Received: 14 February 2020 | Revised: 1 October 2020 | Accepted: 25 October 2020

DOI: 10.1111/csp2.311

CONTRIBUTED PAPER

Conservation Science and Practice
A journal of the Society for Conservation Biology
Open Access

WILEY

To clean or not to clean: Cleaning open-source data improves extinction risk assessments for threatened plant species

Connor T. Panter[1,2] | **Rosemary L. Clegg**[2] | **Justin Moat**[2] | **Steven P. Bachman**[2] | **Bente B. Klitgård**[2] | **Rachel L. White**[1]

[1]Ecology, Conservation and Zoonosis Research and Enterprise, School of Pharmacy and Biomolecular Sciences, University of Brighton, Brighton, East Sussex, United Kingdom

[2]Royal Botanic Gardens, Kew, Richmond, Surrey, United Kingdom

Correspondence
Connor T. Panter, Ecology, Conservation and Zoonosis Research and Enterprise, School of Pharmacy and Biomolecular Sciences, University of Brighton, Brighton, East Sussex BN2 4GJ, United Kingdom.
Email: connorpanter1301@gmail.com

Abstract

Plants are under-represented in conservation efforts, with only 9% of described species published on the IUCN Red List. Biodiversity aggregators including the Global Biodiversity Information Facility (GBIF) and the more recent Botanical Information and Ecology Network (BIEN) contain a wealth of potentially useful occurrence data. We investigate the influence of these data in accelerating plant extinction risk assessments for 225 endemic, near-endemic, and socioeconomic Bolivian plant species. Geo-referenced herbarium voucher specimens verified by taxonomic experts comprised our control data set. Open-source data for 77 species was subjected to a two-stage cleaning protocol (using an automated R package followed by a manual clean) and threat categories were computed based on extent of occurrence thresholds. Accuracy was the highest using cleaned GBIF data (76%) and uncleaned BIEN data (79%). Sensitivity was the highest for cleaned GBIF (73%) and BIEN (80%) data suggesting our cleaning protocol was essential to maximize sensitivity rates. Comparisons between the control, GBIF and BIEN data sets revealed a paucity of occurrence data for 148 species (66%), 72% of which qualified for a threatened category. Balancing data quantity and accuracy must be considered when using open-source data. Filling data gaps for threatened species is a conservation priority to improve the coverage of threatened species within biodiversity aggregators.

KEYWORDS

BIEN, Bolivia, extent of occurrence, extinction risk, GBIF, IUCN red list, open-source data

1 | INTRODUCTION

Despite their important role as the foundation of ecosystems, plants are often under-represented in conservation research (Di Marco et al., 2017). Plant-based conservation initiatives are fewer and receive substantially less funding compared with those focusing on animals (Balding & Williams, 2016; Margulies et al., 2019; Westwood, Cavender, Meyer, & Smith, 2020). However, plants are included in global conservation targets such as the Convention on Biological Diversity's (CBD) Aichi Targets for 2011–2020 and conservation strategies specifically targeted

at plants exist. For example, Target 2 of the CBD's Global Strategy for Plant Conservation (GSPC) aims to achieve "an assessment of the conservation status of all known plant species, as far as possible to guide conservation action" by 2020 (CBD, 2019; https://www.cbd.int/gspc/targets.shtml).

Target 2 also serves as an underpinning target that must be delivered in order to make quantitative responses to other GSPC targets referring to extinction risk, for example, Target 7 "threatened species conserved in situ" and Target 8 "threatened species in ex situ collections". A recent review of progress against Target 2 indicates that although there was growth in coverage of assessments in recent years (+37.2% from 2016 to 2020) for vascular plants with 313,565 assessments representing 28.3% of species (Nic Lughadha et al., 2020), the Target is unlikely to be met by the end of 2020 (Bachman et al. 2017). There is an urgent need to assess the remaining unassessed species as without these, there is a large knowledge gap that limits our ability to make effective evidence-based conservation decisions. A major contribution to GSPC Target 2 can be made by increasing the number of assessments produced using the categories and criteria of the IUCN Red List of Threatened Species (IUCN, 2019; https://www.iucnredlist.org/) (hereafter "Red List").

The Red List is arguably the most robust tool to measure a species' extinction risk at a global scale (Rodrigues, Pilgrim, Lamoreux, Hoffmann, & Brooks, 2006). Finer resolution data may be more accurate and appropriate for use at more local scales (Hermoso & Kennard, 2012) and where conservation policies have differing classifications or criteria than the IUCN (Brito et al., 2010). It has six categories that are used to classify extant species including: critically endangered (CR), endangered (EN) or vulnerable (VU) (collectively defined as "threatened"), near threatened (NT), least concern (LC) (collectively defined as "non-threatened"), or where insufficient information is available to determine a category, data deficient (DD) (IUCN, 2012). Within the IUCN's framework Criterion B, which relates to the geographic range of a species, is the most widely applied criterion for published extinction risk assessments (Collen et al., 2016). This is particularly true for plant assessments which are often calculated using data from natural history collections (see Rivers, Bachman, Meagher, Nic Lughadha, & Brummitt, 2010; Willis, Moat, & Paton, 2003).

Natural history collections, such as herbarium voucher specimens, have been increasingly used to answer scientific questions regarding the diversity and distribution of the world's flora (Nic Lughadha et al., 2018; Roberts, Taylor, & Joppa, 2016). Of particular use is data on the location of specimen collections, either from coordinates written on specimen labels or *post-facto* geo-referencing (Bachman, Moat, Hill, de la Torre, & Scott, 2011). Once verified, these data have been shown to be suitable for estimating extinction risk using IUCN's Criterion B and metrics such as extent of occurrence (EOO) (Willis et al., 2003). The EOO is defined as "the area contained within the shortest continuous imaginary boundary which can be drawn to encompass all the known, inferred or projected sites of present occurrence of a taxon, excluding cases of vagrancy" (IUCN, 2019)—and is calculated using a convex hull (Joppa et al., 2016). The EOO metric can act as both an element of a full Red List assessment (IUCN, 2012) but also when compared against the Criterion B thresholds, can be used to determine a preliminary assessment that contributes towards GSPC Target 2. A continuing debate among conservationists evolves around whether there should be more focus on data quantity or quality when assessing species' extinction risk. Some argue that the availability, quantity, and quality of herbarium specimen data are equally of vital importance for progressing the automation of assessments (Nic Lughadha et al., 2018; Rivers et al., 2011). Others highlight the requirement of accurate measurements of changes in population size, enabling robust estimates of population trends to inform extinction risk assessors (Mace et al., 2008; Taylor, 1995). Conversely, others conclude that in a large majority of cases, high quantity, low quality data produce similar results to that of lower quantity, high quality data (Aceves-Bueno et al., 2017). Poor quality data have also been used to cost-effectively estimate risk levels for data deficient species using a double sampling methodology (see Bland et al., 2015). However, to date, the general consensus within the conservation community favors the use of high quality data (Cayuela et al., 2009; Wood, Sullivan, Lliff, Fink, & Kelling, 2011) especially for extinction risk estimates (Keith, 2001; Nic Lughadha et al., 2018; Rivers et al., 2011). As a result, many botanic gardens have prioritized the accessibility of herbarium material (Westwood et al., 2020; Willis et al., 2017) and have now begun to digitize and geo-reference their collections (Tulig, Tarnowsky, Bevans, Kirchgessner, & Theirs, 2012) which are then often publicly available through online aggregators such as the Global Biodiversity Information Facility (GBIF) (http://www.gbif.org). GBIF is the world's largest open-source biological database (Chandler et al., 2017), collating species occurrence data from collections held in over 800 biodiversity institutions and to date, comprises more than 260 million plant occurrence records. GBIF acknowledges that with broad data accumulation data quality can be impacted. In response GBIF flags potential errors through its data validator (https://www.gbif.org/tools/data-validator). This is an automated process providing information on geographic

