ATIVIDADE FINAL PARTE 1 – Primeira versão do artigo

Aluno: Wesley Lourenco Barbosa

NUSP: 10509976

Título do artigo: Problemas de Qualidade na Aquisição e no Processo de Análise de Dados Bioclimáticos Utilizados em Modelos de Distribuição de Espécies: Uma Revisão Sistemática da Literatura

Seleção do Veículo

O veículo escolhido para submissão do artigo é a PLOS ONE, revista científica multidisciplinar publicada pela Public Library of Science, que possui fator de impacto 3.752 e classificação Qualis A1 na área de Engenharias IV, Computação e Biodiversidade.

Instruções aos autores: Submission Guidelines | PLOS ONE

Problemas de Qualidade na Aquisição e no Processo de Análise de Dados Bioclimáticos Utilizados em Modelos de Distribuição de Espécies: Uma Revisão Sistemática da Literatura

Wesley L. Barbosa[1]*

[1] Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da Universidade de São Paulo (EPUSP), São Paulo - SP, Brazil

* Corresponding Author

Email: wesleyloubar@usp.br (WLB)

Resumo

Os modelos de distribuição de espécies (SDM) são ferramentas importantes para a tomada de decisões em diversas áreas de aplicação, sendo essenciais para a gestão dos ativos da biodiversidade no mundo. A capacidade de representar a realidade desses modelos depende fortemente da adequação dos dados a partir dos quais são gerados. Portanto, entender o contexto de qualidade de dados para a criação do SDM é crucial, mas é um tópico que ainda não foi adequadamente explorado na literatura acadêmica. Assim, este artigo realiza uma revisão sistemática da literatura para examinar os problemas de qualidade de dados observados em dados bioclimáticos aplicados ao contexto SDM e propostas de soluções. A revisão compilou artigos publicados entre 2013 e 2022 para selecionar pesquisas recentes. Seguindo o procedimento metodológico, inicialmente foram selecionados 1793 artigos; após um filtro inicial de pertinência, 526 artigos tiveram os resumos avaliados. Os

231 artigos mais pertinentes tiveram a introdução e conclusão analisados e 59 artigos foram classificados para leitura completa. Os resultados evidenciaram 15 problemas de qualidade que impactam os dados bioclimáticos. As propostas de correção dos problemas de qualidade de dados mais recorrentes na literatura, erro de localização, erro de identificação e vieses, também foram discutidos. Os impactos das questões de qualidade de dados no SDM precisam ser mais detalhados na literatura porque as descrições abstratas focam na aplicação e não na quantificação dos impactos nos modelos. Concluiu-se que os problemas de qualidade dos dados comprometem a capacidade explicativa dos modelos de distribuição, portanto, mais pesquisas devem ser realizadas para melhor caracterizar esses impactos. Uma definição formal de como lidar com problemas de qualidade de dados no contexto bioclimático precisa ser estabelecida. Portanto, ainda há frentes de pesquisas a serem exploradas, principalmente na convergência das áreas da computação com o campo especializado da biogeografia.

Palavras-chave

qualidade de dados, modelos de distribuição de espécies, biodiversidade, clima, aprendizado de máquina, algoritmos estatísticos, dados bioclimáticos.

1 Introdução

Nas últimas duas décadas, houve um aumento significativo no volume de informações armazenadas e disponibilizadas sobre a biodiversidade do planeta (1,2). Entre as razões que têm contribuído para o acúmulo acentuado de dados de distribuição de espécies está a digitalização de coleções de história natural, coletadas ao longo de muitos

séculos de exploração de campo (3,4). Os programas de Citizen science (CS) também contribuíram para um rápido aumento no escopo e no volume dos dados de biodiversidade disponíveis (5). Com o desenvolvimento de novas plataformas computacionais, várias iniciativas surgiram para facilitar o acesso da comunidade científica a esses dados. O Global Biodiversity Information Facility (GBIF[1]) (6), o Data Observation Network for Earth (DataOne[2]), a base de dados Plants Traits (TRY[3]) (7), o Global Biotic Interactions (GloBI[4]) (8) e a base de dados alemã para pesquisa integrativa de biodiversidade sPlot [5](9), são exemplos de esforços para tornar os dados de biodiversidade cada vez mais disponíveis.

O avanço na acessibilidade a grandes coleções de dados bioclimáticos, concomitante ao rápido desenvolvimento de novas técnicas e ferramentas para processá-los, tem permitido análises e interpretações em larga escala (10). Esses dados são, portanto, recursos essenciais para documentar a biodiversidade e sua distribuição no tempo e no espaço com aplicações para pesquisas (11) e formulação de políticas públicas (12).

Na biodiversidade, dados de ocorrência de espécies, integrados com dados espaciais ambientais e climáticos, possibilitam diferentes aplicações, desde estudos sobre aspectos da teoria de nicho ecológico (13) até aplicações de conservação (14), biogeografia (15), agricultura (16), saúde (17), entre outros. De acordo com (18), as mudanças ambientais globais estão afetando rapidamente as distribuições de espécies e a adequação do habitat em todo o mundo, exigindo assim uma demanda contínua por modelos de distribuição de espécies (SDM) atualizados para representar o status mais atualizado do

[1] www.gbif.org
[2] www.dataone.org
[3] www.try-db.org
[4] www.globalbioticinteractions.org
[5] https://www.idiv.de/de/sdiv/arbeitsgruppen/wg_pool/splot.html

cenário da biodiversidade e para orientar decisões eficazes sobre políticas de conservação. No entanto, erros e incertezas nos registros de ocorrência de espécies e dados climáticos colocam em dúvida a exatidão dos modelos gerados a partir desses dados, comprometendo sua utilidade (10).

Trabalhos sobre gestão de qualidade de dados têm sido relatados em diversas áreas de pesquisa: contabilidade (19), gestão (20), segurança pública (21), saúde (22), educação (23), agricultura (24), entre outros. Em estudos de biodiversidade, a qualidade dos dados (DQ) tem sido documentada desde a década de 1970 (25), mas ainda não há um trabalho que consolide a análise e quantificação do reconhecimento da comunidade científica em relação aos problemas de qualidade que afetam os dados utilizados para a criação de modelos de distribuição de espécies. Assim, este artigo apresenta uma revisão sistemática da literatura (SLR) que examina problemas de qualidade observados em dados bioclimáticos usados em SDM, tanto do ponto de vista da aquisição quanto do processo de análise e investiga propostas de soluções para estes problemas. Os resultados desta revisão destacam os desafios e problemas de DQ comumente relatados e as soluções recomendadas. Além disso, alerta-se também para a necessidade de propor soluções formais de correção e tratamento de dados bioclimáticos.

2 Trabalhos Correlatos

Revisões de literatura sobre dados de biodiversidade e SDM já foram publicadas. Embora alguns estudos tenham destacado a importância da DQ, até onde foi possível

pesquisar, não foram encontrados estudos com foco na quantificação dos tipos de problemas podem afetar dados bioclimáticos ou em estratégias para mitigar os problemas. (26) analisaram estudos para entender e descrever como a escala espacial influencia os efeitos dos processos biológicos dos SDMs. Para projetos de CS, (27) revisaram os resultados e benefícios obtidos pelos participantes e voluntários dos projetos. (28) discutiram a influência dos preditores ambientais nas previsões do modelo, destacando os desafios relevantes para melhorar o SDM voltado para aplicações marítimas. (29) estudaram a previsão da variedade geográfica atual e futura e o nicho ambiental de espécies em ambientes marinhos. Eles também sugeriram que os futuros SDMs marítimos deveriam considerar os níveis de incerteza como parte do processo de modelagem.

Dados de ocorrência de espécies são um recurso valioso para pesquisas em ecologia, conservação e biogeografia. No entanto, a qualidade desses dados pode variar muito, dependendo de como foram coletados, registrados e gerenciados. Dados de baixa qualidade podem levar a conclusões incorretas e prejudicar a confiabilidade dos resultados das pesquisas que empregam esses dados. Portanto, é essencial garantir que os dados de ocorrência das espécies tenham qualidade adequada para análises que resultam em tomada de decisão. A preocupação com a qualidade de dados tem sido uma preocupação na literatura científica de ecologia e conservação. Algumas publicações abordaram a qualidade dos dados de diferentes perspectivas no processo de modelagem da distribuição de espécies. Do ponto de vista dos dados brutos, erros de taxonomia (30,31) e imprecisões no georreferenciamento (32,33), são problemas recorrentes em dados de ocorrência de espécies. (30) focaram em mensurar a diferentes regiões geográficas com

abundâncias distintas de espécies. Em (32), os autores destacam a preocupação com a resolução do detalhamento do georreferenciamento dos dados de ocorrência, principalmente quando o local de ocorrência de uma espécie é registrado em uma grande região ou unidade geopolítica, ao contrário de registros precisos com informações de latitude e longitude. Para lidar com esse tipo de inconsistência, 1(33) propõem incorporar as imprecisões das informações de localização da ocorrência nos modelos de distribuição de espécie, o que é uma abordagem diferente da tradicional que buscar tratar desse erro a nível de dados brutos. Nos projetos de *Citizen Science* (CS), além dos problemas mencionados, a experiência (34) e credibilidade (35,36) dos participantes são desafios mencionados na literatura. Propostas de solução como a de (37) visam avaliar a confiabilidade dos dados baseados em protocolos estruturados de CS. Em modelos de distribuição de espécies, a aplicação à conservação e manejo ainda é dificultada por vieses metodológicos espaciais (38) e amostrais (39) significativos, como amostras sobrepostas e interpolações com dados insuficientes, tanto nos dados de ocorrência das espécies quanto no conjunto de variáveis preditivas que representam as condições ambientais. Embora na literatura os problemas de qualidade de dados que afetam dados de ocorrência de espécies venham sendo discutidos, ainda falta um trabalho consolidado que quantifique os esforços que estão sendo direcionados para cada tipo de problema de qualidade. Identifica-se na literatura diversos artigos que abordam de erros de identificação e localização, no entanto, quais outros erros também afetam os dados de ocorrência de espécies? Quais são as propostas para mitigar ou minimizar o impacto dos problemas de

qualidade nos modelos de distribuição de espécies? São essas questões que o artigo proposto almeja responder.

