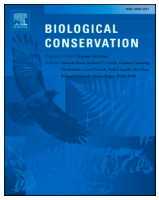
Biological Conservation 273 (2022) 109674

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon



Issues with species occurrence data and their impact on extinction risk assessments

Bruno R. Ribeiro [a,*], Karlo Guidoni-Martins [a], Geiziane Tessarolo [b], Santiago José Elías Velazco [c,d,e], Lucas Jardim [f], Steven P. Bachman [g], Rafael Loyola [h,i]

[a] Programa de Pós-graduação em Ecologia e Evolução, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[b] Universidade Estadual de Goiás, UEG, Campus de Ciências Exatas e Tecnológicas - CCET, Anápolis, Goiás, Brazil
[c] Instituto de Biología Subtropical, Universidad Nacional de Misiones - CONICET, Puerto Iguazú, Misiones, Argentina
[d] Department of Botany and Plant Sciences, University of California - Riverside, Riverside, CA, USA
[e] Programa de Pós-Graduação em Biodiversidade Neotropical, Universidade Federal da Integração Latino-Americana (UNILA), Foz do Iguaçu, Paraná, Brazil
[f] Laboratório de Ecologia Teórica e Síntese, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[g] Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK
[h] Instituto Internacional para Sustentabilidade, Rio de Janeiro, RJ, Brazil
[i] Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Goiás, Brazil

ARTICLE INFO

Keywords:
Biodiversity data
Fitness-for-use
Data quality
GBIF
Plants
Rapid extinction risk assessment

ABSTRACT

Species extinction risk status is critical to support conservation actions. However, full assessments published on the Red List are slow and resource intensive. To tackle assessments for mega-diverse groups, gains can be made through preliminary assessments that can help prioritize efforts toward full assessments. Here, we quantified how incomplete data collation and errors in the taxonomic, spatial, and temporal dimensions of species-occurrence data translate into misclassifications of extinction risk. Using a dataset of >30 million records of terrestrial plants occurring in Brazil compiled from nine databases we conducted preliminary risk assessments for ~94 % of the 6046 species assessed by the Brazilian Red List authority. We found that no unique database contained data sufficient to perform extinction risk assessment of all species; e.g., the risk of 78 % of species can be assessed using data from GBIF. The overall accuracy (66–75 %) and specificity (89–98 %, correct prediction of non-threatened species) were less affected by incomplete data collation and issues in species-occurrence records. Sensitivity rates (correct prediction of threatened species) were commonly low to moderate and strongly affected by incomplete data collation (13–47 %) and spatial issues (38 %). Our results demonstrate that species' preliminary risk assessments have high accuracy in identifying non-threatened species, even when data collection is low and in the presence of issues in species occurrence data highlighting that such an approach can be used to efficiently prioritize species for full Red List assessments. In addition, caution is needed before declaring a species as threatened without considering data collation intensity and quality.

1. Introduction

Recent years have seen an explosion in the availability of species-occurrence data shared in online databases (Canhos et al., 2015; Graham et al., 2004). Such openly accessible biodiversity databases provide a vast and invaluable resource to document species distributions for many research uses (e.g., biogeographic studies, ecological applications, and conservation decision making; Ball-Damerow et al., 2019). However, the accumulation of large datasets may, without scrutiny, propagate errors that might influence outputs (Maldonado et al., 2015). An example of this is the evaluation of extinction risk as applied in systems such as the International Union for Conservation of Nature (IUCN) Red List of Threatened Species, hereafter the Red List (IUCN, 2012).

The Red List is the most comprehensive and authoritative source of information on the global extinction risk of species (Rodrigues et al., 2006). It is used to assign species to categories of extinction risk based on a set of five quantitative criteria (symptoms of extinction risk) associated

* Corresponding author at: Programa de Pós-graduação em Ecologia e Evolução, Universidade Federal de Goiás, Avenida Esperança s/n, Campus Samambaia, CEP 74.690-900 Goiânia, Goiás, Brazil.
E-mail address: ribeiro.brr@gmail.com (B.R. Ribeiro).

https://doi.org/10.1016/j.biocon.2022.109674
Received 13 October 2021; Received in revised form 21 June 2022; Accepted 21 July 2022

with population size, geographic distribution, and rates of decline of both (IUCN, 2012). Due mainly to its rigorous process, full Red List assessments have been proven to be expensive and time-consuming (Miller et al., 2012; Rondinini et al., 2014). Although full Red List assessments have been carried out for several groups of vertebrates, less well-known and species-rich groups such as invertebrates and plants remain largely under-assessed, especially in tropical regions (Nic Lughadha et al., 2019a; Bachman et al., 2019). Rapid, automatically generated assessments have been proven crucial to expediting and assisting the production of full Red List assessments (Nic Lughadha et al., 2019b). Such approaches produce a preliminary assessment of species extinction risk based on digitally available species distribution data and can be used to efficiently prioritize likely threatened species for full Red List assessments (Bachman et al., 2019; Nic Lughadha et al., 2019b; Zizka et al., 2021). They are considered preliminary because other criteria and subcriteria must be met to justify a full Red List assessment (IUCN, 2012).

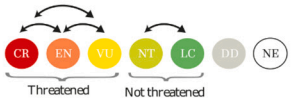
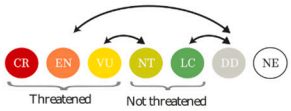
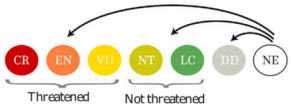
Data availability and quality are crucial for species extinction risk assessments, particularly preliminary automated estimates based on geographical range size (criterion B; Nic Lughadha et al., 2019a; Zizka et al., 2020, 2021; Panter et al., 2020). Issues related to difficulties in standardizing and integrating data from different sources (e.g., GBIF and SpeciesLink; Kissling et al., 2018), discrepancies and errors in taxonomic and nomenclatural data (Nic Lughadha et al., 2019a), and errors and inaccuracies in geographical and temporal information of primary species-occurrence data (e.g., Meyer et al., 2016) may lead to under-or over-estimation of species range size and therefore incorrect classification of extinction risk. These errors, in turn, may trigger inappropriate conservation responses (Brummitt et al., 2008; Nic Lughadha et al., 2019a).

On the one hand, unchecked taxonomy (e.g., misspelled names and synonyms) and the incomplete collation of species records (e.g., data compiled from few data sources) may lead to an under-estimation of a species' range, therefore over-estimating its extinction risk. On the other hand, records with suspect or incorrect coordinates (e.g., within urban areas or geographical outliers) or the use of historical data (where continuing presence of species in a location subject to, for example, habitat conversion, is in doubt), may overestimate the size of the species range, which could be incorrectly assigned a lower risk category (i.e., falsely classify a non-threatened as threatened species). In terms of allocating resources for conservation, incorrectly assessing a threatened species as non-threatened could mean vital resources are deprioritized. In contrast, considering a non-threatened species as threatened represents wasted effort.