discrepancies (e.g., whether coordinates fall within a stated country), meta-data structure and formatting issues, and suggests how the user may improve the quality of the data and associated meta-data before use. However, the GBIF Data Validator is currently in test phase and it is likely that many users are not yet aware of its existence. Furthermore, some flags may not be easy to understand and previous studies have highlighted concerns with the uncritical use of uncleaned GBIF data for conservation purposes (García-Roselló et al., 2015; Maldonado et al., 2015). A recent study estimated that anywhere between 29 and 90% of GBIF records for 18 Neotropical species were potentially erroneous (see Zizka et al., 2020).

More recently, other open-source aggregators have been developed such as the Botanical Information and Ecology Network (BIEN) (Enquist, Condit, Peet, Schildhauer, & Thiers, 2016; http://bien.nceas.ucsb.edu/bien/). The most recent release of the BIEN database provides over 206 million data observations that have undergone a standardized form of data cleaning including the validation of (a) species' names using the Taxonomic Names Resolution Service (Boyle et al., 2013; http://tnrs.iplantcollaborative.org/), (b) geographic coordinates via a geo-referencing workflow, and (c) suggested changes via feedback loop to original source (Enquist et al., 2016). Both data sets provide a potential input source for rapidly generating a large number of extinction risk assessments contributing towards GSPC Target 2. However, a consequence of aggregating data in this way is the potential inclusion of unclean data that can influence subsequent analysis, with measures such as EOO shown to be sensitive to outliers (Burgman & Fox, 2003). Both BIEN and GBIF data have already featured in extinction risk research (Doughty et al., 2016; Robiansyah & Wardani, 2020) but differences between data sources for the estimation of metrics such as EOO and resulting threat categories have yet to be examined.

Here we test for the first time the effects of two levels of data cleaning on computed preliminary threat categories of plant species using real-world data. We compare our results with a control data set verified by taxonomic experts. We calculate threat categories using Criterion B EOO thresholds, as it is the most commonly used metric in plant extinction risk assessments.

As well as using uncleaned data, we adopt a two-stage cleaning protocol using (a) a fast automated cleaning package within R (hereafter "automated clean") and (b) a more time-consuming manual clean (hereafter "manual clean") to assess the accuracy of computed threat categories against our expert verified control data set. Using a confusion matrix, we compare accuracy, sensitivity (correctly predicting threatened categories) and specificity (correctly predicting non-threatened categories) rates between our data sets (control vs. GBIF vs. BIEN) at varying data cleaning stages (uncleaned vs. automated vs. manual). Based on the georeferencing workflow implemented within the BIEN repository (Enquist et al., 2016), we expect overall accuracy to be lower for GBIF data than BIEN. Moreover, based on the findings of Maldonado et al. (2015), we expect uncleaned data to be less accurate than either of the cleaned data sets primarily due to spatial errors. As we apply our two-stage cleaning protocol we predict accuracy, sensitivity, and specificity rates to increase with the level of data cleaning following findings by García-Roselló et al. (2015). Our findings will be of considerable use when applied to the decision-making and data cleaning processes of future extinction risk studies that utilize open-source biological data from online data repositories.

2 | METHODS

We applied a two-stage cleaning protocol using a selected group of plants from our study area (Figure 1). An overview of our methodology provides a step-by-step guide on the processing of our control and open-source data (Figure 2).

2.1 | Control data

Our control data were derived from the Tropical Important Plant Areas (TIPAs) in Bolivia project (https://www.kew.org/science/our-science/projects/tropical-important-plant-areas-bolivia) led by the Royal Botanic Gardens, Kew. A species list was compiled for the Chiquitania ecoregion occurring in the eastern lowlands of Santa Cruz, Bolivia, (Figure 1). The ecoregion consists of a mosaic of Chiquitano dry forest, Cerrado grassland and Pantanal wetland vegetation. Some species' distributions extend into neighbouring Brazil (Figure 1). Species were selected from Bolivia's first plant census, "Catálogo de las Plantas Vasculares de Bolivia" (Jørgensen, Nee, & Beck, 2014) and "NeoTropTree" (NTT) (Oliveira-Filho,-2017), a list of Neotropical woody taxa. The species list was then standardized by taxonomic experts from the Royal Botanic Gardens, Kew (Kew) and Museo de Historia Natural Noel Kempff Mercado (USZ) in Bolivia. From the species list we selected a total of 225 endemic, near-endemic or socio-economic plant species and their synonyms (totaling 339 species names). The associated herbarium specimen data for these species and their synonyms were collated from eight herbaria (see Appendix S1). These specimens are of particular scientific interest as they have not been digitized prior to the TIPAs