3 Processo de Revisão

Uma Revisão Sistemática da Literatura (RSL) permite criar uma base sólida para o avanço do conhecimento científico, fornecendo subsídios para descobrir lacunas de pesquisa (40). O processo empregado nesta RSL foi proposto por (41) que estabelece uma abordagem para revisões de literatura em computação. Este processo também foi utilizado por outras revisões sistemáticas como em (42–44).

Na fase de planejamento da RSL, inclui-se as especificações das questões de pesquisa (RQ) e um conjunto de critérios de inclusão e exclusão criados para selecionar os artigos apropriados. A Tabela 1 mostra as 3 RQs que tentou-se responder nesta RSL, seguidas da motivação de estabelecer a respectiva pergunta.

Tabela 1: Research Questions

QUESTION	MOTIVATION
RQ1: What are the main DQ problems identified in data used to build SDM?	Conducting a survey into the main problems mentioned in the literature on three types of data used in SDM: occurrence, collected by CS projects, and climate.
RQ2: What are the strategies adopted to solve DQ problems?	Determining the main forms of correction adopted for the DQ problems identified.

As bases de dados de pesquisa utilizadas para selecionar os artigos foram: Web of Science (WoS), Scopus e ACM Digital Library. A escolha dessas bases de dados foi

baseada em critérios de qualidade (45,46), relevância para o campo da computação e estudos de biodiversidade, e completude dos metadados retornados por esses repositórios.

Cada uma das bases escolhidas possui diferentes procedimentos de busca; consequentemente, 6 strings de busca semanticamente equivalentes foram criadas para cada RQ. Como exemplo, a Tabela 2 mostra a estrutura da consulta de pesquisa usada na WoS.

Tabela 2: Search Queries

RESEARCH QUESTION	QUERY
Q1.1	TS = (((“species distribution” OR “ecolog* niche”)) AND “data” AND (“quality” OR “error”)) NOT TS = ('marine' OR 'water' OR 'fish' OR 'sea' OR 'ocean*' OR 'river*' OR 'aqua*')
Q1.2	TS = (("climat*" AND (“species distribution” OR “ecolog* niche”)) AND “data” AND (“quality” OR “error”) AND (“assess*” OR “control*” OR “measure*” OR “eval*” OR “estimat*” OR “analysis”)) NOT TS = ('marine' OR 'water' OR 'fish' OR 'sea' OR 'ocean*' OR 'river*' OR 'aqua*')
Q1.3	TS = (“citizen science” AND “data quality” AND (“error*” OR "problem*" OR "challenge"))
Q2	TS = (“species distribution*” AND “data” AND “error*” AND (“solution” OR “solv*” OR “fix” OR “correct*” OR “detect”))

A Tabela 3 apresenta os critérios de inclusão e exclusão definidos. O intervalo de tempo de quase 10 anos foi definido para garantir que os resultados descrevam o cenário mais atual.

Tabela 3: Inclusion and exclusion criteria

		INCLUSION CRITERIA	EXCLUSION CRITERIA
	1	Published between January 2013 and July 2022	Published outside our time span
	2	Full text available	Not written in English
	3	Published in journals or proceedings	Topic related to marine ecosystems
	4	Queried database provided the abstract within metadata (bib file)	Books

A revisão centra-se na distribuição das espécies terrestres. A exclusão das espécies marinhas deve-se à existência de diferenças significativas nos processos ecológicos marinhos e às características intrínsecas deste tipo de ecossistema (47).

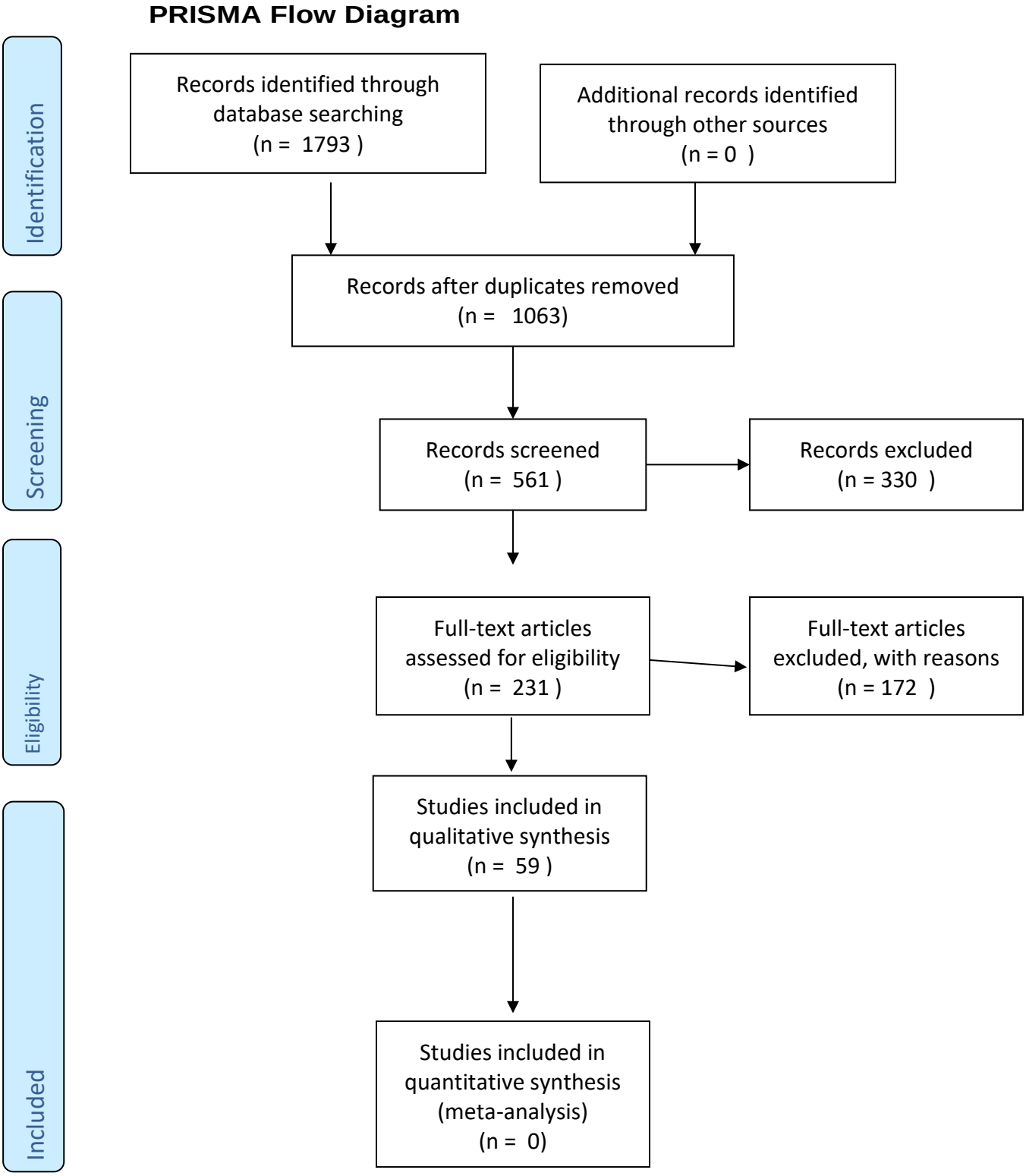
A Tabela 4 mostra o número de resultados retornados por banco de dados para cada consulta de pesquisa (Tabela 3), sem remover duplicatas. Os resultados da Tabela 4 são meramente informativos, não é objetivo desta RSL comparar as bases de dados de artigos científicos. Portanto, o detalhamento pelo banco de dados de pesquisa é apenas uma parte do processo de RSL.

Tabela 4: Number of results per database – initial research

R Q	WOS	Scopus	ACM
RQ1.1	516	392	42
RQ1.2	175	183	0
RQ1.3	65	100	125
RQ2	76	81	38

Aplicando todos os critérios de exclusão, o número total de artigos foi de 1793. Removendo os artigos duplicados entre as bases de dados e entre as RQ, o número de

artigos total foi de 1063 artigos. A partir do volume inicial de artigos, os títulos e palavras-chaves dos artigos foram verificados para excluir aqueles cujo assunto indicava claramente que estavam fora do escopo da revisão pretendida. Essa etapa reduziu a amostra total de artigos para 526. Na etapa seguinte os resumos de todos os artigos foram avaliados. Atribuiu-se um grau de pertinência para os artigos de acordo com a pertinência percebida para cada questão de pesquisa. Os graus de relevância variaram de 1 a 5 (1 indicando relevância limitada e 5, relevância forte). Após a avaliação dos resumos, 231 artigos que foram classificados nos graus de pertinência 4 e 5 foram selecionados para a próxima etapa. A próxima etapa consistiu na análise das seções de introdução e conclusão dos artigos. Nesta fase os artigos foram selecionados de acordo com a relevância de cada artigo para os temas de interesse. Ao final, dos 231 artigos, 59 foram selecionados para leitura e análise na íntegra. O diagrama a seguir apresenta uma representação visual do diagrama do PRISMA (48).