Previous studies evaluating the influence of issues with species

Box 1

Scenarios on how an incomplete data collation and errors, gaps, and inaccuracies in the taxonomic, spatial, and temporal information can influence preliminary Red List assessments (Nic Lughadha et al., 2019a; Zizka et al., 2020, 2021; Panter et al., 2020). Incomplete data collation or issues in species-occurrence data may alter the number of records available to perform Red List assessment, which can change the estimates of Extent of Occurrence (a metric of range size used to assess species under the IUCN's criterion B). Minor issues in species-occurrence records may not change EOO estimates and species' threat category or lead to changes within risk categories (e.g., a threatened species continues to qualify as threatened but under a different category). On the other hand, insufficient data collation or significant issues in species-occurrence records may lead to changes in species extinction risk status (e.g., consider a threatened species as not threatened). Finally, certain species can only be assessed if new data is collated. This may occur, for example, when data from only one data aggregator (e.g., GBIF) is used in Red List assessments. According to the IUCN, species are classified as threatened (critically endangered [CR], endangered [EN], and vulnerable [VU]), not threatened (near threatened [NT] and least concern [LC]), data deficient (DD), or not evaluated (NE). Hypotheses indicated with an asterisk (*) were supported by our findings (Results Sections 3.2 and 3.3).

Scenarios	Examples	Hypotheses	Implications
No change		- No errors found - Minor errors in taxonomy, coordinates, or collection date*	Changes in species' range size are not translated into changes in extinction risk category.
Changes within binary categories		- Minor errors in coordinates or date of collection* - Taxonomy potentially oudated*	Errors lead to changes in detailed extinction risk category, but not in the binary category (e.g., species still qualifies as threatened). Changes in species' range size alter species' detailed extinction risk category. Potential implication for conservation prioritization.
Changes in binary and detailed categories		- Major errors in coordinates* - Presence of historical records* - Outdated taxonomy - Insufficient data collation*	Errors lead to changes in detailed and binary extinction risk status. Potentially threatened species deprioritised for conservation (conservation debt) or conservation efforts misaplied to a non-threatened species (conservation wastage).
Not evaluated vs Evaluated		- New data collation* - Fewer species classified as Data Deficient (DD)*	More species assessed.

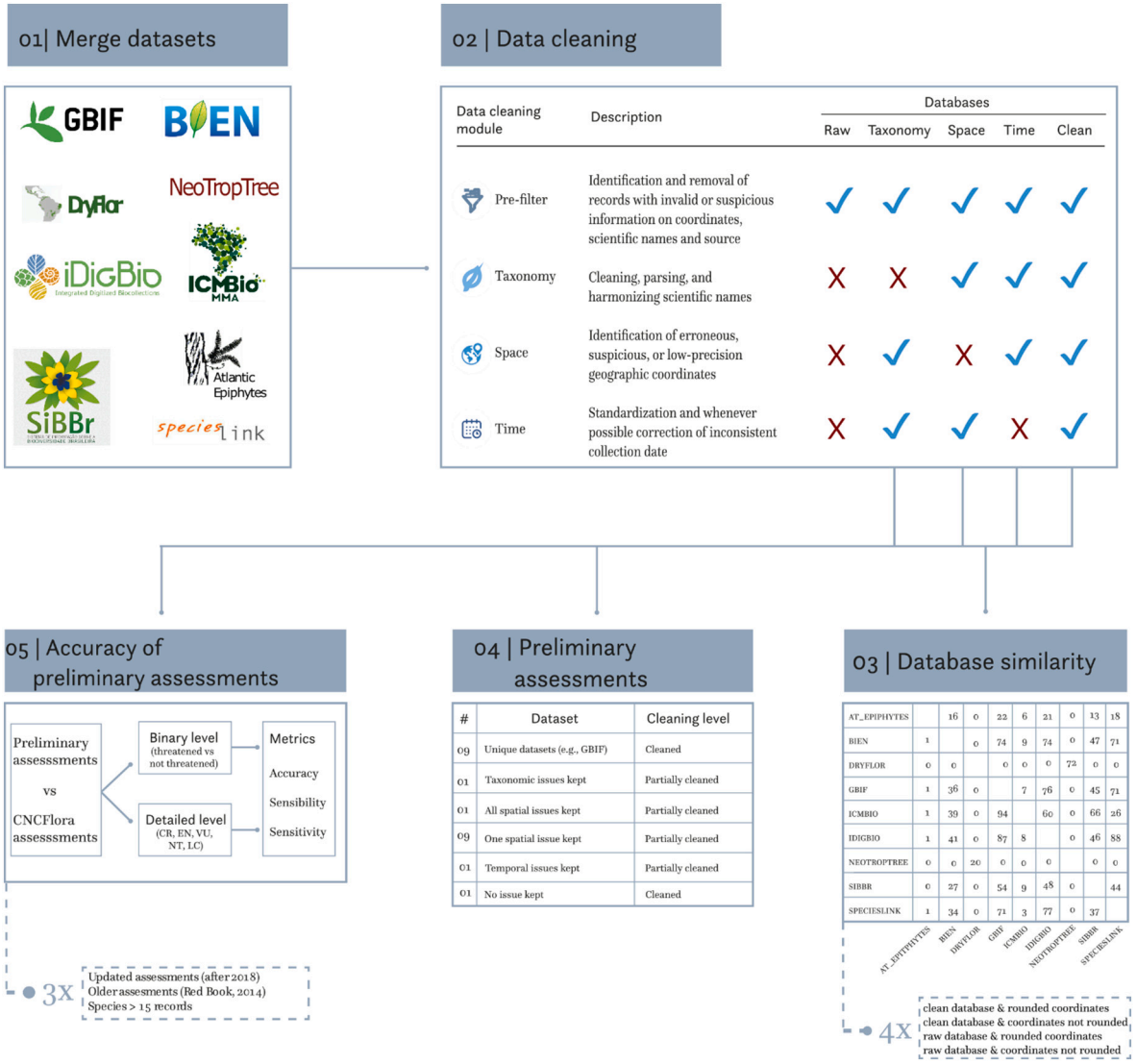


Fig. 1. A schematic representation of the methods used to evaluate the impact of incomplete data collation and issues in species occurrence records on preliminary extinction risk assessments. We obtained ~30 million records of terrestrial plants in Brazil from nine data aggregators. These datasets were merged and standardized to generate a "raw" dataset after the removal of records lacking scientific names, coordinates, or from doubtful sources. We used the "raw" database to generate four partially cleaned databases (taxonomic, spatial, temporal, and clean) containing only one issue (e.g., the taxonomic database only contains taxonomic issues). The "raw" and "clean" databases were used to evaluate the similarity between nine datasets. The "clean" and "partially cleaned" databases were used to carry out a preliminary risk assessment based on IUCN criterion B. We assessed the performance of preliminary assessments in correctly predicting the Red List categories of 5524 species with complete published assessments. The performance was measured in terms of overall accuracy, sensitivity (correct prediction of threatened categories), and specificity (correct prediction of not threatened categories) at both levels, binary and detailed. Preliminary assessments were also carried out using different reference datasets, including species assessed in the Brazilian Red Book published in 2013 ("Older assessments"), assessments carried out in 2018 ("Newer assessments"), and assessments performed considering only species with >15 records (">15 records").

occurrence records on preliminary risk assessment based on criterion B found contrasting results. Some findings showed a more substantial effect of spatial errors on sensitivity rates (i.e., correctly predicting threatened categories) than on specificity rates (i.e., correctly predicting non-threatened categories) (Panter et al., 2020). Such impact varied depending on the data sources (e.g., GBIF vs. BIEN). In contrast, preliminary risk assessment was relatively robust to the presence of records with spatial and taxonomic issues (Nic Lughadha et al., 2019a; Zizka et al., 2021, 2020), mainly due to wide thresholds of criteria used to assess species' extinction risk. Hence, a comprehensive assessment to disentangle the impact of issues related to all biodiversity dimensions (taxonomic, spatial, and temporal) and an incomplete data collation on risk assessment is yet to be examined (Walker et al., 2021). The potential impact of taxonomic, spatial, and temporal issues and an incomplete collation of records on the Red List assessment is summarized in Box 1.

