro das informações dificultam a integração de dados em larga escala, limitando seu potencial de aplicação científica (Ball-Damerow et al., 2019).

Esses dados são utilizados principalmente na modelagem de distribuição de espécies (SDM, do inglês Species Distribution Models), que combinam dados de ocorrência com variáveis ambientais para prever as distribuições das espécies no espaço e no tempo (Elith & Leathwick, 2009). As SDMs fornecem "*insights*" ecológicos e orientam o planejamento de conservação (Guisan & Thuiller, 2005). Novos métodos de modelagem, incluindo abordagens de aprendizado de máquina, têm se mostrado promissores na melhoria das previsões a partir de dados apenas de presença, que são típicos de muitos conjuntos de dados de ocorrência (Elith et al., 2006).

1.1.5. Digitalização

A digitalização e a informatização são processos distintos, mas complementares, na pesquisa sobre biodiversidade. A digitalização envolve a conversão de espécimes físicos e dados associados em formatos digitais, aumentando a acessibilidade e o uso de coleções de história natural (Beaman & Cellinese, 2012; Hedrick et al., 2020). A digitalização oferece diversas vantagens, como a disponibilização das informações para a comunidade de especialistas e para o público em geral, permitindo que pesquisadores priorizem seus esforços em coleções de maior relevância. Eventualmente, a depender de documentos digitais dos exemplares depositados, reduz custos eliminando a necessidade

de visitas presenciais a todas as instituições, o que quebra barreiras geográficas ou ao menos torna as visitas mais focadas. Outra vantagem é a preservação dos dados em caso de tragédias, como incêndios ou inundações, e a maior eficiência na gestão de coleções, facilitando a organização e catalogação dos espécimes (Marinoni et al., 2024).

O processo de digitalização inclui a digitação de informações de etiquetas, livros de tombo e cadernos de campo em planilhas ou formulários digitais, a captura de imagens ou fotografias dos espécimes, e a conversão de sons de fitas ou outras mídias para o formato digital (Marinoni et al., 2024).

Esse processo tem levado à criação de grandes bancos de dados e agregadores de biodiversidade, impactando significativamente a pesquisa em sistemática, ecologia e conservação (Nelson & Ellis, 2019). Por outro lado, a informatização refere-se à transformação mais ampla de processos e estratégias por meio do uso de tecnologias digitais (Gobble, 2018). No contexto da biodiversidade, isso inclui o desenvolvimento de novos fluxos de trabalho exclusivamente digitais e abordagens automatizadas para a descoberta e conservação da biodiversidade (Hedrick et al., 2020). Juntos, esses processos revolucionaram a pesquisa em biodiversidade, melhorando o acesso às informações essenciais, aumentando a relevância das coleções científicas e permitindo avaliações de biodiversidade em larga escala (Beaman & Cellinese, 2012; Nelson & Ellis, 2019).

Além dos projetos de formação e capacitação de pessoal, houve também iniciativas voltadas especificamente para a digitalização, como o INCT Herbário Virtual da Flora e dos Fungos e o REFLORA, que permitiram que as coleções botânicas parceiras mantenham suas coleções digitalizadas. Para as coleções zoológicas, o Ministério da Ciência e Tecnologia (MCTI) coordenou um projeto que fomentou a digitalização em grandes instituições, como o Museu de Zoologia da Universidade de São Paulo (MZUSP), o Museu Paraense Emílio Goeldi (MPEG), o Instituto Nacional de Pesquisas da Amazônia (INPA) e o Museu Nacional do Rio de Janeiro (MNRJ). Outras iniciativas globais incluem a criação de bancos de biodiversidade, alguns para todos os grupos, como o GBIF e o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), que abrangem todos os grupos taxonômicos, além de plataformas específicas, como o JABOT (para botânica) e o microSICol (para microbiologia) (Marinoni et al., 2024).

O progresso da digitalização varia entre as coleções. Por exemplo, a maioria das coleções botânicas já possui grande parte de seus materiais digitalizados, com apenas 8,6% ainda não iniciando o processo. Em contraste, as coleções microbiológicas e

zoológicas estão em estágios semelhantes de digitalização, com um número total de itens digitalizados significativamente menor que em coleções botânicas (Marinoni et al., 2024).

Apenas 12,4% das coleções zoológicas e 8,3% das coleções microbiológicas possuem imagens dos espécimes (Marinoni et al., 2024). Isso é importante porque as imagens dos exemplares potencialmente evitariam a manipulação excessiva dos exemplares, contribuindo com a sua conservação e, como mencionado acima, economizaria os custos associados às visitas às coleções.

Os bancos de dados utilizados nas coleções são fundamentais, pois determinam como essas informações são acessadas tanto por colaboradores internos quanto por pesquisadores externos. Por exemplo, quando os dados são armazenados em planilhas na nuvem, é necessário conceder acesso às informações para pessoas externas. Já em bases de dados públicas, esse acesso é mais direto. Além disso, o acesso pode ser de leitura ou edição, dependendo das permissões concedidas. Embora os bancos de dados sejam ferramentas poderosas para o armazenamento e gestão de informações, eles apresentam desafios, como a menor praticidade em comparação com as planilhas locais ou na nuvem, que são mais familiares para a maioria dos usuários. Por isso, é essencial cursos de capacitação para os curadores e colaboradores, garantindo que possam utilizar essas ferramentas de forma eficiente.

1.2. Padronização

1.2.1. Principais características

Dados padronizados são caracterizados por uniformidade no formato, nomenclatura consistente e a inclusão de metadados claros que contextualizam as informações. Essa padronização permite que diferentes conjuntos de dados sejam interpretados e utilizados de forma integrada, mesmo que tenham sido produzidos por instituições distintas ou em diferentes contextos. A uniformidade no formato envolve o uso de estruturas predefinidas, como tabelas organizadas e campos bem definidos. A nomenclatura consistente assegura que os termos utilizados sejam reconhecidos internacionalmente, como nos nomes científicos das espécies. Já os metadados documentam as condições de coleta, as fontes de dados e outras informações essenciais para garantir a transparência e a reprodutibilidade.

A ausência de padronização nos dados de biodiversidade pode gerar uma série de problemas, incluindo:

- **Redundâncias**: Dados duplicados ou inconsistentes entre diferentes coleções dificultam análises precisas.
- **Erros de Interpretação**: Informações registradas de forma heterogênea podem levar a conclusões incorretas. Por exemplo, a falta de padronização em coordenadas geográficas pode causar erros no mapeamento de espécies.
- **Incompatibilidade**: Dados sem formato uniforme não podem ser integrados com outros sistemas, limitando seu uso em análises globais.

Esses problemas destacam a importância de iniciativas que promovam a padronização e a integração de dados, especialmente no contexto de ciência aberta e colaborativa. Apenas com a adoção de padrões internacionais será possível explorar todo o potencial dos dados de biodiversidade para a conservação e o desenvolvimento sustentável.

A padronização dos dados de biodiversidade é crucial para a integração e interpretação de conjuntos de dados provenientes de diversas fontes. Ela garante a uniformidade no formato, nomenclatura consistente e metadados claros (Zenglein, 2025). Estruturas de dados padronizadas, como a lista de conceitos do Darwin Core (DwC) e o esquema ABCD, aprimoram a acessibilidade, qualidade e usabilidade dos bancos de dados (Willemse, 2008). Convenções de nomenclatura reconhecidas internacionalmente, como os nomes científicos das espécies, são essenciais para alcançar a consistência dos dados. Os metadados desempenham um papel vital na documentação das condições de coleta e das fontes de dados, assegurando transparência e reprodutibilidade (Zenglein, 2025). Organizações como o GBIF e a Biodiversity Information Standards (TDWG) estão liderando esforços para superar barreiras na gestão de dados de biodiversidade, desenvolvendo vocabulários padronizados de metadados e abordando desafios relacionados à ambiguidade de nomes taxonômicos (Lapp et al., 2011). A implementação desses padrões é fundamental para a gestão de dados de biodiversidade a longo prazo, facilitando projetos de pesquisa complexos (Zenglein, 2025).

A falta de padronização nos dados de biodiversidade apresenta desafios significativos para pesquisas e esforços de conservação. Formatos de dados heterogêneos, diferenças no nível de detalhamento dos dados e terminologias inconsistentes dificultam a integração dos dados e limitam a usabilidade em análises dos dados (König et al., 2019). Essas diferenças podem ocorrer na precisão das coordenadas geográficas, na frequência das coletas ao longo do tempo ou na categorização taxonômica, tornando a comparação entre diferentes conjuntos de dados mais complexa.

Além disso, problemas como registros duplicados, datas de coleta ausentes e variações na exatidão das localizações geográficas afetam entre 35% e 55% dos registros em grandes bases de dados, como o GBIF e o Ocean Biodiversity Information System (OBIS) (Moudrý & Devillers, 2020). A incompatibilidade entre sistemas taxonômicos de diferentes bases de dados complica mais os esforços de integração (Feng, et al., 2022).

Esses desafios impactam diretamente a precisão e a eficácia dos estudos de biodiversidade, podendo levar a interpretações equivocadas sobre a distribuição de espécies e a falsos positivos de "*hotspots*" de biodiversidade (Moudrý & Devillers, 2020). Abordar essas questões exige melhorias nos mecanismos de entrada de dados, rotinas de controle de qualidade e fortalecer o intercâmbio de informações entre agregadores (Moudrý & Devillers, 2020). A integração eficaz das bases de dados existentes poderia avançar significativamente o conhecimento sobre biodiversidade, potencialmente aumentando a cobertura taxonômica em bancos de dados (Feng, et al., 2022).

As iniciativas de padronização e integração de dados são cruciais para o avanço da ciência da biodiversidade aberta e colaborativa. A adoção de padrões internacionais aprimora a transparência, a reprodutibilidade e a integração global dos dados de biodiversidade (Zenglein, 2025). Esses esforços desbloqueiam todo o potencial dos dados de biodiversidade para a conservação e o desenvolvimento sustentável, permitindo análises precisas, escaláveis e interdisciplinares (Heberling et al., 2021). Iniciativas globais como GBIF, a Rede Interamericana de Informações sobre a Biodiversidade (IABIN) e o speciesLink promovem a integração de dados por meio de protocolos, padrões e arquiteturas abertas. A acessibilidade às informações de biodiversidade aumentou significativamente, com o uso de dados mediados pelo GBIF, crescendo desde 2007 (Heberling et al., 2021). No entanto, ainda existem desafios para alinhar esforços locais e globais na criação de um repositório comum de dados de biodiversidade. Enfrentar esses desafios por meio de ações coordenadas, como compartilhamento de recursos e harmonização de atividades operacionais, é essencial para realizar todo o potencial dos dados integrados de biodiversidade.

1.2.2. Conceitos e importância da padronização

Em um cenário no qual os dados sobre biodiversidade são gerados por diversas fontes, como coleções biológicas, projetos de monitoramento e registros de observação, a padronização se torna fundamental para possibilitar a integração dessas informações. Estabelecer normas e critérios uniformes para o registro, organização e compartilhamento

de dados permite que pesquisadores, instituições e sistemas computacionais trabalhem de maneira coesa e estruturada (Wieczorek et al., 2012).

Um dos principais objetivos da padronização é minimizar erros e inconsistências, especialmente aqueles decorrentes de registros manuais ou da falta de uniformidade entre diferentes bancos de dados (Chapman, 2005). Ela também amplia a escala e a abrangência das análises científicas ao possibilitar a unificação de dados em sistemas centralizados e ampliando a escala e a abrangência das análises científicas (Zipkin et al., 2021), além de garantir sua reutilização ao longo do tempo, independentemente de mudanças tecnológicas ou institucionais (Wilkinson et al., 2016).

A padronização baseia-se em princípios como clareza, consistência e interoperabilidade, assegurando que o significado dos dados permaneça inequívoco. O uso de metadados bem definidos fornece o contexto necessário para sua compreensão e reaproveitamento, enquanto taxonomias e ontologias padronizadas contribuem para a organização lógica e a compatibilidade com sistemas automatizados (Smith et al., 2013).

No contexto da ciência global, padrões internacionais, como o Darwin Core (DwC), são amplamente utilizados para estruturar dados biológicos, viabilizando a interoperabilidade entre coleções científicas e repositórios, como o GBIF (Wieczorek et al., 2012; GBIF 2025; TDWG 2025). Essa abordagem facilita a colaboração internacional, essencial para iniciativas como o monitoramento da biodiversidade e a modelagem de mudanças climáticas (Hardisty et al., 2019; Pereira et al., 2013).

Com o crescimento exponencial dos dados científicos, a padronização torna-se ainda mais crítica no contexto da *big data*. Repositórios como o GBIF e o SiBBr exemplificam sua importância ao reunir milhões de registros em plataformas unificadas, permitindo a aplicação de algoritmos para identificar padrões ecológicos e apoiar políticas ambientais baseadas em evidências (Hampton et al., 2013; Jetz et al., 2019).

Além de aumentar a confiabilidade e replicabilidade das análises científicas (Borer et al., 2009), a padronização fornece uma base robusta para políticas públicas e estratégias de conservação, permitindo a identificação de espécies ameaçadas e a implementação de medidas de proteção eficazes (Tittensor et al., 2014). Também promove maior eficiência no compartilhamento de dados entre instituições, otimizando recursos e acelerando o avanço do conhecimento (Reichman et al., 2011).

Por fim, ao estabelecer diretrizes claras para documentação e disseminação de dados, a padronização fortalece a ciência aberta, promovendo transparência, colaboração e confiança na pesquisa científica (Tenopir et al., 2011; BRASIL, 2021). Assim,

consolida-se como ferramenta indispensável para enfrentar desafios globais, como a perda de biodiversidade e as mudanças climáticas (IPCC, 2023).

1.2.3. Bancos de dados públicos

Os bancos de dados que utilizam formatos padronizados desempenham um papel crucial no compartilhamento de informações sobre biodiversidade, promovendo a acessibilidade e a integração de dados em escala global. Dentre os principais estão:

1. **GBIF (Global Biodiversity Information Facility)**: É uma das maiores plataformas globais para compartilhamento de dados sobre biodiversidade. O GBIF utiliza padrões como o DwC, que define formatos consistentes para a troca de informações sobre ocorrências de espécies, registros taxonômicos e metadados ambientais. Essa padronização permite que instituições em todo o mundo publiquem dados de maneira uniforme, facilitando análises comparativas e colaborativas (GBIF, 2025).
2. **SiBBr (Sistema de Informação sobre a Biodiversidade Brasileira)**: É o principal repositório de dados sobre biodiversidade no Brasil, alinhado aos padrões do GBIF e do DwC. Ele reúne dados de coleções científicas brasileiras e registros de campo, integrando informações regionais ao panorama global. Além disso, o SiBBr desempenha um papel importante na valorização e visibilidade da biodiversidade brasileira (SiBBr, 2025).
3. **iDigBio (Integrated Digitized Biocollections)**: Localizado nos Estados Unidos, o iDigBio é focado na digitalização e disponibilização de dados de coleções científicas. Ele trabalha para garantir que as informações de espécimes sejam padronizadas e acessíveis para uma ampla gama de usuários, desde pesquisadores até o público em geral (iDigBio, 2025).
4. **SpeciesLink**: Promove o acesso livre e aberto a dados sobre biodiversidade, permitindo que qualquer indivíduo ou grupo utilize suas informações. Os dados compartilhados seguem diretrizes que garantem a confidencialidade, e seu uso é de responsabilidade do usuário. A plataforma facilita a pesquisa, a educação e a formulação de políticas ambientais, conectando coleções regionais a redes globais de conhecimento (SpeciesLink, 2025).

Esses bancos de dados permitem o acesso livre e amplo a informações sobre biodiversidade, conectando coleções regionais e institucionais às redes globais de pesquisa, o que é fundamental para a formulação de políticas ambientais, a conservação de espécies e a avaliação dos impactos das mudanças climáticas. O uso de formatos

padronizados, como o DwC, assegura a interoperabilidade entre diferentes plataformas, permitindo que os dados sejam agregados e analisados em escala global.

A integração de dados provenientes de coleções regionais enfrenta desafios como infraestrutura limitada e recursos financeiros insuficientes, o que dificulta os esforços de digitalização, especialmente em países em desenvolvimento (Monda, 2019; Angeles & Catap, 2022). Essa lacuna se agrava ainda mais quando se considera a concentração dos acervos científicos – sobretudo dados genéticos e morfológicos de fauna – em instituições de países desenvolvidos. Collen et al. (2008) apontam que, enquanto essas nações dispõem de recursos e infraestrutura robusta para a coleta e preservação dos dados, as regiões tropicais, que abrigam aproximadamente 80% da biodiversidade global, carecem de capacidade para conservar e explorar recursos críticos, como linhagens de espécies-chave para estudos ecológicos ou biomédicos. Além disso, o Protocolo de Nagoya, que regula o acesso a recursos genéticos e a repartição justa e equitativa dos benefícios decorrentes de sua utilização, impõe barreiras regulatórias e burocráticas. Essas restrições, que variam entre os países, podem dificultar o acesso e a integração de dados genéticos, limitando a abrangência de grandes bases de dados, como o GBIF e o SiBBr, e, consequentemente, restringindo a aplicação de algoritmos avançados para identificar padrões ecológicos e tendências espaciais e temporais, essenciais para a definição de áreas prioritárias para conservação e para a formulação de políticas ambientais baseadas em evidências (Jetz et al., 2019).

Enfrentar esses desafios exige redes de colaboração, melhorias no gerenciamento de dados e o fortalecimento da infraestrutura local (Jetz et al., 2019; Monda, 2019). Uma abordagem descentralizada, permitindo que comunidades controlem suas próprias infraestruturas locais de dados enquanto coordenam globalmente por meio de acordos de compartilhamento, tem sido proposta para lidar com questões de qualidade de dados e aumentar o envolvimento de especialistas (Sterner et al., 2020). Isso se alinha ao conceito de uma aliança pelo conhecimento da biodiversidade, que busca fortalecer a colaboração e desenvolver soluções compartilhadas para a informática da biodiversidade (Hardisty & Roberts, 2013). Investimentos em infraestrutura, treinamento técnico e colaboração internacional são fundamentais para melhorar o acesso e a integração dos dados. Os esforços para digitalizar e integrar informações sobre espécies e espécimes provenientes de coleções científicas podem aumentar significativamente o valor dos dados de biodiversidade existentes (Nelson & Ellis, 2018).

Apesar dos avanços, incluir dados de coleções regionais ou menores ainda apresenta desafios. Muitas dessas coleções não possuem infraestrutura ou recursos financeiros suficientes para digitalização e padronização, além da falta de treinamento técnico para adesão aos padrões globais. Outro problema é a sub-representação de dados de áreas tropicais e países em desenvolvimento, que abrigam grande parte da biodiversidade mundial, mas enfrentam limitações na publicação e compartilhamento de dados (Marinoni et al., 2024).

Superar esses desafios requer investimentos em infraestrutura, capacitação técnica e colaboração internacional para garantir que dados regionais sejam integrados de forma eficaz às redes globais, enriquecendo a base de informações disponíveis para a ciência e a conservação.

1.2.4. Formato Darwin Core

O Darwin Core (DwC) é um conjunto de padrões amplamente reconhecido na área de informática da biodiversidade, criado com o objetivo de facilitar o compartilhamento e a interoperabilidade de dados relacionados à biodiversidade. Desenvolvido e mantido pelo Biodiversity Information Standards (TDWG), o DwC fornece um vocabulário padronizado que permite a descrição consistente de informações como ocorrência de espécies, taxonomia, eventos de coleta e características ambientais. Sua ampla adoção deve-se à simplicidade e flexibilidade do formato, tornando-o compatível com diversas estruturas de dados, incluindo planilhas, bancos de dados relacionais e arquivos XML.

A padronização promovida pelo Darwin Core é essencial para aprimorar a interoperabilidade entre sistemas e favorecer a compreensão dos padrões globais de biodiversidade. Com o passar do tempo, o padrão evoluiu para lidar com a crescente complexidade dos dados biológicos e permitir a integração de fontes heterogêneas, conforme descrito por Wieczorek et al. (2012). Essa evolução fortalece a confiabilidade, qualidade e usabilidade dos dados, pilares fundamentais para a consolidação da informática da biodiversidade como campo científico e aplicado.

Entre as principais características do Darwin Core, destaca-se o uso de termos padronizados e bem documentados para representar elementos essenciais dos dados biológicos. Por exemplo, no contexto de dados de ocorrência, são utilizados termos como *scientificName*, *decimalLatitude*, *decimalLongitude* e *eventDate*. No âmbito taxonômico, o padrão inclui campos como *kingdom*, *phylum*, *class*, *order*, *family*, *genus* e *specificEpithet*. Para representar eventos de coleta, são definidos termos como *collector*,

samplingProtocol e *habitat*. Essa estrutura facilita a interpretação dos dados por humanos e máquinas, além de permitir validações automáticas e reuso eficiente das informações.

Outro aspecto importante do Darwin Core é sua compatibilidade e extensibilidade. O padrão foi projetado para ser facilmente integrado a outros sistemas e bancos de dados, permitindo a adição de termos personalizados por meio de extensões específicas, quando necessário. Essa característica amplia seu alcance, possibilitando adaptações conforme as necessidades de diferentes projetos e instituições. Além disso, sua facilidade de implementação contribui para a adoção por instituições com variados níveis de infraestrutura técnica.

Na prática, o Darwin Core é amplamente utilizado. Um exemplo notável é o GBIF (Global Biodiversity Information Facility), que adota o formato DwC para integrar dados de biodiversidade provenientes de instituições de todo o mundo, criando um repositório global unificado com milhões de registros. No Brasil, o SiBBr (Sistema de Informação sobre a Biodiversidade Brasileira) também utiliza o padrão para consolidar dados de coleções regionais, permitindo análises e políticas de conservação em escala nacional. Outro caso é o iDigBio, nos Estados Unidos, que centraliza dados de coleções biológicas digitalizadas com base no DwC, organizando informações taxonômicas e de coleta. Além disso, pesquisadores frequentemente utilizam o padrão como base para preparar dados de ocorrência que serão aplicados em algoritmos de modelagem de distribuição de espécies.

Apesar de sua importância e ampla aceitação, o uso do Darwin Core não está isento de limitações e desafios. Uma das principais críticas está relacionada à ausência de um controle de qualidade intrínseco: embora os termos sejam padronizados, o padrão não garante a precisão dos dados subjacentes, como coordenadas geográficas incorretas ou erros na identificação taxonômica. Além disso, certos aspectos da biodiversidade, como interações entre espécies ou metadados ambientais mais complexos, não são totalmente contemplados no escopo original do padrão, exigindo o uso de extensões ou outros padrões complementares. Pequenas instituições, muitas vezes com recursos técnicos e financeiros limitados, podem enfrentar barreiras na adoção completa do DwC, especialmente quando dependem de sistemas legados ou dados armazenados em formatos não padronizados, como registros físicos ou planilhas não estruturadas.