PRISMA Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1793)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 1063)
Records screened (n = 561)
Records excluded (n = 330)
Full-text articles assessed for eligibility (n = 231)
Full-text articles excluded, with reasons (n = 172)
Studies included in qualitative synthesis (n = 59)
Studies included in quantitative synthesis (meta-analysis) (n = 0)

4 Resultados

A Fig. 2 mostra a distribuição temporal dos artigos selecionados. Entre 2013 e 2022 não há uma tendência discernível, oscilando de 1 a 8 trabalhos. O ano de 2018 destaca-se por conter o maior número de trabalhos selecionados. Houve uma queda nos dois anos seguintes, 2019 e 2020, mas em 2021 e 2022 o número de trabalhos selecionados dentro do tópico de interesse desta revisão sistemática aumentou novamente.

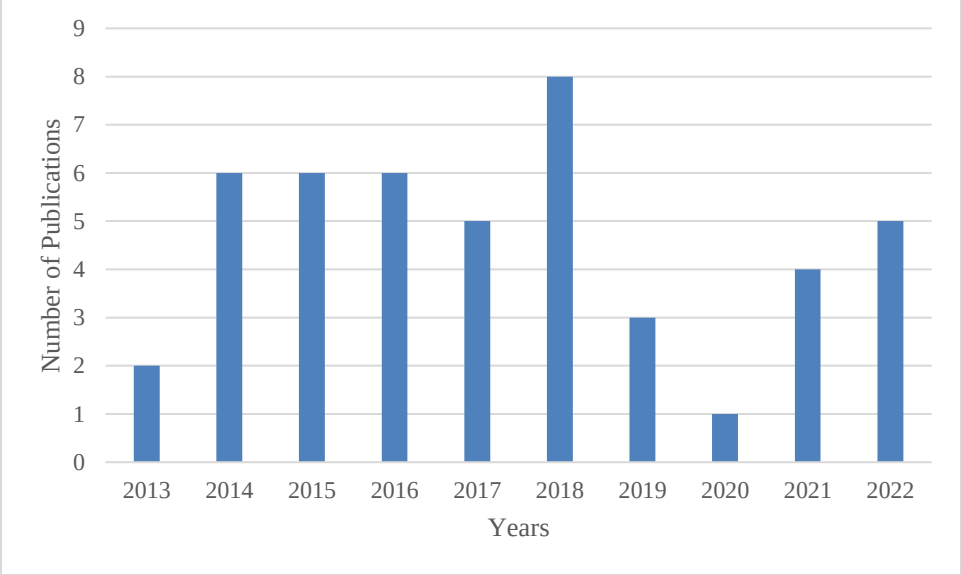


Fig 2: Temporal view of selected papers

Avaliando os tipos de fontes de publicação, foram identificadas 33 fontes diferentes, distribuídas em 3 congressos e 30 periódicos, listados na figura 3.

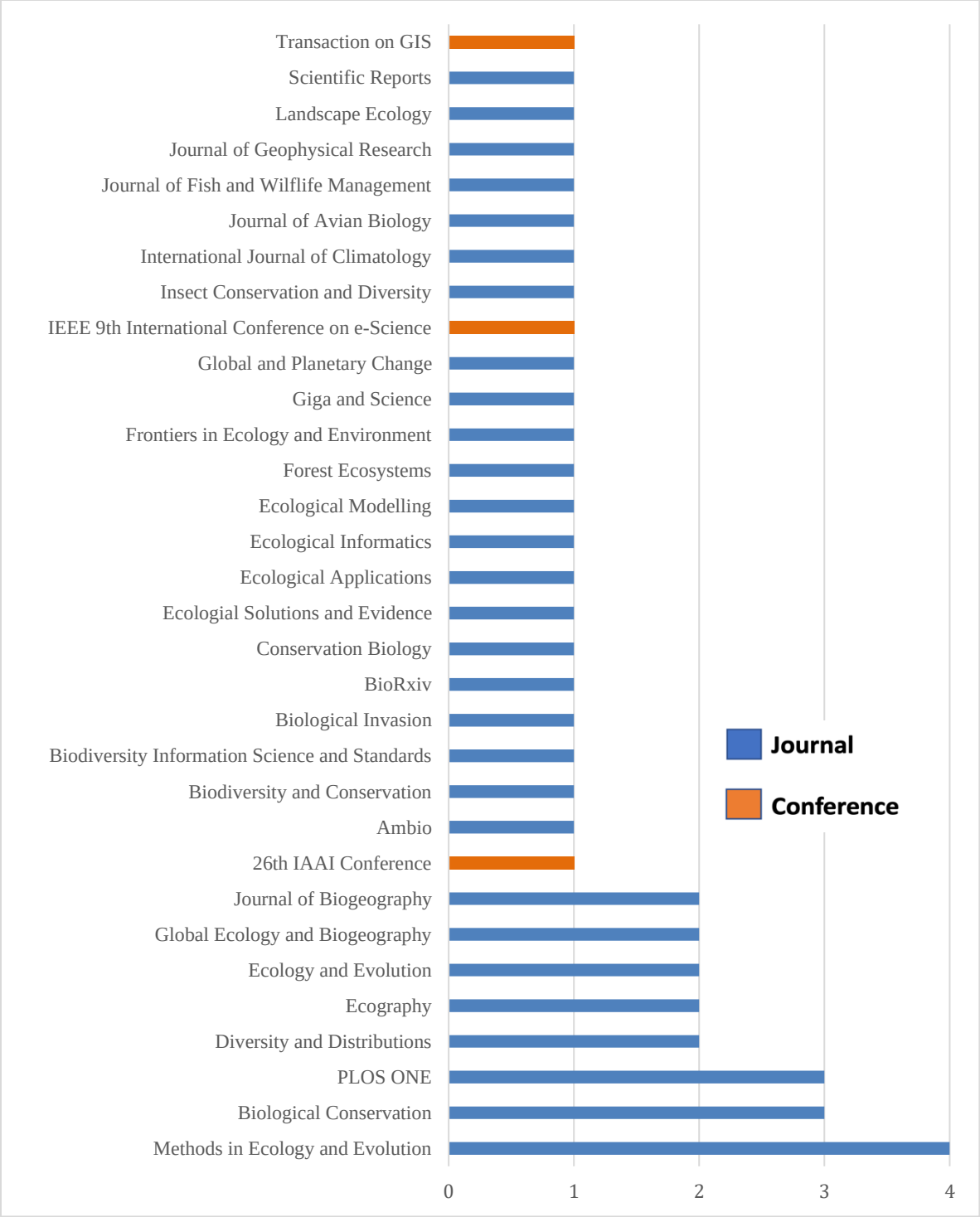


Fig 3: Sources of publications of the papers selected for full reading in the SLR.

A revista Methods in Ecology and Evolution é a fonte com maior número de resultados, representando 8.7% do total de artigos selecionados. A maioria dos periódicos

está relacionada às áreas de conhecimento em ecologia, biogeografia e conservação. Na Figura 3, há apenas 2 fontes, uma conferência sobre Tecnologia da Informação e uma sobre Inteligência Artificial, ligadas à área de computação e qualidade de dados. Isso destaca uma possível lacuna para a cooperação em pesquisa, uma vez que os campos de biodiversidade e climatologia dependem fortemente de algoritmos para análise de dados.

As nuvens de palavras são representações visuais de um documento de texto (49). A nuvem de palavras extrai as palavras mais usadas em um texto específico e as exibe por tamanho, com base em suas frequências (50). No contexto da SLR, a nuvem de palavras foi usada para avaliar o conteúdo e a adequação geral dos artigos selecionados. A nuvem de palavras foi criada a partir da análise dos títulos e resumos dos artigos nos arquivos de metadados do bib. A linguagem R foi utilizada para implementar a nuvem de palavras e o resultado é mostrado na Figura 4.

Fig 4: Word cloud representing the terms most commonly present in the titles and abstracts of selected papers

Algumas palavras da Figura 5 indicam que o procedimento adotado para a revisão sistemática foi capaz de resgatar artigos pertinentes às RQ (Research Questions). O contexto principal da pesquisa foi definido pelas palavras *quality* e *data*. Os dados utilizados nos modelos de distribuição de espécies são representados por *presence*, *absence*, *climate*, *variables*, *measures* e *citizen*. Os problemas de qualidade são expressos por *uncertainty*, *biased*, *error*, *misleading*, e *positional*.

A Tabela 5 apresenta o mapeamento dos 46 artigos selecionados para as questões de pesquisa (Tabela 1). Os artigos foram organizados em ordem alfabética e foi indicado a qual pergunta de pesquisa o respectivo artigo ajudar a responder. Em geral, os artigos que

trazem uma proposta de solução para os problemas de qualidade também mencionam o problema específico que eles desejam solucionar. Assim, dos 46 artigos, todos identificam pelo menos um tipo de problema de qualidade que afetam dados de ocorrência, climáticos ou de Citizen Science; desses 46, 36 artigos também discutem estratégias para mitigar problemas de qualidade.

Tabela 5: Selected papers and the research questions to which they contribute

#	PAPERS	RESEARCH QUESTION	
		1	2
1	(51)	✓	
2	(52)*	✓	
3	(53)*	✓	
4	(54)	✓	✓
5	(55)	✓	✓
6	(56)	✓	✓
7	(57)	✓	✓
8	(58)	✓	✓
9	(59)*	✓	
10	(60)	✓	✓
11	(61)	✓	
12	(62)	✓	✓
13	(63)	✓	✓
14	(64)*	✓	
15	(65)	✓	✓
16	(66)	✓	✓
17	(67)	✓	✓
18	(68)*	✓	✓
19	(69)	✓	✓
20	(70)	✓	✓
21	(71)	✓	✓
22	(72)	✓	✓
23	(73)	✓	✓
24	(74)	✓	✓
25	(75)*	✓	✓
26	(76)	✓	✓

27	(77)	✓	✓
28	(78)	✓	✓
29	(79)*	✓	✓
30	(80)	✓	
31	(81)*	✓	✓
32	(82)	✓	
33	(83)*	✓	
34	(84)*	✓	✓
35	(85)	✓	✓
36	(86)	✓	✓
37	(87)	✓	✓
38	(88)*	✓	✓
39	(89)	✓	✓
40	(90)	✓	✓
41	(91)	✓	✓
42	(33)	✓	✓
43	(92)*	✓	
44	(93)	✓	✓
45	(94)*	✓	✓
46	(95)	✓	✓
	TOTAL	46	36

A Tabela 6 apresenta os tipos de problemas de qualidade de dados identificados nos 46 artigos (Tabela 5) analisados para RQ1. As colunas de problema e descrição caracterizam, respectivamente, o tipo de problema identificado e sua definição. A coluna de categoria de dados indica o tipo de dados afetados (clima, biodiversidade, CS e tudo - quando o problema está relacionado a 3 tipos de dados anteriores). Por fim, a coluna número de artigos apresenta o número de artigos em que o respectivo problema foi mencionado.