Brazil is a mega-diverse nation harboring 35,683 terrestrial plant species, of which 53 % are endemics and many under considerable threat (BFG, 2021). Ongoing attempts to document extinction have resulted in 6046 assessments (CNCFlora, 2021), highlighting the pressing need for full Red List assessments and making Brazil an ideal case study. Here, using a dataset of >30 million records of terrestrial plants occurring in Brazil, we quantified how incomplete data collation and errors in the taxonomic, spatial, and temporal dimensions of occurrence data translate into misclassifications of preliminary extinction risk carried out at national scale. By aggregating large amounts of data from heterogeneous sources across Brazil, we highlight challenges with handling and processing these data. Specifically, we ask the following questions: 1) How similar are online databases regarding species-occurrence records they share? 2) What is the impact of gaps generated by an incomplete collation of occurrence data – e.g., use of data

from a single database (e.g., GBIF) – on estimates of species ranges and preliminary extinction risk assessments? 3) What proportion of preliminary extinction risk assessments are potentially over-or underestimated due to inaccuracies and errors in taxonomic, geographical, and temporal dimensions of biodiversity data? 4) Which of these issues contributes most to misclassifying species extinction risk?

2. Materials and methods

2.1. Data compilation and cleaning

The dataset underlying this study's analyses was compiled in Ribeiro et al. (2022). The dataset includes >30 million records for terrestrial plant species in Brazil accessed via nine public, freely, and openly available online databases (Fig. 1). The compilation included data from eight biodiversity repositories: Botanical Information and Ecological Network version 4.1 (BIEN, bien.nceas.ucsb.edu/bien); Global Biodiversity Information Facility (www.gbif.org), Integrated Digitized Biocollections (iDigBio, www.idigbio.org/), speciesLink network (SpeciesLink, splink.cria.org.br), Brazilian Biodiversity Information System (SiBBr, www.sibbr.gov.br), Tree flora of the Neotropical Region (NEOTROPTREE, www.neotroptree.info), The Latin American Seasonally Dry Tropical Forest Floristic Network (www.dryflor.info), Brazilian Biodiversity Portal (ICMBio, portaldabiodiversidade.icmbio.gov.br/portal), and from the data paper Atlantic forest epiphytes (Ramos et al., 2019).

We used the Biodiversity Data Cleaning (*bdc*) package for assessing the quality of species-occurrence data, considering their taxonomic, spatial, and temporal dimensions (Ribeiro et al., 2022; https://cran.r-project.org/web/packages/bdc/index.html). The package contains functions to clean and assess the quality of biodiversity data grouped in the following thematic modules: 1) Merge datasets: standardization and integration of different databases in a standardized format; 2) pre-filter: flagging and removal of invalid or non-interpretable information; 3) taxonomy: cleaning, parsing, and standardizing scientific names; 4) space: flagging of erroneous, suspect, and low-precision geographic coordinates; and 5) time: flagging and, whenever possible, correcting inconsistent collection dates (Fig. 1).

Each module contains a series of tests to assert data quality. By executing each test, original data are retained, and the result is appended in a different field as TRUE (accurate records) or FALSE (records flagged as erroneous or suspect). We created a "raw" dataset after excluding records flagged as incorrect in the "pre-filter" step, including records missing coordinates or species names, in the ocean, with out-of-range coordinates, outside Brazil, and from distrustful sources (e.g., from drawing and photographs, among others). These records are commonly not fit for the Red List assessment without prior amendments (Fig. 1). We also excluded records of non-native species (cultivated and naturalized), algae, and fungi species (Table S1).

We used the "raw" dataset to generate four databases of species occurrence with different levels of data curation (i.e., three "partially cleaned" and one "clean" dataset) (Panter et al., 2020). In each "partially cleaned" dataset, all issues were corrected except the problem to be tested (Fig. 1). Thus, the "taxonomic" dataset contains only taxonomic issues, i.e., synonyms, nomenclatural variants, and misspelled names. The "spatial" dataset contains only geographical issues, e.g., records assigned to country capital, in urban areas, with low-precision coordinates, and geographical outliers (Fig. S1; Zizka et al., 2019). The "temporal" dataset contains only temporal issues, including records collected before 1970 or containing illegitimate information (e.g., collection date in the future). Occurrence data collected in the last 30–40 years are more likely to be geo-referenced using GPS, and, therefore, more accurate (Boitani et al., 2011; Graham et al., 2004); in comparison, records collected before the year 1970 are generally associated with a moderate to a high level of inaccuracies (Tessarolo et al., 2017). Finally, we created a "clean", well-curated, and near-comprehensive dataset in which all issues were removed or corrected (Fig. 1). The "taxonomic", "spatial", "temporal", and "clean" datasets were used to perform the downstream analysis (Fig. 1).

2.2. Database similarity

To answer the question about the similarity of online databases regarding the species-occurrence records they share (Question 1), we compared the proportion of redundant information shared between them. To do so, we built a similarity matrix with the number of unique records (i.e., those with equal species name, latitude, and longitude data) shared between databases (Fig. 1). Low similarity values indicate the uniqueness of a database, i.e., few records are shared with other databases.

We evaluated the similarity considering the "raw" and "clean" datasets separately (Fig. 1). Before assessing the similarity between original databases, we removed records missing coordinates or scientific names and authority names and annotations from scientific names (but kept terms denoting taxonomic uncertain and intraspecific levels such as "cf.", "var"). These procedures were necessary to avoid falsely considering records as duplicated. Since databases contain coordinates with different precision, we also evaluated the similarity after rounding the geographical coordinates to four decimals degrees. Our approach is based upon the amount of potentially duplicate records shared between databases since comprehensive and rich meta-data is needed to detect actual duplicate records, but such data is seldom available.

2.3. Preliminary risk assessments

We used the R package *rCAT* (Moat, 2017) to generate preliminary extinction risk assessments for each species based on extent of occurrence (EOO; Red List criteria B), a range size metric commonly used for extinction risk assessment for plants due to scarcity of population data (Brummitt et al., 2008, 2015). The EOO, the minimum area encompassing all species records, was calculated as a minimum convex polygon (IUCN, 2019). Based on EOO size estimates, *rCAT* classifies species in one of the following IUCN categories: critically endangered (CR; EOO < 100 km^2), endangered (EN; EOO ≥ 100 and <5000 km^2), vulnerable (VU; EOO ≥ 5000 and $<20{,}000$ km^2), near threatened (NT; EOO $\geq 20{,}000$ and $<30{,}000$ km^2), or least concern (LC; EOO $\geq 30{,}000$ km^2).