Em suma, o Darwin Core constitui um dos pilares fundamentais para a gestão, análise e compartilhamento de dados de biodiversidade em escala global. Sua adoção promove a uniformidade e a interoperabilidade necessárias para estudos colaborativos e análises abrangentes, fundamentais em um contexto de crescente demanda por dados

científicos acessíveis e confiáveis. Apesar dos desafios que ainda persistem, o padrão continua sendo um elemento estratégico em iniciativas voltadas à conservação, pesquisa e políticas ambientais baseadas em evidências.

1.3. Modelagem de distribuição de espécie

1.3.1. Histórico e evolução

A modelagem de distribuição de espécies (SDM, do inglês *Species Distribution Modeling*) é uma abordagem que prevê a distribuição geográfica de uma espécie com base em dados ambientais, dados de ocorrência e modelos matemáticos. Esses modelos são fundamentais para entender como as espécies interagem com seu ambiente e para prever mudanças em sua distribuição devido a fatores como mudanças climáticas, perda de habitat e invasões biológicas (Peterson, et al., 2011).

Os primeiros métodos de SDM surgiram na década de 1970, com técnicas computacionais que analisavam o impacto das variáveis ambientais na distribuição das espécies. Um dos métodos pioneiros foi o BIOCLIM, desenvolvido para modelar o nicho climático das espécies. O BIOCLIM utiliza uma abordagem de envelope climático, definindo um hipervolume n-dimensional baseado nas condições ambientais dos locais de ocorrência conhecidos (Booth et al., 2014). Esse método, embora simples, foi amplamente utilizado em estudos paleoecológicos e biogeográficos, mas apresentou limitações em previsões sobre mudanças climáticas, devido à sua sensibilidade a outliers e à falta de flexibilidade para capturar relações não-lineares (Elith, et al., 2006).

Um marco significativo na evolução dos SDMs foi o desenvolvimento do Maxent (Máxima Entropia), um algoritmo que se tornou popular por sua eficiência e facilidade de uso. Diferente dos métodos baseados em envelope, o Maxent utiliza dados de presença e "*background*" (pseudo-ausências) para estimar a distribuição de probabilidade de ocorrência de uma espécie. Ele é particularmente útil quando dados de ausência são escassos ou indisponíveis (Phillips, Anderson & Schapire, 2006). O Maxent se destaca por sua capacidade de lidar com grandes volumes de dados ambientais e por sua flexibilidade em modelar relações complexas entre variáveis preditoras e a ocorrência de espécies (Merow, Smith, & Silander Jr, 2013).

Com o avanço da tecnologia, métodos baseados em aprendizado de máquina ("*machine learning*") ganharam destaque na modelagem de distribuição de espécies. Algoritmos como *Random Forest* (RF), *Support Vector Machines* (SVM) e *Artificial Neural Network* (ANN) permitem capturar padrões complexos e não-

lineares nos dados, superando muitas limitações dos métodos tradicionais (Elith, Leathwick & Hastie, 2008). Além disso, a integração de "*big data*" e sensoriamento remoto tem ampliado a capacidade de coletar e processar grandes volumes de dados ambientais e de ocorrência, permitindo modelagens em escalas globais e com alta resolução (Guisan, Thuiller & Zimmermann, 2017).

Apesar da diversidade de algoritmos disponíveis, o Maxent permanece como o método mais utilizado na modelagem de distribuição de espécies. Sua popularidade se deve à sua facilidade de implementação, à disponibilidade de pacotes computacionais como o *dismo* em R, e à sua eficácia em lidar com dados de presença única (presença-background) (Merow, Smith & Silander Jr, 2013). Outros métodos, como RF e SVM, embora poderosos, exigem maior esforço computacional e conhecimento técnico para ajuste de hiperparâmetros e validação, o que pode limitar sua aplicação em estudos com recursos limitados (Elith, Leathwick & Hastie 2008).

Principais Marcos Históricos

- **Década de 1970**: Primeiro método de SDM, o BIOCLIM (Booth et al., 2014).
- **Década de 2000**: Popularização do Maxent como ferramenta principal para modelagem de nicho ecológico (Phillips, Anderson, & Schapire, 2006).
- **Década de 2010**: Integração de "*big data*", sensoriamento remoto e aprendizado de máquina na modelagem de distribuição de espécies (Guisan, Thuiller & Zimmermann, 2017).
- **Década de 2020:** Avanço na supercomputação e modelos de "*ensemble*" para previsões globais, como o uso de plataformas integradas e abordagens multi-algoritmos (Zurell et al., 2020; Hao et al., 2020).

1.3.2. Relevância do SDM

A modelagem de distribuição de espécies (SDM) tem se mostrado uma ferramenta indispensável para a conservação da biodiversidade, oferecendo "*insights*" valiosos sobre a distribuição geográfica de espécies e suas respostas às mudanças ambientais. Esses modelos são particularmente úteis para espécies ameaçadas de extinção ou consideradas extintas, pois podem identificar áreas potenciais de ocorrência com base em condições ambientais adequadas. Um exemplo emblemático é o caso da perereca *Phyllomedusa ayeaye*, uma espécie rara e ameaçada no sudeste do Brasil. Com base em apenas três registros de ocorrência conhecidos, pesquisadores aplicaram a modelagem de nicho ecológico para gerar mapas preditivos da distribuição dessa espécie, direcionando novos

levantamentos para áreas indicadas pelo modelo (Giovanelli et al., 2008). Esse tipo de aplicação demonstra como os SDMs podem direcionar esforços de conservação e subsidiar políticas públicas voltadas para a proteção de espécies ameaçadas.

Os SDMs são amplamente utilizados para identificar áreas prioritárias para conservação, especialmente em regiões com alta biodiversidade e endemismo. Eles permitem mapear habitats adequados para espécies ameaçadas, auxiliando na criação de unidades de conservação e corredores ecológicos (Guisan, et al., 2013). Além disso, os modelos podem prever como as mudanças climáticas afetarão a distribuição de espécies, permitindo a adoção de medidas proativas para mitigar impactos (Bellard et al., 2012).

Os SDMs são ferramentas essenciais para estudar os impactos das mudanças climáticas sobre a biodiversidade. Eles permitem projetar como a distribuição das espécies pode mudar em cenários futuros de aquecimento global, identificando áreas que podem se tornar inóspitas ou, ao contrário, novas áreas de potencial colonização (Thuiller et al., 2005). Por exemplo, modelos de distribuição têm sido utilizados para prever o deslocamento de espécies de aves e mamíferos em direção a latitudes mais altas ou altitudes maiores, em resposta ao aumento das temperaturas (Parmesan & Yohe, 2003).

Além de suas aplicações práticas, os SDMs são amplamente utilizados em estudos teóricos, como biogeografia, ecologia evolutiva e dinâmica de populações. Eles permitem testar hipóteses sobre os fatores que limitam a distribuição das espécies e como esses fatores interagem ao longo do tempo (Soberon & Peterson, 2005). Por exemplo, estudos utilizando SDMs têm contribuído para entender como eventos históricos, como glaciações, moldaram a distribuição atual de espécies (Nogués-Bravo, 2009).

1.3.3. Seleção de variáveis ambientais

A seleção criteriosa de variáveis ambientais é um dos aspectos fundamentais para a construção de SDMs, pois influencia diretamente a precisão das predições e a interpretação ecológica dos resultados. A escolha inadequada de variáveis pode levar a inferências equivocadas sobre os fatores que determinam a distribuição das espécies, reduzindo a aplicabilidade dos modelos para a conservação e o manejo da biodiversidade (Guisan & Zimmermann, 2000).

As variáveis ambientais refletem os fatores bióticos e abióticos que influenciam a ocorrência das espécies. Entre as mais utilizadas em SDMs, destacam-se variáveis climáticas, como temperatura e precipitação, que afetam os processos fisiológicos e a tolerância climática dos organismos (Austin, 2007), e variáveis topográficas, como altitude, que influencia padrões de distribuição ao longo de gradientes ambientais

(McCain, 2007). A seleção dessas variáveis deve ser guiada pelo conhecimento ecológico das espécies modeladas, garantindo que os fatores determinantes para sua distribuição sejam adequadamente representados (Guisan & Thuiller, 2005).

Além da escolha das variáveis, a definição das escalas espaciais e temporais dos dados ambientais também é um fator crítico. Modelos construídos com variáveis em escalas inadequadas podem não refletir as condições reais experimentadas pelas espécies, comprometendo sua capacidade preditiva (Pearson & Dawson, 2003). Da mesma forma, a presença de multicolinearidade entre variáveis ambientais pode inflacionar a importância de alguns preditores e prejudicar a robustez do modelo. Para mitigar esse problema, são empregadas técnicas estatísticas, como Análise de Componentes Principais (PCA), que reduz a redundância entre variáveis correlacionadas (Dormann, et al., 2013).

A correta seleção de variáveis ambientais, aliada ao conhecimento ecológico das espécies e ao uso de técnicas estatísticas apropriadas, é essencial para garantir que os modelos de distribuição de espécies sejam biologicamente informativos e possam ser utilizados para prever padrões de biodiversidade e apoiar estratégias de conservação.

1.3.4. Tipos de algoritmos

A SDM emprega diversos algoritmos, que podem ser classificados com base em sua complexidade, abordagem estatística ou fundamentação em aprendizado de máquina. Esses métodos variam desde técnicas simples baseadas em distância até modelos complexos de "*machine learning*", cada um apresentando vantagens, desvantagens e aplicações específicas (Guisan & Zimmermann, 2000).

Classificação dos Algoritmos de SDM

Os algoritmos de SDM podem ser classificados de acordo com diferentes critérios (Elith et al., 2006):

Baseados na complexidade computacional:

- **Métodos baseados em distância**: mais simples, utilizam cálculos matemáticos diretos.
- **Métodos estatísticos**: empregam modelagem estatística para determinar a probabilidade de ocorrência.
- **Métodos de aprendizado de máquina**: utilizam inteligência artificial para capturar padrões complexos nos dados.

Baseados no tipo de dado necessário:

- **Modelos baseados apenas em presença**: utilizam apenas registros confirmados da espécie (Brotons et al., 2004).

- **Modelos baseados em presença/ausência**: requerem tanto registros de presença quanto de ausência (ex.: GLM, Random Forest).

Baseados na abordagem metodológica:

- **Modelos correlativos**: relacionam a ocorrência da espécie a variáveis ambientais (ex.: GLM, Random Forest, MaxEnt).
- **Modelos mecanísticos**: consideram processos biológicos explícitos, como fisiologia e dispersão (Kearney & Porter, 2009).

A seguir, são detalhadas as principais categorias de algoritmos utilizados.

1. Métodos Baseados em Distância

Os métodos baseados em distância são os mais simples e amplamente utilizados na modelagem de distribuição de espécies. Eles se baseiam na distância matemática entre cada ponto de um mapa e um ponto de referência comum, como a média, moda ou mediana dos pontos de ocorrência.

Um dos métodos mais conhecidos dessa categoria é o **Bioclim**, um modelo de envelope climático. O Bioclim define os valores máximos e mínimos das variáveis ambientais nos locais de ocorrência da espécie e considera que qualquer local dentro desse intervalo é potencialmente adequado (Booth et al., 2014). Embora seja fácil de implementar, o Bioclim apresenta limitações, principalmente quando há relações ambientais não-lineares e em cenários de mudanças climáticas (Elith, et al., 2006).

2. Métodos Estatísticos

Os métodos estatísticos utilizam cálculos matemáticos para determinar a probabilidade de ocorrência de uma espécie em um determinado local. Entre os mais utilizados estão:

GLM (Generalized Linear Models): Modelos lineares generalizados que estabelecem relações entre variáveis ambientais e a presença da espécie. São amplamente utilizados devido à sua flexibilidade e fácil interpretação (Guisan & Thuiller, 2005).

GAM (Generalized Additive Models): Semelhante ao GLM, mas permite modelar relações não-lineares, utilizando funções de suavização, tornando-se útil para capturar padrões ambientais mais complexos (Elith, Leathwick & Hastie, 2008).

MARS (Multivariate Adaptive Regression Splines): Método que utiliza *splines* e modela relações não-lineares e interações entre variáveis ambientais (Friedman, 1991).

Esses modelos são estatisticamente robustos, mas exigem dados de presença e ausência, o que pode ser um desafio, visto que a ausência real de uma espécie nem sempre pode ser confirmada (Guisan & Zimmermann, 2000).

3. Métodos de Aprendizado de Máquina

Os algoritmos de aprendizado de máquina são cada vez mais aplicados na modelagem de distribuição de espécies devido à sua capacidade de capturar padrões complexos nos dados. Alguns dos principais métodos incluem:

RF (Random Forest): Método baseado em árvores de decisão que combina múltiplas árvores para melhorar a precisão e reduzir o sobreajuste. É robusto e adequado para grandes volumes de dados (Breiman, 2001).

ANN (Artificial Neural Network): Redes neurais artificiais são modelos inspirados no cérebro humano, capazes de aprender relações não-lineares entre variáveis ambientais e a ocorrência de espécies. São robustas e eficazes para capturar padrões complexos em dados (Lek & Guégan, 1999).

SVM (Support Vector Machines): Algoritmo que busca encontrar um hiperplano ótimo para separar classes (presença/ausência), sendo eficaz em conjuntos de dados de alta dimensionalidade (Cortes, 1995).

MaxEnt (Maximum Entropy): Um dos métodos mais populares para dados de presença única, baseado no princípio da máxima entropia para estimar a distribuição de ocorrência (Phillips, Anderson, & Schapire, 2006).

XGBoost: Implementação eficiente de *gradient boosting*, que combina múltiplos modelos para aumentar a precisão. Apesar de ser menos comum em SDMs, tem mostrado bons resultados em cenários complexos (Chen & Guestrin, 2016).

4. Presença vs. Presença/Ausência e Uso de Pseudoausências

Métodos estatísticos e de aprendizado de máquina frequentemente necessitam de dados de presença e ausência. No entanto, a ausência real de uma espécie pode ser difícil de determinar. O fato de não encontrar uma espécie em campo não significa necessariamente que ela não ocorre naquele local. Para contornar essa limitação, utilizam-se pseudoausências, ou seja, pontos aleatórios gerados para representar locais onde a espécie potencialmente não ocorre.

O uso de pseudoausências pode introduzir viés nos modelos, uma vez que esses pontos não refletem ausências reais. A forma de selecionar essas pseudoausências influencia diretamente os resultados e a acurácia do modelo (Elith et al., 2006).

5. Comparação de Algoritmos: Vantagens e Desvantagens

Cada tipo de algoritmo apresenta vantagens e desvantagens que devem ser consideradas na escolha do método mais adequado para um estudo específico. A tabela a seguir resume essas características:

Tabela 1: Comparação de algoritmos utilizados em modelagem ecológica.

Algoritmo	Vantagens	Desvantagens
Bioclim	Simples e fácil de implementar; útil para estudos introdutórios.	Limitado a relações lineares; pouco eficaz em cenários de mudanças climáticas.
Euclidean	Simples e amplamente utilizado; mede a menor distância reta entre pontos.	Não leva em conta correlações entre variáveis ambientais.
Manhattan	Funciona bem em grades regulares; mede distância baseada em eixos ortogonais.	Menos eficiente em espaços contínuos e irregulares.
Mahalanobis	Considera a correlação entre variáveis; adequado para dados multivariados.	Computacionalmente mais complexo; exige boas estimativas da matriz de covariância.
GLM	Flexível e de fácil interpretação; amplamente utilizado.	Exige dados de presença/ausência; limitado a relações lineares.
GAM	Modela relações não-lineares; útil para dados complexos.	Pode sofrer com sobreajuste; exige ajuste cuidadoso de parâmetros.
MARS	Capaz de capturar interações complexas entre variáveis.	Pode ser computacionalmente intenso e menos interpretável.
RF	Robusto; lida bem com grandes volumes de dados; reduz sobreajuste.	Computacionalmente intensivo; difícil de interpretar.
ANN	Capaz de capturar relações não-lineares complexas; bom desempenho com grandes volumes de dados.	Requer grande poder computacional; difícil de interpretar; risco de sobreajuste.
MaxEnt	Eficaz para dados de presença única; amplamente utilizado.	Sensível à seleção de variáveis; pode superestimar áreas de ocorrência.
XGBoost	Alta precisão; eficiente em conjuntos de dados grandes.	Complexo de implementar; exige ajuste de hiperparâmetros.

6. Escolha do Algoritmo

A escolha do algoritmo ideal depende de fatores como a quantidade e tipo de dados disponíveis, complexidade das relações ecológicas e objetivos do estudo. Exemplos de recomendações incluem:

- **Para estudos exploratórios ou com dados limitados**, Bioclim ou GLM podem ser opções viáveis (Booth et al., 2014).
- **Para dados com relações não-lineares**, GAM e Random Forest são mais apropriados (Elith, Leathwick & Hastie, 2008).
- **Para dados de presença única**, MaxEnt é a escolha mais comum (Phillips, Anderson & Schapire, 2006).

No entanto, ao construir modelos de distribuição de espécies (SDMs), é importante estar atento a dois desafios frequentes: o *overfitting* e o *underfitting*. O *overfitting* (ou sobreajuste) ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, capturando inclusive ruídos ou variações aleatórias, o que reduz sua capacidade de generalização para novos dados. Já o *underfitting* (ou subajuste) acontece quando o modelo é incapaz de representar adequadamente os padrões existentes nos dados, geralmente por ser demasiado simples. Ambos os casos comprometem a qualidade preditiva do modelo.

Estudos indicam que a quantidade ideal de pontos de presença varia conforme o algoritmo utilizado, o que influencia diretamente o risco de ocorrência desses problemas. Modelos mais complexos, como Random Forest e MaxEnt, geralmente exigem um maior número de pontos para evitar o sobreajuste, enquanto métodos mais simples, como Bioclim, podem funcionar bem com conjuntos de dados menores (Guisan & Zimmermann, 2000).

1.3.5. Processos de validação e avaliação dos modelos

A validação e avaliação dos modelos de distribuição de espécies (SDM) são etapas essenciais para garantir a confiabilidade e a capacidade preditiva das modelagens ecológicas. Esses processos envolvem técnicas estatísticas robustas e métricas de desempenho que permitem quantificar a precisão dos modelos e evitar problemas como sobreajuste (Guisan & Zimmermann, 2000).

Técnicas de Validação

Particionamento de Dados: Uma das abordagens mais comuns é a divisão dos dados em conjuntos de treinamento e teste. O modelo é ajustado utilizando o conjunto de treinamento e avaliado com o conjunto de teste, permitindo estimar sua capacidade de generalização (Fielding & Bell, 1997). No entanto, essa técnica pode ser limitada em amostras pequenas, pois reduz ainda mais a quantidade de dados disponíveis para treinamento.

***Cross-Validation* (Validação Cruzada):** A validação cruzada é uma técnica mais robusta, especialmente útil para conjuntos de dados pequenos. Nesse método, os dados são divididos em *k* subconjuntos (*folds*), e o modelo é treinado *k* vezes, utilizando cada *fold* como conjunto de teste uma vez. Isso resulta em uma avaliação mais confiável do desempenho do modelo (Hastie, Tibshirani & Friedman, 2009).

***Bootstrap*:** O método de *bootstrap* consiste em criar múltiplos subconjuntos de dados a partir de amostras aleatórias com reposição. O modelo é ajustado e validado em

cada subconjunto, permitindo gerar uma distribuição de métricas de desempenho que reflete a variabilidade do modelo e reduz a dependência do conjunto de dados original (Tibshirani & Efron, 1993).

Métricas de Avaliação de Desempenho

AUC (Área Sob a Curva ROC): Uma das métricas mais utilizadas para avaliar a capacidade discriminatória do modelo. Mede a área sob a curva ROC (*Receiver Operating Characteristic*), que relaciona a taxa de verdadeiros positivos (sensibilidade) com a taxa de falsos positivos (1-especificidade). Valores de AUC próximos a 1 indicam um modelo altamente preditivo, enquanto valores próximos a 0,5 sugerem um desempenho equivalente ao acaso (Fielding & Bell, 1997).

TSS (*True Skill Statistic*): Métrica que considera tanto a sensibilidade quanto a especificidade do modelo. Seus valores variam de -1 a 1, onde valores acima de 0,6 são considerados bons e acima de 0,8 indicam excelente desempenho (Allouche, Tsoar, & Kadmon, 2006).

Kappa: O índice Kappa mede a concordância entre as previsões do modelo e os dados observados, ajustando para a concordância esperada ao acaso. Valores de Kappa acima de 0,4 são considerados aceitáveis, enquanto valores acima de 0,8 indicam uma concordância quase perfeita (Landis & Koch, 1977).

O *overfitting* ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, capturando padrões irrelevantes e reduzindo sua capacidade de prever novos dados. Para minimizar esse problema, algumas estratégias incluem:

- Utilizar técnicas de validação robustas, como *cross-validation* e *bootstrap* (Hastie, Tibshirani & Friedman, 2009).
- Regularizar modelos complexos, como *Random Forest* e *MaxEnt*, para reduzir a complexidade e evitar sobreajuste (Phillips, Anderson, & Schapire, 2006).
- Selecionar variáveis ambientais relevantes e evitar multicolinearidade, que pode inflacionar a importância de algumas variáveis e prejudicar a interpretação dos resultados (Dormann, et al., 2013).

A generalização dos modelos é fundamental para garantir que as previsões sejam aplicáveis em diferentes contextos ecológicos e escalas espaciais. Para isso, é necessário avaliar o desempenho do modelo em conjuntos de dados independentes e considerar as incertezas associadas às previsões (Guisan & Thuiller, 2005).

2. Objetivos

2.1. Objetivos gerais

Desenvolver e implementar metodologias de padronização e automação de dados taxonômicos e de biodiversidade, visando à digitalização e integração da Coleção Acarológica da UFMG-AC a plataformas internacionais de biodiversidade, como o Global Biodiversity Information Facility (GBIF) e o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr). Paralelamente, aplicar a modelagem de distribuição de espécies (SDM) para inferir a potencial distribuição geográfica dos ácaros catalogados, com o intuito de subsidiar estratégias de conservação da biodiversidade e ampliar o conhecimento científico sobre a ecologia desses organismos.

2.2. Objetivos específicos

1. **Padronizar os dados da Coleção Acarológica da UFMG-AC:** Revisar e padronizar os dados taxonômicos e geográficos da coleção no formato Darwin Core (DwC), garantindo conformidade com protocolos internacionais de qualidade e interoperabilidade, para viabilizar a integração em plataformas globais, como GBIF e SiBBr.
2. **Desenvolver softwares e fluxos de trabalho automatizados:** Criar ferramentas computacionais e pipelines automatizados para correção, padronização e validação de dados, com o objetivo de otimizar a eficiência, escalabilidade e confiabilidade dos processos, minimizando a ocorrência de erros associados a intervenções manuais.
3. **Unificar e disponibilizar os dados da coleção em bancos de dados públicos:** Consolidar os dados revisados em uma base de dados integrada e publicá-los em repositórios internacionais de biodiversidade, como GBIF e SiBBr, visando ampliar a visibilidade da coleção, facilitar o acesso à informação científica e promover a transparência no compartilhamento de dados.
4. **Aplicar modelagem de distribuição de espécies SDM:** Selecionar variáveis ambientais bioclimáticas relevantes (e.g., temperatura, precipitação) e aplicar algoritmos de modelagem preditiva (e.g., Random Forest, GAM) para inferir áreas potenciais de ocorrência dos ácaros catalogados, contribuindo para o planejamento da conservação e estudos ecológicos.

