ATIVIDADE 4 - Resultados

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

Objetivo do artigo: Identificar os problemas de qualidade observados nos dados bioclimáticos utilizados em modelos de distribuição de espécies, tanto do ponto de vista da aquisição quanto do processo de análise, e investigar seus efeitos nos modelos.

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.

Resultados

Avaliação Bibliométrica

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. 2018 destaca-se por ser o ano com 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.

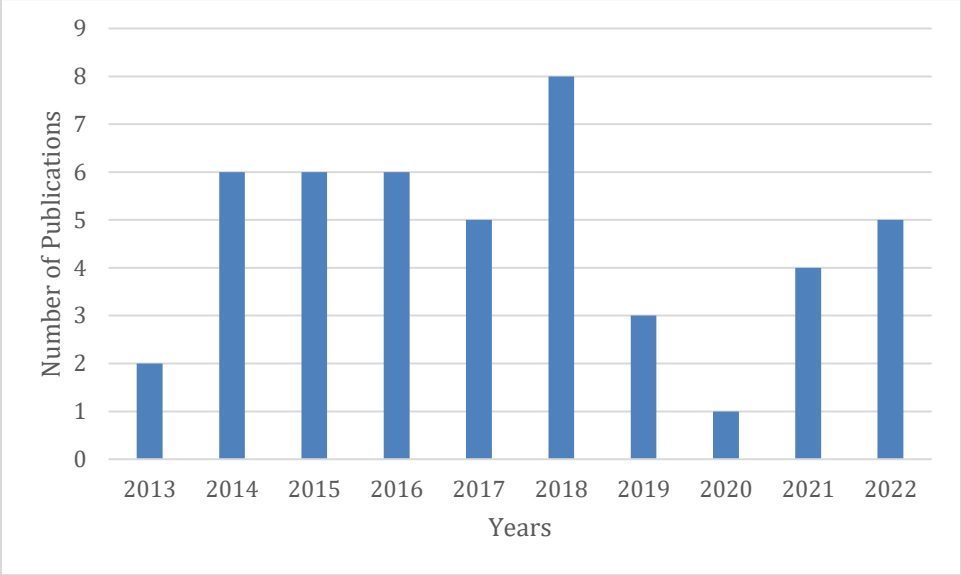

Fig 2: Visão temporal dos artigos selecionados

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

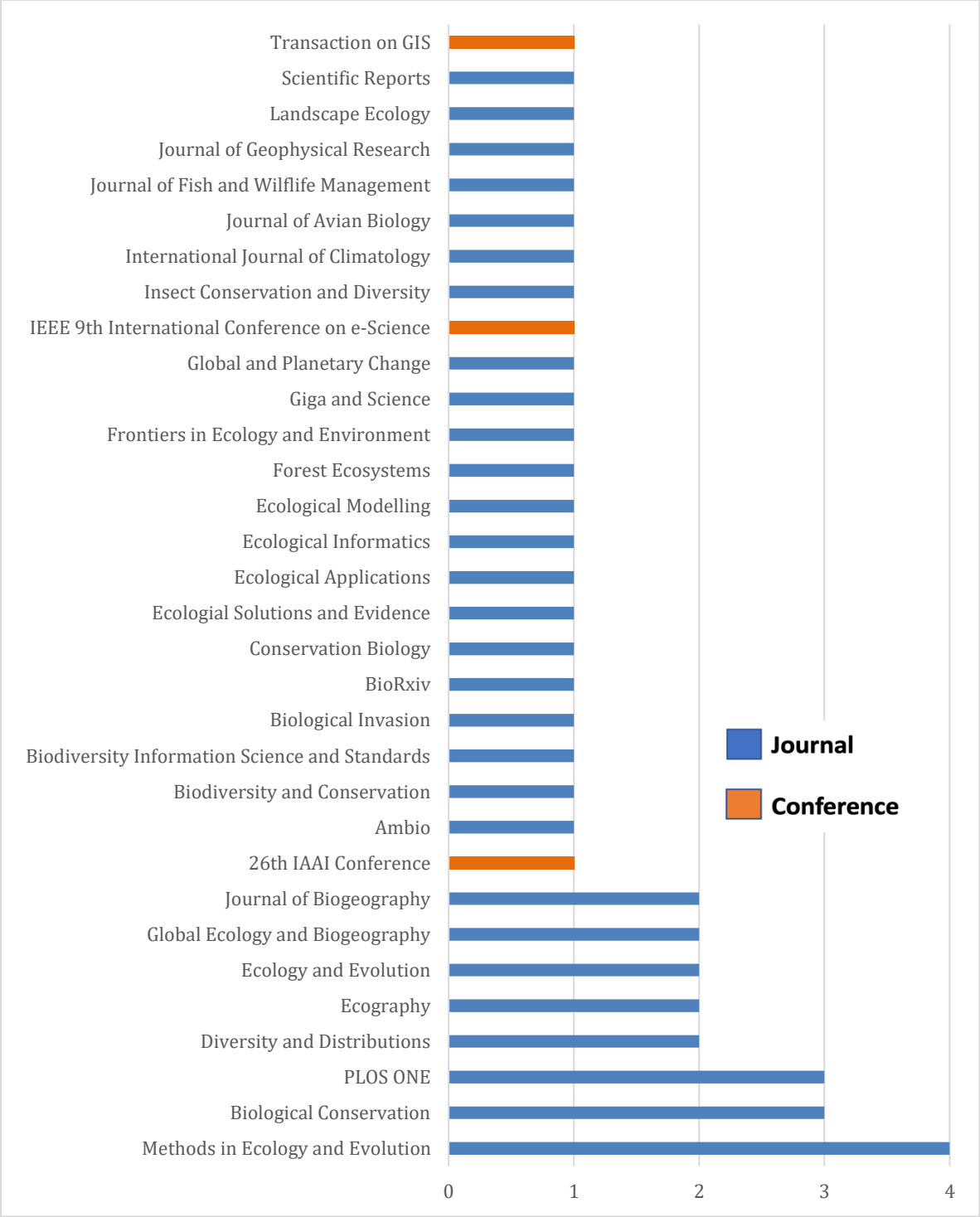

Fig 3: Fontes de publicações dos artigos selecionados para leitura final da SLR.

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 (KAPTEIN; HIEMSTRA; KAMPS, 2010). A nuvem de palavras extrai as palavras mais usadas em um texto específico e as exibe por tamanho, com base em suas frequências (DEPAOLO; WILKINSON, 2014). 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: Nuvem de palavras representando os termos mais comumente presente nos títulos e resumos dos artigos selecionados.

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 qualidade e dados. 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 6 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.

Table 1: Artigos selecionados e as questões de pesquisa que eles ajudam a responder.