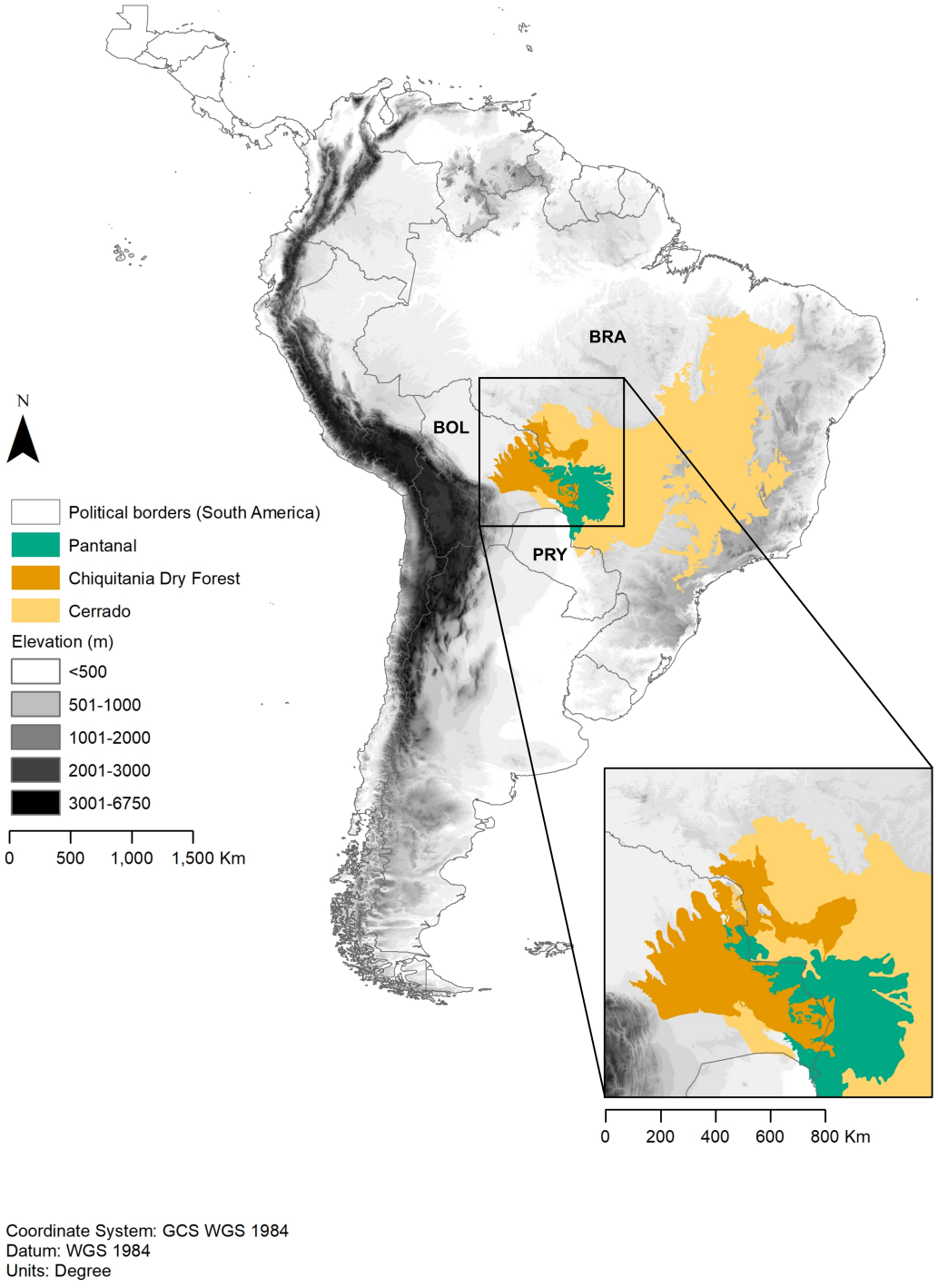

FIGURE 1 Map of the study area including the Chiquitania Dry Forest, Pantanal and Cerrado habitat extending across Bolivia and Brazil (BOL = Bolivia, BRA = Brazil and PRY = Paraguay). Map created using ArcMap v10.7.1 (ERSI 2011)

project and thus not used in any other large-scale study and our control records for these remained unavailable from open-source aggregators such as GBIF and BIEN. Specimen identification in the control data set was verified by taxonomic specialists at Kew and USZ prior to inclusion in the study. Where coordinate data was missing, we retrospectively geo-referenced using the label data on each specimen, adopting a point-radius method (see Wieczorek, Guo, & Hijmans, 2004) with a minimum diameter of 2 km in Google Earth Pro version 7.3.2.5 (Figure 2). Otherwise, coordinates were validated from the "Locality notes," "Habitat description," and "Altitude" information present on specimen labels, where available. Species that had less than three occurrence

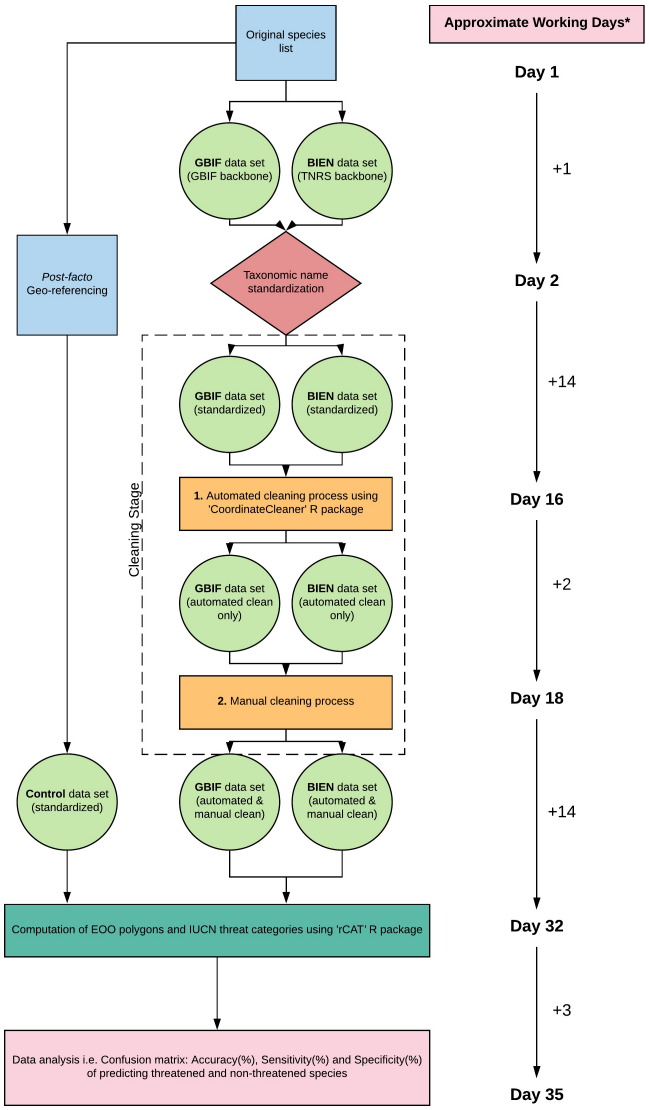

FIGURE 2 Flow diagram depicting the methods used within data collection, cleaning and preparation prior to analyses. *Method timeline accounts only for processes involving the open-source data sets

records were subsequently removed from further analysis ($N = 41$) as this is the minimum number of points required to compute a convex hull when measuring EOO. Such species may be range restricted, or wider ranging and under-sampled—either way, they require further detailed assessment.

2.2 | GBIF and BIEN data

We used the same species list to query the GBIF and BIEN databases for occurrence records (Figure 2). Using the R software v.3.5.1. (R Core Team, 2018), GBIF data were downloaded on November 27, 2018, with the package "rgbif" (Chamberlain & Boettiger, 2017). We used the public version of the BIEN 3.4.5 database and downloaded data on December 10, 2018, using the R package "BIEN" (Maitner et al., 2017). All GBIF meta-data were downloaded. Records with nonnumerical values in the coordinate columns (e.g., "NA") were subsequently removed from the data set before analysis (Figure 3a,b) (hereafter referred to as the "basic coordinate check").

2.3 | Taxonomic standardization

We standardized the species names of the control data set according to the accepted names, and noted their synonyms, gleaned from the NeoTropTree (Oliveira-Filho, 2017; http://www.neotroptree.info) and The Plant List databases (http://www.theplantlist.org) (Figure 2). Species names from GBIF and BIEN were matched using exact genus and species names from those in the control data set. Unmatched species names (including varieties and subspecies) were recorded but removed prior to analysis (Figure 3a,b). For example, *Paspalum plicatulum* var. *supectinatum* Döll (a variety from Bolivia) was included as part of the original species list and any subsequent unmatched species names, for example, *Paspalum plicatulum* Michx. (a globally widespread species), were removed (Figure 3).