3. Materiais e métodos

3.1. Padronização

3.1.1. Revisão dos dados de ocorrência

O ponto de partida deste projeto foram as planilhas de catálogo criadas ao longo dos anos por pesquisadores associados à Coleção Acarológica UFMG-AC. Essas planilhas, armazenadas no Google Drive, permitiam edição colaborativa por membros do laboratório. A coleção estava organizada por ano, com os dois primeiros dígitos de cada número de catálogo, correspondendo ao ano de registro, a partir de 2012.

Entretanto, a prática de criar uma nova planilha para cada ano e permitir edições por vários membros gerou diversos problemas, como erros tipográficos, ortográficos e informações registradas em colunas inadequadas. Além disso, surgiram inconsistências entre as planilhas, incluindo nomes de campos variados e diferenças na ordem das colunas. Diante dessas incongruências, uma revisão abrangente dos dados tornou-se essencial para garantir a qualidade e a integridade das informações.

O processo de revisão manual iniciou-se com uma análise detalhada de todos os registros para identificar e corrigir erros frequentes, como grafias incorretas, falta de acentuação e adição ou supressão indevida de letras. Ferramentas como o OpenRefine e os filtros avançados das planilhas do Google Sheets foram utilizados para auxiliar na identificação e correção de padrões e erros. Por exemplo, termos como “Mar Bálitco” foram corrigidos para “Mar Báltico”. Outra etapa importante foi a uniformização dos nomes das colunas, eliminando discrepâncias entre as planilhas. Essa padronização foi fundamental para assegurar compatibilidade e consistência entre os registros.

Outro problema recorrente envolvia espaços extras no início, meio ou final das células, o que dificultava a execução de scripts e algoritmos. Esses espaços foram sistematicamente removidos para facilitar o processamento posterior dos dados. Além disso, houve casos em que as informações estavam alocadas em colunas inadequadas. Por exemplo, um registro sobre um ácaro coletado no Mar Báltico estava na coluna “Município” quando deveria estar em “*locality*” ou “*waterBody*”, conforme os padrões do *Darwin Core* (DwC).

A falta de padronização nos formatos de datas e coordenadas também representou um desafio significativo. As datas apareciam em diferentes formatos, como “01/I/2001”, “1/I/2001”, “01/01/2001”, “01.01.2001”, “01.01/01”, enquanto as coordenadas utilizavam diversos sistemas de referência, incluindo Sirgas 2000, graus, decimais e

WGS84. Durante a revisão, todas as datas foram convertidas para um único formato padronizado, e as coordenadas foram unificadas em um sistema de referência consistente, garantindo maior integração e precisão nos dados.

Por fim, quebras de linha desnecessárias, espaços excessivos e outros problemas relacionados à formatação foram resolvidos diretamente nas planilhas. Esses ajustes não apenas melhoraram a qualidade dos dados, mas também prepararam a coleção para a aplicação de scripts e algoritmos desenvolvidos nas etapas subsequentes.

A informatização dos metadados da coleção, além de padronizar e corrigir os dados, contribui para a preservação da informação. A digitalização garante que os registros da coleção não sejam perdidos em caso de acidentes, uma vez que estarão armazenados em servidores na nuvem ou em plataformas globais de dados, como o SiBBr (Sistema de Informação sobre a Biodiversidade Brasileira) e o GBIF (Global Biodiversity Information Facility).

O processo de padronização seguiu o formato Darwin Core (DwC), um padrão internacional para dados de coleções biológicas (Wieczorek, et al., 2012). Esse formato foi adotado para viabilizar a publicação da coleção em plataformas de biodiversidade e promover sua acessibilidade, garantindo que os dados estejam alinhados às exigências globais de interoperabilidade.

3.1.2. Scripts para correção

Para garantir a padronização dos dados da Coleção Acarológica UFMG-AC e corrigir erros recorrentes, foi desenvolvido um script em Python, utilizando a biblioteca Pandas para manipulação eficiente de grandes volumes de dados. O objetivo principal foi alinhar os registros ao padrão DwC, garantindo conformidade com requisitos de interoperabilidade e integração com plataformas globais de biodiversidade.

Funcionalidades principais do script

1. Correção de nomes Geográficos: O script identifica e corrige inconsistências na coluna “Estado”, convertendo siglas (ex.: “AC”) para a forma completa (“Acre”) e corrigindo erros de grafia (ex: “Espirito Santos” ou Espirito Santo → “Espírito Santo”). Informações contextuais ausentes, como “country” (Brasil), “countryCode” (BRA) e “continent” (América do Sul), foram inseridas automaticamente para completude dos metadados.
2. Segmentação de colunas combinadas: Campos com múltiplas informações, como “Sexo/Estágio”, foram desmembrados em colunas específicas (“sex” e

"lifeStage"), seguindo os termos do DwC. Essa abordagem eliminou ambiguidades e facilitou a interpretação dos dados.

3. Padronização via dicionários de mapeamento: Para a coluna "*recordedBy*" (coletores), um dicionário associou nomes originais (ex: "Almir", "A pepato", "Almir R Pepato") a versões padronizadas (ex: "Pepato, AR"), minimizando variações. Essa técnica foi estendida a outras colunas, como "*identifiedBy*" e "*institutionCode*".
4. Preenchimento de campos taxonômicos: Colunas como "*taxonRank*" e "*scientificName*" foram preenchidas com base no nível taxonômico mais específico disponível. Por exemplo, para um ácaro identificado como "Halacaridae", "*taxonRank*" foi definido como "família", e "*scientificName*" como "Halacaridae". A coluna "*higherClassification*" recebeu a hierarquia completa (ex: "Animalia | Arthropoda | Arachnida | Trombidiformes | Prostigmata | Anystides | Halacaroidea | Halacaridae")

O processo foi estruturado em seis etapas:

1. **Carregamento do arquivo**: Importação dos dados para processamento.
2. **Criação de colunas ausentes**: Adição de campos obrigatórios do DwC (ex: "*basisOfRecord*").
3. **Correção de dados**: Aplicação de rotinas para corrigir grafias, formatos e inconsistências.
4. **Atualização de campos dependentes**: Preenchimento automático de colunas inter-relacionadas (ex: "*higherClassification*").
5. **Renomeação de colunas**: Adequação aos termos do DwC (ex: "Município" → "*municipality*").
6. **Exportação de dados**: Salvamento em formato CSV para integração em plataformas de biodiversidade.

A automação proporcionada pelo script, demonstrou impactos positivos mensuráveis. Além de corrigir inconsistências e erros comuns, assegurou a conformidade dos dados com padrões internacionais como o DwC, promovendo a interoperabilidade das informações em plataformas globais de biodiversidade, como o GBIF e o SiBBr. A utilização estratégica de dicionários para padronização de nomenclaturas e a segmentação de colunas combinadas resultaram em uma estrutura taxonômica mais robusta, reduzindo ambiguidades e elevando a qualidade dos metadados.

A abordagem automatizada também otimizou a eficiência do processamento reduzindo substancialmente o esforço manual e viabilizando a escalabilidade das análises. Com a exportação de dados padronizados em formato CSV, estabeleceu-se uma base confiável para estudos taxonômicos, ecológicos e de modelagem de distribuição de espécie (SDM). A reprodutibilidade metodológica, assegurada pela documentação detalhada dos scripts, permite a adaptação do processo a outras coleções biológicas, reforçando seu potencial de impacto científico.

3.1.3. Pipeline

Para lidar com erros relacionados à coerência e inconsistência dos dados, particularmente em informações taxonômicas, foi desenvolvido um pipeline em Python. Essa ferramenta foi projetada para validar e padronizar informações taxonômicas, garantindo maior confiabilidade dos dados de biodiversidade. Exemplos de inconsistências incluem casos como a classificação incorreta de um ácaro da família Macronyssidae na ordem Trombidiformes em vez de Mesostigmata, o que demonstra a necessidade de uma abordagem robusta para assegurar a precisão das informações.

O pipeline utiliza arquivos de entrada no formato CSV, preferencialmente organizados de acordo com o formato (DwC), e processa nomes científicos armazenados na coluna intitulada "*scientificName*". Bases de dados externas, como o *NCBI Taxonomy*, e ferramentas como a *API Taxallnomy* (Sakamoto & Ortega, 2021), são integradas para recuperar, validar e complementar informações taxonômicas. Essa abordagem garante a consistência e a completude das informações entre diferentes níveis taxonômicos.

O processo inicia com consultas à base de dados do NCBI Taxonomy, que fornece informações detalhadas, incluindo o nome científico, o nível taxonômico (e.g., reino, filo, ordem), *taxonID* (identificador único para cada táxon) e níveis taxonômicos superiores. Os dados recuperados são armazenados em um banco de dados local, formando uma estrutura organizada para validações subsequentes. Em caso de erros na consulta, como ausência de correspondências ou informações incompletas, o pipeline realiza até três tentativas de recuperação, aumentando a confiabilidade do processo.

Após obter um *taxonID* válido, os dados são enviados para a *API Taxallnomy*, que complementa as informações preenchendo lacunas nos níveis taxonômicos e atribuindo nomes substitutos temporários quando necessário. Essa etapa assegura uma representação completa e consistente da hierarquia taxonômica. Em seguida, o pipeline verifica a existência de táxons superiores no banco de dados local, reiniciando o processo de busca caso sejam identificadas inconsistências ou lacunas.

O pipeline foi estruturado em etapas bem definidas, como descrito a seguir:

1. **Entrada de Dados**: Leitura de arquivos CSV contendo nomes científicos, com a identificação da coluna principal (e.g., "*scientificName*").
2. **Busca no NCBI Taxonomy**: Consulta dos nomes científicos à base do NCBI para recuperar informações taxonômicas hierárquicas.
3. **Adição de Dados Taxonômicos**: Armazenamento das informações recuperadas no banco de dados local, incluindo nível taxonômico, *taxonID* e hierarquias superiores.
4. **Tratamento de Erros**: Repetindo o processo de busca em caso de erros, como ausência de dados ou correspondências múltiplas, com até três tentativas.
5. **Integração com a *API Taxallnomy***: Complementação das informações taxonômicas, preenchendo lacunas e adicionando classificações temporárias.
6. **Verificação de Táxons Superiores**: Confirmação de consistência dos níveis taxonômicos superiores, reiniciando o processo caso necessário.
7. **Construção de Dados Hierárquicos**: Geração de um banco de dados estruturado, incluindo relações entre *taxonID* de pais e filhos, garantindo a coerência taxonômica.

O pipeline enfrentou desafios como ambiguidades nos nomes científicos. Por exemplo, o nome “Porolohmanella violaceae” foi erroneamente associado à família de plantas. Violaceae em vez da espécie pretendida *Porolohmanella violacea*. Esses erros foram registrados para revisão manual posterior. Além disso, variações regionais ou erros de digitação nos nomes científicos foram detectados e corrigidos automaticamente, reduzindo ambiguidades.

Erros mais complexos, como a ausência de dados hierárquicos completos, foram parcialmente resolvidos por meio de consultas adicionais à *API Taxallnomy*, enquanto outros exigiram revisão manual. A detecção de inconsistências entre os campos taxonômicos e geográficos também foi tratada, assegurando a qualidade dos dados.

O pipeline representa um avanço significativo na automação da validação e padronização de dados taxonômicos, promovendo a coerência e a precisão das informações. Ele oferece uma estrutura taxonômica abrangente, integrando bases de dados como o NCBI Taxonomy e a API Taxallnomy para solucionar lacunas e inconsistências.

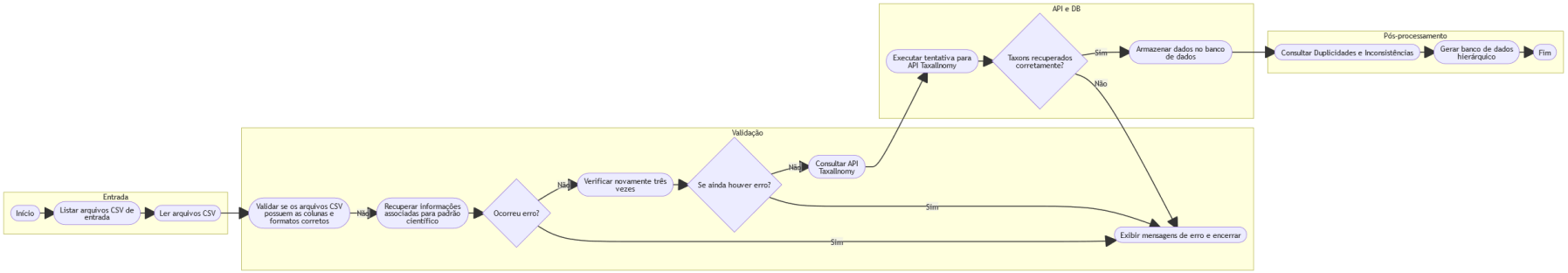

Figura 1: Fluxograma do pipeline de validação taxonômica.

Figura 2: QR Code para os códigos escritos durante a etapa de padronização.

Essa ferramenta não apenas melhora a qualidade dos dados, mas também facilita a integração com plataformas de biodiversidade, permitindo análises mais confiáveis em estudos taxonômicos e ecológicos. Ao reduzir esforços manuais e aprimorar a eficiência do processamento, o pipeline se torna uma solução essencial para a gestão e o estudo de dados de biodiversidade.

3.2. SDM

3.2.1. EcoDistrib

A modelagem de distribuição de espécie (SDM do inglês Species Distribution Modeling) requer dados de ocorrência e variáveis ambientais para prever de forma eficiente a distribuição potencial de uma espécie. A qualidade desses dados é essencial para garantir resultados precisos, sendo necessário tratá-los adequadamente desde a sua obtenção até a aplicação nos modelos. Os dados ambientais podem ser obtidos de diversas fontes, cada uma cobrindo aspectos específicos do meio ambiente. Entre as mais utilizadas estão: o Wordclim, que fornece dados climáticos terrestres, como temperatura média, máxima, mínima, além, de pluviosidade totalizando 19 variáveis; o Biooracle e o Marspec, voltados para variáveis ambientais de ecossistemas marinhos; e o EarthEnv e o UFZ, que abrangem outras características ambientais, como cobertura vegetal, luz solar incidente e dados espeleológicos.

Como parte deste trabalho, desenvolvi a biblioteca EcoDistrib, implementada em Python, com o objetivo de automatizar e simplificar todas as etapas do fluxo de modelagem de distribuição de espécies. A biblioteca está disponível no repositório do github. A EcoDistrib foi projetada para ser uma ferramenta prática, modular e flexível, permitindo desde o download automatizado de variáveis ambientais até a aplicação de algoritmos de modelagem, incluindo técnicas estatísticas e de aprendizado de máquina.

O fluxo da biblioteca inicia com o download dos dados ambientais, que pode ser feito com um único comando. Após isso, o usuário pode definir a área de interesse utilizando shapefiles personalizados ou gerados automaticamente a partir de nomes de estados ou países, bastando fornecer a sigla do estado ou o nome do país para que o EcoDistrib realize o corte das camadas *raster*.

Com as camadas ambientais delimitadas, inicia-se o processo de seleção e tratamento das variáveis. Nesta etapa, busca-se verificar quais variáveis realmente contribuem para o modelo, utilizando cálculos estatísticos para evitar redundâncias e minimizar o risco de "*overfitting*". O cálculo de correlação entre as variáveis é uma

abordagem empregada, sendo que variáveis altamente correlacionadas são removidas. Caso o número de variáveis seja elevado, aplica-se a Análise de Componentes Principais (PCA, em inglês *Principal Component Analysis*) para reduzir a dimensionalidade dos dados, facilitando visualizações e cálculos subsequentes.

Com as variáveis ambientais tratadas, é possível aplicar os algoritmos de SDM, que foram organizados no EcoDistrib em três classes principais, algoritmos de Distância (Bioclim, Mahalanobis, Euclidiana, Canberra, Chebyshev, Cosseno, Minkowski, Manhattan) algoritmos Estatísticos (GLM e GAM) algoritmos de Machine Learning (RF, ANN e SVM).

Na SDM, a inclusão de pseudoausências é essencial para treinar modelos que utilizam pontos de presença e ausência, como os algoritmos baseados em estatística ou aprendizado de máquina. Pseudoausências são pontos gerados artificialmente em locais onde não há registro de ocorrência da espécie, para a comparação com pontos de presença.

Figura 3: QR Code para o repositório no GitHub da biblioteca EcoDistrib.

No processo implementado, as pseudoausências são geradas aleatoriamente dentro dos limites espaciais definidos pelos dados de *raster*, garantindo que as coordenadas geradas estejam em áreas válidas (sem valores NaN) e não coincidam com os pontos de ocorrências existentes. Para evitar redundância e melhorar a precisão dos modelos as coordenadas de pseudoausência passam por uma validação que assegura a integridade dos valores ambientais associados. Essa abordagem balanceia os dados e minimiza o risco de sobreajuste ao fornecer informações mais completas sobre a distribuição ambiental em áreas sem registros conhecidos da espécie.

Ao final do processo, a biblioteca gera mapas preditivos no formato TIFF para cada algoritmo utilizado, representando as distribuições potenciais da espécie. Esses mapas são ferramentas valiosas para a conservação, manejo de espécies e estudo da biodiversidade, reforçando a necessidade de uma modelagem precisa e eficiente.

3.2.2. Modelagem de distribuição de espécies do gênero *Whartonia*

Foram selecionadas duas espécies para a modelagem de distribuição: *Whartonia* (*Whartonia*) *nudosetosa* (Wharton, 1938) e *W.* (*W.*) *pachywhartoni* Vercammen-Grandjean, 1966, ambas pertencentes à epifamília Trombiculoidae, proposta por Costa et al (2024). Esse gênero de ácaros possui larvas parasitas de morcegos, raramente outros mamíferos (Takahashi et al. 2006), enquanto os demais estágios do ciclo de vida (deutoninfa e adultos) são predadores de vida livre. A distinção morfológica entre as fases larvais e adultos representa um desafio para a taxonomia do grupo, já que não são morfologicamente associáveis sem a criação em laboratório ou por associação molecular,

No Brasil, foram registradas três espécies deste gênero, *W. nudosetosa*, *W. pachywhartoni* e *W. parauapebensis* Bassini-Silva & Jacinavicius, 2022. *W.* (*W.*) *nudosetosa* é uma espécie amplamente distribuída, ocorrendo do México e América Central ao estado de Minas Gerais; *W.* (*W.*) *pachywhartoni* ocorre apenas no Brasil, originalmente descrita a partir de um espécime coletado no morcego *Micronycteris megalotis* (Gray, 1842) no século XIX e posteriormente relatada no morcego *Carollia perspicillata* (Linnaeus, 1758) (Silveira et al. 2015); e *W.* (*W.*) *parauapebensis* descrita recentemente, relatada em solo de cavernas no estado do Pará (Bassini-Silva, et al., 2025).

Whartonia (*Whartonia*) *nudosetosa* (Wharton, 1938), W. (W.) *pachywhartoni* correspondem às duas espécies da modelagem, com maior número de ocorrências. As larvas são associadas aos locais de coleta dos seus hospedeiros morcegos e os estágios predadores têm sido somente encontrados em cavernas.

Whartonia nudosetosa possui distribuição de coleta no estado de Minas Gerais e no Pará, enquanto *W. pachywhartoni* apresenta uma distribuição em Minas Gerais. Ambas as espécies foram descritas com base em morfologia detalhada das larvas. No entanto, recentemente, Gomes-Almeida et al (2023) descreveram pela primeira vez deutoninfa e adultos das duas espécies.

A escolha dessas espécies se justifica pela sua relevância taxonômica e ecológica, além da disponibilidade de dados de ocorrência confiáveis. Por representarem linhagens pouco conhecidas dentro de um grupo subamostrado no Brasil, essas espécies oferecem oportunidade para análises comparativas de distribuição geográfica em função de variáveis ambientais. Seu modo de vida, intimamente ligado a hospedeiros vertebrados, também sugere que fatores bióticos e abióticos influenciam sua distribuição espacial.

Os dados de ocorrência foram obtidos exclusivamente a partir de coletas realizadas pela coleção UFMG-AC, com registros provenientes de levantamentos de

campo. Os dados utilizados nesta etapa foram extraídos após o processo de padronização descrito previamente na dissertação. Foram considerados apenas os pontos de ocorrência georreferenciados, com coordenadas válidas, informações completas de coleta e espécimes devidamente identificados por especialistas do laboratório (2008-2021). Excluindo registros com coordenadas inconsistentes.

Após o processamento e filtragem dos dados, foram utilizados 24 registros para *W. nudosetosa* e 78 registros para *W. pachywhartoni*. As duas espécies foram submetidas aos mesmos critérios de seleção e tratamento, assegurando uma abordagem comparável entre os modelos. A distribuição espacial dos pontos está representada na Figura 16.

A área de estudo compreendeu todo o território brasileiro, de forma a abranger plenamente o gradiente ambiental relevante para as espécies. A delimitação espacial inicialmente considerada englobava uma escala global; no entanto, devido a limitações de capacidade computacional, optou-se por restringir a modelagem ao território nacional. Embora uma abordagem regionalizada pudesse oferecer detalhamentos adicionais, a escala adotada representou adequadamente o padrão de distribuição das espécies.

4. Resultados e discussão

4.1. Padronização

A padronização do banco de dados da coleção UFMG-AC foi realizada com base no formato DwC, promovendo maior organização e conformidade dos dados internacionais. O processo de padronização assegurou a conformidade dos dados com esse padrão, facilitando a integração com plataformas globais de dados como o GBIF e SiBBr, permitindo um compartilhamento mais eficiente e acessível das informações.

O processo de padronização do banco de dados da coleção UFMG-AC resultou em uma reorganização significativa das colunas, visando garantir conformidade com o padrão Darwin Core e melhorar a granularidade dos dados. Essas transformações incluíram a renomeação, divisão e criação de colunas, além da tradução de termos para adequar as informações às terminologias do padrão.