2.4. Accuracy of preliminary risk assessments

We calculated the accuracy of our preliminary assessments in correctly predicting the Red List categories of 5524 species with complete published assessments carried out by the Brazilian Center for Flora Conservation (hereafter, CNCFlora assessments; CNCFlora, 2021). We compared the performance of preliminary evaluations at the binary level of threatened (Red List categories CR, EN, or VU) vs. not threatened (NT or LC) as well as a more detailed level where predicted Red List categories had to match the published CNCFlora categories (Fig. 1; Zizka et al., 2021). Performance of preliminary assessments was calculated as overall accuracy, sensitivity (correct prediction of threatened categories), and specificity (correct prediction of not threatened categories; Nic Lughadha et al., 2019a).

2.5. Impact of incomplete collation of species-occurrence data on preliminary risk assessments

To evaluate the impact of incomplete collation of occurrence data on risk assessment (Question 2), we performed preliminary extinction risk assessments using records from each database separately (e.g., only data from GBIF) after removing suspect or erroneous records (Fig. 1). The impact of an incomplete records collation on the Red List assessment was summarized in four scenarios described in Box 1. No change: no change in species risk category; Change within binary categories:

changes in risk occur within threat (CR, EN, VU) and not threat (NT and LC) binary categories; Changes in binary and detailed categories: a threatened species becomes not threatened and vice-versa; Not evaluated vs. Evaluated: extinction risk of species cannot be evaluated because species are not present in such database.

2.6. Impact of taxonomic, spatial, and temporal issues on preliminary risk assessments

To address questions 3 and 4, we used each "partially cleaned" dataset separately to assess the effect of taxonomic, spatial, and temporal issues on preliminary extinction risk assessments (Fig. 1). To quantify the effects of taxonomic and nomenclatural errors, we used the "taxonomic" dataset to compare changes in the risk category due to non-standardized record names. Similarly, to quantify the impact of the spatial-related issues on risk assessment, we fixed all problems. Still, we kept records with one spatial issue (e.g., outliers) to quantify its relative impact on the risk category. This process was repeated for the nine spatial issues (Table S1). Finally, to assess the impact of early collections or legacy data on risk assessments, we removed all records collected before 1970 or containing illegitimate information. It is worth noting that changes in the risk category result from changes in the number of locality data available to each record due to spatial, taxonomic, or temporal issues not corrected or due to an incomplete data collation.

We also assessed the impact of taxonomic, spatial, and temporal issues on preliminary assessments using three additional reference datasets representing subsets of the CNCFlora data, including 1) species with up-to-date assessments carried out post-2018, 2) species with older assessments presented in the Red Book of the Brazilian Flora (Martinelli and Moraes, 2013), and 3) species with >15 occurrence records, a number suggested as the minimum for reliable automated assessments (Fig. S1; Rivers et al., 2011).

We used R (v. 4.02; R Core Team, 2020) to perform all analyses, the package *bdc* for data-cleaning (Ribeiro et al., 2022; https://cran.r-project.org/web/packages/bdc/index.html), and *ggplot2* (Wickham, 2016) to create all figures, except the Sankey diagram, which was generated using *networkD3* package (Allaire et al., 2017).

3. Results

3.1. Data cleaning

From >30 million records in the original databases, only ~3.9 million records (13 %) passed all data-cleaning tests. The number of records detected and removed in each module of the data-cleaning workflow can be found in Table S1. As expected, taxonomic, spatial, and temporal errors changed the number of occurrences and species available for carrying out preliminary assessments. Without harmonizing species names (the "taxonomic" dataset), 63,112 specimens were recognized. After the taxonomic harmonization, the number of species was reduced to 38,690 in the "spatial" dataset and 38,207 in the "temporal" dataset. The "clean" dataset contains data of 37,519 species (Table S2). We found a similar pattern of species reduction in species' availability for conducting preliminary assessment after cross-referencing species from our dataset and CNCFlora data (Table S2).

3.2. Similarity between databases

Overall, we found a higher similarity of data shared between large data aggregators than with regional and taxon-specific databases (Fig. 1). Regarding the proportion of species-occurrence records shared between "raw" databases, we found higher similarity levels (75–95 % of similarity) between BIEN, GBIF, iDigBio, and SpeciesLink databases after rounding coordinates to four decimals degrees (Figs. S1 and S2). Rounding coordinates to four decimals degrees and removing suspect or erroneous records (i.e., the similarity between "clean" databases) increased similarity levels (80–98 %; Figs. S3–S4). Most data from the DRYFLOR are available in NEOTROPTREE, but only after rounding coordinates, and species-occurrence data from both databases are not shared with other databases (Figs. S1 and S3). Considering the uniqueness of each database weighted by its number of records, databases containing fewer records (e.g., AT_EPIPHYTES, DRYFLOR, and NEOTROPTREE) had the highest proportion of unique records (Table S3).

3.3. Incomplete collation of occurrence data

We found a strong impact of an incomplete collation of occurrence data on the performance of preliminary risk assessments in correctly classifying species in the binary categories threatened and not threatened. From the 5524 species presented in the "clean" database (representing 94 % of species assessed by CNCFlora), BIEN, SiBBr and GBIF contained data for assessing the extinction risk of 69 % (n = 3806), 70 % (n = 3903) and 78 % (n = 4299) of species, respectively (Fig. 2, Table S4). The overall binary accuracy of preliminary assessments performed using data from only a single database ranged from 61 % (iDigBio) to 72 % (DRYFLOR). Interestingly, while the sensitivity rates were often low (13 to 48 %), specificity rates ranged from 84 % (ICMBio) to 94 % (DRYFLOR), highlighting that data from single databases can be sufficient to estimate species' range and accurately identify not threatened species (Table S4).

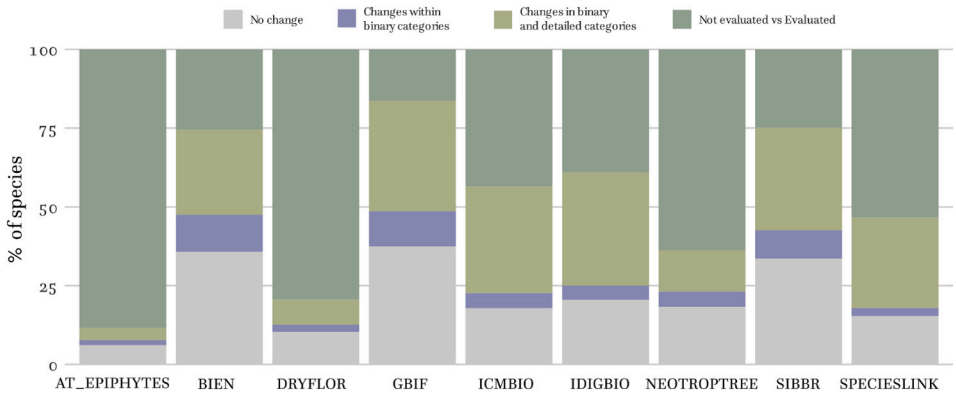


Fig. 2. Impact of an incomplete collation of species-occurrence records (i.e., the use of records from only a single database) on the performance of preliminary species extinction risk assessments. Changes in species category were evaluated by comparing the CNCFlora assessments (n = 5524 species) against the category derived from preliminary assessments using a single database after removing erroneous or suspect records. The impact of an incomplete records collation on the Red List assessment is summarized in four scenarios. *No change*: no change in species risk category; *Change within binary categories*: changes in risk occur within threat (CR, EN, VU) and not threat (NT and LC) binary categories; *Changes in binary and detailed categories*: a threatened species becomes not threatened and vice-versa; Not evaluated vs. Evaluated: extinction risk of species cannot be evaluated because species are not present in such database.