Tabela 6: Issues identified in the papers analyzed.

	ISSUE	DESCRIPTION	DATA CATEGORY	NUMBER OF PAPERS
1	**Incomplete or missing data**	Records with data or metadata attribute not filled.	All	6
2	**Duplicated Records**	Multiple records that represent the same entity in the real world.	All	4
3	**Outside Range**	Measurement values of climatic data inconsistent with the local reality.	Climate	2
4	**Temporality**	Data in a restricted and insufficient time interval for analyzing long-term distribution changes.	All	3
5	**Imprecise Location (Location Error)**	Data with incorrect or inaccurate location records.	All	11
6	**Misidentification**	Detections of a species that are mistaken for another.	Biodiversity and CS	8
7	**Observer Skill**	Ability or experience of a non-professional observer in detecting the species occurrence.	CS	6
8	**Geographical Bias**	The tendency towards recording species observations in regions that are easier to access.	All	5
9	**Detection Bias**	Observer's tendency towards registering certain species at the expense of others, causing over and under-sampling of occurrences.	CS	4
10	**Temporally Bias**	The trend in seasonal records of species occurrences.	CS	2
11	**Positional Error**	Incorrect location record due to the observer's position in relation to the specimen.	Biodiversity	1
12	**Region Uncertainty**	Spatial climate data sets in which some regions are consistently overestimated while others are underestimated.	Climate	1
13	**Sampling Bias**	This occurs when data are not randomly distributed across the ecological niche or are unbalanced.	Biodiversity	1

A localização imprecisa é o erro mais recorrentes nos artigos analisados. Além disso, é um tipo de erro que pode afetar tanto os dados de ocorrência como os coletados por CS e dados climáticos. O segundo tipo de erro mais comum identificado é o erro de identificação que acometem os dados de ocorrência. Em seguida os autores destacam as habilidades do observador e incompletude no preenchimento dos registros de informações de interesse.

5 Discussão

Até onde se sabe, esta é a primeira revisão sistêmica realizada para identificar os problemas de qualidade que podem afetar dados utilizados na modelagem de distribuição de espécies, e também as propostas de soluções para os problemas identificados. Na área de ecologia relacionada aos modelos de distribuição de espécies, o trabalho de (96) examinou as tendências e lacunas relacionadas ao uso de CS como fonte de dados para SDMs por meio de uma revisão quantitativa da literatura, analisando 207 artigos para medir como a representação de diferentes táxons, regiões e tipos de dados mudou nas publicações SDM desde a década de 2010. Outra revisão quantitativa foi o trabalho de (97) que tinha como objetivo fornecer orientação prática sobre a aplicação de abordagens estatísticas para prever a abundância de espécies e identificar os fatores que mais afetam o desempenho preditivo de SDMs. Em revisões sistemáticas, o trabalho de (98) identificou que, de 236 artigos analisados, 94% falham em relatar as incertezas derivadas de deficiências de dados e parâmetros de modelo. No entanto essa revisão foca somente espécies marinhas [marine-based]. Outra revisão sistemática na área de ecologica é a de (99) que tem foco na estrutura da ecologia espacial para apoiar o manejo florestal na qual os autores revisaram conjuntos de dados de ocorrência de espécies disponíveis, dados ambientais, algoritmos de modelagem, processos de avaliação e projeções espaciais, discutindo as implicações das descobertas para ciência florestal e silvicultura.

A qualidade de dados é comumente descrita na literatura como relacionada ao contexto de uso (100–102). Assim, a compreensão do conceito de qualidade é conduzida pela circunstância em que os dados são usados. Examinando os tipos de problemas

identificados nos artigos analisados, dois padrões de percepção de erro foram observados: um puramente relacionado aos dados e outro relacionado ao processo de modelagem de SDM. Os problemas puramente relacionados a dados são os gerais, ou seja, que não são exclusivos do contexto SDM, podendo impactar os dados em diversos contextos de aplicação como saúde [healthcare] (103–105), transportes (106,107) e computação (108,109) . Na Tabela 7, os problemas 1 a 4 se enquadram nessa categoria. Erros relacionados ao processo de modelagem são dependentes do contexto do SDM e não são facilmente resolvidos com um simples processo de limpeza de dados. Esses tipos de erros exigem um estudo mais profundo do contexto da aplicação para propor possíveis soluções; os problemas 5 a 15 (Tabela 7) podem ser classificados nesta categoria.

5.1 RQ1 – Principais problemas de qualidade identificados em dados utilizados em modelos de distribuição de espécies

Os problemas de localização imprecisa foram os mais citados na literatura, presente em 23.9% dos artigos analisados. (58) aborda o problema de localização imprecisa presente em metadados de observações de sons de animais. Segundo os autores, neste cenário específico, o problema ocorre devido a campos de geolocalização de metadados não preenchidos, ou quando a descrição da localização real é vaga, como quando o observador atribui o estado, ou país, como local de aquisição. Essa descrição generalista evita que os dados sejam usados para criar SDMs que requerem resolução geográfica fina. (68) descrevem o mesmo problema, mas para dados coletados por voluntários de CS. Segundo os autores, o erro de localização é um problema recorrente em projetos de CS porque o observador tende a registrar os locais de ocorrência como a estrada mais próxima, cidade

ou algum outro ponto de referência. Os dados climáticos também podem ter inconsistências na localização geográfica. Como afirmado em (75), na ausência de coordenadas geográficas precisas, as variáveis climáticas são frequentemente atribuídas a centróides de regiões geopoliticamente definidas, como municípios. Com base em (87), os problemas de localização podem ocorrer devido a uma entrada incorreta do sistema de coordenadas, mudança de longitude e latitude, confusão de sinais numéricos ou, ainda, por uma tentativa de aproximar o local real de ocorrência da espécie a uma referência apontar. A digitalização de dados históricos de biodiversidade também pode adicionar incerteza à localização das ocorrências porque, em muitos casos, dados geográficos não estão disponíveis (10).

O segundo problema mais comum identificado foram erros de identificação de espécies. Embora as pesquisas de (10,58,110) não atribuam diretamente a ocorrência desse tipo de problema a projetos de CS, (57,70,81,96) afirmam que problemas de identificação taxonômica são característicos de dados coletados por observadores voluntários.

De acordo com (111), o maior impacto das iniciativas de CS ocorre em pesquisas de biologia, conservação e ecologia, sendo utilizadas como metodologia para coleta e classificação de dados. No entanto, erros taxonômicos nos dados coletados comprometem o potencial desses dados para descrever padrões corretos sobre a distribuição da biodiversidade. (70,81,96,112), indicam que erros na identificação de espécies geralmente são causados pela falta de experiência do observador. Assim, um voluntário inexperiente pode ter dificultado a identificação de espécimes animais à distância, e também na diferenciação de espécies semelhantes (54).

Vários problemas relacionados ao viés de dados são abordados na literatura. Conforme relatado por (113), os vieses impactam a validade e a confiabilidade dos resultados de um estudo. Do ponto de vista da modelagem de distribuição de espécies, não reconhecer esses problemas pode comprometer a interpretação dos modelos e pode ter consequências importantes nos resultados. Erros de viés podem ocorrer em dois pontos no ciclo de vida dos dados: no processo de aquisição de dados ou no procedimento metodológico de seleção de dados para análise. Iniciativas para coletar dados de ocorrência de espécies por meio de CS tendem a apresentar esses tipos de erros (54,70,112). Segundo (59), os vieses geográficos ocorrem porque a facilidade de acesso a algumas áreas torna o registro de ocorrência nessas regiões muito maior do que em áreas mais remotas. O registro da ocorrência da espécie tende a ocorrer em locais convenientes, próximos a estradas, rodovias e trilhas, o que permite a subamostragem de áreas de interesse. Além dos vieses geográficos, outro problema é a sazonalidade em que os registros acontecem. Há uma diferença no volume de dados que os voluntários contribuem dependendo da hora do dia, dia da semana ou mesmo entre as estações (70). Alguns tipos de análise, como a estimativa das tendências de acasalamento das espécies em determinadas épocas do ano, podem ser comprometidos por esse desequilíbrio.

No processo de aquisição de dados, o viés de detecção pode estar presente. Esse problema acontece quando os voluntários começam a documentar a ocorrência de espécies que acham mais fáceis de identificar. Em espécies de difícil reconhecimento, isso pode resultar na subnotificação de ocorrências reais. Conforme explicado por (114), esse tipo de erro pode levar a conclusões equivocadas sobre a abundância de algumas espécies em

detrimento de outras, quando na verdade o volume de indivíduos pode não ser tão diferente na prática.