Renomeação de Colunas

Diversas colunas foram renomeadas para alinhar seus nomes aos termos estabelecidos pelo DwC, garantindo maior consistência e clareza dos dados. Exemplos notáveis incluem:

- Tombo → *catalogNumber*

- Autor e ano → *scientificNameAuthorship*
- Código → *otherCatalogNumbers*
- Determinador → *identifiedBy*
- U.F./Província → *stateProvince*
- Coletor(es) → *recordedBy*

Divisão e Reestruturação de Colunas

Colunas com informações agregadas foram desmembradas em campos mais específicos, proporcionando maior detalhamento e adequação ao padrão. Por exemplo:

- Sexo/Estágio → *sex* e *lifeStage*
- Exempl. → *individualCount* e *preparations*
- Altitude → *minimumElevationInMeters* e *maximumElevationInMeters*
- Coletado em → *associatedTaxa* e *habitat*

Criação de Novas Colunas

O processo de padronização também envolveu a criação de colunas que não existem no banco de dados original, mas que foram fundamentais para capturar informações de forma mais estruturada e precisa. Por exemplo:

- A coluna *subgenus* foi criada a partir do desmembramento de informações presentes em gênero, que originalmente incluía dados no formato "Gênero (Subgênero)".
- *countryCode*
- *continent*

Essas transformações foram essenciais para eliminar inconsistências, como variações na escrita de táxons, e para garantir que as informações fossem apresentadas em campos dedicados e compatíveis com o padrão DwC. Detalhes específicos sobre as alterações realizadas podem ser consultados na Tabela 3 apresentada no Apêndice.

O processo de padronização de dados da coleção UFMG-AC resultou em um banco de dados detalhado e bem organizado, formatado de acordo com o padrão DwC. Após os ajustes, o conjunto de dados passou a conter 89 colunas, incluindo 12 colunas auxiliares que, embora não façam parte do padrão DwC, são fundamentais para facilitar a inserção de dados e apoiar operações internas.

- Graus, minutos, segundos e N/S para latitude.
- Graus.1, minutos.1, segundos.1, e W/E para longitude.

- Subordem, Supercoorte, Coorte e Subcoorte.

Essas colunas foram configuradas de forma a espelhar informações em campos padronizados do DwC, como "*higherClassification*", "*decimalLatitude*", "*decimalLongitude*" e "*verbatimCoordinates*", garantindo compatibilidade com o padrão e que informações não sejam perdidas.

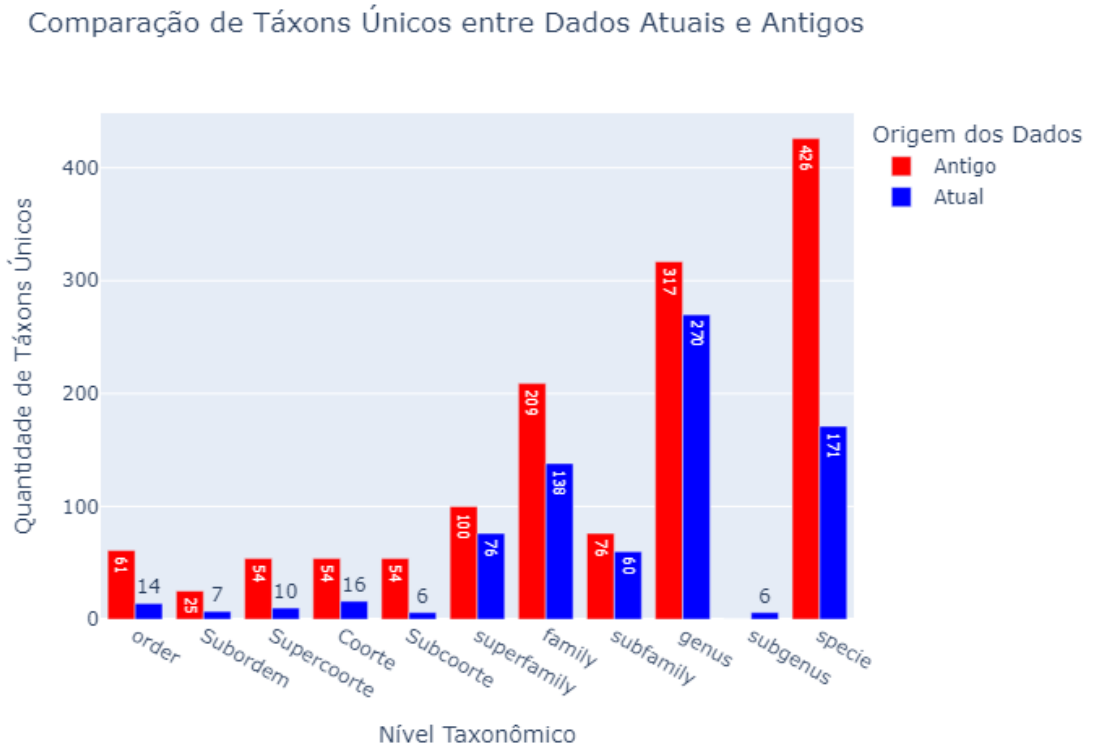

Figura 4: Comparação de táxons únicos entre dados atuais e antigos. A categoria "subgenus", ausente no conjunto antigo, foi incluída após a integração com bases externas. O nível "Coorte ou Supercoorte" foi desagregado em "Supercoorte", "Coorte" e "Subcoorte".

Foi realizada uma análise detalhada para identificar a quantidade de itens únicos nas colunas relacionadas aos táxons, comparando os valores antes e após a padronização do banco de dados. Os resultados demonstram uma redução significativa de duplicidades e variações.

Durante a padronização, foram identificadas inconsistências relevantes, como a presença de táxons não-pertencentes aos ácaros, incluindo "Aranae", "Copepoda", "Tardigrada" e outros que serão detalhados no tópico sobre a cobertura taxonômica. Esses táxons foram identificados como espécimes pertencentes a outras ordens que não são ácaros, porém, presentes na coleção. Mesmo considerando essas ordens, a discrepância era evidente. Por exemplo, a coluna ordem apresentava 61 táxons únicos antes da padronização, enquanto o total esperado de ordens para ácaros é de apenas 6 (ou 14, incluindo os táxons externos).

As colunas Subordem, Supercoorte, Coorte, Subcoorte não possuem equivalentes diretos no DwC. Na coleção UFMG-AC, essas informações estavam originalmente reunidas na coluna denominada “Coorte e Subcoorte”. Esse agrupamento contribuiu para a inconsistência na quantidade de itens registrados, uma vez que os dados frequentemente apresentavam grafias incorretas, duplicidades ou termos não padronizados.

Cobertura Temporal

A coleção UFMG-AC teve início em 2012 com o Professor Dr. Almir Rogério Pepato, cuja contribuição inicial foi composta principalmente por ácaros marinhos de sua coleção pessoal. Desde então, por meio de parcerias, redes de colaboração e do esforço contínuo de professores e estudantes no laboratório, outras famílias de ácaros foram incorporadas. Materiais de diferentes origens foram adicionados à coleção ao longo dos anos, contribuindo para a sua expansão.

A coleção tem crescido de forma consistente desde a sua criação, com uma média de 1.339 novos registros de catálogo por ano. No entanto, a pandemia de COVID-19 impactou significativamente esse ritmo, com apenas 571 registros em 2020 e 473 registros em 2021, os números mais baixos do período analisado.

A Figura 3 apresenta a quantidade de registros catalogados anualmente, destacando o pico em 2017, que foi o ano com o maior número de novos catálogos. Por outro lado, o maior número de espécimes coletados ocorreu em 2016, indicando uma discrepância entre o momento da coleta e a catalogação. Essa diferença pode ser atribuída a processos que demandam tempo, como extração, identificação, preparação e etiquetagem de lâminas. O atraso médio observado entre a coleta e a identificação dos espécimes foi de 9 anos.

Outro aspecto analisado foi o intervalo entre a coleta e o registro oficial de material tipo, como holótipos e parátipos, ou tempo de prateleira (em inglês, “*shelf-life*”). A análise revelou um atraso médio de 1,95 anos, refletindo a complexidade dos processos de descrição taxonômica.

A distribuição das coletas ao longo da semana também foi avaliada. O gráfico a seguir mostra que a segunda-feira concentra o maior número de registros de coleta. Esse padrão pode ser explicado pelo início das atividades de campo no início da semana, após o planejamento logístico realizado nos dias anteriores.

Embora fatores como condições climáticas e cronograma específicos de expedições possam influenciar essa concentração, os dados reforçam a importância do planejamento estratégico para maximizar os esforços de coleta e os recursos disponíveis.

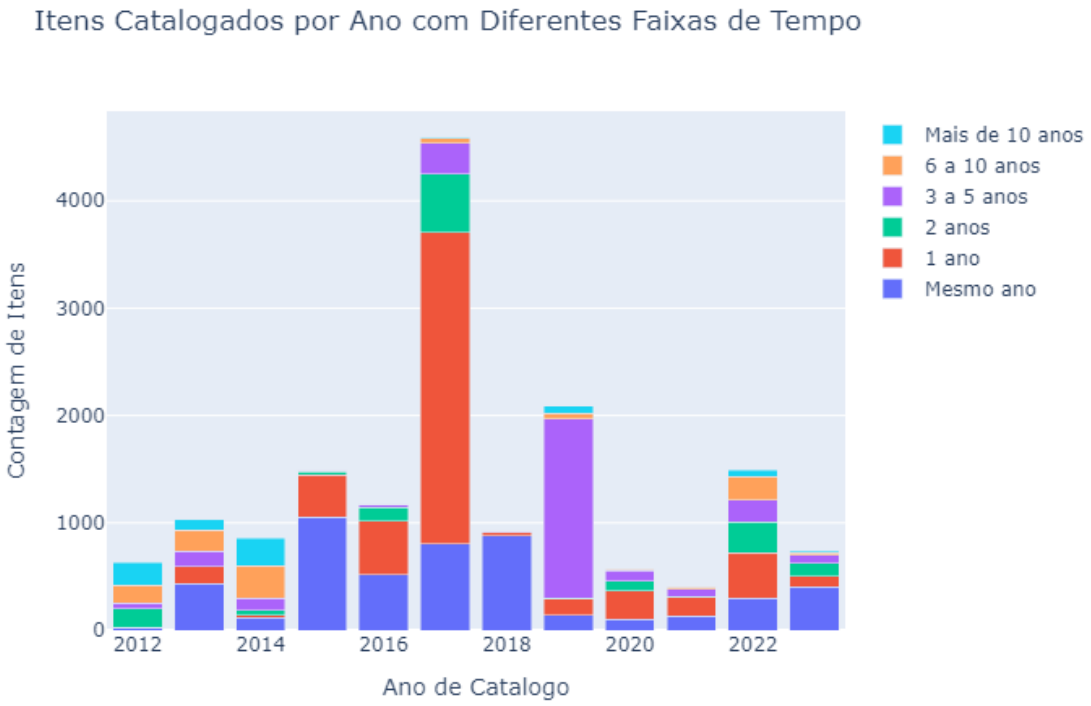

Figura 5: Itens catalogados por ano com diferentes faixas de tempo.

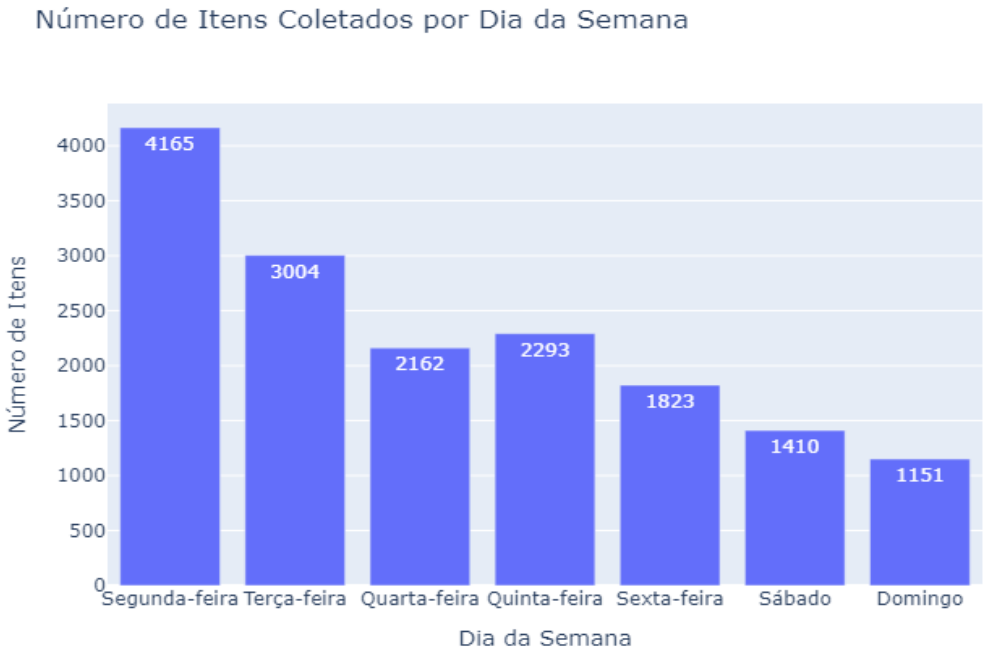

Figura 6: Número de itens coletados por dia da semana.

Por fim, o ano de 2017 apresentou um comportamento atípico no número de registros catalogados. Uma análise mais detalhada utilizando a coluna "*recordedBy*"

identificou que a Equipe Carste contribuiu significativamente para os números de coleta em anos como 2016. No entanto, mesmo considerando o número de coletores únicos por ano, que atingiu o pico em 2013 (com 64 coletores), não foi possível estabelecer uma relação direta entre a quantidade de coletores e a quantidade de itens coletados que explicasse a alta quantidade de catálogos finalizados em 2017.

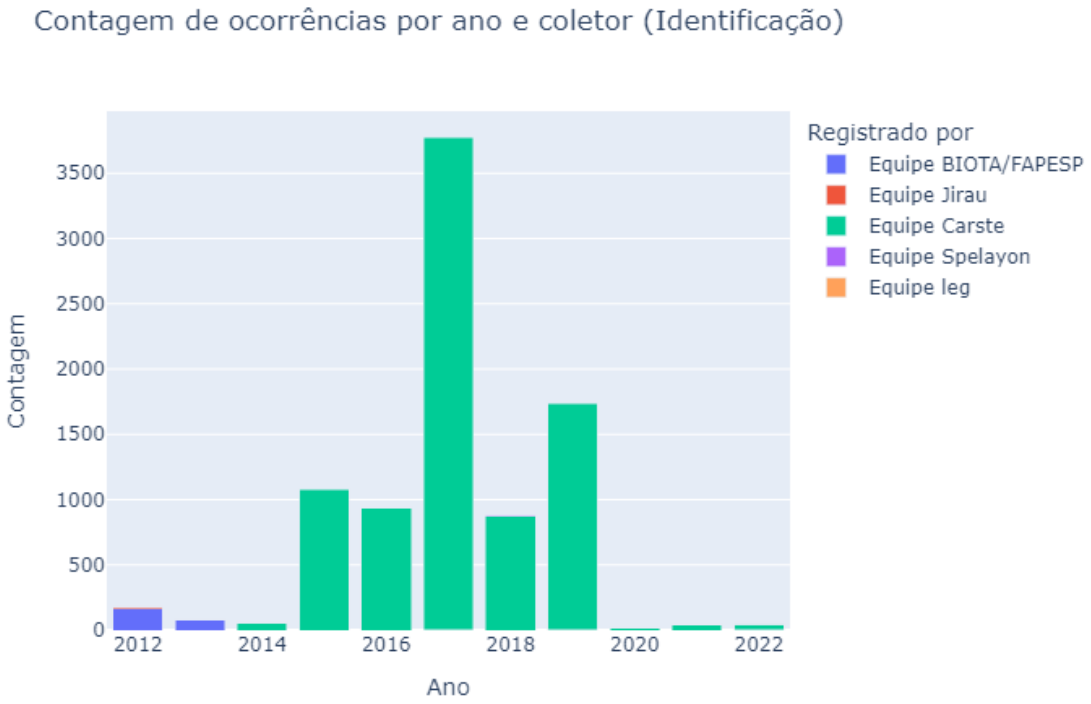

Figura 7: Contagem de ocorrências por ano e coletor (Identificação).

Esses resultados indicam que a dinâmica de crescimento da coleção está associada a múltiplos fatores, como as colaborações entre equipes, a logística das expedições e a complexidade dos processos necessários para a identificação e catalogação de espécimes.

Cobertura Espacial

A coleção UFMG-AC reúne registros de ácaros coletados em diversas regiões do Brasil e do mundo, refletindo a atuação contínua de pesquisadores em diferentes contextos geográficos e ecológicos. A distribuição espacial dos espécimes depositados na coleção é influenciada por múltiplos fatores, incluindo a localização institucional, os objetivos científicos das coletas, parcerias regionais e condições logísticas.

Tabela 2: Distribuição geográfica da UFMG-AC por país e continente.

Continente	País	Quantidade tombos	Porcentagem Total (%)	Porcentagem exterior (%)
América do Sul	Brasil	15372	91.68	--------
	Chile	62	0.37	6.47
	Bolívia	12	0.07	1.29
	Peru	6	0.04	0.65
	Guiana Francesa	2	0.01	0.22
	Equador	1	0.01	0.11
América Central	Panamá	12	0.07	1.29
	Cuba	2	0.01	0.22
	Honduras	2	0.01	0.22
América do Norte	EUA	72	0.43	7.76
Europa	Espanha	200	1.19	21.57
	Alemanha	53	0.32	5.72
Ásia	Rússia	219	1.31	23.62
	Azerbaijão	39	0.23	4.21
	Irã	28	0.17	3.02
	Myanmar	5	0.03	0.54
	Tajiquistão	2	0.01	0.22
Oceania	Austrália	117	0.70	12.62
	Nova Zelândia	94	0.56	10.14
Antártica		1	0.01	0.11
NI	NI	465	2.77	--------
Total		16766	100.00	100.00

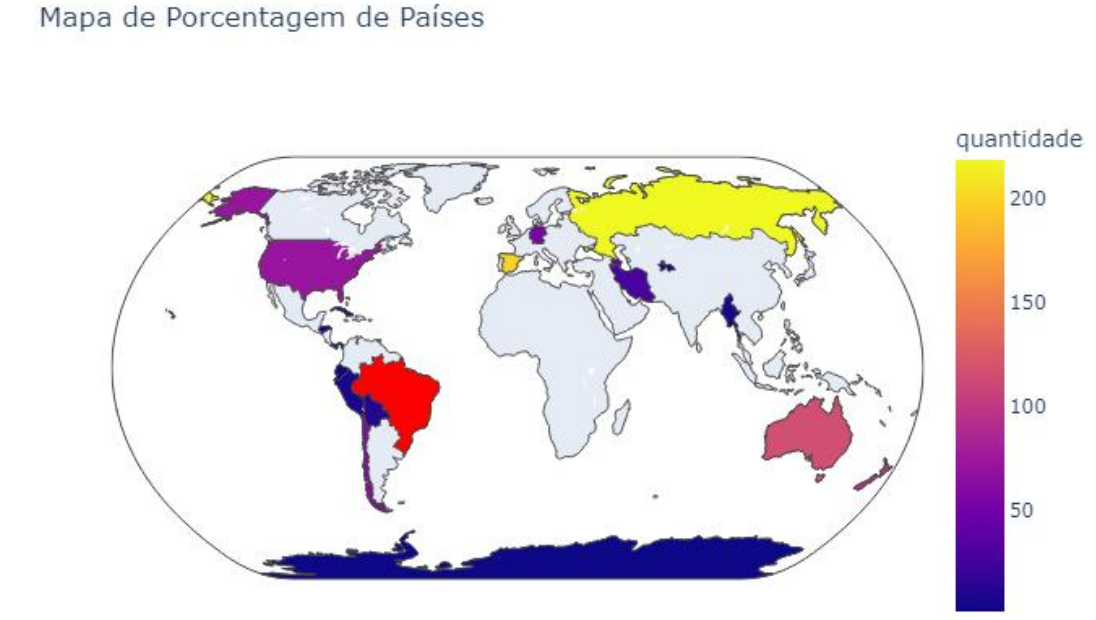

Figura 8: Mapa de distribuição geográfica por país com base na quantidade de ocorrências. O Brasil está fora da escala devido a quantidade muito superior à dos outros países.

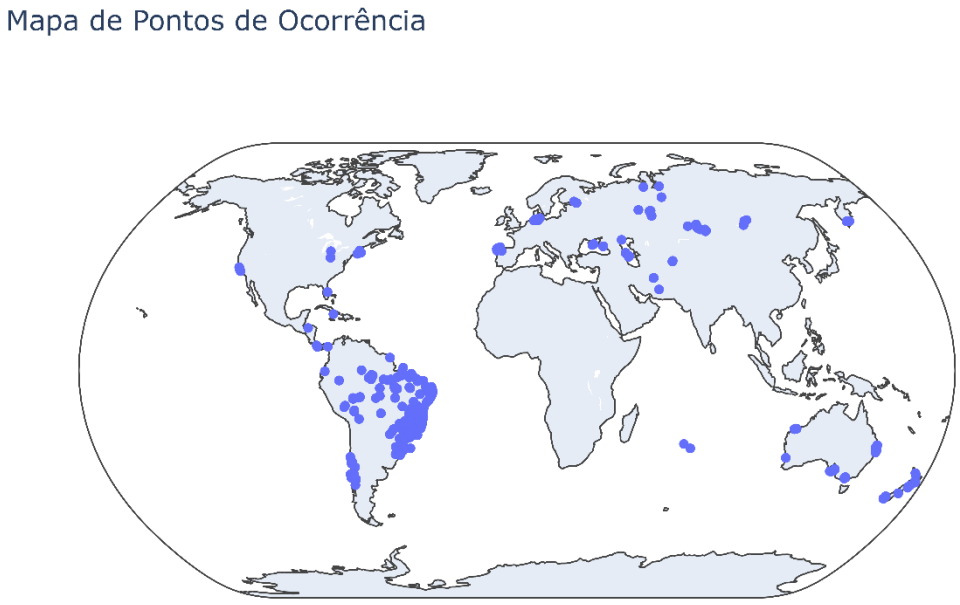

Figura 9: Mapa de distribuição espacial dos pontos de ocorrência.

Podemos observar que a maior parte dos registros está concentrada no Brasil (91,68%), como esperado, mas há também amostras oriundas de outros continentes, reforçando o caráter colaborativo da coleção em projetos internacionais.

No Brasil, a região Sudeste é a mais representada, correspondendo a 88,17% dos itens catalogados, seguida pelas regiões Nordeste (4,72%), Norte (4,26%), Sul (1,85%) e Centro-Oeste (0,40%). Em nível estadual, Minas Gerais lidera com 76,68% dos registros, seguida por São Paulo (8,28%) e Pará (3,17%). Esse predomínio de Minas Gerais é esperado, dado que a UFMG-AC está localizada neste estado, o que facilita as coletas locais devido aos menores custos logísticos.

Tabela 3: Distribuição dos tombos por estado, região e número de municípios no Brasil.