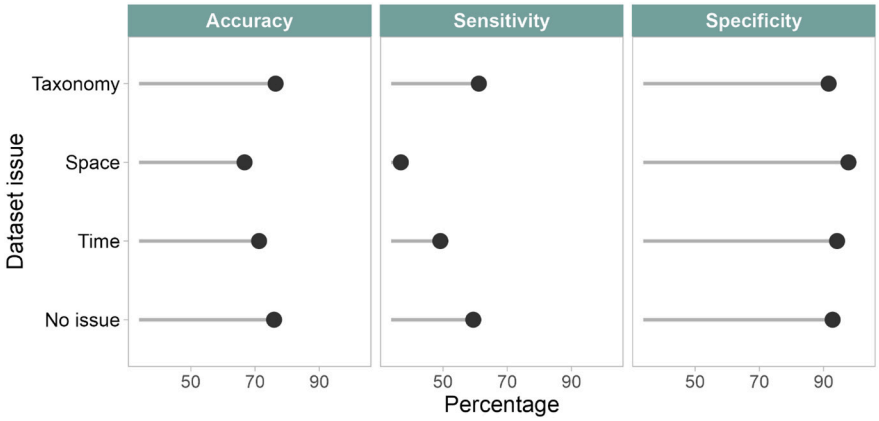


Fig. 3. The performance of preliminary extinction risk assessments carried out based on "partially cleaned" datasets (containing taxonomic, spatial, or temporal issues) and on a "clean" dataset in correctly classifying species as threatened or not threatened compared with CNCFlora assessments. The performance was measured in terms of overall accuracy, sensitivity (correct prediction of species as threatened), and specificity rates (correct prediction of species as not threatened).

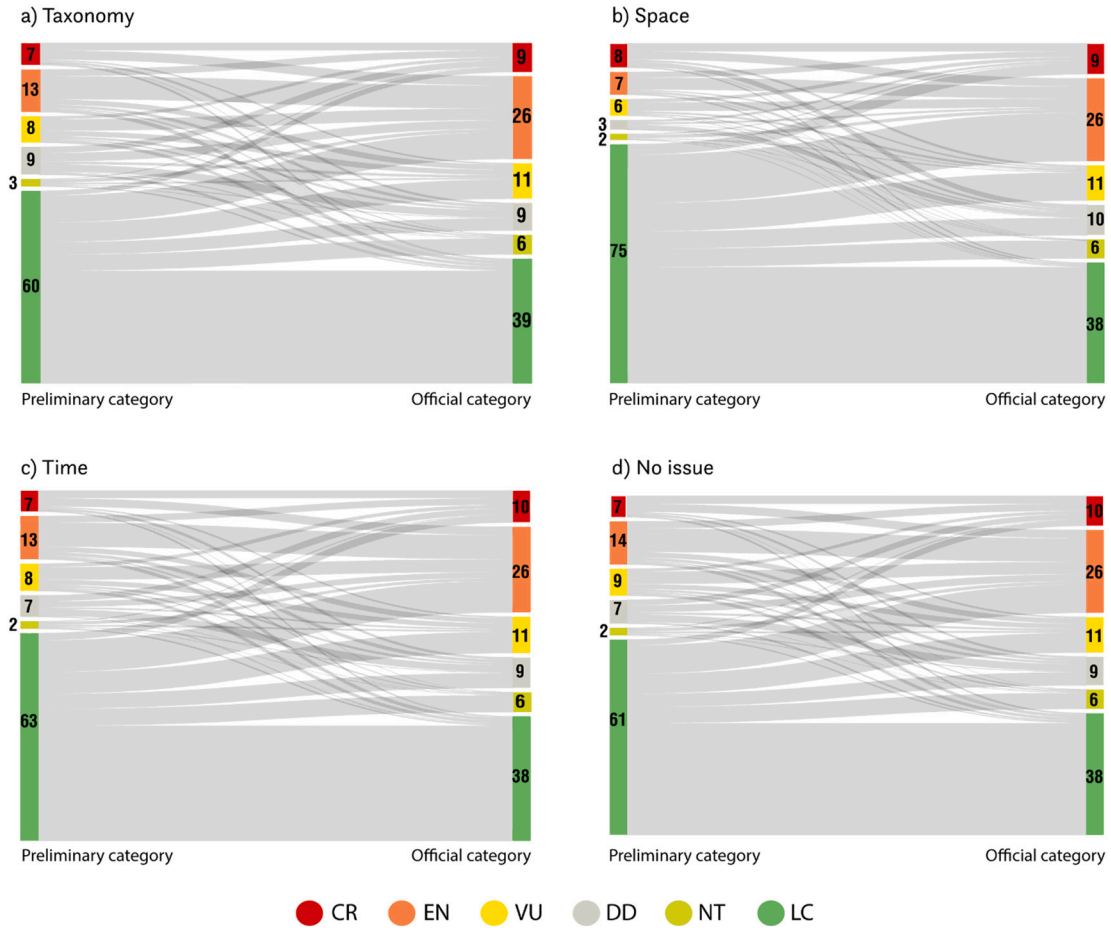


Fig. 4. Impact of taxonomic, spatial, and temporal issues on preliminary species extinction risk assessments. Changes in the preliminary assessment were evaluated by comparing CNCFlora category against the category derived from datasets with different data curation, including a) non-standardized taxonomy, b) spatial issues, c) temporal issues, d) no issue. Numbers within the bar represent the proportion of species in the category. Acronyms refer to IUCN Red List categories: CR, critically endangered; EN, endangered; VU, vulnerable; NT, near threatened; LC, least concern; DD, data deficient.

3.4. Taxonomic, spatial, and temporal issues

The accuracy of preliminary risk assessments calculated at binary level (i.e., correctly classifying species as threatened and not threatened) was marginally affected by unchecked taxonomy, errors in geo-referenced information, and the use of early collected or legacy data (Fig. 3). While specificity rates were often high, ranging from 89 % ("taxonomy" dataset) to 92 % ("spatial" dataset; Fig. 4), sensitivity rates were more affected by spatial errors (38 %) and temporal issues (48 %), and slightly affected by the presence of taxonomic issues (61.5 %) (Fig. 3).