3 | CLEANING GBIF AND BIEN DATA

3.1 | Stage 1: Automated clean

For the purpose of time efficiency we applied an automated clean to the GBIF and BIEN data sets using the "CoordinateCleaner" R package (Zizka et al., 2019) as it was the most up-to-date and comprehensive package available (Figure 2). Nine default tests were applied to the GBIF and BIEN data to flag erroneous records (Table 1). All automated tests can be customized by the user based on individual requirements, and we used default options to standardize the methodology between data sets. Flagged records were subsequently removed and a manual clean was applied to the remaining occurrence records (Figure 2).

3.2 | Stage 2: Manual clean

A visual assessment was performed by plotting the occurrence records in ArcMap v10.7.1 (ESRI, 2011). This was

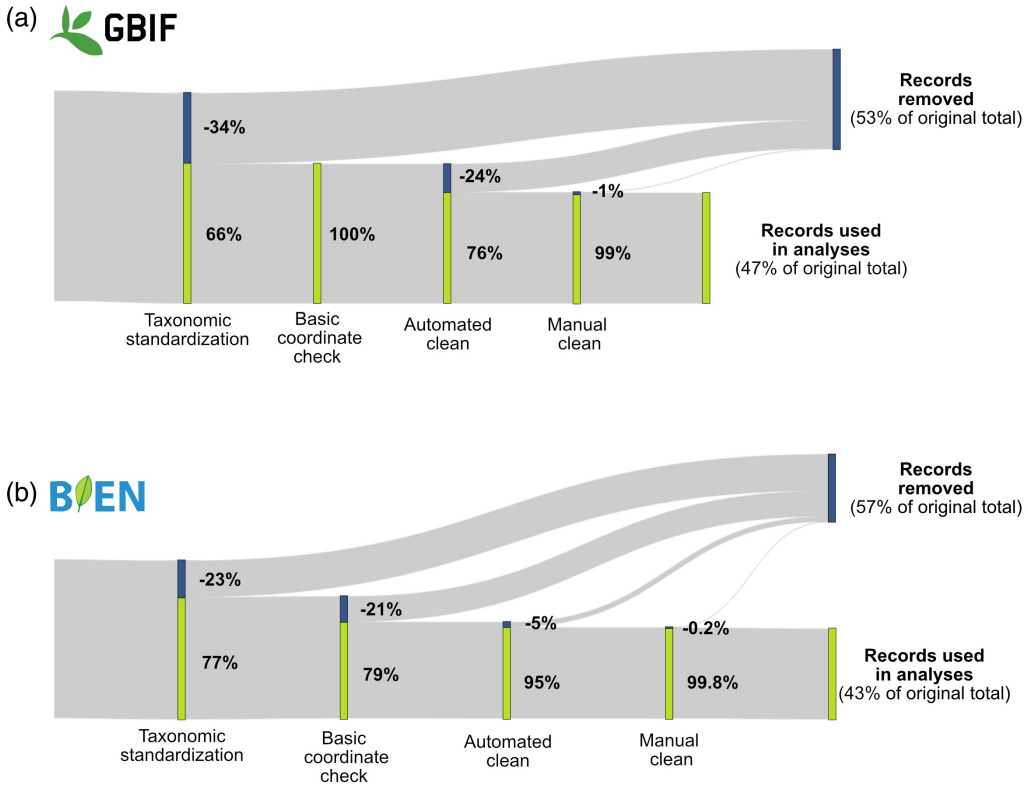

FIGURE 3 Percentages of occurrence record loss and records used during each cleaning stage for (a) GBIF and (b) BIEN. Percentages are calculated based on decrement number of records taken forward to each cleaning stage. Records taken through to each step GBIF versus BIEN: initial occurrence records (9,593 vs. 11,923), taxonomic standardization (6,366 vs. 9,181), basic coordinate check (6,366 vs. 7,246), automated clean (5,073 vs. 6,851), and manual clean (5,046 vs. 6,823)

done to check for erroneous records that were missed by the automated clean. Erroneous records were investigated further using species protologues, peer-reviewed articles and published geographic floras. Literature searches were also used to verify species with disjunct distributions and to trap potential identification errors. Records suspected to be derived from identification errors were removed from the resulting data set.

3.3 | Computation of EOO and threat categories

Before computing EOOs, all remaining occurrence records were projected using an equal area cylindrical projection, using the WGS84 datum, around a central point (Latitude/Longitude −12.179, −59.074) located in the center of South America. The R package "rCAT" (Moat, 2017) was then used to compute EOOs based on a convex hull for each species from (a) the control, GBIF, and BIEN data sets and (b) for each cleaning stage (uncleaned, automated and manual). "Uncleaned" data refers to records that remain uncleaned by either the automated or manual cleaning stages. Following the computation of EOOs, species were then cross-referenced and only those with sufficient quantities of data for both GBIF and BIEN were used in our analyses.

3.4 | The confusion matrix

Similar to Nic Lughadha et al. (2018), we adopted a confusion matrix to examine differences and similarities between threat categories comparing (a) accuracy rates (overall, how often was the classifier correct when predicting threat categories), (b) sensitivity rates (rate of correctly predicting threatened categories [VU, EN and CR]) and (c) specificity rates (rate of correctly predicting nonthreatened categories (LC, NT, and DD).

4 | RESULTS

4.1 | Number of species and occurrence records

Occurrence data for 225 species with more than three occurrence records formed the control data set. A total of

339 species names (210 accepted names and their 129 synonyms) were queried against the GBIF and BIEN repositories which returned 9,593 records for 198 species and 11,923 records for 166 species, respectively (Table 2; Figure 3a,b).

4.2 | Spatial issues detected during the automated clean

In total, 87% of flagged BIEN records (404 records for 150 species) failed the automated test for geographic outliers (see Appendix S2). The following tests: "Centroids" ($N = 1{,}188$), "Equal coordinates" ($N = 1{,}125$), "Sea" ($N = 1{,}144$) and "Zero" coordinates ($N = 1{,}124$) were the highest test failures for GBIF records (Appendix S2).

4.3 | Record removal and threatened species

For GBIF and BIEN, records for 148 species were excluded from our analyses (31 species had less than three occurrence records post-data cleaning) and totaled 66% of species within the control data set. Cross-referencing of species between GBIF and BIEN data sets (where data was only available for one of these data sets but not the other) resulted in the removal of 13 species from our analyses. Of all excluded species 72% qualified for a threatened category. Of the remaining 77 species, 29% qualified for a threatened category (Table 2). Taxonomic standardization removed 34 and 23% of GBIF and BIEN records (Figure 3a,b), and a further 24 and 5% were removed by the automated clean, respectively. Our

TABLE 1 Spatial and temporal tests applied to GBIF and BIEN data during the automated cleaning process

Test number	Test name	Description
1	General Coordinate Validity	Flags records that do not correspond to lat/long coordinate reference system
2	Equal Coordinates	Flags records with matching lat/long coordinates
3	Zero Coordinates	Flags records with zero coordinates (0/0)
4	Capitals	Flags records ≤10 km radius of capital city
5	Centroids	Flags records ≤1 km radius of country centroids
6	Sea Coordinates	Flags records that fall on non-terrestrial surfaces
7	Geographic Outliers	Flags records that are outside a multiple of the interquartile range of minimum distances to the nearest neighbour occurrence point of the same species
8	GBIF Headquarters	Flags records that fall ≤1° radius around GBIF headquarters in Copenhagen, Denmark
9	Biodiversity Institutions	Flags records ≤100 m of recognized biodiversity institutions

Note: Descriptions adapted from Zizka et al. (2019).