Região	**Estado**	**Quantidade tombos**	**Porcentagem Total (%)**	**Porcentagem parcial (%)**	**Número de cidades**
Sudeste	Minas Gerais	11801	76.76	--------	96
	São Paulo	1298	8.44	36.91	23
	Espírito Santo	369	2.40	10.49	17
	Rio de Janeiro	125	0.81	3.55	10
Norte	Pará	485	3.15	13.79	18
	Amazonas	126	0.82	3.58	3
	Rondônia	33	0.22	0.92	1
	Acre	8	0.05	0.23	1
Centro-oeste	Mato Grosso do Sul	51	0.33	1.45	2
	Mato Grosso	11	0.07	0.31	3

Sul	Santa Catarina	122	0.79	3.47	11
	Rio Grande do Sul	119	0.77	3.38	4
	Paraná	49	0.32	1.39	4
Nordeste	Bahia	292	1.90	8.30	20
	Pernambuco	113	0.74	3.21	8
	Piauí	88	0.57	2.50	2
	Alagoas	61	0.40	1.73	4
	Rio Grande do Norte	52	0.34	1.48	2
	Ceará	49	0.32	1.39	2
	Paraíba	41	0.27	1.17	4
	Sergipe	14	0.09	0.40	3
	Maranhão	9	0.06	0.26	1
Não informado		56	0.36	--------	
Minas Gerais | São Paulo		2	0.01	0.06	
Total		15374	100.00	100.00	

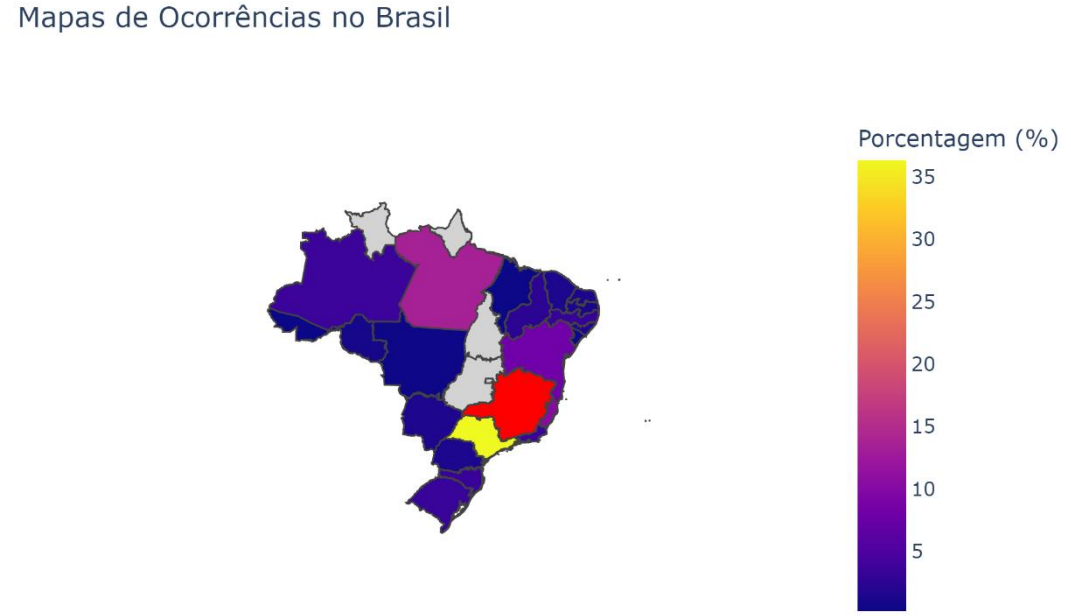

Figura 10: Mapa percentual das ocorrências por unidade federativa no Brasil. Minas Gerais está fora da escala de cor, e os estados em cinza não possuem coletas (Amapá, Distrito Federal, Goiás, Roraima e Tocantins).

A forte representação da região Sudeste e, em particular, de Minas Gerais, reflete não apenas fatores logísticos, mas também a atuação de parceiros estratégicos. No caso de Minas Gerais, a expressiva quantidade de espécimes está associada majoritariamente às coletas realizadas pela equipe da Carste Consultoria, e não diretamente à proximidade com a sede da UFMG-AC. Isso evidencia como a contribuição de instituições parceiras pode influenciar significativamente a cobertura espacial da coleção, complementando os esforços próprios da universidade e ampliando a representatividade geográfica por meio de colaborações especializadas.

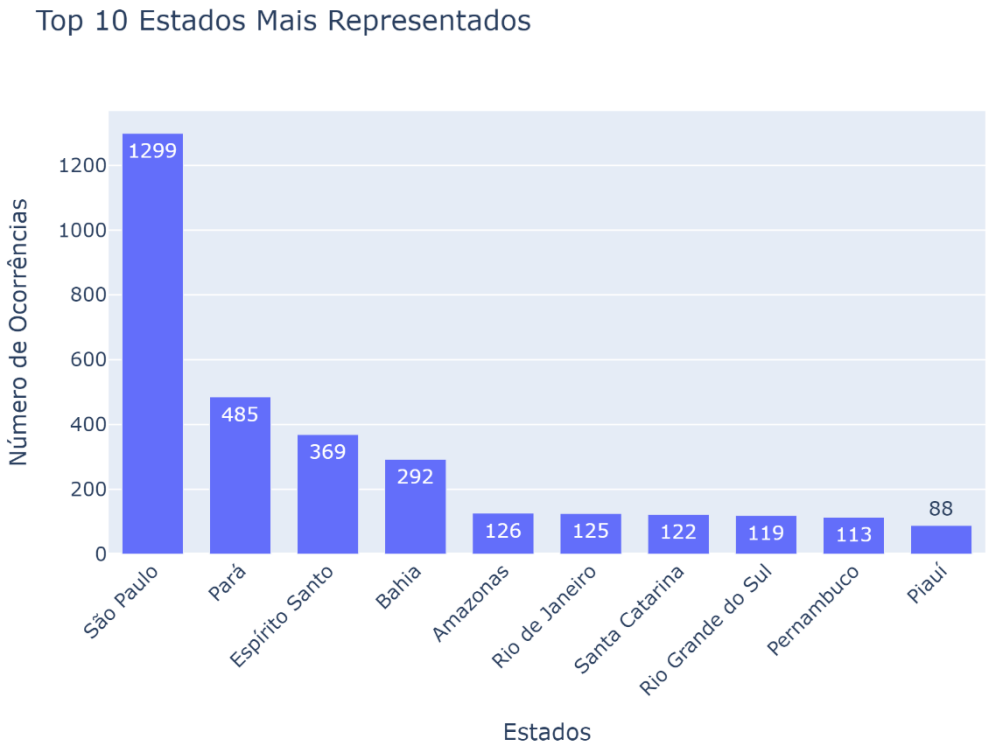

Figura 11: Top 10 estados brasileiros com maior número de ocorrências registradas, exceto Minas Gerais.

Cobertura Taxonômica

A coleção UFMG-AC demonstra uma notável diversidade taxonômica. Atualmente, a coleção contém espécimes distribuídos em 14 ordens, conforme ilustrado na figura 4. Embora os ácaros sejam o principal foco, com representantes de seis ordens — Trombidiformes, Mesostigmata, Ixodida, Opilioacariformes, Sarcoptiformes e Holothyrida —, outros grupos também estão representados, como Araneae, Copepoda, Opiliones, Ostracoda, Palpigradi, Schizomida, Pseudoscorpiones e Tanaidacea.

A inclusão desses grupos não-pertencentes aos ácaros ocorre por conveniência, devido ao seu pequeno tamanho e à associação com os mesmos eventos de coleta. No entanto, esses espécimes serão eventualmente transferidos para coleções mais adequadas, já que o CCT-UFMG conta com coleções aracnológicas e de outros invertebrados.

No que diz respeito à identificação taxonômica, cerca de dois terços da coleção foram classificados ao nível de família, um marco essencial para o progresso do processo taxonômico. A classificação em níveis inferiores, como gênero e espécie, depende da identificação prévia no nível de família.

Uma análise detalhada das famílias representadas na UFMG-AC revela uma diversidade expressiva, com 138 famílias. Entre essas, destacam-se famílias como Halacaridae (ácaros marinhos), que refletem os interesses de pesquisa do laboratório e

demonstram o foco direcionado da coleção em grupos específicos. O histograma da Figura 13, que apresenta as famílias com mais de 100 espécimes, ilustra de forma clara essa riqueza e amplitude taxonômica, destacando a UFMG-AC como um recurso de referência para a acarologia e estudos de biodiversidade no Brasil.

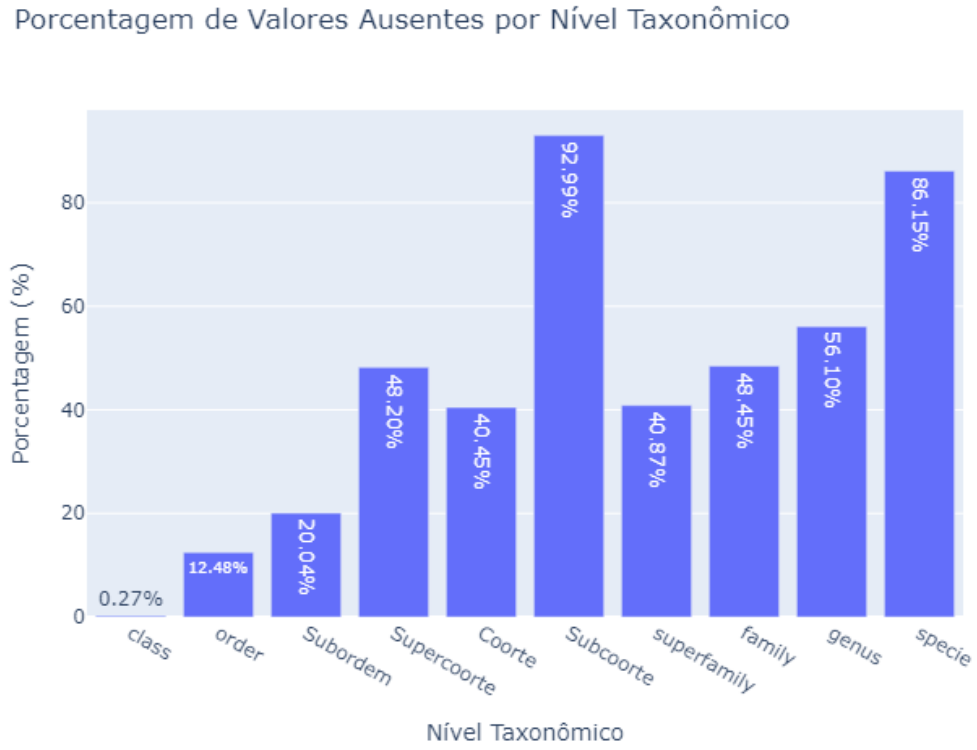

Figura 12: Porcentagem de valores ausentes por nível taxonômico.

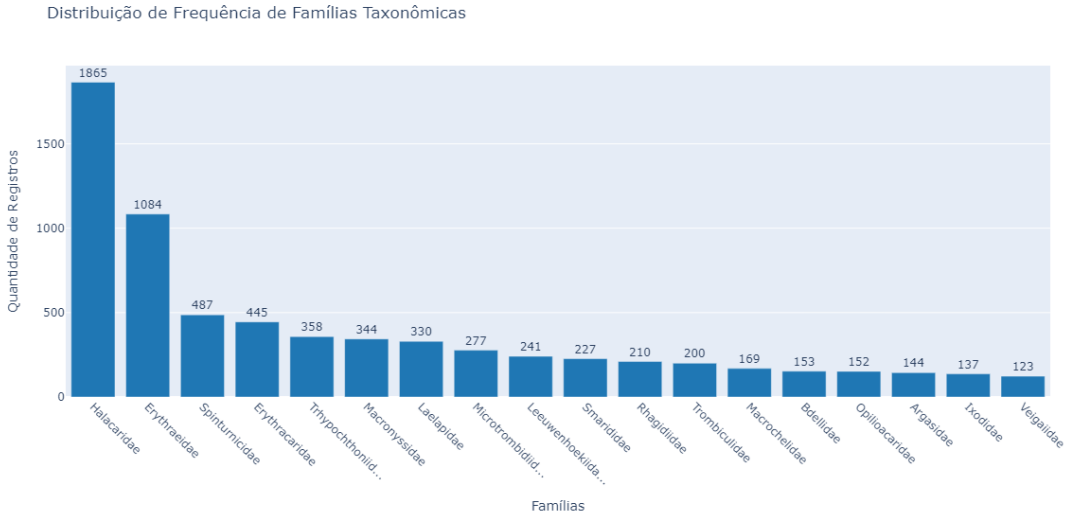

Figura 13: Distribuição de frequência de famílias taxonômicas.

Automação e Integração dos Dados

O conjunto de dados padronizado da coleção UFMG-AC não apenas está em conformidade com os padrões do Darwin Core (DwC), mas também incorpora

ferramentas adicionais que aprimoram a gestão e automação dos dados. Uma planilha geral foi criada para consolidar todos os registros da coleção UFMG-AC de 2012 a 2023, organizada de acordo com os termos do DwC. Essa planilha inclui fórmulas que facilitam a gestão dos dados e um script personalizado do Google Apps Script que atualiza automaticamente a data de modificação sempre que alterações são realizadas.

Além disso, uma planilha separada foi desenvolvida especificamente para o ano de 2024, destinada ao registro de ácaros coletados, triados, identificados e catalogados durante o período. Seguindo o mesmo formato DwC, essa planilha também utiliza as mesmas fórmulas e o script para atualização automática da data de modificação. Após sua finalização, a planilha de 2024 será integrada ao conjunto de dados geral, garantindo consistência e completude no banco de dados.

Embora as colunas auxiliares sejam essenciais para uso interno, elas não são publicadas em plataformas como o SiBBr, pois não fazem parte dos padrões DwC. Apenas os campos centrais, compatíveis com o formato DwC, são compartilhados por meio de protocolo Integrated Publishing Toolkit (IPT). Isso garante que os dados estejam disponíveis em plataformas como o SiBBr, atendendo aos padrões internacionais para dados de biodiversidade.

Figura 14: QR Code para a publicação dos dados UFMG-AC no GBIF.

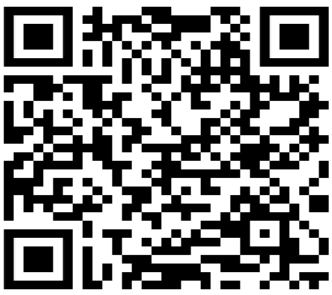

Figura 15: QR Code para a publicação dos dados UFMG-AC no SiBBr.

As melhorias introduzidas na estrutura, conformidade e automação representam avanços significativos na gestão da coleção UFMG-AC. Essas atualizações não apenas

promovem eficiência no manejo interno, mas também asseguram que o conjunto de dados atenda aos padrões globais, incentivando a acessibilidade e a integração com iniciativas globais de biodiversidade.

O conjunto de dados padronizado da coleção UFMG-AC foi publicado com sucesso nas plataformas globais de biodiversidade GBIF e SiBBr. Essa publicação torna os registros da coleção acessíveis à comunidade científica e ao público global, promovendo a visibilidade da coleção e contribuindo para o avanço de estudos sobre biodiversidade. A disponibilização dos dados nessas plataformas também demonstra o compromisso da UFMG-AC com a transparência e os padrões internacionais, fortalecendo sua relevância em iniciativas globais de pesquisa e conservação.

4.2. SDM

Para esse estudo, foram utilizadas duas espécies do gênero *Whartonia*: *Whartonia nudosetosa* (Wharton, 1938) e *Whartonia pachywhartoni Vercammen-Grandjean, 1966*. Essas espécies foram coletadas e catalogadas pela coleção UFMG-AC e estão disponíveis no GBIF e no SiBBr. A base de dados contou com 24 ocorrências de *W. nudosetosa* e 78 ocorrências de *W. pachywhartoni*, distribuídas nos estados Minas Gerais e Pará. A distribuição geográfica dessas ocorrências é ilustrada na Figura 16, que mostra a concentração dos registros em áreas específicas do país.

Figura 16: Mapa de ocorrências das espécies no Brasil.

Para a modelagem, utilizamos as 19 variáveis bioclimáticas e a elevação disponíveis no Wordclim. Inicialmente, aplicamos três métodos de análise de correlação – Pearson, Spearman e Kendall – para avaliar a interdependência entre as variáveis figura 23. A Figura 17 mostra os resultados dessas análises, indicando que oito variáveis

apresentaram correlação abaixo de 0,75 não sendo significativa entre si utilizando a correlação de Pearson.

Com base nessa análise, selecionamos as variáveis que não apresentaram multicolinearidade e realizamos um recorte espacial delas para o território brasileiro, a fim de restringir a análise à área de interesse. Esse recorte é apresentado na Figura 18, que destaca a variação espacial das condições ambientais no país.

Essas variáveis foram selecionadas para a análise de componentes principais (PCA), que reduziu a dimensionalidade dos dados, resultando em três novas camadas ambientais. O resultado do PCA é ilustrado na Figura 18, que destaca as áreas com maior influência das variáveis ambientais selecionadas.

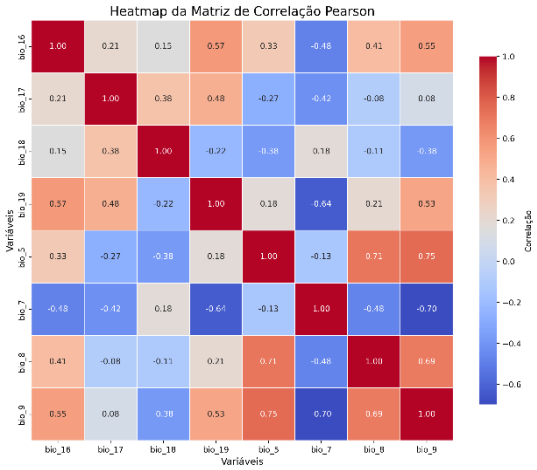

Figura 17: Heatmap da matriz de correlação de Pearson.

Com base nessas camadas, realizamos a modelagem de distribuição para cada uma das espécies, utilizando todos os algoritmos das três classes definidas anteriormente. Os mapas finais de distribuição potencial para *W. pachywhartoni* e *W. nudosetosa*, gerados por esses algoritmos serão apresentados a seguir.

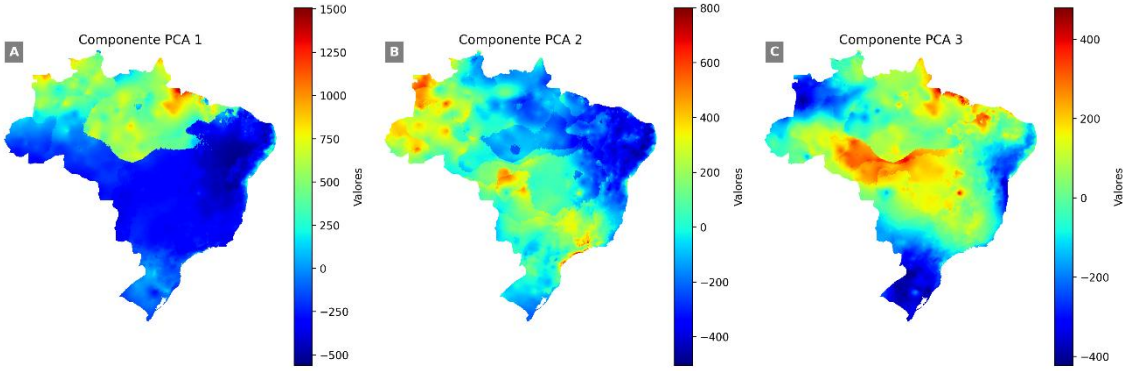

Figura 18: Mapa das três principais componentes da PCA aplicadas às variáveis ambientais do Brasil.

Além disso, foram incluídos mapas de ausência, que representam áreas onde as espécies não foram registradas, mas que foram consideradas no processo de modelagem. Foi gerado uma quantidade de 30% do total de espécimes, resultando em 23 pseudoausências para *W. pachywhartoni* e 7 pseudoausências para *W. nudosetosa*. Essas pseudoausências, geradas aleatoriamente, foram utilizadas nos algoritmos das classes Estatísticas e Machine learning.

Whartonia nudosetosa

A Figura 19 apresenta o mapa com os pontos de ocorrência e pseudoausência para a espécie *Whartonia nudosetosa*. Após a obtenção dos dados de pseudoausência, precedeu-se à aplicação dos modelos de distribuição de espécies (SDMs). Para os modelos de distância (MDI), utilizamos apenas os dados de presença, enquanto para os modelos estatísticos (MES) e os modelos de machine learning (MML), empregamos tanto os dados de presença quanto os de pseudoausência. Utilizando a biblioteca **EcoDistrib**, aplicamos os modelos por meio de uma função específica, obtendo como resultado os mapas de distribuição potencial apresentados na Figura 20.

Figura 19: Mapa com os pontos de ocorrência e pseudoausência para a espécie *W. Nudosetosa*.

A Figura 20 exibe os mapas de distribuição potencial gerados para *W. nudosetosa*. Esses mapas ilustram as áreas com maior probabilidade de ocorrência da espécie, conforme previsto pelos diferentes modelos aplicados. A análise visual dos mapas permite identificar padrões espaciais e comparar as previsões entre os modelos.

Nos mapas gerados pela biblioteca (Fig. 20) podemos observar os resultados dos modelos, para todos os mapas a cor vermelha significa presença e o azul é a ausência.

Devido a variações nos modelos, podemos ver que nem todos correspondem às áreas, porém é possível observar que existe uma sobreposição em todos eles. Algo importante de se reparar é que para os métodos de distância a escala de cor apresentada ao lado do mapa é invertida comparado aos outros mapas. Isso ocorre devido ao modo como o método funciona, já que nesses métodos de distância quanto menor o valor, menor a distância ao ponto central indicando uma maior similaridade do lugar de presença.