Os erros de amostragem têm consequências para a integridade dos resultados do SDM. Existem três erros principais identificados na literatura compondo o viés da amostra: desequilíbrio de classe (Robinson et al., 2018), representatividade e completude (115). O desequilíbrio de classe ocorre quando o número de amostras de uma classe excede significativamente a amostra de outra classe, por exemplo, dados de ausência e presença. (Robinson et al., 2018) aborda o problema do viés de amostragem sob a perspectiva do desequilíbrio de classes para espécies raras. (115) buscam avaliar quantitativamente a qualidade dos dados de distribuição das espécies. Nessa análise, os autores indicam que a representatividade dos dados fica comprometida quando a amostra não é representativa da extensão do nicho ecológico da região. Assim, a amostra apresenta o erro de completude quando não abrange toda a extensão do nicho ecológico no espaço ambiental.

A maioria dos artigos abordou problemas de DQ específicos para o contexto SDM. Os problemas de qualidade que podem estar presentes nos dados de diferentes áreas de pesquisa, apesar de mencionados, não receberam uma descrição detalhada. Os artigos tendem a se concentrar nos desafios de qualidade dos dados bioclimáticos que são mais difíceis de compreender sem uma compreensão mais profunda do contexto dos dados. Como resultado, as publicações tendem a explorar dados cujas características são bem conhecidas pelos autores ou pelo grupo de pesquisa da publicação. Como a contribuição de dados de ocorrência de espécies por voluntários para projetos de SC é o cenário mais viável para expandir os bancos de dados sobre biodiversidade no mundo, os autores se

preocuparam em estudar os erros que ocorrem nesse tipo de processo de aquisição. É importante conhecer os possíveis erros que podem estar presentes nos dados bioclimáticos para que as incertezas dos dados não produzam resultados enviesados. Os pesquisadores podem escolher conscientemente estudar regiões ricas em certas espécies e pobres em outras. Assim, é necessário esclarecer o objetivo e as características dos dados de interesse da pesquisa, a fim de não se chegar a resultados enviesados quando o método de aquisição per se já está distorcido. Lacunas geográficas e temporais também podem contribuir para a deterioração dos dados. Essas características precisam ser levadas em consideração em pesquisas que visam descrever processos de mudança ecológica, mudança climática, uso da terra, perda de habitat ou migração de espécies. Os vieses de dados podem influenciar o planejamento da política de conservação devido à super ou subestimação da distribuição das espécies. Para inferir padrões de distribuição de espécies, é necessário considerar os fatores de incerteza nos dados, para que problemas de qualidade amostral não sejam interpretados como um cenário real para a distribuição de espécies. Assim, reconhecer erros nas informações sobre biodiversidade é de grande importância.

5.2 RQ2 - Estratégias sugeridas para lidar com problemas de DQ em dados utilizados em modelos de distribuição de espécies

Existem diferentes abordagens para melhorar a qualidade dos dados bioclimáticos, desde métodos estatísticos clássicos até técnicas modernas de aprendizado de máquina. Analisando a Tabela 7, para os problemas de qualidade de dados 1 a 4, os autores não

descrevem formas específicas de lidar com este tipo de dados. Por serem mais generalistas, as estratégias de gestão de DQ adotadas em outras áreas também podem ser aplicadas aos tipos de dados avaliados nesta RSL (116).

O erro de imprecisão de localização (problema 5 - Tabela 7) pode ser bastante complexo de resolver dependendo de sua causa (origem). Muitos autores tentam identificar as inconsistências nos dados e então verificar se há solução para o problema. Identificar as inconsistências consiste em escanear os dados para verificar se as coordenadas de localização são consistentes com as espécies correspondentes. A partir desta etapa, os autores propõem diferentes caminhos de solução. (58) descrevem o cruzamento da distribuição geográfica esperada da espécie com os dados de coordenadas geográficas dos metadados do registro. Caso sejam identificadas inconsistências, uma ferramenta de referenciamento geográfico pode ser utilizada para encontrar uma aproximação do local real da ocorrência, com base em outras informações encontradas nos metadados, como país, estado e cidade. Para dados históricos, esta é uma solução viável, pois em muitas coletas as coordenadas exatas de ocorrência não estão disponíveis; assim, metadados podem ser usados para inferir áreas de ocorrência aproximadas. Para esse mesmo tipo de problema, (68) sugerem o uso de estatísticas espaciais aplicando mudança de suporte (COS) para superar as limitações da calibração de regressão em dados apenas de presença. Segundo os autores, esta técnica pode ser utilizada para corrigir erros de localização quando as localizações reais são pontos com coordenadas desconhecidas contidos em polígonos de forma arbitrária.

A correção de dados de projetos de CS é descrita na literatura sob dois aspectos: a metodologia de aquisição de dados e o tratamento dos dados. Avaliar dados em projetos de CS é uma tarefa complexa, pois envolve uma série de dimensões que podem comprometer sua qualidade. A preocupação com a qualidade dos dados deve estar presente desde o início da concepção do projeto. Adotar protocolos que previnam ou minimizem erros desde o início do ciclo de vida dos dados é a estratégia mais prudente. Nesse sentido, (70) defendem a criação de protocolos simplificados e ferramentas de verificação para geração de dados de alta qualidade. A formação de observadores voluntários participantes dos projetos é o principal ponto em comum entre os autores. (70) propõem o treinamento de voluntários por meio de oficinas. (81) examinam o efeito de experiências anteriores dos voluntários combinados com três métodos de treinamento (flayers, apresentação de slides e treinamento direto). Ambos os trabalhos sugerem a aplicação de testes de aptidão aos voluntários. Esses métodos podem mitigar os problemas de habilidade de observadores inexperientes e reduzir a identificação incorreta de espécies.

Em relação ao viés de aquisição, as soluções tendem a ser mais complexas. Problemas relacionados às preferências de aquisição dos observadores (8, 9 e 10 na Tabela 7) são difíceis de prevenir ou corrigir. Os autores tendem a indicar a detecção e quantificação desses problemas para levá-los em conta no SDM. Também houve abordagens em que não há ação realizada diretamente sobre os dados. Eles sugeriram quantificar os problemas ou marcar os dados em que a precisão é incerta. Assim, cabe ao pesquisador que utilizará esses dados definir a melhor forma de abordá-los no contexto de interesse, excluindo-os, limpando-os ou incluindo incertezas nos modelos. Essa abordagem é indicada para pesquisas que utilizam informações de bancos de dados públicos, pois, na

maioria das vezes, os autores não têm acesso aos metadados ou aos detalhes do processo de aquisição dos dados. Isso pode dificultar significativamente ou mesmo impossibilitar a correção dos dados. Assim, para usar dados públicos, é importante qualificar ou quantificar as incertezas da amostra. A esse respeito, (57,66) mencionam a sinalização como uma estratégia válida para observações anômalas, seja para revisão em etapas de validação subsequentes ou como informação para futuros usuários desses dados. Para qualificar as observações, (86) adota um sistema de avaliação sequencial composto por 10 etapas. Os registros de ocorrência seguem um fluxo de trabalho em que cada tarefa trata de um possível aspecto de incongruência. Ao final do fluxo de trabalho, é atribuída uma nota ao registro de acordo com a qualidade identificada.

(74) propõem o uso da estatística K para quantificar a associação espacial local ao local de ocorrência das espécies. (33) argumentam que em determinados cenários, como quando o tamanho da amostra é pequeno, ou os erros de referência geográfica são grandes, ou ainda, quando os erros são regulares, o uso de modelos de erro de medição tem um resultado melhor do que melhoria em métodos que funcionam nos níveis de dados. (BIRD et al., 2014) fornecem orientações sobre como aplicar técnicas de análise estatística e aprendizado de máquina, como Modelos Aditivos Generalizados (GAM), Modelo Linear Generalizado (GLM), Regressão Geograficamente Ponderada (GWR), Árvores de Regressão e Maxent, para lidar com vieses presentes em bancos de dados com dados coletados por projetos de CS. (54) reconhecem que os dados gerados pelo CS tendem a ter grande variabilidade em relação aos dados coletados por cientistas e instrumentos, indicando assim as técnicas analíticas como as mais adequadas para diferentes tipos de erros e vieses.

Resolver problemas bioclimáticos de DQ é uma tarefa desafiadora. Conhecimentos específicos sobre o contexto de aquisição e características fundamentais dos dados são necessários para o tratamento dos dados. Em alguns casos, a correção pode não ser possível, tornando a caracterização qualitativa dos dados importante para futuros usuários. A publicação de fontes de dados em bases científicas é importante para a divulgação da ciência, oportunizando o acesso da comunidade científica a esses dados. Os principais repositórios de dados de biodiversidade e clima, como GBIF e WorldClim, realizam uma ampla gama de verificações automatizadas dos dados alimentados nessas plataformas, sinalizando, por exemplo, registros cujas coordenadas estão fora do país declarado de observação, ou algumas outras inconsistências. No entanto, devido à natureza heterogênea dos dados, não é possível criar métodos de verificação automática que detectem todos os tipos de erros. Assim, pela necessidade de conhecimento especializado e pela ampla gama de possíveis erros, cabe às iniciativas que fornecem dados para esses repositórios, detectar, corrigir sempre que possível, e indicar os problemas de qualidade existentes nos dados.

5.3 Perspectivas para Direções Futuras

Um dos efeitos mais óbvios da revolução digital observada no campo da biodiversidade foi a possibilidade de criar grandes volumes de dados biológicos sistemáticos e fornecer diferentes bancos de dados para a comunidade de pesquisa. A análise e derivação do conhecimento biogeográfico dependem da adequação dos dados às tarefas pretendidas e de técnicas computacionais capazes de extrair padrões.