TABLE 2 Number of species with and without data during the initial download (uncleaned) and following the two-stage clean (cleaned)

Stage	Data source	No. species with data (%)[a]	No. species without data (%)[a]	No. threatened species (%)	No. non-threatened species (%)
Uncleaned	GBIF	198 (58.4)	141 (41.6)	94 (47.5)	104 (52.5)
Uncleaned	BIEN	166 (49)	173 (51)	73 (32.4)	152 (67.6)
Cleaned	GBIF	196 (49.1)	143 (35.8)	112 (57.1)	84 (42.9)
Cleaned	BIEN	150 (44.2)	189 (55.8)	87 (58)	63 (42)

Note: "Threatened species" = vulnerable, endangered, and critically endangered; "Non-threatened species" = least concern and near-threatened. From the species remaining in the cleaned stage, only 77 were used in our analysis as a result of cross-referencing species from the GBIF and BIEN data sets.
[a]Values calculated from the 339 species names derived from the original species list.

manual clean removed the least records removing 1 and 0.2% of records, respectively (Figure 3a,b). After data cleaning, 5,046 GBIF and 6,843 BIEN occurrence records remained and were used in the analyses (Table 2; Figure 3a,b).

4.4 | EOO size differences between data sets

On average, uncleaned GBIF EOOs were four times larger than those derived from BIEN data and were reduced to two times as large as EOOs from BIEN data after the automated clean and remained approximately two times larger following the manual clean (Appendix S3). Compared with the control, uncleaned GBIF and BIEN EOOs were on average approximately seven times and twice the size larger for the same species, respectively. The legume *Arachis rigonii* Krapov. & W.C. Greg. had the largest EOO size difference of all the species, with uncleaned GBIF and BIEN EOOs more than 30,000 times and nearly 1,500 times larger than the control (see Appendix S3). The automated clean had a positive effect reducing the average GBIF and BIEN EOO size difference to twice the size and approximately the same size as average control EOOs (Figure 4). The manual clean had a negligible effect on average EOO size differences, however, GBIF EOOs remained two times larger than the control.

5 | CONFUSION MATRIX RESULTS

5.1 | Accuracy rates

The GBIF data were influenced more by the manual clean compared to BIEN, as evidenced by a larger increase in overall accuracy rate (Figure 4a). Overall, accuracy rates differed between GBIF and BIEN (77 vs. 79%, respectively) when using uncleaned data. Both automated data sets had the lowest accuracy rates (76 vs. 74%) (Figure 5a). Overall accuracy was higher for the manual (79%) compared with the automated clean (76%). Accuracy rates were highest for GBIF data after the manual clean but highest for uncleaned BIEN data (Figure 4a). This suggests that the GBIF data requires more cleaning and subsequent user effort to achieve a higher accuracy rate than BIEN.

5.2 | Sensitivity rates

Data cleaning had a positive effect on sensitivity rate (Figure 4b). Sensitivity was the lowest when using uncleaned GBIF and BIEN data (27 vs. 67%, respectively), this improved once the automated clean was applied (60 vs. 73%). Once the manual clean was applied, sensitivity rate increased to the highest percentages for both data sets (73 vs. 80%) (Figure 4b). From a conservation perspective, users should expect to invest more time resources in data cleaning to minimize the risk of underestimating extinction risk for threatened species.

5.3 | Specificity rates

Specificity rates decreased for GBIF and BIEN when applying the automated clean and did not change in response to the manual clean (Figure 4c). Specificity was highest using uncleaned GBIF and BIEN data (91 vs.82%, respectively) and declined after the automated and manual clean were applied (80 vs. 75%) (Figure 4c). These

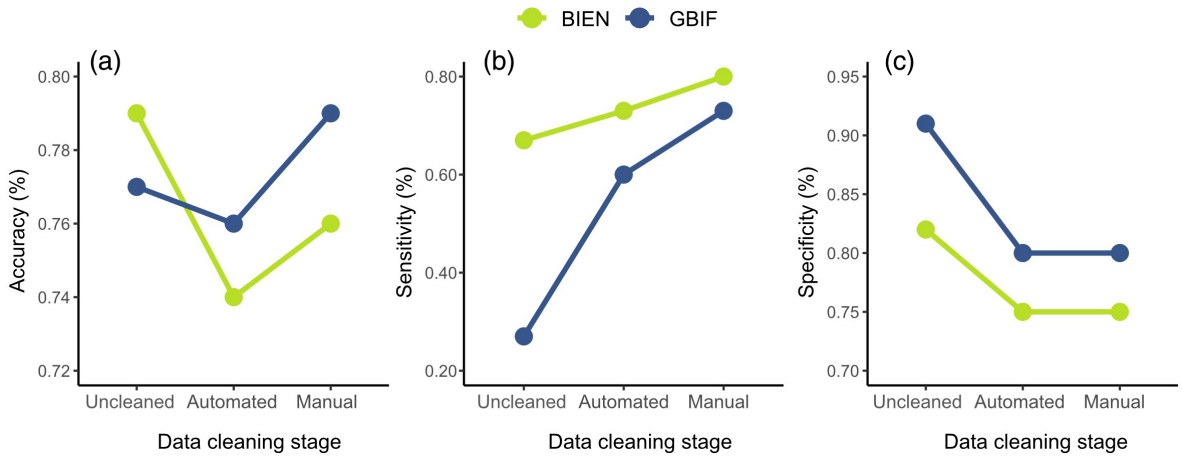

FIGURE 4 Confusion matrix results comparing GBIF and BIEN-derived threat categories against control threat categories at each data cleaning stage: (a) accuracy rate, (b) sensitivity rate, and (c) specificity rate given as percentages