Using records with non-standardized taxonomy slightly affected the

performance of preliminary risk assessment in correct classifying species in risk categories compared with database with no issue (Fig. 4a). In contrast, the presence of spatial issues and old records tended to overestimate the number of least concern species (Fig. 4c). Compared to species correctly classified as threatened and not threatened using the "clean" dataset, geographic outliers and records in urban areas were the spatial issues that most contributed to the misclassification of species category (Figs. 4b and S5).

As expected, the detailed accuracy (i.e., correctly placing species into six IUCN categories) was often low (51–53 %) and more affected by spatial issues (Fig. S6). Regarding the assessment carried out using different CNCFlora reference datasets, the binary accuracy (64–78 %) and specificity (78–99 %) were the highest by using as reference the dataset containing only species with >15 records with no issue (Fig. S7). Overall, the sensitivity rate was the lowest when the preliminary assessment was carried out based on a dataset containing both species with >15 records and spatial issues (Fig. S7). The sensitivity of preliminary risk assessment in predicting the category of species assessed after 2018 ("newer assessments") had the highest sensitivity rates (69 %) when based on datasets with no issue or only taxonomic issues (Fig. S7).

4. Discussion

We investigated the impact of incomplete data collation and taxonomic, spatial, and temporal issues in species' occurrence data on preliminary risk assessments. We found that the overall accuracy and specificity (correct prediction of non-threatened species) of preliminary risk assessments are less affected by incomplete data collation and issues in all three dimensions of species occurrence data. However, sensitivity rates (correct predicting threatened species) were often low to moderate and strongly affected by incomplete data collation and spatial issues, and marginally affected by temporal issues. Our results demonstrate that species' preliminary risk assessments accurately identify non-threatened species, even when data collection is low and in the presence of issues in species occurrence data. Such an approach can be used to efficiently prioritize likely threatened species for full Red List assessments, saving time and resources needed for such effort. In addition, caution is needed before declaring a species as threatened without considering data collation intensity and quality.

Data from natural history collections have errors that could lead to misclassification of extinction risk and undesirable conservation outputs (Maldonado et al., 2015; Panter et al., 2020). Our results showed that data cleaning is a fundamental process to improve the sensitivity of preliminary risk assessment and that specificity is marginally affected by both issues in occurrence records and incomplete collation of data (Panter et al., 2020). As the extent of occurrence (EOO) is purely a measure of range size, the presence of spatial issues (e.g., outliers and records in urban areas) often overestimates species range size, resulting in misclassification of many threatened species as not threatened. In contrast, the presence of spatial issues marginally improved specificity compared to specificity rates of preliminary assessment carried out using the "clean" dataset as reference. This may be due to the cleaning process that excluded records flagged as outliers or within urban areas. We stressed that our results and conclusion regard preliminary risk assessments based on IUCN's criterion B. Machine-learning methods used to predict Red List categories are commonly more robust to data availability and spatial issues (Zizka et al., 2021).

In the era of Big Data, many biodiversity data today are increasingly digitized and made available online through a wide range of heterogeneous databases (Kissling et al., 2018). By assessing the commonalities of databases regarding the number of species-occurrence they share, we found an overall high similarity between large data aggregators (e.g., GBIF, BIEN, and SiBBr) and an often-lower similarity between such databases and regional and taxon-specific databases. Our results showed that many records from small or local datasets are not shared with large data aggregators, highlighting the importance of such databases for estimating species' Red List category, especially for threatened species.

An incomplete collation of species records affects preliminary risk assessments' sensitivity more than specificity rates. On the one hand, our results showed that preliminary assessments carried out based on data from single databases had moderate accuracy (61–72 %) and high specificity (83–93 %). These results highlight that even a reduced number of occurrence records are likely sufficient to represent the range of not threatened species, offering a helpful first step before investing in further collation of data (Nic Lughadha et al., 2019b; Panter et al., 2020; Rivers et al., 2011). Nevertheless, assessments based on records from single databases can only be performed on a reduced set of species compared to the total species pool assessed when compiling data from several sources. On the other hand, we found that the overall sensitivity of species' preliminary risk was commonly low and strongly affected by an incomplete data collation and spatial issues and moderately affected by temporal issues. An insufficient collation of species-occurrence records is likely only to represent a fraction of species' range, underestimate thus EOO and resulting in lower sensitivity rates. In contrast, spatial issues and older or legacy data often result in larger estimates of EOO, thereby decreasing sensitivity rates.

Unchecked taxonomy and the removal of early collected or legacy records also led species to be classified as data deficient in a smaller proportion than an incomplete data collation and spatial issues. The negligible impact of unchecked taxonomy can indicate that synonyms and misspelled names had a small impact on the sensitivity of preliminary risk assessments or that databases had a relatively well-curated and updated taxonomy. Unchecked taxonomy, in most cases, did not decrease EOO to below the threshold to classify a threatened species as not threatened (but see Nic Lughadha et al., 2019b). Outdated taxonomy, however, can result in species that cannot be assessed due to taxonomic remodeling (lumping and splitting; Nic Lughadha et al., 2019b). Similarly, several species could not be assessed if older or legacy data were removed since these records constitute the best and single information to support the Red List assessment for some species (IUCN, 2019).

As shown by our results, drawbacks in taxonomic, spatial, and temporal dimensions of species-occurrence data can result in a range of possible estimates of EOO and conservation categories, particularly for rare species in which the exclusion of records could lead to extensive modification on the EOO estimates (Rivers et al., 2011; Zizka et al., 2020). When assessing the extinction risk of many species, it is challenging to determine whether a record assigned as suspicious is de facto erroneous. As a single category must be defined when assessing IUCN Red Lists status, it is highly recommended to adopt a precautionary but realistic approach by informing the risk category from the best available information (IUCN, 2019).

There is a clear need for a drastic increase in the production rate of species risk assessment, especially in tropical countries, where species diversity and threats to plants are greatest. The robustness of preliminary risk assessments based on criterion B, especially to identify non-threatened species highlights the importance of preliminary risk assessments to efficiently prioritize potentially threatened species for full Red List assessments (Bachman et al., 2020). As the sensitivity of preliminary risk assessments was low and more affected by incomplete data collation and spatial issues, continuous efforts to ensemble and generate high-quality primary biodiversity data should be a priority in parallel with the increasing use and improvement of methods to facilitate and expedite Red List assessments.

CRediT authorship contribution statement

BRR, SJEV, KGM, LJ, GT, SPB, and RL conceived the ideas and designed the methodology. BRR, SJEV, KGM, LJ, and GT collected the data. BRR analyzed the data. BRR led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for

publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

BRR and KGM are supported by a CAPES scholarship. RL research is funded by CNPq (grant #306694/2018-2). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This paper is a contribution of the INCT in Ecology, Evolution and Biodiversity Conservation founded by MCTIC/CNPq (grant #465610/2014-5) and FAPEG (grant #201810267000023). SJEV was supported by the National Science Foundation (Award 1853697) and CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas). GT was supported by PNPD/CAPES postdoc fellowship no. 20132984-52012018005P7. LJ was supported by CNPq's PDJ postdoctoral fellowship (nº 165615/2020-6).

Appendix A

The scripts used to carry out the analyses are available at https://doi.org/10.6084/m9.figshare.20402076.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biocon.2022.109674.