Em relação à qualidade, vários tipos de problemas podem comprometer os dados utilizados em SDM. Resolver esses problemas continua sendo um importante desafio de

pesquisa devido às características específicas dos tipos de dados envolvidos. Ao contrário dos erros mais comuns que aparecem em diferentes aplicações, como duplicação e falta de dados, os dados empregados em SDM precisam de informações contextuais adicionais para que qualquer solução seja proposta. Como a disponibilidade ou mesmo a existência de tal informação contextual é rara; em muitos casos, os pesquisadores não têm uma maneira eficiente de resolver tais problemas. Portanto, métodos e padrões formais devem ser adotados na etapa de aquisição de dados para que a DQ seja garantida desde o início do ciclo de vida dos dados, principalmente em projetos de CS. Para o caso de dados comprometidos existentes, seria importante propor instruções indicando como lidar com esses dados e como expressar incertezas nos resultados da análise. Pesquisadores ou grupos de pesquisa estão adotando atualmente procedimentos de soluções heterogêneas específicas para seu ambiente de estudo. Para o futuro, é importante criar uma sistematização para problemas de qualidade no processo de modelagem, resultando em um conjunto formal de instruções para padronizar a forma como esses desafios são tratados, sugerindo abordagens mais generalistas que funcionem em uma ampla gama de aplicações.

O fortalecimento e implementação de processos de controle de qualidade na aquisição de dados primários da biodiversidade são fundamentais para a construção do conhecimento sobre a diversidade macrobiológica. Os dados coletados sobre a ocorrência das espécies são processados por algumas técnicas algorítmicas para gerar resultados que são os pilares para investigar características, mudanças e evolução dos padrões de distribuição de espécies no mundo. Além disso, a análise de dados de biodiversidade é vital para a tomada de decisões sobre a conservação e exploração sustentável da biodiversidade e dos serviços ecossistêmicos. Os desafios relacionados ao DQ e às técnicas de modelagem

exigirão a integração consistente de diferentes áreas de pesquisa, especialmente nas áreas de computação e biologia. O campo da computação deve atuar para automatizar e integrar tarefas que, de outra forma, exigiriam um esforço manual substancial. A disponibilidade de grandes quantidades de dados relacionados à biodiversidade também destaca problemas associados a questões como qualidade de dados e técnicas de análise, que precisam ser cuidadosamente abordadas e implementadas. É necessário captar o conhecimento humano sobre o campo da biodiversidade para tratar e automatizar, na medida do possível, a resolução dos problemas de DQ. As distribuições de nicho ecológico possuem propriedades particulares envolvendo uma complexa interação entre componentes bióticos e climáticos. Por esta razão, a criação de ferramentas computacionais para análise e modelagem de dados depende de conhecimento especializado sobre o contexto de distribuição das espécies. Consequentemente, há necessidade de convergência de ações e colaboração de áreas multidisciplinares para implementação e validação de técnicas de modelagem dedicadas ao SDM e soluções para problemas de qualidade de dados.

6 Conclusão

Os dados de biodiversidade e climáticos são recursos essenciais para a compreensão do padrão de distribuição geográfica e ocupação do hábitat das espécies. O conhecimento potencial que pode ser obtido a partir desses dados constitui base fundamental para a tomada de decisões sobre estratégias de mitigação dos impactos da atividade humana sobre a biodiversidade. No entanto, problemas de qualidade que comprometem esses dados interferem na capacidade explicativa do SDM e, consequentemente, na eficácia das

estratégias de preservação. Portanto, mais pesquisas são necessárias para caracterizar o impacto desses problemas de qualidade no SDM.

Este trabalho realizou um levantamento sobre as características dos dados de biodiversidade e climáticos que podem comprometer os resultados do SDM. Foram identificados 15 problemas de DQ e dois deles, erros de localização e identificação, foram os mais discutidos na literatura. Resolver esses problemas é um desafio a ser superado. Algumas soluções têm sido propostas, mas, em alguns casos, é preciso aceitar a inevitável ocorrência desses problemas e apontar incertezas nos modelos gerados a partir deles.

Problemas de qualidade com dados de biodiversidade e climáticos podem ter uma consequência direta na capacidade preditiva dos modelos de distribuição de espécies e podem impactar as decisões de preservação baseadas em modelos tendenciosos. Assim, é necessário aprimorar os processos de aquisição de dados para minimizar a ocorrência de erros e, também, aprimorar as técnicas de modelagem para criar modelos cada vez mais robustos e menos suscetíveis a problemas de DQ.

Este artigo contribui para o corpo de conhecimento sobre qualidade de dados em estudos de distribuição de espécies em dois aspectos principais: resumindo sistematicamente os problemas de qualidade em dados bioclimáticos e seus impactos nos modelos de distribuição de espécies gerados, propondo as principais direções futuras para investigação.

7 Referências Bibliográficas

1. Escribano N, Galicia D, Ariño AH. The tragedy of the biodiversity data commons: a data impediment creeping nigher? Database (Oxford). 2018;2018.

2. Stephenson P, Brooks TM, Butchart SH, Fegraus E, Geller GN, Hoft R, et al. Priorities for big biodiversity data. Front Ecol Environ. abril de 2017;15(3):124–5.

3. Smith VS, Blagoderov V. Bringing collections out of the dark. Zookeys. 2012;(209):1–6.

4. Nelson G, Ellis S. The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B: Biological Sciences. 7 de janeiro de 2019;374(1763):20170391.

5. Muki Haklay M, Mazumdar S, Wardlaw J. Citizen Science for Observing and Understanding the Earth. Em: Earth Observation Open Science and Innovation. Cham: Springer International Publishing; 2018. p. 69–88.

6. Edwards JL. The Global Biodiversity Information Facility : An International Network of Interoperabel Biodiversity Databases. Joho Chishiki Gakkaishi. 2001;10(4):58–61.

7. Kattge J, DÍAZ S, LAVOREL S, PRENTICE IC, LEADLEY P, BÖNISCH G, et al. TRY - a global database of plant traits. Glob Chang Biol. 1º de setembro de 2011;17(9):2905–35.

8. Poelen JH, Simons JD, Mungall CJ. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecol Inform. 1º de novembro de 2014;24:148–59.

9. Bruelheide H, Dengler J, Jiménez-Alfaro B, Purschke O, Hennekens SM, Chytrý M, et al. sPlot – A new tool for global vegetation analyses. Chiarucci A, organizador. Journal of Vegetation Science. 8 de março de 2019;30(2):161–86.

10. Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, et al. Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? Global Ecology and Biogeography. 2015;24(8):973–84.

11. Carmel Y, Kent R, Bar-Massada A, Blank L, Liberzon J, Nezer O, et al. Trends in Ecological Research during the Last Three Decades – A Systematic Review. Nunes Amaral LA, organizador. PLoS One. 24 de abril de 2013;8(4):e59813.

12. Santamaría L, Méndez PF. Evolution in biodiversity policy - current gaps and future needs. Evol Appl. fevereiro de 2012;5(2):202–18.

13. Anderson RP. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. Ann N Y Acad Sci. julho de 2012;1260(1):66–80.

14. Geijzendorffer IR, Regan EC, Pereira HM, Brotons L, Brummitt N, Gavish Y, et al. MODEL-ASSISTED MONITORING OF BIODIVERSITY Bridging the gap between biodiversity data and policy reporting needs: An Essential Biodiversity Variables perspective. Journal of Applied Ecology. 2016;53:1341–50.

15. Carstensen DW, Dalsgaard B, Svenning JC, Rahbek C, Fjeldså J, Sutherland WJ, et al. The functional biogeography of species: biogeographical species roles of birds in Wallacea and the West Indies. 2013;

16. Lin C, Trianingsih D, Lin C, Trianingsih D. Identifying forest ecosystem regions for agricultural use and conservation. Sci Agric. fevereiro de 2016;73(1):62–70.

17. Alho CJR. Importância da biodiversidade para a saúde humana: uma perspectiva ecológica. Estudos Avançados. 2012;26(74):151–66.

18. Arenas-Castro S, Gonçalves J, Alves P, Alcaraz-Segura D, Honrado JP. Assessing the multi-scale predictive ability of ecosystem functional attributes for species distribution modelling. Joseph S, organizador. PLoS One. 18 de junho de 2018;13(6):e0199292.

19. Peters J, Krishnan R, Padman R, Kaplan D. On Data quality Assessment in Accounting Information Systems. 1998.

20. Hartl K. The Role of Data Quality in Business Intelligence-An empirical study in German medium-sized and large companies. Vol. 4. 2016.

21. Mitar M. Data Quality as a Challenge to Modern Policing and Criminal Justice. Policing in Central and Eastern Europe: Dilemmas of Contemporary Criminal Justice. 2004.

22. Sukumar SR, Ramachandran N, Ferrell RK. Data Quality Challenges in Healthcare Claims Data : Experiences and Remedies. 2016;(April 2014):15.

23. Mark. Improving Data Quality for Title I Standards, Assessments, and Accountability Reporting Guidelines For States, LEAs, and Schools. 2006.

24. Gonzales Malaverri JE, Medeiros CB. Data Quality in Agriculture Applications. 2012.

25. Busby JR. Australian Biotaxonomic Information System : introduction and data interchange standards. Canberra; 1979.

26. Record S, Strecker A, Tuanmu MN, Beaudrot L, Zarnetske P, Belmaker J, et al. Does scale matter? A systematic review of incorporating biological realism when predicting changes in species distributions. Bosso L, organizador. PLoS One. 13 de abril de 2018;13(4):e0194650.

27. Peter M, Diekötter T, Kremer K, Peter M, Diekötter T, Kremer K. Participant Outcomes of Biodiversity Citizen Science Projects: A Systematic Literature Review. Sustainability. 15 de maio de 2019;11(10):2780.

28. Robinson L, Elith J, Hobday AJ, Pearson RG, Kendall BE, Possingham HP, et al. Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. Global Ecology and Biogeography. 1º de novembro de 2011;20(6):789–802.

29. Robinson N, Nelson WA, Costello MJ, Sutherland JE, Lundquist CJ. A Systematic Review of Marine-Based Species Distribution Models (SDMs) with Recommendations for Best Practice. Front Mar Sci. 18 de dezembro de 2017;4:421.