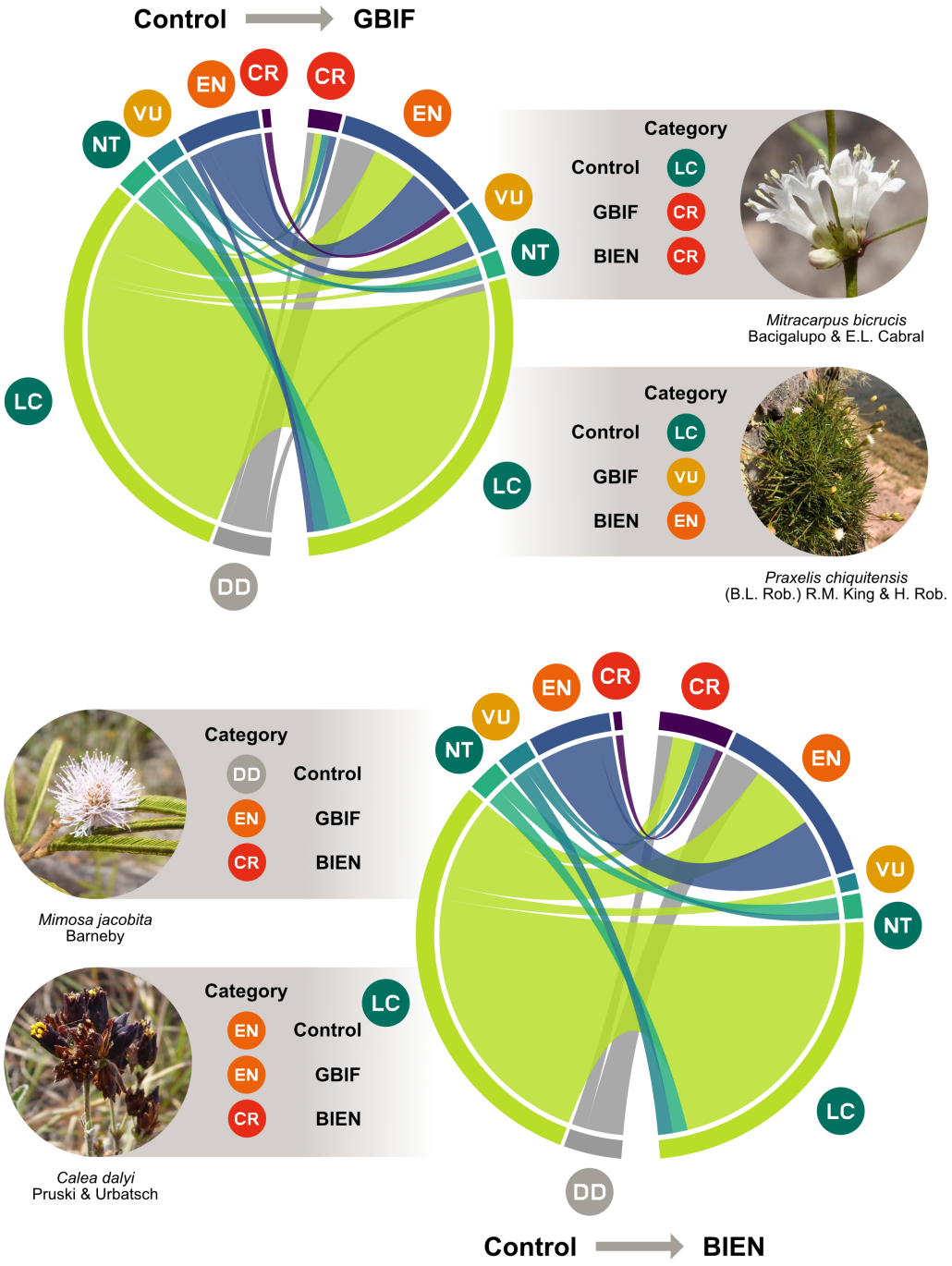

FIGURE 5 Computed IUCN category changes (DD, data deficient; LC, least concern; NT, near-threatened; VU, vulnerable; EN, endangered; CR, critically endangered) by data source with species examples. Category changes from control data to GBIF (top) and BIEN data (bottom) after the two-stage clean. Percentage of species which had different threat categories when GBIF and BIEN data were used compared to the control data, respectively are as follows: GBIF versus BIEN; DD 9.1 versus 9.1%, LC 15.6 versus 18.2%, NT 3.9 versus 2.6%, VU 5.2 versus 5.2%, EN 5.2 versus 2.6% and CR 1.3 versus 0%. Percentages of species that had the same threat categories as the control when GBIF and BIEN data were used, respectively are as follows: DD 0 versus 0%, LC 50.6 versus 48.1%, NT 1.3 versus 2.6%, VU 0 versus 0%, EN 7.8 versus 10.4% and CR 0 versus 1.3%. Species sample size = 77. All photograph credits (R. Clegg)

findings indicate an unintentional overestimation of extinction risk when using uncleaned data likely a result of outlier points or those located in country or departmental centroids. The user should be aware of these before deducing extinction risk estimates from uncleaned data.

5.4 | Over- and underestimating extinction risk

Overall, sensitivity rates for the BIEN data were higher than GBIF, however, there was a larger increase in sensitivity rate for GBIF data following the automated clean (GBIF sensitivity rate increase of 46 vs. 13% for BIEN). Uncleaned GBIF data was more likely to underestimate extinction risk than BIEN. However, uncleaned BIEN data showed a similar pattern and had a lower sensitivity rate compared to automated and manually cleaned BIEN data (Figure 4b). The classifier results showed that GBIF and BIEN were more likely to correctly predict non-threatened species prior to data cleaning (Figures 4d and 5). This is expected as outlier points increase EOO size and for the purposes of this study, any EOO >20,000 km^2 will be classified as non-threatened. The decrease in specificity rate was more notable in GBIF than BIEN (GBIF specificity rate decrease of 11 vs. 7%) (Figures 4c and 5). For the full extent of threat category changes see Appendix S4 and Figure 5 for a visual summary.

6 | DISCUSSION

Our study is one of the first to explore the use of GBIF and BIEN plant occurrence data when estimating preliminary species extinction risk status and demonstrates the influence of a two-stage cleaning protocol on threat category accuracy. We found data cleaning improved the rate of correctly predicting threatened species and automation provides a time efficient solution to process batch extinction risk assessments, however, manual cleaning is essential to achieve accurate results. Between data sets, there was a paucity of occurrence records for 148 (66%) species, 72% of which would qualify for a threatened category highlighting an important conservation concern. Our methods demonstrate the need to consider a balance between data quantity and accuracy when using GBIF and BIEN data to accelerate extinction risk assessments in order to meet future extinction risk targets set by the GSPC.

Achieving such targets requires an acceleration in the rate of extinction risk assessments whilst upholding assessment quality (Darrah, Bland, Bachman, Clubbe, & Trias-Blasi, 2017). Our study builds on existing research investigating novel methods to accelerate the rate of extinction risk assessments (Nic Lughadha et al., 2018). We demonstrate that non-threatened species are less sensitive to data quality than threatened species; which generally require higher quality data derived from data cleaning. Our findings support those of Hjarding, Tolley, and Burgess (2015) who recommend manual cleaning of open-source data to prevent compromised extinction risk assessments.

Occurrence data coverage for the studied species was satisfactory in the GBIF and BIEN data sets, with occurrence records available for 58 and 49% of species, respectively. As expected, GBIF returned data for more species than BIEN (198 vs. 166), but fewer occurrence records (9,593 vs. 11,923), resulting in fewer records per species for GBIF compared to BIEN (48 vs. 72%). BIEN was a good source of plant occurrence records and previous research has used BIEN data as a template to improve the accuracy of GBIF data (Pender et al., 2019). However, an initial bias towards the Americas during the early stages of the BIEN data repository may explain why a higher number of occurrence records were observed for our list of Bolivian plant species compared to GBIF which has a greater global coverage.