#	PAPERS	RESEARCH QUESTION	
		1	2
1	(ARENAS-CASTRO et al., 2022)	✓	
2	(AUBRY; RALEY; MCKELVEY, 2017)*	✓	
3	(BEDIA; HERRERA; GUTIÉRREZ, 2013)*	✓	
4	(BIRD et al., 2014)	✓	✓
5	(BLOOM; FLOWER; DECHAINE, 2018)	✓	✓
6	(BOYD et al., 2021)	✓	✓
7	(CLARE et al., 2019)	✓	✓
8	(CUGLER et al., 2013)	✓	✓
9	(DORAZIO, 2014)*	✓	
10	(FEI; YU, 2016)	✓	✓
11	(FELDMAN et al., 2021)	✓	
12	(FÜHRDING-POTSCHKAT; KREFT; ICKERT-BOND, 2022)	✓	✓
13	(GELDMANN et al., 2016)	✓	✓
14	(GOMES et al., 2018)*	✓	
15	(GRAHAM; HAINES-YOUNG; FIELD, 2015)	✓	✓
16	(GUETA; CARMEL, 2016)	✓	✓
17	(HEFLEY et al., 2014)	✓	✓

18	(HEFLEY; BROST; HOOTEN, 2017)*	✓	✓
19	(KELLING et al., 2015)	✓	✓
20	(KOSMALA et al., 2016)	✓	✓
21	(LEWANDOWSKI; SPECHT, 2015)	✓	✓
22	(LIN et al., 2015)	✓	✓
23	(MALDONADO et al., 2015)	✓	✓
24	(NAIMI et al., 2014)	✓	✓
25	(PARK; DAVIS, 2017)*	✓	✓
26	(PEARSON, 2018)	✓	✓
27	(PERRY; DICKSON, 2018)	✓	✓
28	(PETERSEN et al., 2021)	✓	✓
29	(QUILLFELDT et al., 2017)*	✓	✓
30	(RADOVIĆ et al., 2018)	✓	
31	(RATNIEKS et al., 2016)*	✓	✓
32	(RIBEIRO et al., 2022)	✓	
33	(ROBERTS; WOOD; MARSHALL, 2019)*	✓	
34	(ROBINSON; RUIZ-GUTIERREZ; FINK, 2018)*	✓	✓
35	(SALIM et al., 2022)	✓	✓
36	(SERRA-DIAZ et al., 2017)	✓	✓
37	(SIMÕES; PETERSON, 2018)	✓	✓
38	(STOKLOSA et al., 2015)*	✓	✓
39	(TARR; BENSON; RUBINO, 2022)	✓	✓
40	(TORRE et al., 2019)	✓	✓
41	(VAN EUPEN et al., 2021)	✓	✓
42	(VELÁSQUEZ-TIBATÁ; GRAHAM; MUNCH, 2016)	✓	✓
43	(WATLING et al., 2014)*	✓	
44	(YU; WONG; KELLING, 2014)	✓	✓
45	(ZHANG et al., 2018)*	✓	✓
46	(ZIZKA et al., 2020)	✓	✓
	TOTAL	46	36