30. Egli L, LeVan KE, Work TT. Taxonomic error rates affect interpretations of a national-scale ground beetle monitoring program at National Ecological Observatory Network. Ecosphere [Internet]. 1º de abril de 2020 [citado 28 de março de 2023];11(4):e03035. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/ecs2.3035

31. Stribling JB, Leppo EW. Relationship of taxonomic error to frequency of observation. PLoS One [Internet]. 1º de novembro de 2020 [citado 28 de março de 2023];15(11):e0241933. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241933

32. Smith AB, Murphy SJ, Henderson D, Erickson KD. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth: Don't let the perfect be the enemy of the good. bioRxiv [Internet]. 18 de março de 2022 [citado 28 de março de 2023];2021.06.10.447988. Disponível em: https://www.biorxiv.org/content/10.1101/2021.06.10.447988v2

33. Velásquez-Tibatá J, Graham CH, Munch SB. Using measurement error models to account for georeferencing error in species distribution models. Ecography [Internet]. 1º de março de 2016 [citado 3 de junho de 2019];39(3):305–16. Disponível em: http://doi.wiley.com/10.1111/ecog.01205

34. Earp H, Vye S, Bohn K, Burrows M, Chenery J, Dickens S, et al. Do You See What I See? Quantifying Inter-Observer Variability in an Intertidal Marine Citizen Science Experiment. Citiz Sci [Internet]. 4 de maio de 2022 [citado 28 de março de 2023];7(1):1–13. Disponível em: http://theoryandpractice.citizenscienceassociation.org/articles/10.5334/cstp.483/

35. Meschini M, Machado Toffolo M, Marchini C, Caroselli E, Prada F, Mancuso A, et al. Reliability of Data Collected by Volunteers: A Nine-Year Citizen Science Study in the Red Sea. Front Ecol Evol. 24 de junho de 2021;9:395.

36. Santos-Fernandez E, Mengersen K. Understanding the reliability of citizen science observational data using item response models. Methods Ecol Evol [Internet]. 1º

de agosto de 2021 [citado 28 de março de 2023];12(8):1533–48. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13623

37. Leocadio JN, Ghilardi-Lopes NP, Koffler S, Barbiéri C, Francoy TM, Albertini B, et al. Data reliability in a citizen science protocol for monitoring stingless bees flight activity. Insects [Internet]. 1º de setembro de 2021 [citado 28 de março de 2023];12(9):766. Disponível em: /pmc/articles/PMC8467663/

38. Bowler DE, Callaghan CT, Bhandari N, Henle K, Benjamin Barth M, Koppitz C, et al. Temporal trends in the spatial bias of species occurrence records. Ecography [Internet]. 1º de agosto de 2022 [citado 28 de março de 2023];2022(8):e06219. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.06219

39. Baker DJ, Maclean IMD, Goodall M, Gaston KJ. Correlations between spatial sampling biases and environmental niches affect species distribution models. Global Ecology and Biogeography [Internet]. 1º de junho de 2022 [citado 28 de março de 2023];31(6):1038–50. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/geb.13491

40. Webster J, Watson RT. Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly [Internet]. 2002;26(2):xiii–xxiii. Disponível em: http://www.jstor.org/stable/4132319

41. Kitchenham B, Chartes S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Durham; 2007 jul.

42. Schneider S, Torkar R, Gorschek T. Solutions in global software engineering: A systematic literature review. Int J Inf Manage [Internet]. 1º de fevereiro de 2013 [citado 25 de junho de 2019];33(1):119–32. Disponível em: https://www.sciencedirect.com/science/article/pii/S0268401212000989

43. Rekik R, Kallel I, Casillas J, Alimi AM. Assessing web sites quality: A systematic literature review by text and association rules mining. Int J Inf Manage [Internet]. 1º de fevereiro de 2018 [citado 25 de junho de 2019];38(1):201–16. Disponível em: https://www.sciencedirect.com/science/article/pii/S0268401215302875#bib0115

44. Sizo A, Lino A, Reis LP, Rocha Á. An overview of assessing the quality of peer review reports of scientific articles. Int J Inf Manage [Internet]. 1º de junho de 2019 [citado 25 de junho de 2019];46:286–93. Disponível em: https://www.sciencedirect.com/science/article/pii/S0268401218304857

45. Šubelj L, Bajec M, Mileva Boshkoska B, Kastrin A, Levnajić Z. Quantifying the Consistency of Scientific Databases. 2015;

46. Martín-Martín A, Orduna-Malea E, Thelwall M, Delgado López-Cózar E. Scopus: a systematic comparison of citations in 252 subject categories. J Informetr. 2018;12(4):1160–77.

47. Carr MH, Neigel JE, Estes JA, Andelman S, Warner RR, Largier JL. COMPARING MARINE AND TERRESTRIAL ECOSYSTEMS: IMPLICATIONS FOR THE DESIGN OF COASTAL MARINE RESERVES. Ecological Applications. 1º de fevereiro de 2003;13(sp1):90–107.

48. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev [Internet]. 1º de dezembro de 2021 [citado 5 de abril de 2023];10(1):1–11. Disponível em: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-021-01626-4

49. Kaptein R, Hiemstra D, Kamps J. How Different Are Language Models andWord Clouds? Em Springer, Berlin, Heidelberg; 2010. p. 556–68.

50. DePaolo CA, Wilkinson K. Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data. TechTrends. 21 de maio de 2014;58(3):38–44.

51. Arenas-Castro S, Regos A, Martins I, Honrado J, Alonso J. Effects of input data sources on species distribution model predictions across species with different distributional ranges. J Biogeogr [Internet]. 1º de julho de 2022 [citado 11 de abril de 2023];49(7):1299–312. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/jbi.14382

52. Aubry KB, Raley CM, McKelvey KS. The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. Baldwin RF, organizador. PLoS One. 22 de junho de 2017;12(6):e0179152.

53. Bedia J, Herrera S, Gutiérrez JM. Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections. Glob Planet Change. 2013;107:1–12.

54. Bird TJ, Bates AE, Lefcheck JS, Hill NA, Thomson RJ, Edgar GJ, et al. Statistical solutions for error and bias in global citizen science datasets. Biol Conserv. 2014;173:144–54.

55. Bloom TDS, Flower A, DeChaine EG. Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. Ecol Evol. 2018;8(1):765–77.

56. Boyd RJ, Powney G, Carvell C, Pescott OL. occAssess: An R package for assessing potential biases in species occurrence data. bioRxiv [Internet]. 10 de agosto de 2021 [citado 11 de abril de 2023];2021.04.19.440441. Disponível em: https://www.biorxiv.org/content/10.1101/2021.04.19.440441v3

57. Clare JDJ, Townsend PA, Anhalt-Depies C, Locke C, Stenglein JL, Frett S, et al. Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved? Ecological Applications. 2019;29(2).

58. Cugler DC, Medeiros CB, Shekhar S, Toledo LF. A geographical approach for metadata quality improvement in biological observation databases. Proceedings - IEEE 9th International Conference on e-Science, e-Science 2013. 2013;212–20.

59. Dorazio RM. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography. 2014;23(12):1472–84.

60. Fei S, Yu F. Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. Landsc Ecol. 2016;31(1):31–42.

61. Feldman MJ, Imbeau L, Marchand P, MazerolleI MJ, Darveau M, Fenton1 NJ. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. PLoS One [Internet]. 1º de março de 2021 [citado 11 de abril de 2023];16(3):e0234587. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0234587

62. Führding-Potschkat P, Kreft H, Ickert-Bond SM. Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models. Ecol Evol [Internet]. 1º de agosto de 2022 [citado 11 de abril de 2023];12(8):e9168. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.9168

63. Geldmann J, Heilmann-Clausen J, Holm TE, Levinsky I, Markussen B, Olsen K, et al. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Divers Distrib. 2016;22(11):1139–49.

64. Gomes VHF, IJff SD, Raes N, Amaral IL, Salomão RP, de Souza Coelho L, et al. Species Distribution Modelling: Contrasting presence-only models with plot abundance data. Sci Rep. 17 de dezembro de 2018;8(1):1003.

65. Graham LJ, Haines-Young RH, Field R. Using citizen science data for conservation planning: Methods for quality control and downscaling for use in stochastic patch occupancy modelling. Biol Conserv. 2015;192:65–73.

66. Gueta T, Carmel Y. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. Ecol Inform. 2016;34:139–45.

67. Hefley TJ, Baasch DM, Tyre AJ, Blankenship EE. Correction of location errors for presence-only species distribution models. Methods Ecol Evol. 2014;5(3):207–14.

68. Hefley TJ, Brost BM, Hooten MB. Bias correction of bounded location errors in presence-only data. Methods Ecol Evol. 2017;8(11):1566–73.

69. Kelling S, Fink D, Sorte FA La, Johnston A, Bruns NE, Hochachka WM. Taking a "Big Data" approach to data quality in a citizen science project. Ambio. 2015;44:11.

70. Kosmala M, Wiggins A, Swanson A, Simmons B. Assessing data quality in citizen science. Front Ecol Environ. 1º de dezembro de 2016;14(10):551–60.

71. Lewandowski E, Specht H. Influence of volunteer and project characteristics on data quality of biological surveys. Conservation Biology. 1º de junho de 2015;29(3):713–23.

72. Lin YP, Deng D, Lin WC, Lemmens R, Crossman ND, Henle K, et al. Uncertainty analysis of crowd-sourced and professionally collected field data used in species distribution models of Taiwanese moths. Biol Conserv. 2015;181:102–10.

73. Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, et al. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Global Ecology and Biogeography. 1º de agosto de 2015;24(8):973–84.