References

Allaire, J.J., Gandrud, C., Russell, K., Yetman, C.J., 2017. networkD3: D3 JavaScript Network Graphs From R.

Bachman, S., Walker, B., Barrios, S., Copeland, A., Moat, J., 2020. Rapid least concern: towards automating red list assessments. Biodivers. Data J. 8 https://doi.org/10.3897/BDJ.8.e47018.

Bachman, S.P., Field, R., Reader, T., Raimondo, D., Donaldson, J., Schatz, G.E., Lughadha, E.N., 2019. Progress, challenges and opportunities for red listing. Biol. Conserv. 234, 45–55. https://doi.org/10.1016/j.biocon.2019.03.002.

Ball-Damerow, J.E., Brenskelle, L., Barve, N., Soltis, P.S., Sierwald, P., Bieler, R., LaFrance, R., Ariño, A.H., Guralnick, R.P., 2019. Research applications of primary biodiversity databases in the digital age. PLoS One 14, 1–26. https://doi.org/10.1371/journal.pone.0215794.

BFG (The Brazil Flora Group), 2021. Flora do Brasil 2020. Jardim Botânico do Rio de Janeiro, Rio de Janeiro.

Boitani, L., Maiorano, L., Baisero, D., Falcucci, A., Visconti, P., Rondinini, C., 2011. What spatial data do we need to develop global mammal conservation strategies? Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 366, 2623–2632. https://doi.org/10.1098/rstb.2011.0117.

Brummitt, N., Bachman, S.P., Moat, J., 2008. Applications of the IUCN red list: towards a global barometer for plant diversity. Endanger. Species Res. 6, 127–135. https://doi.org/10.3354/esr00135.

Brummitt, N.A., Bachman, S.P., Griffiths-Lee, J., Lutz, M., Moat, J.F., Farjon, A., Donaldson, J.S., Hilton-Taylor, C., Meagher, T.R., Albuquerque, S., Aletrari, E., Andrews, A.K., Atchison, G., Baloch, E., Barlozzini, B., Brunazzi, A., Carretero, J., Celesti, M., Chadburn, H., Cianfoni, E., Cockel, C., Coldwell, V., Concetti, B., Contu, S., Crook, V., Dyson, P., Gardiner, L., Ghanim, N., Greene, H., Groom, A., Harker, R., Hopkins, D., Khela, S., Lakeman-Fraser, P., Lindon, H., Lockwood, H., Loftus, C., Lombrici, D., Lopez-Poveda, L., Lyon, J., Malcolm-Tompkins, P., McGregor, K., Moreno, L., Murray, L., Nazar, K., Power, E., Tuijtelaars, M.Q., Salter, R., Segrott, R., Thacker, H., Thomas, L.J., Tingvoll, S., Watkinson, G., Wojtaszekova, K., Lughadha, E.M.N., 2015. Green plants in the red: a baseline global assessment for the IUCN sampled red list index for plants. PLoS One 10, 1–22. https://doi.org/10.1371/journal.pone.0135152.

Canhos, D.A.L., Sousa-Baena, M.S., de Souza, S., Maia, L.C., Stehmann, J.R., Canhos, V. P., De Giovanni, R., Bonacelli, M.B.M., Los, W., Peterson, A.T., 2015. The importance of biodiversity E-infrastructures for megadiverse countries. PLoS Biol. 13, e1002204 https://doi.org/10.1371/journal.pbio.1002204.

CNCFlora, 2021. Centro Nacional de Conservação da Flora. http://cncflora.jbrj.gov.br/portal.

Graham, C., Ferrier, S., Huettman, F., Moritz, C., Peterson, A., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19, 497–503. https://doi.org/10.1016/j.tree.2004.07.006.

IUCN, 2019. Guidelines for using the IUCN red list categories and criteria. Version 14. Prepared by the Standards and Petitions Committee. http://www.iucnredlist.org/documents/RedListGuidelines.pdf.

IUCN, 2012. IUCN Red List Categories and Criteria: Version 3.1, Second edi. Gland, Switzerland and Cambridge, UK.

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick, R.P., Isaac, N.J.B., Kelling, S., Los, W., McRae, L., Mihoub, J.B., Obst, M., Santamaria, M., Skidmore, A.K., Williams, K.J., Agosti, D., Amariles, D., Arvanitidis, C., Bastin, L., De Leo, F., Egloff, W., Elith, J., Hobern, D., Martin, D., Pereira, H.M., Pesole, G., Peterseil, J., Saarenmaa, H., Schigel, D., Schmeller, D.S., Segata, N., Turak, E., Uhlir, P.F., Wee, B., Hardisty, A.R., 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biol. Rev. 93, 600–625. https://doi.org/10.1111/brv.12359.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J., Chilquillo, E., Rønsted, N., Antonelli, A., 2015. Estimating species diversity and distribution in the era of big data: to what extent can we trust public databases? Glob. Ecol. Biogeogr. 24, 973–984. https://doi.org/10.1111/geb.12326.

Martinelli, G., Moraes, M.Á., 2013. Livro Vermelho da Flora do Brasil, 1 st. ed. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rio de Janeiro.

Meyer, C., Weigelt, P., Kreft, H., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19, 992–1006. https://doi.org/10.1111/ele.12624.

Miller, J.S., Porter-Morgan, H.A., Stevens, H., Boom, B., Krupnick, G.A., Acevedo-Rodríguez, P., Fleming, J., Gensler, M., 2012. Addressing target two of the global strategy for plant conservation by rapidly identifying plants at risk. Biodivers. Conserv. 21, 1877–1887. https://doi.org/10.1007/s10531-012-0285-3.

Moat, J., 2017. rCAT: Conservation Assessment Tools Version 0.1.6.

Nic Lughadha, E.M., Graziele Staggemeier, V., Vasconcelos, T.N.C., Walker, B.E., Canteiro, C., Lucas, E.J., 2019a. Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. Conserv. Biol. 33, 511–522. https://doi.org/10.1111/cobi.13289.

Nic Lughadha, E.M., Walker, B.E., Canteiro, C., Chadburn, H., Davis, A.P., Hargreaves, S., Lucas, E.J., Schuiteman, A., Williams, E., Bachman, S.P., Baines, D., Barker, A., Budden, A.P., Carretero, J., Clarkson, J.J., Roberts, A., Rivers, M.C., 2019b. The use and misuse of herbarium specimens in evaluating plant extinction risks. Philos. Trans. R. Soc. B Biol. Sci. 374, 20170402 https://doi.org/10.1098/rstb.2017.0402.

Panter, C.T., Clegg, R.L., Moat, J., Bachman, S.P., Klitgård, B.B., White, R.L., 2020. To clean or not to clean: cleaning open-source data improves extinction risk assessments for threatened plant species. Conserv. Sci. Pract. 2 https://doi.org/10.1111/csp2.311.

R Core Team, 2020. R: A Language and Environment for Statistical Computing.