A Tabela 7 apresenta os tipos de problemas de qualidade de dados identificados nos 46 artigos (Tabela 6) 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.

Table 7: 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

Referências somente da atividade

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

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

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

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

BLOOM, T. D. S.; FLOWER, A.; DECHAINE, E. G. Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. **Ecology and Evolution**, v. 8, n. 1, p. 765–777, 2018.

BOYD, R. J. et al. occAssess: An R package for assessing potential biases in species occurrence data. **bioRxiv**, p. 2021.04.19.440441, 10 ago. 2021.

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

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

DEPAOLO, C. A.; WILKINSON, K. Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data. **TechTrends**, v. 58, n. 3, p. 38–44, 21 maio 2014.

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

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

FELDMAN, M. J. et al. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. **PLOS ONE**, v. 16, n. 3, p. e0234587, 1 mar. 2021.

FÜHRDING-POTSCHKAT, P.; KREFT, H.; ICKERT-BOND, S. M. Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models. **Ecology and Evolution**, v. 12, n. 8, p. e9168, 1 ago. 2022.

GELDMANN, J. et al. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. **Diversity and Distributions**, v. 22, n. 11, p. 1139–1149, 2016.

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

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

GUETA, T.; CARMEL, Y. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. **Ecological Informatics**, v. 34, p. 139–145, 2016.

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

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

KAPTEIN, R.; HIEMSTRA, D.; KAMPS, J. How Different Are Language Models andWord Clouds? Em: [s.l.] Springer, Berlin, Heidelberg, 2010. p. 556–568.

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

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

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

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

MALDONADO, C. et al. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? **Global Ecology and Biogeography**, v. 24, n. 8, p. 973–984, 1 ago. 2015.

NAIMI, B. et al. Where is positional uncertainty a problem for species distribution modelling? **Ecography**, v. 37, n. 2, p. 191–203, 2014.

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

PEARSON, K. D. Rapid enhancement of biodiversity occurrence records using unconventional specimen data. **Biodiversity and Conservation**, v. 27, n. 11, p. 3007–3018, 2018.

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

PETERSEN, T. K. et al. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. **Ecological Solutions and Evidence**, v. 2, n. 1, 1 jan. 2021.

QUILLFELDT, P. et al. Influence of device accuracy and choice of algorithm for species distribution modelling of seabirds: a case study using black-browed albatrosses. **Journal of Avian Biology**, v. 48, n. 12, p. 1549–1555, 2017.

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

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

RIBEIRO, B. R. et al. Issues with species occurrence data and their impact on extinction risk assessments. **Biological Conservation**, v. 273, p. 109674, 2022.

ROBERTS, D. R.; WOOD, W. H.; MARSHALL, S. J. Assessments of downscaled climate data with a high-resolution weather station network reveal consistent but predictable bias. **International Journal of Climatology**, v. 39, n. 6, p. 3091–3103, 2019.

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

SALIM, J. A. et al. Data standardization of plant-pollinator interactions. v. 11, p. 1–15, 2022.

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

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

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

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**, v. 6, 24 ago. 2022.

TORRE, M. et al. Producing knowledge by admitting ignorance: Enhancing data quality through an "I don't know" option in citizen science. **PLoS ONE**, v. 14, n. 2, p. 1–15, 2019.

VAN EUPEN, C. et al. The impact of data quality filtering of opportunistic citizen science data on species distribution model performance. **Ecological Modelling**, v. 444, p. 109453, 2021.

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

WATLING, J. I. et al. Assessing Effects of Variation in Global Climate Data Sets on Spatial Predictions from Climate Envelope Models. **Journal of Fish and Wildlife Management**, v. 5, n. 1, p. 14–25, 2014.

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

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

ZIZKA, A. et al. No one-size-fits-all solution to clean GBIF. **PeerJ**, v. 8, p. e9916, 28 set. 2020.