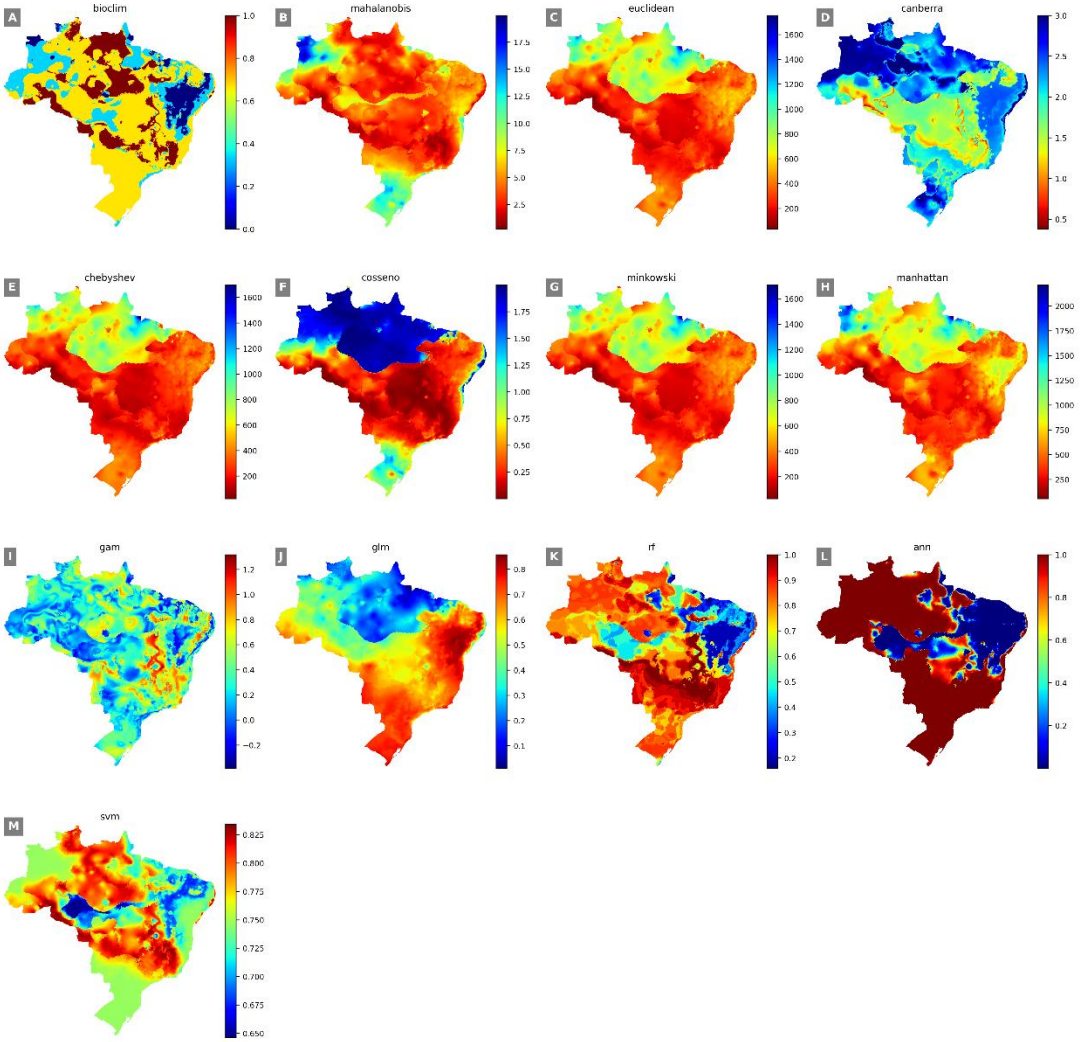

Figura 20: Mapas da distribuição potencial para *W. nudosetosa* gerado pela biblioteca EcoDistrib.

Tabela 4: Desempenho dos modelos utilizados na previsão de ocorrência para *W. nudosetosa*.

Model	aucroc	acura	precis	recall	f1	tss	vp	vn	fp	fn
Bioclim	0.24	0.40	0.33	0.21	0.26	-0.21	5	14	10	19
Mahalanobis	0.39	0.50	0.50	0.21	0.29	0.00	5	19	5	19
Euclidean	0.28	0.42	0.36	0.21	0.26	-0.17	5	15	9	19
Canberra	0.29	0.38	0.31	0.21	0.25	-0.25	5	13	11	19
Chebyshev	0.30	0.40	0.33	0.21	0.26	-0.21	5	14	10	19

Cosseno	0.27	0.46	0.42	0.21	0.28	-0.08	5	17	7	19
Minkowski	0.27	0.35	0.29	0.21	0.24	-0.29	5	12	12	19
Manhattan	0.37	0.54	0.62	0.21	0.31	0.08	5	21	3	19
GAM	0.28	0.44	0.38	0.21	0.27	-0.12	5	16	8	19
GLM	0.33	0.48	0.45	0.21	0.29	-0.04	5	18	6	19
RF	0.29	0.46	0.42	0.21	0.28	-0.08	5	17	7	19
ANN	0.20	0.40	0.33	0.21	0.26	-0.21	5	14	10	19
SVM	0.41	0.52	0.56	0.21	0.30	0.04	5	20	4	19

A modelagem de distribuição de espécies (SDM) realizada para *Whartonia nudosetosa* baseou-se na utilização de pontos de presença obtidos pela coleção UFMG-AC, sem a incorporação de registros provenientes da literatura. A distribuição desses pontos apresentou certa concentração nos estados de Minas Gerais e no Pará, o que pode ter influenciado diretamente as áreas preditas como adequadas no mapa de distribuição gerado, quanto aos pontos de pseudoausência foram obtidos de forma aleatória, contendo pontos que estão em áreas que pelo sdm foi considerada como presença. As regiões identificadas como de presença refletem, portanto, mais a distribuição dos dados coletados do que uma projeção abrangente do nicho ecológico real da espécie.

As regiões preditas como adequadas para a ocorrência de *Whartonia nudosetosa* mostraram uma concentração na região centro-oeste e ao sul da região norte, além de parte da região nordeste e de Minas Gerais, abrangendo principalmente os biomas Cerrado, Amazônia e Caatinga, com transições para Mata Atlântica em Minas. Tais padrões podem refletir tanto a verdadeira preferência ambiental da espécie quanto limitações impostas pela distribuição espacial dos dados coletados. Foi observado que algumas áreas marcadas como presença incluem regiões onde a espécie não foi registrada e cuja adequabilidade ecológica é incerta, o que sugere a necessidade de estudos adicionais para validar essas predições em campo. Tais resultados são relevantes para orientar esforços de amostragem futuros, além de contribuir para estratégias de conservação da espécie, especialmente frente a possíveis ameaças ambientais.

A ausência de dados de ocorrência secundários, especialmente aqueles descritos em estudos anteriores, impõe uma limitação importante à modelagem. A não-inclusão de registros históricos reduz a abrangência espacial e ambiental do modelo, podendo subestimar ou distorcer o real potencial de distribuição da espécie. Algumas áreas em que foram feitas novas coletas e a princípio são considerados como pontos de ausência nos

modelos, porém ainda estão em desenvolvimento estudos para se afirmar que esses pontos de coletas são da espécie em questão. Dessa forma, os resultados obtidos devem ser interpretados com cautela, reconhecendo que não representam um modelamento completo da distribuição da espécie.

Além disso, é importante destacar que, devido à quantidade de pontos de ocorrência disponíveis, alguns métodos não são adequados para a realização do SDM, por exemplo, foi utilizado o método BIOCLIM, que é mais adequado para situações com poucos dados. Embora o BIOCLIM seja eficiente em contextos de baixa amostragem, ele possui limitações metodológicas importantes, como a incapacidade de modelar relações complexas entre variáveis ambientais e a presença da espécie. Não somente o Bioclim, mas os outros métodos também possuem uma faixa ótima para a predição da distribuição de espécies. Por esse motivo, a escolha do método, aliada ao número reduzido de pontos, pode ter impactado a acurácia e a capacidade preditiva do modelo gerado.

As variáveis ambientais utilizadas na modelagem apresentaram contribuições distintas para a predição da distribuição de *Whartonia nudosetosa*. Apesar da limitação no número de pontos de ocorrência, foi possível observar que variáveis relacionadas à precipitação parecem exercer maior influência sobre a definição das áreas de presença. Esse padrão sugere que a espécie pode estar associada a regimes específicos de precipitação, reforçando a importância de incluir variáveis hidrológicas e climáticas em estudos futuros que investiguem a ecologia e a distribuição potencial de *W. nudosetosa*.

Embora o modelo tenha conseguido projetar áreas potenciais de ocorrência, existem diversas possibilidades para aprimorar futuras modelagens de distribuição. Uma abordagem promissora seria o aumento da base de dados de ocorrência, incorporando registros de literatura, novos levantamentos de campo e dados sobre os hospedeiros, permitindo modelar interações não-lineares entre variáveis. Além disso, a utilização de validação cruzada e a inclusão de variáveis ambientais mais específicas, como dados de vegetação, altitude detalhada ou estrutura do solo, poderiam refinar ainda mais a predição do nicho ecológico de *W. nudosetosa*, proporcionando resultados mais confiáveis para esforços de conservação.

Whartonia pachywhartoni

A Figura 21 apresenta o mapa com os pontos de ocorrência e pseudoausência utilizados na modelagem da distribuição da espécie *Whartonia pachywhartoni*. Assim como realizado para *W. nudosetosa*, os dados de pseudoausência foram gerados previamente e, em seguida, os modelos de distribuição de espécies (SDMs) foram

aplicados. Os modelos foram executados por meio da biblioteca EcoDistrib, que gerou os mapas de distribuição potencial conforme ilustrado na Figura 22.

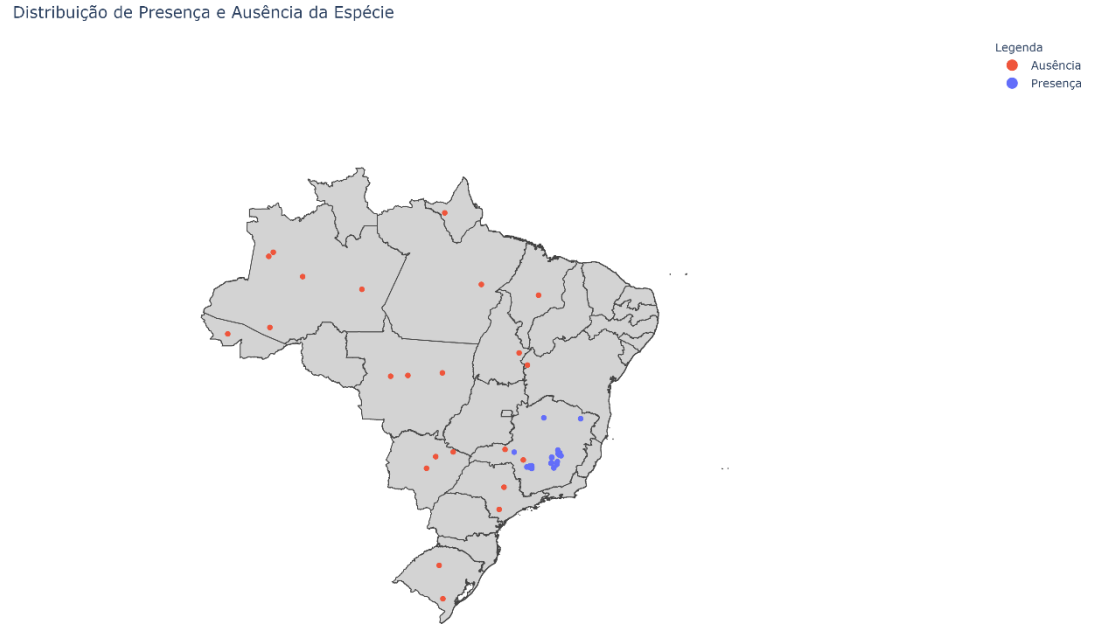

Figura 21: Mapa com os pontos de ocorrência e pseudoausência para a espécie *Whartonia pachywhartoni*.

A Figura 22 mostra os mapas preditivos de distribuição potencial para *W. pachywhartoni*, onde áreas em vermelho indicam maior probabilidade de ocorrência da espécie, e áreas em azul indicam ausência, de acordo com a saída padrão da biblioteca. Como observado anteriormente, os modelos apresentam variações em suas predições espaciais, mas há uma clara sobreposição nas áreas de maior adequabilidade ambiental, sugerindo um padrão coerente entre os diferentes algoritmos. Vale destacar que, nos modelos de distância, a escala de cores é invertida: valores mais baixos (mais próximos do vermelho) indicam maior similaridade com os pontos centrais de ocorrência.

Nos mapas gerados, observa-se que os resultados para os modelos de distância, estatísticos e de *machine learning* indicam padrões espaciais relativamente consistentes, ainda que com variações de abrangência e intensidade. No entanto, diferentemente de *W. nudosetosa*, as métricas obtidas para *W. pachywhartoni* (Tabela 5) demonstraram valores muito baixos em termos de AUC-ROC, acurácia, precisão, recall, F1 e TSS.

Tabela 5: Desempenho dos modelos utilizados na previsão de ocorrência *W. pachywhartoni*.

model	aucroc	acura	precis	recall	f1	tss	vp	vn	fp	fn
Bioclim	0.26	0.34	0.0	0.0	0.0	-0.32	0	53	25	78
Mahalanobis	0.24	0.35	0.0	0.0	0.0	-0.31	0	54	24	78
Euclidean	0.16	0.24	0.0	0.0	0.0	-0.53	0	37	41	78

Canberra	0.19	0.32	0.0	0.0	0.0	-0.36	0	50	28	78
Chebyshev	0.22	0.31	0.0	0.0	0.0	-0.37	0	49	29	78
Cosseno	0.21	0.29	0.0	0.0	0.0	-0.42	0	45	33	78
Minkowski	0.27	0.26	0.0	0.0	0.0	-0.47	0	41	37	78
Manhattan	0.23	0.25	0.0	0.0	0.0	-0.5	0	39	39	78
GAM	0.22	0.28	0.0	0.0	0.0	-0.44	0	44	34	78
GLM	0.28	0.35	0.0	0.0	0.0	-0.31	0	54	24	78
RF	0.22	0.29	0.0	0.0	0.0	-0.41	0	46	32	78
ANN	0.23	0.31	0.0	0.0	0.0	-0.37	0	49	29	78
SVM	0.25	0.34	0.0	0.0	0.0	-0.32	0	53	25	78

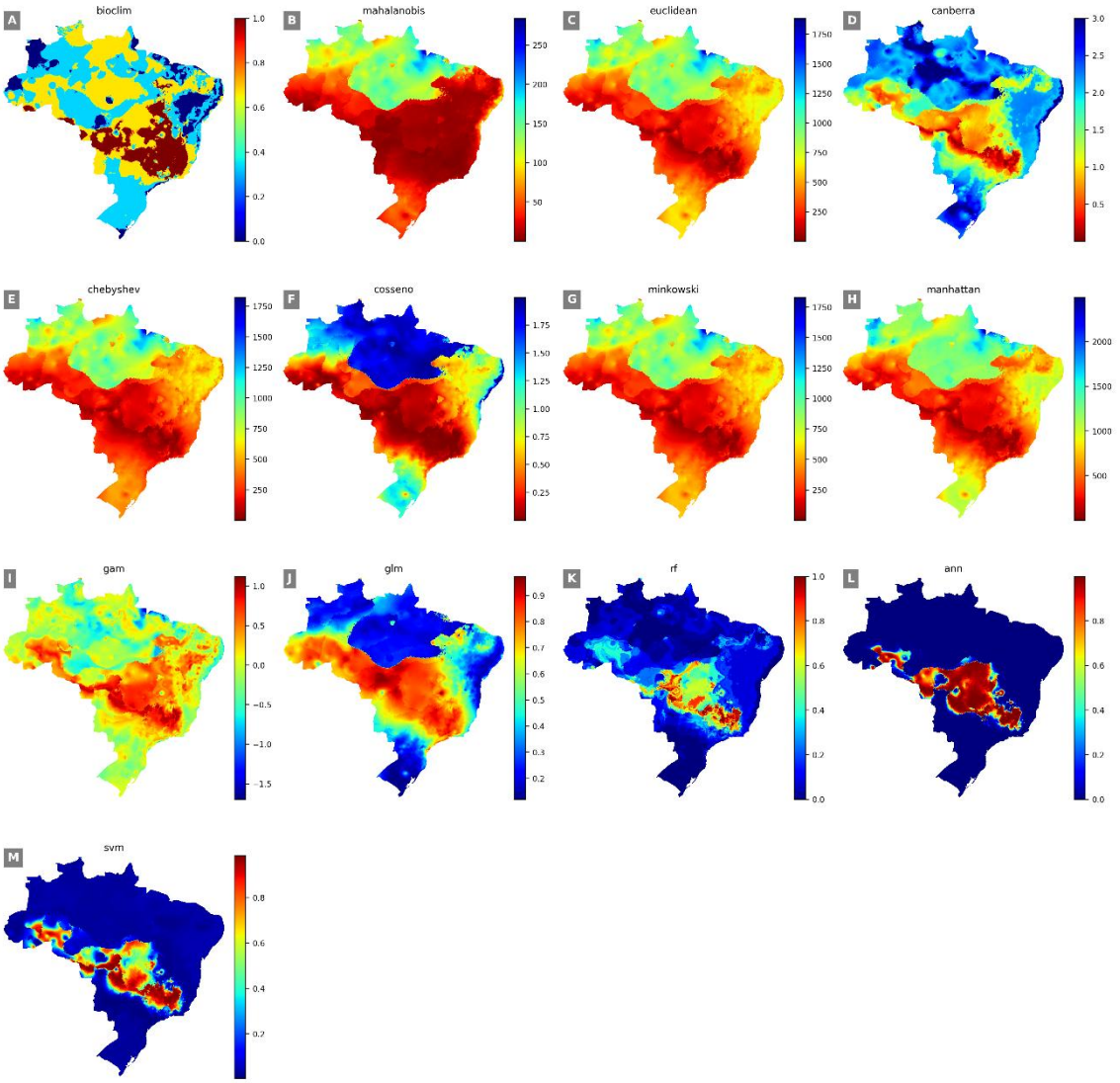

Figura 22: Mapas da distribuição potencial para *W. pachywhartoni*.

A modelagem de distribuição de espécies (SDM) para *Whartonia pachywhartoni* também foi baseada apenas nos dados de presença oriundos da coleção UFMG-AC, sem

complementação de registros da literatura, o que limita a abrangência da modelagem. A distribuição dos pontos de presença apresentou uma concentração geográfica bastante restrita, o que provavelmente influenciou a configuração espacial das predições obtidas. Além disso, a variedade limitada nos pontos de pseudoausência, muitos deles sobrepostos ou próximos, pode ter prejudicado a capacidade dos modelos em distinguir adequadamente entre condições de presença e ausência.

As áreas preditas como adequadas para *W. pachywhartoni* apresentaram uma forte sobreposição com aquelas preditas para *W. nudosetosa*, embora com extensão espacial reduzida. Esse padrão sugere que ambas as espécies compartilham características ecológicas semelhantes em relação às variáveis ambientais consideradas no modelo, mas que *W. pachywhartoni* pode ocupar um nicho mais restrito dentro desse espaço ambiental. Entretanto, essa limitação também pode ser decorrente da similaridade entre os próprios pontos de ocorrência utilizados, que não apresentam grande variação espacial ou ambiental. Essa falta de diversidade nas condições dos pontos de presença dificulta a capacidade dos modelos em capturar de forma precisa os limites ecológicos da espécie, o que reforça a necessidade de validação em campo para confirmar as áreas preditas e aprimorar a modelagem com dados mais variados.

A ausência de registros adicionais, especialmente aqueles históricos ou oriundos de outras fontes, impôs uma limitação severa ao modelo gerado. Sem uma diversidade maior de dados, o SDM produzido para *W. pachywhartoni* pode subestimar ou distorcer o verdadeiro potencial de distribuição da espécie. Ressalta-se ainda que alguns pontos considerados como ausência no modelo correspondem a áreas de novas coletas que ainda estão sob avaliação taxonômica, e podem futuramente ser confirmados como registros válidos da espécie.

Para aprimorar futuras modelagens de distribuição para *W. pachywhartoni*, recomenda-se o aumento expressivo da base de dados de ocorrência, a incorporação de registros de literatura, a realização de novos levantamentos em campo e a consideração de características específicas do ambiente dos hospedeiros. Além disso, a utilização de técnicas de modelagem que permitam trabalhar com poucos dados de forma robusta e a realização de validações cruzadas seriam estratégias importantes para aumentar a confiabilidade dos resultados.

O desenvolvimento da biblioteca EcoDistrib surgiu da necessidade de integrar, em um único fluxo automatizado, as etapas essenciais da modelagem de distribuição de espécies (SDM), desde a obtenção de variáveis ambientais até a aplicação e comparação

de múltiplos algoritmos de modelagem. A ferramenta foi projetada para ser prática, modular e flexível, contemplando uma ampla gama de funcionalidades: download automático de dados ambientais, recorte espacial personalizado, tratamento e seleção de variáveis para redução de multicolinearidade, geração de pseudoausências, aplicação de algoritmos de três grandes classes metodológicas (distância, estatística e aprendizado de máquina) e exportação de mapas preditivos padronizados.

A utilização do EcoDistrib no estudo com *Whartonia nudosetosa* e *Whartonia pachywhartoni* permitiu avaliar a eficiência da biblioteca em cenários com bases de dados relativamente pequenas e espacialmente restritas. Mesmo com desafios inerentes, como a concentração dos pontos de ocorrência em áreas específicas e a baixa variabilidade ambiental capturada, a ferramenta demonstrou robustez na execução dos diferentes algoritmos, fornecendo resultados comparáveis e com boa organização visual.

Um diferencial importante da biblioteca é o tratamento padronizado dos resultados, com mapas de distribuição usando escalas consistentes e ajustes automáticos para corrigir interpretações específicas de alguns algoritmos baseados em distância. A inclusão automatizada da Análise de Componentes Principais (PCA) foi fundamental para reduzir a dimensionalidade dos dados ambientais sem perder a capacidade de discriminar zonas com diferentes aptidões ambientais, especialmente quando o número inicial de variáveis era elevado.

Apesar dos avanços alcançados, algumas limitações também foram observadas. O desempenho final dos modelos gerados é fortemente influenciado pela qualidade dos dados de entrada. No caso das espécies estudadas, a forte similaridade espacial dos pontos de ocorrência, aliada à limitação no número de registros, impactou a capacidade dos modelos de generalizar para outras áreas. Assim, embora o EcoDistrib ofereça suporte a toda a cadeia de modelagem, a qualidade das previsões ainda depende criticamente de uma boa representatividade espacial e ambiental dos dados iniciais.

Outro ponto de atenção refere-se ao processo de geração de pseudoausências. Apesar do cuidado implementado para evitar sobreposição com pontos de ocorrência e ambientes inadequados, a geração aleatória ainda pode não capturar totalmente a complexidade ecológica real da ausência de uma espécie, o que pode influenciar o desempenho dos algoritmos estatísticos e de aprendizado de máquina.

Finalmente, cabe destacar que a atual versão do EcoDistrib prioriza a facilidade de uso e a modularidade, mas abre possibilidades para expansões futuras, como: integração de técnicas mais avançadas de seleção de variáveis, validação cruzada

automática, balanceamento dinâmico de pseudoausências, e conectividade direta com bancos de dados públicos de biodiversidade. Dessa forma, a biblioteca representa um passo inicial sólido para democratizar o acesso a modelagens de distribuição de espécies mais completas e reprodutíveis.

5. Conclusões

Este trabalho representa uma contribuição relevante para a sistematização e modernização da curadoria de dados biológicos, ao abordar de forma integrada a digitalização de uma coleção zoológica especializada e o desenvolvimento de uma ferramenta computacional voltada à modelagem de distribuição de espécies. A partir da Coleção Acarológica UFMG-AC, foi possível identificar e resolver diversos problemas estruturais e taxonômicos nos dados, utilizando técnicas automatizadas de padronização e validação baseadas no padrão Darwin Core. O processo revelou falhas comuns, como homônimos taxonômicos, ausência de hierarquia taxonômica completa, e inconsistências geográficas, que comprometem não apenas a qualidade dos dados, mas também sua reusabilidade por outras instituições e pesquisadores.