Uncleaned GBIF and BIEN data achieved 77 and 79% overall accuracy, respectively. Accuracy decreased for both data sets when the automated clean was applied but improved for the manual clean. Numerically there was a small difference in our accuracy results for GBIF versus BIEN, however, these differences have serious ramifications for threat category calculations. Compared with Nic Lughadha et al. (2018), who assessed accuracy of extinction risk assessments using data derived from various plant groups, our overall accuracy rates were low. However, this is expected due to known errors and limitations, such as spatial biases (Beck, Böller, Erhardt, & Schwanghart, 2014), associated with open-source biological data. However, using such data in extinction risk assessments has its benefits including reduced time and financial cost investments and as shown in our results, can be used to accurately predict extinction risk for some species. Furthermore, much of the data loss experienced as a result of data cleaning was due to duplication, whereby records occurring within the outer boundaries of the EOO were removed and therefore did not affect the resulting extinction risk category.

BIEN data were comparatively cleaner than GBIF, where a wider range of errors were present (Appendix S2), likely due to differential data cleaning strategies within the repositories. Specificity was highest using uncleaned GBIF and BIEN data and overall accuracy decreased after the automated clean was applied in both data sets. Legitimate records may have been mistakenly removed by the automated clean therefore overestimating extinction risk in some species. Despite this, sensitivity rates for both data sets performed better following the two-stage cleaning protocol demonstrating manual data cleaning is essential when assessing species of conservation concern.

A consequence of our method was that of the 225 control data set species, there was a paucity of data for 148 (66%), the majority of which (72%) qualified for threatened categories using our control data, posing a challenge to conservation when attempting to accelerate extinction risk assessments using open-source data. Gaps between the control and open-source data sets were observed, for example, for *Plantago pyrofila* Villarroel and J.R.I. Wood, likely derived from specimen data yet to be uploaded onto the GBIF and BIEN repositories and/or shared between herbaria—a global issue that many herbaria are addressing (Tulig et al., 2012). However, it is likely the case that endemic and range-restricted species are poorly sampled and remain under-represented in herbarium collections. Therefore continued targeted field collections, uploading of specimen data with high precision location data, and up to date specimen identifications onto open-source repositories is required to facilitate effective conservation of threatened plant species.

6.1 | Study limitations

We used EOO as it is a good predictor of extinction risk (Collen et al., 2016), and population-level data required for other Red List criteria are rarely available for plants (Brummitt et al., 2015). Another Criterion B metric, area of occupancy (AOO), has been shown to be calculable with occurrence data and future studies should test this metric in conjunction with EOO. We chose not to include AOO as it is mostly applied when species are well-sampled (Moat, 2017), which was not the case here.

Our use of endemic, near-endemic, and range restricted species may have limited our sample sizes due to the reduction in data availability. However, our decision to use poorly-sampled species is unlikely to have repercussions for data cleaning results if our methods are to be repeated. This is due to the fact that common errors within open-source repositories (e.g., sampling and spatial biases) are not species-dependent and are likely to occur for well-sampled species just as frequently as for poorly-sampled species. To improve time efficiency for future studies, there is a need for clean data sets to be fed back to the source data repositories and any erroneous records highlighted and removed. It is recommended that data repositories such as GBIF and BIEN facilitate this by providing user-friendly platforms for data owners and researchers to feedback cleaned data back into the database. However, this is not without limitations and standardization processes may need to be developed to ensure consistency between cleaned data sets.

Identification errors may have affected our results (Mackay-Smith & Roberts, 2019), but are unlikely to have had the same influence as spatial inconsistencies (Maldonado et al., 2015) such as sampling and specimen collection biases (e.g., along roadsides) (Oliveira et al., 2016). There is yet to be an automated method to detect identification errors within occurrence data due to the complex science of plant names (Thomson et al., 2018) therefore manual checks are still required. Bystriakova, De Melo, Moat, Nic Lughadha, and Monro (2019) report incongruences between taxonomies when using BIEN data, with some species being treated as multiple different species which may result in inaccurate extinction risk assessments. Species with disjunct distributions or introduced populations pose further complications for automated methods and are difficult to detect but are vital for accurate extinction risk assessments. These species remain vulnerable to erroneous removal by automated methods mistakenly interpreted as outliers (de Castro Peña, Kamino, Rodrigues, Mariano-Neto, & de Siqueira, 2014). We recommend that users should identify disjunct species within initial lists of species names prior to data cleaning to prevent the loss of valuable data; albeit at a time cost. We provide a basic timeline of methodological events and recommend an in-depth account of the costs and time investments required for each methodological step; allowing for costs in accuracy to be balanced with time savings. Future research should also investigate the impact of EOO spatial variations when computing threat categories (Appendix S5), as these may affect accuracy and pose challenges when quantifying real world threats during full extinction risk assessments.

7 | IMPLICATIONS

Finding a balance between data quantity and quality must be considered when using open-source data in extinction risk assessments. We recommend the use of cleaned open-source data to support batch preliminary extinction risk assessments, as these data performed better when correctly predicting threatened species than non-threatened species. There were considerable gaps between data sets which pose a conservation concern. Continued *in situ* collection of herbaria material and digitization of herbarium voucher specimens is fundamental for reliable future research. Our methods can be applied to other open-source and citizen science repositories such as iNaturalist (https://www.inaturalist.org/), however, inclusion of additional data sources will likely increase the need for data cleaning. As demonstrated by this study, cleaned open-source data from repositories such as GBIF and BIEN if used carefully, can be mobilized and included within the framework of the IUCN Red Listing process. Now that Target 2 is likely to be adapted into

future recommendations, there is a need for new and achievable conservation goals to ensure the future protection of biodiversity and open-source data has the potential to accelerate the rate of extinction risk assessments highlighting species in most need of protection.

ACKNOWLEDGMENTS

We are indebted to S. Zmartzy and A. Haigh for their contributions to the TIPAs project without which this study would not have been possible. We would further like to thank D.M. Scott, F. Fabriani, N. Biggs, S. Edwards, S. Frisby, N. Hind, A. Monro, R. Schley, M. Martinez U, D Villarroel, J.R.I. Wood, all the TIPAs Bolivia volunteers and Kew's Plant Assessment Unit. The authors are also grateful to B. Walker and A. Zizka for providing technical assistance. Finally, the authors would like to thank two anonymous reviewers and C. Cook for their constructive comments which greatly improved the first version of the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Connor T. Panter, Rosemary L. Clegg, and Bente B. Klitgård originally conceptualized the study. Rosemary L. Clegg and Bente B. Klitgård provided the control data and Connor T. Panter collected the open-source data. Justin Moat, Steven P. Bachman, and Rachel L. White helped develop the methodology and analysis. Connor T. Panter wrote the original draft and all authors contributed to the manuscript review and editing.

DATA AVAILABILITY STATEMENT

Example R code and data for 20 randomly selected species are available from: https://github.com/ConnorPanter/Example-code-for-Panter-et-al.git

ORCID

Connor T. Panter https://orcid.org/0000-0002-4017-158X
Justin Moat https://orcid.org/0000-0002-5513-3615
Steven P. Bachman https://orcid.org/0000-0003-1085-6075

REFERENCES

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017). The accuracy of citizen science data: A quantitative review. *Bulletin of the Ecological Society of America*, *98*, 278–290.

Bachman, S., Moat, J., Hill, A. W., de la Torre, J., & Scott, B. (2011). Supporting red list threat assessments with GeoCAT: Geospatial conservation assessment tool. *Zookeys*, *150*, 117–126.