Ramos, F.N., Mortara, S.R., Monalisa-Francisco, N., Elias, J.P.C., Neto, L.M., Freitas, L., Kersten, R., Amorim, A.M., Matos, F.B., Nunes-Freitas, A.F., Alcantara, S., Alexandre, M.H.N., Almeida-Scabbia, R.J., Almeida, O.J.G., Alves, F.E., Oliveira Alves, R.M., Alvim, F.S., Andrade, A.C.S., Andrade, S., Aona, L.Y.S., Araujo, A.C., Araújo, K.C.T., Ariati, V., Assis, J.C., Azevedo, C.O., Barbosa, B.F., Barbosa, D.E.F., Barbosa, F.dos R., Barros, F., Basilio, G.A., Bataghin, F.A., Bered, F., Bianchi, J.S., Blum, C.T., Boelter, C.R., Bonnet, A., Brancalion, P.H.S., Breier, T.B., Buzatto, C.R., Cabral, A., Cadorin, T.J., Caglioni, E., Canêz, L., Cardoso, P.H., Carvalho, F.S., Carvalho, R.G., Catharino, E.L.M., Ceballos, S.J., Cerezini, M.T., César, R.G., Cestari, C., Chaves, C.J.N., Citadini-Zanette, V., Coelho, L.F.M., Coffani-Nunes, J.V., Colares, R., Colletta, G.D., Brion, C.de T., Corrêa, N.de M., Costa, A.F., Costa, G.M., Costa, L.M.S., Costa, N.G.S., Couto, D.R., Cristofolini, C., Cruz, A.C.R., Del Neri, L.A., Pasquo, M., Santos Dias, A., Dias, L.do C.D., Dislich, R., Duarte, M.C., Fabricante, J. R., Farache, F.H.A., Faria, A.P.G., Faxina, C., Ferreira, M.T.M., Fischer, E., Fonseca, C.R., Fontoura, T., Francisco, T.M., Furtado, S.G., Galetti, M., Garbin, M.L., Gasper, A.L., Goetze, M., Gomes-da-Silva, J., Gonçalves, M.F.A., Gonzaga, D.R., Silva, A.C.G.E., Guaraldo, A.de C., Guarino, E.de S.G., Hudson, L.B., Jardim, J.G., Jungbluth, P., Guislon, A.V., Kaeser, S.dos S., Kessous, I.M., Koch, N.M., Kuniyoshi, Y.S., Labiak, P.H., Lapate, M.E., Santos, A.C.L., Leal, R.L.B., Leite, F.S., Leitman, P., Liboni, A.P., Liebsch, D., Lingner, D.V., Lombardi, J.A., Lucas, E., Mai, P., Mania, L.F., Mantovani, W., Maragni, A.G., Marques, M.C.M., Marquez, G., Martins, C., Luzzi, J.dos R., Martins, L.do N., Martins, P.L.S.S., Mazziero, F.F.F., Melo, C.de A., Melo, M.M.F., Mendes, A.F., Mesacasa, L., Morellato, L.P.C., Moreno, V.de S., Muller, A., Murakami, M.M.da S., Cecconello, E., Nardy, C., Nervo, M.H., Neves, B., Nogueira, M.G.C., Nonato, F.R., Oliveira-Filho, A.T., Oliveira, C.P.L., Overbeck, G.E., Marcusso, G.M., Paciencia, M.L.B., Padilha, P., Padilha, P.T., Pereira, A.C.A., Pereira, L.C., Pereira, R.A.S., Pincheira-Ulbrich, J., Pires, J.S.R., Pizo, M.A., Pôrto, K.C., Rattis, L., Reis, J.R.de M., Reis, S.G.dos, Rocha-Pessôa, T.C., Rocha, C.F.D., Rocha, F.S., Rodrigues, A.R.P., Rodrigues, R.R., Rogalski, J.M., Rosanelli, R.L., Rossado, A., Rossatto, D.R., Rother, D.C., Ruiz-Miranda, C.R., Saiter, F.Z., Sampaio, M.B., Santana, L.D., Santos, J.S.dos, Sartorello, R., Sazima, M., Schmitt, J.L., Schneider, G., Schroeder, B.G., Sevegnani, L., Júnior, V.O.S., Silva, F.R., Silva, M.J., Silva, M.P.P., Silva, R.G., Silva, S.M., Singer, R.B., Siqueira, G., Soares, L.E., Sousa, H.C., Spielmann, A., Tonetti, V.R., Toniato, M.T.Z., Ulguim, P.S.B., Berg, C., Berg, E., Varassin, I.G., Silva, I.B.V., Vibrans, A.C., Waechter, J.L., Weissenberg, E.W., Windisch, P.G., Wolowski, M., Yañez, A., Yoshikawa, V.N., Zandoná, L.R., Zanella, C.M., Zanin, E.M., Zappi, D.C., Zipparro, V.B., Zorzanelli, J.P.F., Ribeiro, M.C., 2019. ATLANTIC EPIPHYTES: a data set of vascular and non-vascular epiphyte plants and lichens from the Atlantic Forest. Ecology 100, e02541. https://doi.org/10.1002/ecy.2541.

Ribeiro, B.R., Velazco, S.J.E., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P., Loyola, R., 2022. Bdc : a toolkit for standardizing, integrating, and cleaning biodiversity data. Methods Ecol. Evol. 2022, 1–8. https://doi.org/10.1111/2041-210x.13868.

Rivers, M.C., Taylor, L., Brummitt, N.A., Meagher, T.R., Roberts, D.L., Lughadha, E.N., 2011. How many herbarium specimens are needed to detect threatened species? Biol. Conserv. 144, 2541–2547. https://doi.org/10.1016/j.biocon.2011.07.014.

Rodrigues, A.S.L., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M., Brooks, T.M., 2006. The value of the IUCN red list for conservation. Trends Ecol. Evol. 21, 71–76. https://doi.org/10.1016/j.tree.2005.10.010.

Rondinini, C., Di Marco, M., Visconti, P., Butchart, S.H.M., Boitani, L., 2014. Update or outdate: long-term viability of the IUCN red list. Conserv. Lett. 7, 126–130. https://doi.org/10.1111/conl.12040.

Tessarolo, G., Ladle, R., Rangel, T., Hortal, J., 2017. Temporal degradation of data limits biodiversity research. Ecol. Evol. 7, 6863–6870. https://doi.org/10.1002/ece3.3259.

Walker, B.E., Leão, T.C.C., Bachman, S.P., Lucas, E., Nic, E., 2021. Evidence-based Guidelines for Developing Automated Assessment Methods Supplementary Materials.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., Gondim Lambert Moreira, F., Santos, N.M.C., Santos, T.A.B., dos Santos-Costa, R.C., Serrano, F.C., Alves da Silva, A.P., de Souza Soares, A., Cavalcante de Souza, P.G., Calisto Tomaz, E., Vale, V.F., Vieira, T.L., Antonelli, A., 2020. No one-size-fits-all solution to clean GBIF. PeerJ 8, e9916. https://doi.org/10.7717/peerj.9916.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., Antonelli, A., 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 744–751. https://doi.org/10.1111/2041-210X.13152.

Zizka, A., Silvestro, D., Vitt, P., Knight, T.M., 2021. Automated conservation assessment of the orchid family with deep learning. Conserv. Biol. 35, 897–908. https://doi.org/10.1111/cobi.13616.