A informatização da UFMG-AC mostra-se essencial em um contexto nacional de vulnerabilidade das coleções biológicas. Casos recentes de perdas irreparáveis em acervos de relevância histórica, como o Museu Nacional e o Instituto Butantan, reforçam a necessidade urgente de estratégias de digitalização e disponibilização online. Nesse sentido, a publicação da coleção em repositórios como o GBIF e o SiBBr não apenas assegura sua preservação digital, mas também amplia seu potencial de uso por pesquisadores de todo o mundo. Além disso, a diversidade taxonômica e ecológica da coleção — com registros que incluem ácaros parasitas, marinhos, de vida livre, aquáticos e fósseis — fortalece seu papel como infraestrutura de pesquisa interdisciplinar.

A segunda grande frente deste trabalho foi o desenvolvimento da biblioteca **EcoDistrib**, uma ferramenta modular e acessível para modelagem de distribuição de espécies (SDM), voltada especialmente para contextos de dados escassos ou coleções não-padronizadas. A biblioteca foi projetada com foco na automação do fluxo completo de modelagem — desde o download e corte espacial de variáveis ambientais, até a aplicação de algoritmos, geração de pseudoausências, validação dos modelos e produção de mapas preditivos. O uso da EcoDistrib nas modelagens de *Whartonia nudosetosa* e *W. pachywhartoni* demonstrou sua funcionalidade e flexibilidade, mesmo diante de

limitações como o número de registros e a baixa variabilidade ambiental dos pontos de ocorrência.

Apesar dos resultados promissores, a aplicação prática evidenciou desafios inerentes na modelagem de espécies com poucos dados disponíveis. Entre os principais desafios, destacam-se o risco de sobreajuste (*overfitting*), a dificuldade em capturar a variabilidade ambiental real da espécie, e a influência da distribuição espacial dos dados de entrada sobre os resultados finais. A dependência exclusiva de registros primários, sem o apoio de dados da literatura ou repositórios adicionais, compromete a representatividade dos modelos. Nesse sentido, uma das principais recomendações metodológicas que emergem deste trabalho é a integração sistemática com bases de dados públicas e literatura científica para fortalecer a base de ocorrência.

A EcoDistrib, além de oferecer uma alternativa viável às ferramentas existentes, apresenta um diferencial importante ao unir métodos de distância, estatísticos e de aprendizado de máquina em uma única interface, com suporte à análise exploratória, redução de dimensionalidade por PCA, e validação cruzada com métricas robustas como AUC, TSS e F1. O desenvolvimento futuro da biblioteca, conforme proposto nas perspectivas deste trabalho, contempla a integração com o GBIF e SiBBr, a tradução para R, melhorias na interface e o benchmark frente a outras bibliotecas, consolidando-a como uma ferramenta confiável para a comunidade científica.

Em síntese, ao articular curadoria de dados, tecnologia da informação e ecologia computacional, esta dissertação contribui com uma abordagem metodológica replicável e adaptável, tanto para instituições com acervos biológicos subutilizados quanto para pesquisadores que buscam ferramentas acessíveis para explorar padrões espaciais de biodiversidade. Acredita-se que os resultados aqui apresentados servirão como base para novos projetos de modelagem, digitalização de coleções e desenvolvimento de ferramentas abertas que contribuam para a conservação e o entendimento da diversidade biológica em tempos de mudanças ambientais aceleradas.

6. Perspectivas

As perspectivas futuras para este trabalho incluem uma série de ações que visam aprimorar a biblioteca de SDM, ampliar sua acessibilidade e garantir a qualidade dos dados utilizados. Essas ações foram organizadas em etapas, desde as mais simples e de curto prazo até as mais complexas e de longo prazo.

A primeira etapa consiste em melhorar a padronização e a qualidade dos dados de ocorrência. Para isso, será desenvolvido um *script* que visa corrigir erros de coerência nos dados geográficos, garantindo que as coordenadas estejam consistentes e dentro dos limites esperados. Além disso, os dados de ocorrência de 2024 serão adicionados às plataformas do GBIF e SiBBr, após uma rigorosa conferência e validação desses dados.

Impacto Esperado

- A correção de erros geográficos aumentará a confiabilidade dos dados utilizados nas análises, reduzindo a ocorrência de resultados distorcidos ou incorretos.
- A inclusão dos dados de 2024 no GBIF e SiBBr ampliará a disponibilidade de informações atualizadas para a comunidade científica, facilitando pesquisas futuras e promovendo a integração de dados.

Outra perspectiva importante é a divulgação da biblioteca de SDM e a capacitação de usuários. Para isso, estão planejados:

- A escrita e publicação de artigos científicos que descrevam a biblioteca, suas funcionalidades e aplicações, visando ampliar sua visibilidade na comunidade acadêmica.
- A ministração de um curso de curta duração, com o objetivo de capacitar pesquisadores e estudantes no preenchimento correto de planilhas de dados e no uso do formato DwC, que é amplamente utilizado para padronização de dados de biodiversidade.

Impacto esperado

- A publicação de artigos científicos aumentará o reconhecimento da biblioteca, atraindo mais usuários e colaboradores.
- Curso de capacitação garantirá que os usuários utilizem a biblioteca de forma eficiente e correta, reduzindo erros no preenchimento de dados e promovendo a padronização.

No âmbito técnico, as principais perspectivas envolvem o refinamento e a expansão da biblioteca de SDM. Isso inclui:

- A correção de "*bugs"* no código, garantindo maior estabilidade e confiabilidade.
- A implementação de novas funcionalidades, como a integração direta com as plataformas GBIF e SiBBr, permitindo que os usuários obtenham dados de ocorrência diretamente dessas fontes.

- A tradução dos códigos do EcoDistrib para a linguagem R, com o objetivo de tornar a biblioteca mais acessível a pesquisadores da área biológica, que frequentemente utilizam R para análises ecológicas.

Impacto Esperado

- A correção de bugs melhorará a experiência do usuário, aumentando a confiança na biblioteca.
- A integração com GBIF e SiBBr simplificará o processo de coleta de dados, tornando a biblioteca mais eficiente e atrativa.
- A tradução para R ampliará o público-alvo da biblioteca, permitindo que mais pesquisadores da área biológica utilizem as ferramentas disponíveis.

Será conduzido um *benchmark* para comparar o desempenho do EcoDistrib com outras ferramentas consolidadas, como *MaxEnt*, *biomod2* e *sdm* (R). Serão utilizados dados padronizados e métricas reconhecidas, em diferentes cenários e volumes de dados. O objetivo é avaliar a acurácia, eficiência computacional e usabilidade da biblioteca.

Impacto Esperado

- A comparação objetiva com outras ferramentas destacará os pontos fortes e as limitações do EcoDistrib, orientando ajustes e melhorias futuras.
- O benchmark fornecerá evidências concretas da eficácia da biblioteca, aumentando sua credibilidade na comunidade científica.
- Os resultados poderão ser utilizados na elaboração de artigos comparativos, contribuindo para a disseminação e validação da ferramenta.

Por fim, as ações de longo prazo incluem o desenvolvimento contínuo da biblioteca, com o acréscimo de funcionalidades avançadas e a adaptação às necessidades emergentes da comunidade científica. Isso envolve:

- A incorporação de novas técnicas de modelagem e análise de dados, mantendo a biblioteca atualizada com as tendências da área.
- A criação de uma comunidade de usuários e colaboradores, que possam contribuir com sugestões, relatos de bugs e desenvolvimento de novas funcionalidades.

Impacto Esperado

- A incorporação de técnicas avançadas manterá a biblioteca relevante e útil para a comunidade científica, atendendo às demandas por métodos inovadores.
- A criação de uma comunidade de usuários e colaboradores promoverá o crescimento sustentável da biblioteca, com contribuições contínuas que garantirão sua evolução e adaptação.

7. Referências bibliográficas

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of applied ecolog, 43*(6), pp. 1223-1232.

Angeles, N. A. C., & Catap, E. S. (2023). Challenges on the development of biodiversity biobanks: The living archives of biodiversity. *Biopreservation and Biobanking*, *21*(1), 5-13.

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in ecology & evolution, 22*(1), pp. 42-47.

Arengo, F., Porzecanski, A. L., Blair, M. E., Amato, G., Filardi, C., & Sterling, E. J. (2017). *The essential role of museums in biodiversity conservation.*

Austin, M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological modelling, 200*(1-2), pp. 1-19.

Ball-Damerow, J. E., Brenskelle, L., Barve, N., Soltis, P. S., Sierwald, P., Bieler, R., . . . Guralnick, R. P. (2019). Research applications of primary biodiversity databases in the digital age. *PloS one, 14*(9), p. e0215794.

Basílio, D. S., Petters-vandresen, D., Henriques, D. K., Morais, G. D., Milhorini, S. D., & Marinoni, L. (2024). Identificação e Caracterização. Em L. MARINONI, D. S. BASÍLIO, & A. L. GASPER, *Coleções biológicas científicas brasileiras: diagnóstico, prioridades e recomendações.* (pp. 17-48). Curitiba: Sociedade Brasileira de Zoologia (SBZ).

Bassini-Silva, R., Zampaulo, R. D. A., Welbourn, C., Ochoa, R., Brescovit, A. D., Barros-Battesti, D. M., & Jacinavicius, F. D. C. (2022). A new genus and two new species of chigger mites (Trombidiformes: Leeuwenhoekiidae) from Brazilian caves with notes about the genus Whartonia Ewing, 1944. *Journal of Natural History*, *56*(29-32), 1297-1313.

Bassini-Silva, R., de Almeida, B. R., Lourenço, E. C., Welbourn, C., Ochoa, R., Famadas, K. M., ... & de Castro Jacinavicius, F. (2025). The rediscovery of the types of Whartonia pachywhartoni Vercammen-Grandjean, 1966 (Trombidiformes: Leeuwenhoekiidae) with new records in Brazil. *International Journal of Acarology*, 1-7.

Bauer, A. M., & Wahlgren, R. (2013). On the Linck collection and specimens of snakes figured by Johann Jakob Scheuchzer (1735)–the oldest fluid-preserved herpetological collection in the world. *Bonn Zoological Bulletin, 62*, pp. 220-252.

Bauer, A. M., Ceregato, A., & Delfino, M. (2013). The oldest herpetological collection in the world: the surviving amphibian and reptile specimens of the Museum of Ulisse Aldrovandi. *Amphibia-Reptilia, 31*, pp. 305-321.

Beaman, R. S., & Cellinese, N. (2012). Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology letters, 15*(4), pp. 365-377.

Benham, P. M., & Bowie, R. C. (2023). Natural history collections as a resource for conservation genomics: Understanding the past to preserve the future. *Journal of Heredity, 114*(4), pp. 367-384.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MAXENT studies. *Diversity and Distributions, 20*(1), pp. 1-9.

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some simple guidelines for effective data management. *Bulletin of the Ecological Society of America*, *90*(2), 205-214.

Brasil. Ministério do Meio Ambiente. (2021). *Política Nacional de Biodiversidade: Diretrizes para a Gestão de Dados*.

Breiman, L. (2001). Random forests. *Machine learning, 45*, pp. 5-32.

Brotons, L., Thuiller, W., Araújo, M. B., & Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, *27*(4), 437-448.

BUTANTAN. (s.d.). *Histórico.* Acesso em 07 de janeiro de 2025, disponível em https://butantan.gov.br/institucional/historico

Castillo-Figueroa, D. (2018). Beyond specimens: linking biological collections, functional ecology and biodiversity conservation. *Revista peruana de biología, 25*(3), pp. 343-348.

Cavallin, E. K., Munhoz, C. B., Harris, S. A., Villarroel, D., & Proença, C. E. (2016). Influence of biological and social-historical variables on the time taken to describe an angiosperm. *American Journal of Botany, 103*(113), pp. 2000-2012.

CCT-UFMG. (s.d.). *Centro de Coleções Taxonômicas*. Acesso em 07 de janeiro de 2025, disponível em https://www2.icb.ufmg.br/cct/

Chapman, A. D. (2005). *Principles and methods of data cleaning*. GBIF.

Chen, T., & Guestrin, C. (August de 2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75-88.

Cortes, C. (1995). Support-Vector Networks. *Machine Learning.*

Costa, S. G., Tolstikov, A., Saboori, A., Batista-Ribeiro, D., Noei, J., Harvey, M. S., ... & Pepato, A. R. (2024). A comprehensive molecular phylogeny of the terrestrial Parasitengona (Acariformes, Prostigmata) provides insights into the evolution of their metamorphosis, invasion into aquatic habitats and classification. *Molecular Phylogenetics and Evolution*, *199*, 108147.

Da Silveira, P. S. A., de Oliveira Bernardi, L. F., & Pepato, A. R. (2015). New records of the genus Whartonia (Acari, Leeuwenhoekiidae) associated with the bat Carollia perspicillata from southeastern Brazil. *Check List*, *11*(6), 1793-1793.

De Vivo, M., Silveira, L. F., & Nascimento, F. O. (2014). Reflexões sobre coleções zoológicas, sua curadoria e a inserção dos Museus na estrutura universitária brasileira. *Arquivos de Zoologia, 45*, pp. 105-113.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., & .... Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), pp. 27-46.

Elith*, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., . . . E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*(2), pp. 129-151.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics, 40*(1), pp. 677-697.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of animal ecology, 77*(4), pp. 802-813.

Feng, X., Enquist, B. J., Park, D. S., Boyle, B., Breshears, D. D., Gallagher, R. V., . . . López-Hoffman, L. (2022). A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. *Global Ecology and Biogeography, 31*(7), pp. 1242-1260.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation, 24*(1), pp. 38-49.

Fontaine, B., Perrard, A., & Bouchet, P. (2012). 21 years of shelf life between discovery and description of new species. *Current Biology, 22*(22), pp. R943-R944.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics, 19*(1), pp. 1-67.

GBIF. (s.d.). *Quick guide to publishing data through GBIF.org*. Acesso em 07 de janeiro de 2025, disponível em https://www.gbif.org/

Giovanelli, J. G., Araujo, C. O., Haddad, C. F., & Alexandrino, J. (2008). Modelagem do nicho ecológico de Phyllomedusa ayeaye (Anura: Hylidae): previsão de novas áreas de ocorrência para uma espécie rara. *Neotropical Biology and Conservation, 3*(2), pp. 59-65.

Glienke, C., Petters-vandresen, D. A., Souto, A. D., Marinoni, L., & Da Silva, M. (2024). Microbiological Collections in Brazil: Current Status and Perspectives. *Diversity, 16*(2), p. 116.

Gobble, M. M. (2018). Digitalization, digitization, and innovation. *Research-Technology Management, 61*(4), pp. 56-59.

Gomes-Almeida, B. K., Costa, S. G., Ribeiro, D. B., Bernardi, L. F., & Pepato, A. R. (2023). First multi-instar descriptions of cave-dwelling Whartonia Ewing, 1944 (Parasitengona, Leeuwenhoekiidae) from Brazil through integrative taxonomy. *Systematic and Applied Acarology*, *28*(3), 568-606.

Guedes, J. J., Feio, R. N., Meiri, S., & Moura, M. R. (2020). Identifying factors that boost species discoveries of global reptiles. *Zoological Journal of the Linnean Society, 190*(4), pp. 1274-1284.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology letters, 8*(9), pp. 993-1009.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological modelling, 135*(2-3), pp. 147-186.

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with applications in R.* Cambridge University Press.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., . . . Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology letters, 16*(12), pp. 1424-1435.

Guralnick, R., & Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics, 25*(4), pp. 421-428.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., ... & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156-162.

Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, *43*(4), 549-558.

Hardisty, A. R., Michener, W. K., Agosti, D., García, E. A., Bastin, L., Belbin, L., ... & Kissling, W. D. (2019). The Bari Manifesto: An interoperability framework for essential biodiversity variables. *Ecological informatics*, *49*, 22-31.

Hardisty, A., Roberts, D., & Biodiversity Informatics Community Biodiversity-community-list@ nhm. ac. uk. (2013). A decadal view of biodiversity informatics: challenges and priorities. *BMC ecology*, *13*, 1-23.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer.

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences, 118*(6).

Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Grassa, C. J., Park, D. S., . . . Davis, C. C. (2020). Digitization and the future of natural history collections. *BioScience, 70*(3), pp. 243-251.

Hilton, E. J., Watkins-Colwell, G. J., & Huber, S. K. (2021). The expanding role of natural history collections. *Ichthyology & Herpetology, 109*(2), pp. 379-391.

iDigBio. (s.d.). Acesso em 07 de janeiro de 2025, disponível em https://www.idigbio.org/

IPCC. (2023). *Climate Change 2023: Synthesis Report*. IPCC. Disponível em https://www.ipcc.ch/report/ar6/syr/

JBRJ. (s.d.). *História*. Acesso em 07 de janeiro de 2025, disponível em https://www.gov.br/jbrj/pt-br/assuntos/299

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., ... & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature ecology & evolution*, *3*(4), 539-551.

Kellner, A. W. (2024). Biological collections in danger? *Anais da Academia Brasileira de Ciências, 96*(1).

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology letters*, *12*(4), 334-350.

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J., & Kreft, H. (2019). Biodiversity data integration—the significance of data resolution and domain. *PLoS biology, 17*(3), p. e3000183.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pp. 159-174.

Lapp, H., Morris, R. A., Catapano, T., Hobern, D., & Morrison, N. (2011). Organizing our knowledge of biodiversity. *Bulletin of the American Society for Information Science and Technology, 37*(4), pp. 38-42.

Lek, S., & Guégan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological modelling*, *120*(2-3), 65-73.

Lima, A. R., & Faleiro, B. T. (2020). Coleções Biológicas Científicas. Em C. B. Oswald, C. A. Dias, G. S. Garbino, & J. C. Oliveira, *Princípios de Sistemática Zoológica* (pp. 69-77). Belo Horizonte.

Luján, M., Lemos, R. M., Lucas, E., Michelangeli, F. A., Prance, G. T., Pennington, T. D., . . . Zuntini, A. R. (2024). Trials and tribulations of Neotropical plant taxonomy: pace of tree species description. *Plants, People, Planet, 6*(2), pp. 515-527.

Marinoni, L., Gasper, A. L., ..., Chiquito, E. A., Glienke, C., Fonseca, C. B., . . . Vicente, V. A. (2024). *Introdução e orientações às boas práticas para as Coleções Biológicas Científicas Brasileiras.* Sociedade Brasileira de Zoologia.

McCain, C. M. (2007). Could temperature and water availability drive elevational species richness patterns? A global case study for bats. *Global Ecology and biogeography, 16*(1), pp. 1-13.

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography, 36*(10), pp. 1058-1069.

Miller, S. E., Barrow, L. N., Ehlman, S. M., Goodheart, J. A., Greiman, S. E., Lutz, H. L., . . . Light, J. E. (2020). Building natural history collections for the twenty-first century and beyond. *BioScience, 70*(8), pp. 674-687.

MNRJ. (s.d.). *O Museu | Museu Nacional - UFRJ.* Acesso em 07 de janeiro de 2025, disponível em https://www.museunacional.ufrj.br/dir/omuseu/omuseu.html

Monda, L. (2019). Biodiversity Data Management: Regional challenges. *Biodiversity Information Science and Standards*.

Monfils, A. K., Powers, K. E., Marshall, C. J., Martine, C. T., Smith, J. F., & Prather, L. A. (2017). Natural history collections: Teaching about biodiversity across time, space, and digital platforms. *Southeastern Naturalist, 16*(10), pp. 47-57.

Moudrý, V., & Devillers, R. (2020). Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecological Informatics, 56*, p. 101051.

MZUSP. (s.d.). *História – Museu de Zoologia da USP*. Acesso em 07 de janeiro de 2025, disponível em https://mz.usp.br/pt/museu/historia/

Nelson, G., & Ellis, S. (2019). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B, 374*(1763).

Nogués-Bravo, D. (2009). Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography, 18*(5), pp. 521-531.

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *nature, 421*(6918), pp. 37-42.

Pearson, K. D. (2018). Rapid enhancement of biodiversity occurrence records using unconventional specimen data. *Biodiversity and Conservation, 27*(11), pp. 3007-30018.

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global ecology and biogeography, 12*(5), pp. 361-371.

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H., Scholes, R. J., ... & Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*(6117), 277-278.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49).* Princeton University Press.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling, 190*(3-4), pp. 231-259.

Rafael, J. A., Aguiar, A. P., & Amorim, D. D. (2009). Knowledge of insect diversity in Brazil: challenges and advances. *Neotropical Entomology, 38*, pp. 565-570.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703-705.

Sakamoto, T., & Ortega, J. M. (2021). Taxallnomy: an extension of NCBI Taxonomy that produces a hierarchically complete taxonomic tree. *BMC bioinformatics, 22*, pp. 1-23.

Santos, C. R., Aviz, D., & Albuquerque, E. Z. (2019). Coleções biológicas do Museu Paraense Emílio Goeldi: 150 anos de história. Estado Atual e Perspectivas Futuras. Em A. V. GALÚCIO, & A. L. PRUDENTE, *Museu Goeldi: 150 anos de ciências na amazônia.* (pp. 248-272). Belém: Instituto Brasileiro de Informação em Ciência e Tecnologia.

SiBBr. (s.d.). *O que é o SiBBr*. Acesso em 07 de janeiro de 2025, disponível em https://sibbr.gov.br/

Silva, L. A. E. D., Fraga, C. N. D., Almeida, T. M. H. D., Gonzalez, M., Lima, R. O., Rocha, M. S. D., ... & Forzza, R. C. (2017). Jabot-Sistema de Gerenciamento de Coleções Botânicas: a experiência de uma década de desenvolvimento e avanços. *Rodriguésia*, *68*, 391-410.

Smith, V., Georgiev, M. T., Stoev, P., Biserkov, M. J., Miller, J., Livermore, M. L., ... & Penev, L. (2013). Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal. *Biodiversity data journal*, (1), e995.

Snow, N. (2005). Successfully curating smaller herbaria and natural history collections in academic settings. *BioScience, 55*(9), pp. 771-779.

Soberon, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics, 2*, pp. 1-10.

SpeciesLink. (s.d.). Acesso em 07 de janeiro de 2025, disponível em https://specieslink.net/

Speed, J. D. (2018). Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PLoS One, 13*(4), p. e0196417.

Sterner, B. W., Gilbert, E. E., & Franz, N. M. (2020). Decentralized but globally coordinated biodiversity data. *Frontiers in Big Data*, *3*, 519133.

Suarez, A. V., & Tsutsui, N. D. (2004). The value of museum collections for research and society. *BioScience, 54*(1), pp. 66-74.