74. Naimi B, Hamm NAS, Groen TA, Skidmore AK, Toxopeus AG. Where is positional uncertainty a problem for species distribution modelling? Ecography. 2014;37(2):191–203.

75. Park DS, Davis CC. Implications and alternatives of assigning climate data to geographical centroids. J Biogeogr. 2017;44(10):2188–98.

76. Pearson KD. Rapid enhancement of biodiversity occurrence records using unconventional specimen data. Biodivers Conserv. 2018;27(11):3007–18.

77. Perry GLW, Dickson ME. Using Machine Learning to Predict Geomorphic Disturbance: The Effects of Sample Size, Sample Prevalence, and Sampling Strategy. J Geophys Res Earth Surf. 2018;123(11):2954–70.

78. Petersen TK, Speed JDM, Grøtan V, Austrheim G. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. Ecological Solutions and Evidence. 1º de janeiro de 2021;2(1).

79. Quillfeldt P, Engler JO, Silk JRD, Phillips RA. Influence of device accuracy and choice of algorithm for species distribution modelling of seabirds: a case study using black-browed albatrosses. J Avian Biol. 2017;48(12):1549–55.

80. Radović A, Schindler S, Rossiter D, Nikolić T. Impact of biased sampling effort and spatial uncertainty of locations on models of plant invasion patterns in Croatia. Biol Invasions. 2018;20(12):3527–44.

81. Ratnieks FLW, Schrell F, Sheppard RC, Brown E, Bristow OE, Garbuzov M. Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. Methods Ecol Evol. 2016;7(10):1226–35.

82. Ribeiro BR, Guidoni-Martins K, Tessarolo G, José S, Velazco E, Jardim L, et al. Issues with species occurrence data and their impact on extinction risk

assessments. Biol Conserv [Internet]. 2022 [citado 11 de abril de 2023];273:109674. Disponível em: https://doi.org/10.1016/j.biocon.2022.109674

83. Roberts DR, Wood WH, Marshall SJ. Assessments of downscaled climate data with a high-resolution weather station network reveal consistent but predictable bias. International Journal of Climatology. 2019;39(6):3091–103.

84. Robinson OJ, Ruiz-Gutierrez V, Fink D. Correcting for bias in distribution modelling for rare species using citizen science data. Divers Distrib. 2018;24(4):460–72.

85. Salim JA, Saraiva AM, Zermoglio PF, Agostini K, Wolowski M, Drucker DP, et al. Data standardization of plant-pollinator interactions. 2022 [citado 11 de abril de 2023];11:1–15. Disponível em: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac043/6593428

86. Serra-Diaz JM, Enquist BJ, Maitner B, Merow C, Svenning JC. Big data of tree species distributions: how big and how good? For Ecosyst. 2017;4(1).

87. Simões MVP, Peterson AT. Utility and limitations of climate-matching approaches in detecting different types of spatial errors in biodiversity data. Insect Conserv Divers. 2018;11(5):407–14.

88. Stoklosa J, Daly C, Foster SD, Ashcroft MB, Warton DI. A climate of uncertainty: Accounting for error in climate variables for species distribution models. Methods Ecol Evol. 2015;6(4):412–23.

89. Tarr N, Benson A, Rubino M. Wildlife Wrangler: A high-level data processing framework that supports the utilization of species occurrence data for biogeographical analyses. Biodiversity Information Science and Standards. 24 de agosto de 2022;6.

90. Torre M, Nakayama S, Tolbert TJ, Porfiri M. Producing knowledge by admitting ignorance: Enhancing data quality through an “I don’t know” option in citizen science. PLoS One. 2019;14(2):1–15.

91. Van Eupen C, Maes D, Herremans M, Swinnen KRR, Somers B, Luca S. The impact of data quality filtering of opportunistic citizen science data on species distribution model performance. Ecol Modell [Internet]. 2021 [citado 11 de abril de 2023];444:109453. Disponível em: http://creativecommons.org/licenses/by/4.0/

92. Watling JI, Fletcher RJ, Speroterra C, Bucklin DN, Brandt LA, Romañach SS, et al. Assessing Effects of Variation in Global Climate Data Sets on Spatial Predictions from Climate Envelope Models. J Fish Wildl Manag. 2014;5(1):14–25.

93. Yu J, Wong WK, Kelling S. Clustering species accumulation curves to identify skill levels of citizen scientists participating in the eBird project. Twenty-sixth IAAI Conference. 2014;3017–23.

94. Zhang G, Zhu AX, Huang ZP, Xiao W. A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. Transactions in GIS. 2018;22(1):202–16.

95. Zizka A, Carvalho FA, Calvente A, Baez-Lizarazo MR, Cabral A, Ramos Coelho JF, et al. No one-size-fits-all solution to clean GBIF. PeerJ [Internet]. 28 de setembro de 2020 [citado 11 de abril de 2023];8:e9916. Disponível em: https://peerj.com/articles/9916

96. Feldman MJ, Imbeau L, Marchand P, MazerolleI MJ, Darveau M, Fenton1 NJ. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. PLoS One [Internet]. 1º de março de 2021 [citado 17 de abril de 2023];16(3):e0234587. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0234587

97. Waldock C, Stuart-Smith RD, Albouy C, Cheung WWL, Edgar GJ, Mouillot D, et al. A quantitative review of abundance-based species distribution models. Ecography [Internet]. 1º de janeiro de 2022 [citado 17 de abril de 2023];2022(1). Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.05694

98. Robinson NM, Nelson WA, Costello MJ, Sutherland JE, Lundquist CJ. A systematic review of marine-based Species Distribution Models (SDMs) with recommendations for best practice. Front Mar Sci. 18 de dezembro de 2017;4(DEC):421.

99. Pecchi M, Marchi M, Burton V, Giannetti F, Moriondo M, Bernetti I, et al. Species distribution modelling to support forest management. A literature review. Ecol Modell [Internet]. 1º de novembro de 2019 [citado 17 de abril de 2023];411. Disponível em: https://www.researchgate.net/publication/335856189_Species_distribution_modelling_to_support_forest_management_A_literature_review

100. Strong DM, Lee YW, Wang RY. Data Quality In Context. 1997 [citado 17 de abril de 2023];40(5). Disponível em: http://web.mit.edu/tdqm.

101. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Source: Journal of Management Information Systems. 1996;12(4):5–33.

102. Ramasamy A, Chowdhury S. BIG DATA QUALITY DIMENSIONS: A SYSTEMATIC LITERATURE REVIEW. JISTEM - Journal of Information Systems and Technology Management [Internet]. 13 de julho de 2020 [citado 17 de abril de 2023];17:e202017003. Disponível em: http://www.scielo.br/j/jistm/a/dz8CSmJT3MXC7J5dw8qPfss/

103. Almutiry O, Wills G, Crowder R. DIMENSION-ORIENTED TAXONOMY OF DATA QUALITY PROBLEMS IN ELECTRONIC HEALTH RECORD. IADIS International Journal on WWW/Internet. 2015;13(2):98–114.

104. Miot HA. Anomalous values and missing data in clinical and experimental studies Valores anômalos e dados faltantes em estudos clínicos e experimentais. Journal Vascular Brasileiro [Internet]. 2019 [citado 17 de abril de 2023]; Disponível em: https://doi.org/10.1590/1677-5449.190004

105. Miguel DI, Solange N. AS, Marcia I, Da Silva Suzana Alves. Data Quality in health records: A literature review. Iberian Conference on Information Systems and Technologies, CISTI. 23 de junho de 2021;

106. Wang A, Ye Y, Song X, Zhang S, Yu JJQ. Traffic Prediction With Missing Data: A Multi-Task Learning Approach. IEEE Transactions on Intelligent Transportation Systems. 9 de janeiro de 2023;1–14.

107. Chan RKC, Lim JMY, Parthiban R. Missing Traffic Data Imputation for Artificial Intelligence in Intelligent Transportation Systems: Review of Methods, Limitations, and Challenges. IEEE Access [Internet]. 2023 [citado 17 de abril de 2023];11:34080–93. Disponível em: https://ieeexplore.ieee.org/document/10091533/

108. Maragatharajan M, Prequiet L. Removal of duplicate data from encrypted cloud storage. Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017. 26 de fevereiro de 2018;2018-February:1–5.

109. Zhang P. Similar Duplicate Record Detection of Big Data Based on Entropy Grouping Clustering. Proceedings - 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering, AEMCSE 2022. 2022;646–50.

110. Gomes VHF, Ijff SD, Raes N, Amaral IL, Salomão RP, Coelho LDS, et al. Species Distribution Modelling: Contrasting presence-only models with plot abundance data. Scientific Reports 2018 8:1 [Internet]. 17 de janeiro de 2018 [citado 10 de janeiro de 2023];8(1):1–12. Disponível em: https://www.nature.com/articles/s41598-017-18927-1

111. Kullenberg C, Kasperowski D. What is citizen science? - A scientometric meta-analysis. PLoS One. 1º de janeiro de 2016;11(1).

112. Lewandowski E, Specht H. Influence of volunteer and project characteristics on data quality of biological surveys. Conservation Biology [Internet]. 1º de junho de 2015 [citado 1º de dezembro de 2018];29(3):713–23. Disponível em: http://doi.wiley.com/10.1111/cobi.12481

113. Smith J, Noble H. Bias in research. Evidence Based Nursing. 2014;17(4):100–1.

114. Robinson O, Ruiz-Gutierrez V, Fink D. Correcting for bias in distribution modelling for rare species using citizen science data. Heikkinen R, organizador. Divers Distrib. 1º de abril de 2018;24(4):460–72.

115. Fei S, Yu F. Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. Landsc Ecol. 2016;31(1):31–42.

116. Silvola R, Harkonen J, Vilppola O, Vehkapera HK, Haapasalo H. Data quality assessment and improvement. Int J Bus Inf Syst. 2016;22(1):62.