Takahashi, M., Takahashi, H., & Kikuchi, H. (2006). Whartonia (Fascutonia) natsumei (Acari: Trombiculidae): a new bat chigger collected from Plecotus auritus (Chiroptera: Vespertilionidae) in Japan, with host and distribution records of the genus Whartonia. *Journal of medical entomology*, *43*(2), 128-137.

TDWG. (s.d.). *Darwin Core*. Acesso em 07 de janeiro de 2025, disponível em https://dwc.tdwg.org/

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one*, *6*(6), e21101.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., & Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences, 102*(23), pp. 8245-8250.

Tibshirani, R. J., & Efron, B. (1993). *An introduction to the bootstrap.* Chapman & Hall.

Tittensor, D. P., Walpole, M., Hill, S. L., Boyce, D. G., Britten, G. L., Burgess, N. D., ... & Ye, Y. (2014). A mid-term analysis of progress toward international biodiversity targets. *Science*, *346*(6206), 241-244.

Wen, J., Ickert-bond, S. M., Appelhans, M. S., Dorr, L. J., & Funk, V. A. (2015). Collections-based systematics: Opportunities and outlook for 2050. *Journal of Systematics and Evolution, 53*(6), pp. 477-488.

Wieczorek, J. B., Guralnick, R., Blum, S., Döring, M., Giovanni, R., ..., & Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PloS one, 7*(1), p. e29715.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

Willemse, L. P. (2008). Standardisation in data-entry across databases: Avoiding Babylonian confusion. *Taxon, 57*(2), pp. 343-345.

Zaher, H., & Young, P. S. (2003). ZAHER, Hussam; YOUNG, Paulo S. *Ciência e Cultura, 55*(3), pp. 24-26.

Zamudio, K. R., Kellner, A., Serejo, C., De Britto, M. R., Castro, C. B., Buckup, P. A., . . . Rocha, L. A. (2018). Lack of science support fails Brazil. *Science, 361*(6409), pp. 1322-1323.

Zenglein, F. (2025). The standardization of biodiversity: how politicization changes standardization for corporate sustainability reporting. *Frontiers in Sustainability*, p. 1433799.

Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., ... & Tingley, M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, *19*(1), 30-38.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., ... & Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261-1277.

8. Apêndices

Tabela 6: Descrição das 19 variáveis bioclimáticas do *Wordclim* utilizados na modelagem de distribuição de espécies (SDM). As variáveis representam parâmetros térmicos e pluviométricos derivados de médias anuais, sazonais e extremos mensais. **Fonte**: Dados bioclimáticos disponibilizados por *WorldClim*.

Código Bioclimático	**Descrição da Variável**
BIO1	Temperatura Média Anual
BIO2	Amplitude Térmica Diária (Média mensal (temp máx- temp mín))
BIO3	Isotermalidade (BIO2/BIO7 × 100)
BIO4	Sazonalidade da Temperatura (desvio padrão × 100)
BIO5	Temperatura Máxima do Mês Mais Quente
BIO6	Temperatura Mínima do Mês Mais Frio
BIO7	Amplitude Térmica Anual (BIO5 - BIO6)
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 da Precipitação (Coeficiente de Variação)
BIO16	Precipitação do Trimestre Mais Úmido
BIO17	Precipitação do Trimestre Mais Seco
BIO18	Precipitação do Trimestre Mais Quente
BIO19	Precipitação do Trimestre Mais Frio

Tabela 7: Valores para cada ponto de ocorrência presença e ausência, nas camadas após o filtro da correlação para *W. nudosetosa*.

Lat	Lon	pres	bio16	bio17	bio18	bio19	bio5	bio7	bio8	bio9
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-18.32	-42.14	1	609	45	424	45	31.16	18.08	24.47	20.08
-6.46	-50.24	1	808	111	304	808	33.81	15.87	24.74	25.46
-6.46	-50.24	1	808	111	304	808	33.81	15.87	24.74	25.46
-16.15	-44.63	1	620	19	325	33	31.71	20.26	23.49	20.31
-16.15	-44.63	1	620	19	325	33	31.71	20.26	23.49	20.31
-19.22	-43.4	1	805	41	600	41	28.34	18.91	22.04	16.95
-6.46	-50.24	1	808	111	304	808	33.81	15.87	24.74	25.46
-6.46	-50.24	1	808	111	304	808	33.81	15.87	24.74	25.46
-3.32	-52.27	1	982	119	149	937	32.36	12.49	25.02	26.36
-19.56	-43.97	1	751	35	556	35	29.17	18.44	22.74	18.49
-19.56	-43.97	1	751	35	556	35	29.17	18.44	22.74	18.49
-19.56	-43.97	1	751	35	556	35	29.17	18.44	22.74	18.49
-19.56	-43.97	1	751	35	556	35	29.17	18.44	22.74	18.49
-19.56	-43.97	1	751	35	556	35	29.17	18.44	22.74	18.49
-19.56	-43.97	1	751	35	556	35	29.17	18.44	22.74	18.49
-19.56	-43.97	1	751	35	556	35	29.17	18.44	22.74	18.49
-12.6	-46.41	0	753	12	157	32	33.69	16.31	25.14	24.72
-3.95	-62.9	0	850	320	599	544	31.7	9.94	26.37	26.6
-2.86	-46.04	0	1012	123	192	925	31.48	10.88	25.64	26.47
-11.37	-38.64	0	257	106	198	207	30.32	14.05	23.77	22.99
-12.24	-60.53	0	876	48	424	87	32.92	18.51	24.68	23.49
-29.16	-55.12	0	468	413	421	413	30.87	21.92	16.73	14.63
3.81	-60.7	0	855	169	254	809	33.32	11.43	26.23	27.61

Tabela 8: Valores para cada ponto de ocorrência presença e ausência, nas camadas após o filtro da correlação para *W. pachywhartoni*.

Lat	Lon	pres	bio16	bio17	bio18	bio19	bio5	bio7	bio8	bio9
-16.22	-41.48	1	469	26	300	26	29.86	17.21	23.53	19.52
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-18.91	-43.43	1	784	42	588	42	28.49	19.24	22.21	16.93
-18.91	-43.43	1	784	42	588	42	28.49	19.24	22.21	16.93
-16.15	-44.63	1	620	19	325	33	31.71	20.26	23.49	20.31
-19.22	-43.39	1	805	41	600	41	28.34	18.91	22.04	16.95
-19.17	-43.27	1	804	43	595	43	28.19	18.98	21.89	16.74
-19.22	-43.37	1	776	35	575	35	29.55	19.56	22.99	17.78
-19.17	-43.27	1	794	43	589	43	28.74	19.07	22.37	17.20
-19.22	-43.39	1	805	41	600	41	28.34	18.91	22.04	16.95
-19.17	-43.28	1	804	43	595	43	28.19	18.98	21.89	16.74
-19.09	-43.36	1	799	42	595	42	28.43	19.06	22.12	16.91
-19.09	-47.17	1	825	28	696	52	27.60	15.98	22.40	18.45
-19.22	-43.39	1	805	41	600	41	28.34	18.91	22.04	16.95
-19.22	-43.39	1	805	41	600	41	28.34	18.91	22.04	16.95
-19.22	-43.39	1	805	41	600	41	28.34	18.91	22.04	16.95
-19.09	-43.37	1	799	42	595	42	28.43	19.06	22.12	16.91
-19.17	-43.27	1	794	43	589	43	28.74	19.07	22.37	17.20
-19.17	-43.27	1	794	43	589	43	28.74	19.07	22.37	17.20
-19.22	-43.39	1	805	41	600	41	28.34	18.91	22.04	16.95
-18.94	-43.41	1	779	41	583	41	28.91	19.53	22.53	17.15
-20.35	-46.07	1	757	56	630	56	28.61	20.58	22.56	16.93
-20.35	-46.08	1	757	56	630	56	28.61	20.58	22.56	16.93
-20.15	-43.51	1	864	65	681	65	24.15	17.30	18.26	13.92
-19.33	-43.31	1	790	36	586	36	29.04	19.21	22.54	17.46
-19.4	-43.14	1	796	43	589	43	28.12	18.73	21.76	16.75
-20.44	-43.77	1	806	50	649	50	25.98	17.54	19.90	15.59
-19.9	-43.47	1	830	62	635	62	26.80	17.81	20.62	16.16
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-20.42	-45.69	1	745	59	589	59	28.47	20.31	22.10	16.94
-20.34	-45.78	1	745	56	591	56	28.90	20.74	22.44	17.18
-20.37	-45.66	1	741	56	579	56	28.62	20.28	22.22	17.11
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.29	-45.85	1	745	55	596	55	28.93	20.85	22.42	17.09
-20.42	-45.77	1	746	59	596	59	28.72	20.55	22.31	17.07
-20.42	-45.77	1	746	59	596	59	28.72	20.55	22.31	17.07
-20.42	-45.77	1	746	59	596	59	28.72	20.55	22.31	17.07
-20.25	-45.67	1	738	52	582	52	29.32	21.10	22.71	17.39
-20.31	-45.68	1	740	55	582	55	28.89	20.61	22.39	17.22
-20.31	-45.68	1	740	55	582	55	28.89	20.61	22.39	17.22
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.31	-45.68	1	740	55	582	55	28.89	20.61	22.39	17.22
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86

-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.29	-45.85	1	745	55	596	55	28.93	20.85	22.42	17.09
-20.29	-45.85	1	745	55	596	55	28.93	20.85	22.42	17.09
-20.29	-45.85	1	745	55	596	55	28.93	20.85	22.42	17.09
-20.47	-45.66	1	748	60	592	60	28.38	20.27	22.01	16.86
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-19.54	-43.94	1	744	33	554	33	29.42	18.54	22.94	18.59
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.03	-43.99	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.01	-43.98	1	891	49	691	49	26.26	17.18	20.13	16.24
-20.48	-54.68	0	635	130	569	201	30.63	16.26	25.13	21.24
-18.86	-47.95	0	822	25	704	43	28.08	16.27	22.14	19.10
-12.59	-57.74	0	1008	38	709	73	32.45	18.39	24.98	23.24
-19.08	-52.39	0	711	77	688	111	30.10	17.92	24.41	20.48
-1.99	-67.81	0	920	659	742	873	31.19	10.49	25.64	26.25
-10.59	-46.74	0	653	10	329	36	34.24	19.42	24.71	23.82
-31.63	-53.27	0	405	309	346	396	28.27	20.61	13.27	21.62
-5.16	-60.22	0	934	253	447	934	32.76	11.15	26.61	27.21
-22.10	-48.04	0	665	89	596	89	27.53	17.54	22.48	17.11
-28.77	-53.58	0	454	405	423	406	29.66	21.34	18.65	15.02
-11.64	-46.02	0	661	10	317	32	32.92	19.40	23.32	22.36
-12.31	-53.32	0	871	16	667	41	34.67	20.95	25.35	23.60
-8.97	-71.69	0	698	139	613	139	30.86	13.30	25.28	23.85
-4.07	-65.28	0	857	341	482	463	32.00	10.66	26.43	26.54
-23.98	-48.45	0	517	157	511	157	27.11	18.59	21.87	14.99
-8.42	-68.09	0	860	132	686	132	31.58	14.42	25.37	23.98
-12.53	-56.26	0	1055	35	764	35	32.32	18.16	24.69	23.16
-5.67	-45.07	0	595	33	138	94	33.98	15.99	25.40	26.25
-2.32	-68.20	0	926	676	695	843	31.10	10.42	25.82	25.66
-4.73	-49.97	0	951	100	153	951	32.39	12.36	25.19	26.24
1.39	-53.11	0	936	210	210	777	31.68	11.20	24.95	26.06
-19.47	-53.90	0	681	103	631	151	30.31	17.30	24.55	20.72
-19.76	-46.38	0	829	55	692	55	26.07	17.76	20.53	16.04

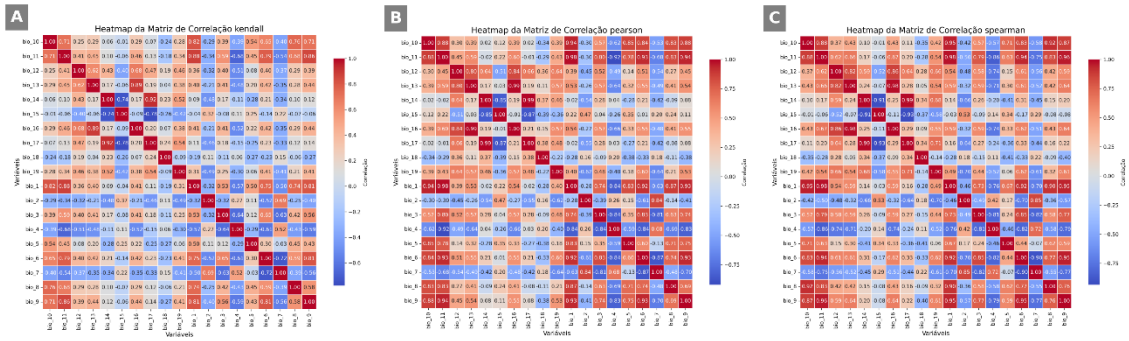

Figura 23: Matriz de correlação das variáveis ambientais. (A) Correlação de Kendall, (B) Correlação de Pearson e (C) Correlação de Spearman.

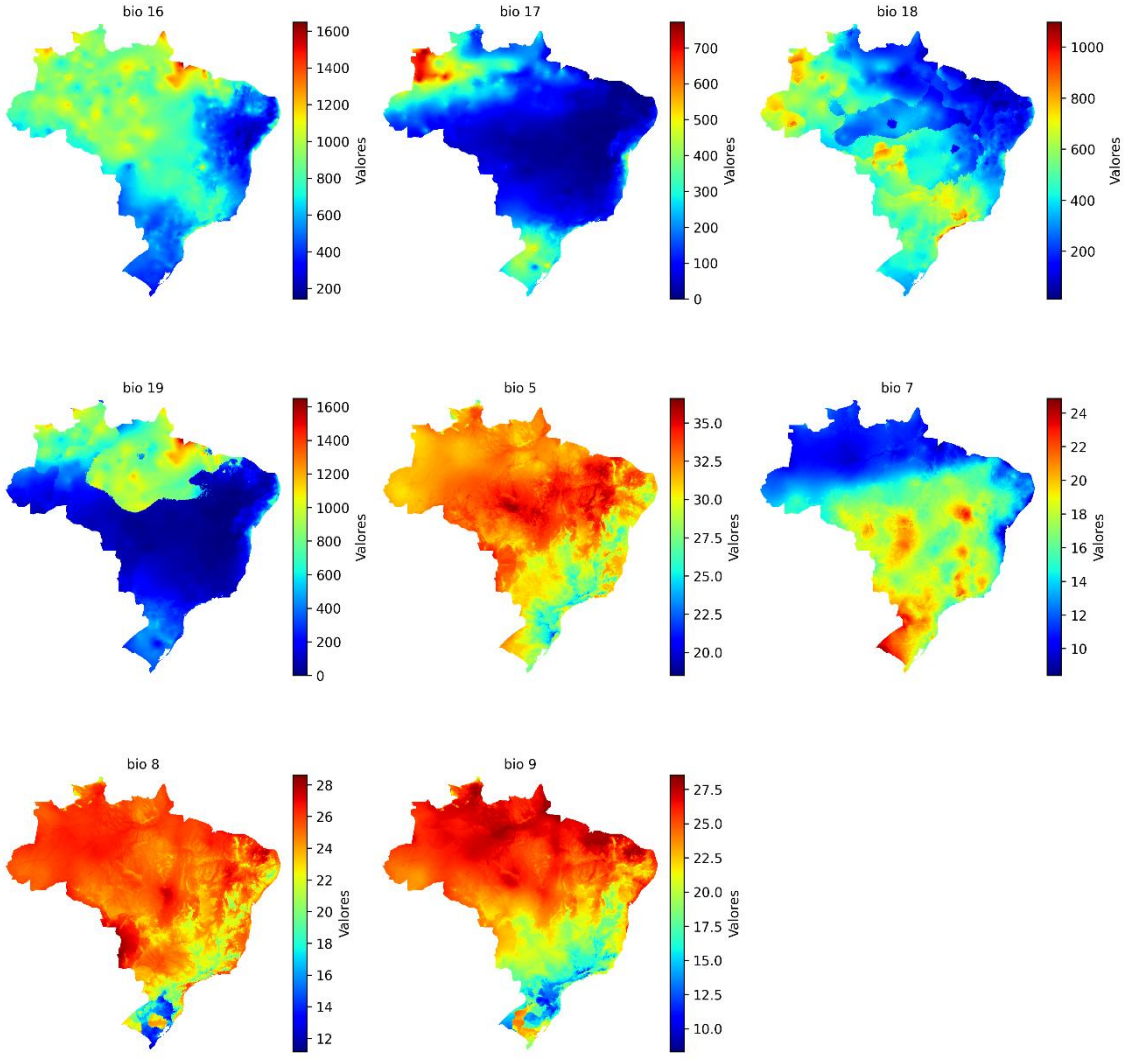

Figura 24: Mapas das variáveis após o filtro da correlação de Pearson.

EcoDistrib

A **EcoDistrib** é uma biblioteca Python para modelagem de distribuição de espécies (SDM - Species Distribution Modeling). Ela oferece ferramentas para baixar variáveis ambientais, pré-processar dados (como corte de áreas de interesse, aplicação de matriz de correlação e PCA), e realizar modelagem de distribuição de espécies usando diversos métodos (distância, estatísticos e machine learning). Além disso, a biblioteca calcula métricas de avaliação para os modelos gerados.

Instalação

Para instalar a biblioteca, utilize o seguinte comando: Dependências A biblioteca depende dos seguintes pacotes Python:

- pandas
- numpy
- rasterio
- scikit-learn
- matplotlib
- geopandas
- shapely

Certifique-se de que todas as dependências estão instaladas antes de usar a biblioteca.

Funcionalidades Principais

1. **Download de Dados Ambientais**
   Baixa variáveis ambientais de fontes públicas (por exemplo, WorldClim).
   Classe: DataDownloader
   Método: download_data

2. **Manipulação de Shapefiles**
   Cria shapefiles para países ou estados específicos.
   Classe: ShapefileHandler
   Métodos:

   - create_shapefile_countries: Cria um shapefile para um ou mais países.
   - create_shapefile_states: Cria um shapefile para um ou mais estados.

3. **Manipulação de Rasters**
   Recorta rasters com base em bounding boxes ou polígonos de shapefiles.

Classe: RasterHandler
Método: crop_raster

4. **Análise de Correlação**
   Calcula e exibe matrizes de correlação (Spearman, Kendall, Pearson) entre variáveis ambientais.
   Classe: CorrelationAnalyzer
   Métodos:
   - calculate_tiffs_correlation: Calcula a matriz de correlação.
   - display_correlation_heatmap: Exibe um heatmap da matriz de correlação.
   - calculate_filter_display_heatmap: Filtra variáveis com base na correlação e exibe o heatmap.

5. **Aplicação de PCA**
   Aplica Análise de Componentes Principais (PCA) nas variáveis ambientais.
   Classe: PCAProcessor
   Método: apply_pca

6. **Preparação de Dados para Modelagem**
   Gera pseudo-ausências e prepara os dados para modelagem.
   Classe: ModelDataPrepare
   Método: generate_pseudo_absence

7. **Extração de Valores de Rasters**
   Extrai valores de variáveis ambientais para coordenadas específicas.
   Classe: RasterDataExtract
   Método: get_values

8. **Modelagem de Distribuição de Espécies**
   Implementa diversos métodos de modelagem:
   - **Métodos de Distância:** Bioclim, Mahalanobis, Euclidiana, Canberra, Chebyshev, Cosseno, Minkowski, Manhattan.
   - **Métodos Estatísticos:** GLM (Modelo Linear Generalizado), GAM (Modelo Aditivo Generalizado).
   - **Métodos de Machine Learning:** Random Forest, ANN (Redes Neurais Artificiais), SVM (Máquinas de Vetores de Suporte).
   - **MaxEnt:** Modelo de entropia máxima.

   Classes:
   - DistanceModeling
   - StatisticalModeling
   - MLModeling
   - MaxentModeling

9. **Avaliação de Modelos**
   Calcula métricas de avaliação (por exemplo, AUC, TSS, Kappa).
   Classe: ModelEvaluator
   Método: compute_metrics

Exemplos de Uso

1. **Download de Dados Ambientais**

```
from EcoDistrib.utils.data_download import DataDownloader

DataDownloader().download_data(destination_dir='../wordclim', output_dir='')
```

2. **Criação de Shapefiles**

```
from EcoDistrib.preprocessing.shapefile_operations import ShapefileHandler

# Cria um shapefile para o Brasil
ShapefileHandler().create_shapefile_countries(["Brazil"], "shapefile/brasil.shp")

# Cria um shapefile para os estados de SP e AC
ShapefileHandler().create_shapefile_states(["SP", "AC"], "shapefile/estados_selecionados.shp")
```

3. **Recorte de Rasters**

```
from EcoDistrib.utils import RasterHandler

# Recorta rasters usando um bounding box
RasterHandler().crop_raster(
    raster_path='downloaded_data/wordclim_data',
    output_path='raster_bound',
    method="bounding_box",
    bounding_box=[-46, -20, -42, -16]  # Exemplo de bounding box
)

# Recorta rasters usando um polígono de shapefile
RasterHandler().crop_raster(
    raster_path='downloaded_data/wordclim_data',
    output_path='raster_estados',
    method="polygon",
    shapefile_path="shapefile/estados_selecionados.shp"
)
```

4. **Análise de Correlação**

```
from EcoDistrib.preprocessing import CorrelationAnalyzer

# Calcula e exibe a matriz de correlação de Spearman
corr_matrix_spearman, variables = CorrelationAnalyzer().calculate_tiffs_correlation('raster_pa
CorrelationAnalyzer().display_correlation_heatmap(corr_matrix_spearman, variables, title='Heat
```

5. **Aplicação de PCA**

```
from EcoDistrib.preprocessing import PCAProcessor

PCAProcessor().apply_pca(input_folder="variaveis_filtradas/", output_folder="raster_pca/", n_c
```

6. **Modelagem de Distribuição de Espécies**

```
from EcoDistrib.modeling import DistanceModeling, ModelEvaluator

# Modelagem usando Bioclim
model = DistanceModeling()
map_bioclim = model.sdm_bioclim(df_for_sdm, 'raster_pca', save=True, output_save="sdm/mapa_res

# Avaliação do modelo
evaluator = ModelEvaluator(model=model, occurrence_data=df_for_sdm, tiff_paths='raster_pca')
metrics = evaluator.compute_metrics(save=True, output_save='sdm/metrics.csv')
```

Métricas de Avaliação

- **AUC:** Área sob a curva ROC.
- **TSS:** True Skill Statistic.
- **Kappa:** Coeficiente Kappa.

Contribuição

Contribuições são bem-vindas! Siga as diretrizes de contribuição no repositório.

Licença