Bachman, S. P., Nic Lughadha, E., & Rivers, M. C. (2017). Quantifying progress toward a conservation assessment for all plants. *Conservation Biology*, *32*, 516–524.

Balding, M., & Williams, K. J. H. (2016). Plant blindness and the implications for plant conservation. *Conservation Biology*, *30*, 1192–1199.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15.

Bland, L. M., Orme, C. D. L., Bielby, J., Collen, B., Nicholson, E., & McCarthy, M. A. (2015). Cost-effective assessment of extinction risk with limited information. *Journal of Applied Ecology*, *52*, 861–870.

Boyle, B., Hopkins, N., Lu, Z., Garay, J. A. R., Mozzherin, D., Rees, T., ... Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*, 16.

Brito, D., Ambal, R. G., Brooks, T., De Silva, N., Foster, M., Hao, W., ... Rodríguez, J. V. (2010). How similar are national red lists and the IUCN red list? *Biological Conservation*, *143*, 1154–1158.

Brummitt, N., Bachman, S. P., Aletrari, E., Chadburn, H., Griffiths-Lee, J., Lutz, M., ... Nic Lughadha, E. M. (2015). The sampled red list index for plants, phase II: Ground-truthing specimen-based conservation assessments. *Philosophical Transactions of the Royal Society B*, *370*, e20140015.

Burgman, M. A., & Fox, J. C. (2003). Bias in species range estimates from minimum convex polygons: Implications for conservation and options for improved planning. *Animal Conservation forum*, *6*, 19–28.

Bystriakova, N., De Melo, P. H. A., Moat, J., Nic Lughadha, E., & Monro, A. K. (2019). A preliminary evaluation of the karst flora of Brazil using collections data. *Scientific Reports*, *9*, 17037. https://doi.org/10.1038/s41598-019-53104-6

Cayuela, L., Golicher, D. J., Newton, A. C., Kolb, M., de Alburquerque, F. S., Arets, E. J. M. M., ... Pérez, A. M. (2009). Species distribution modelling in the tropics: Problems, potentialities, and the role of biological data for effective species conservation. *Tropical Conservation Science*, *2*, 319–352.

CBD. (2019). *Global strategy for plant conservation: Targets*. Montreal, Canada. Available from: Convention on Biological Diversity. https://www.cbd.int/gspc/targets.shtml

Chamberlain, S., & Boettiger, C. (2017). R python, and ruby clients for GBIF species occurrence data. *PeerJ*, *5*, e3304v1.

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., ... Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294.

Collen, B., Dulvy, N. K., Gaston, K. J., Gärdenfors, U., Keith, D. A., Punt, A. E., ... Akçakaya, H. R. (2016). Clarifying misconceptions of extinction risk assessment with the IUCN red list. *Biology Letters*, *12*, 20150843.

Darrah, S. E., Bland, L. M., Bachman, S. P., Clubbe, C. P., & Trias-Blasi, A. (2017). Using coarse-scale species distribution data to predict extinction risk in plants. *Diversity and Distributions*, *23*, 435–447.

de Castro Peña, J., Kamino, L. H. Y., Rodrigues, M., Mariano-Neto, E., & de Siqueira, M. F. (2014). Assessing the conservation status of species with limited available data and disjunct distribution. *Biological Conservation*, *170*, 130–136.

Di Marco, M., Chapman, S., Althor, G., Kearney, S., Besancon, C., Butt, N., … Watson, J. E. M. (2017). Changing trends and persisting biases in three decades of conservation science. *Global Ecology and Conservation*, *10*, 32–42.

Doughty, C. E., Wolf, A., Morueta-Holme, N., Jørgensen, P. M., Sandel, B., Violle, C., … Galetti, M. (2016). Megafauna extinction, tree species range reduction, and carbon storage in Amazonian forests. *Ecography*, *39*, 194–203.

Enquist, B. J., Condit, R. R., Peet, R. K., Schildhauer, M., & Thiers, B. M. (2016). Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ*, *4*, e2615v2.

ESRI. (2011). *ArcGIS desktop: Release 10*. Redlands, CA: Environmental Systems Research Institute.

García-Roselló, E., Guisande, C., Manjarrés-Hernández, A., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., … Lobo, J. M. (2015). Can we derive macroecological patterns from primary global biodiversity information facility data? *Global Ecology and Biogeography*, *24*, 335–347.

Hermoso, V., & Kennard, M. J. (2012). Uncertainty in coarse conservation assessments hinders the efficient achievement of conservation goals. *Biological Conservation*, *147*, 52–59.

Hjarding, A., Tolley, K. A., & Burgess, N. D. (2015). Red list assessments of east African chameleons: A case study of why we need experts. *Oryx*, *49*, 652–658.

IUCN. (2012). *IUCN red list categories and criteria: Version 3.1* (2nd ed.). Gland, Switzerland and Cambridge, UK: IUCN Iv + 32 pp.

IUCN. (2019). The IUCN red list of threatened species. Version 2019-1. Available from http://www.iucnredlist.org.

Joppa, L. N., Butchart, S. H. M., Hoffmann, M., Bachman, S. P., Akçakaya, H. R., Moat, J. F., … Hughes, A. (2016). Impact of alternative metrics on estimates of extent of occurrence for extinction risk assessment. *Conservation Biology*, *30*, 362–370.

Jørgensen, P. M., Nee, M. H., & Beck, S. G. (Eds.) (2014). Catálogo de plantas vasculares de Bolivia. In *Monographs in systematic botany* (pp. 1–1744). USA: Missouri Botanical Garden Press.

Keith, D. A. (2001). An evaluation and modification of world conservation union red list criteria and classification of extinction risk in vascular plants. *Conservation Biology*, *12*, 1076–1090.

Mace, G. M., Collar, N. J., Gaston, K. J., Hilton-Taylor, C., Akçakaya, H. R., Leader-Williams, N., … Stuart, S. N. (2008). Quantification of extinction risk: IUCN's system for classifying threatened species. *Conservation Biology*, *22*, 1424–1442.

Mackay-Smith, T. H., & Roberts, D. L. (2019). Accuracy in the identification of orchids of the genus *Angraecum* by taxonomists and non-taxonomists. *Kew Bulletin*, *74*, 27.

Maitner, B. S., Boyle, B., Casler, N., Condit, R., Donoghue, J., II, Durán, S. M., … Enquist, B. J. (2017). The BIEN R package: A tool to access the botanical information and ecology network (BIEN) database. *Methods in Ecology and Evolution*, *9*, 373–379.

Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., … Antonelli, A. (2015). Estimating species diversity and distribution in the era of big data: To what extent can we trust public databases? *Global Ecology and Biogeography*, *24*, 973–984.

Margulies, J. D., Bullough, L.-A., Hinsley, A., Ingram, D. J., Cowell, C., Goettsch, B., … Phelps, J. (2019). Illegal wildlife trade and the persistence of "plant blindess". *Plants People Planet*, *1*, 173–182.

Moat, J. (2017). rCAT: Conservation assessment tools. R Package Version 0.1.5. Available from https://CRAN.R-project.org/package=rCAT

Nic Lughadha, E., Bachman, S. P., Leão, T. C. C., Forest, F., Halley, J. M., Moat, J., … Walker, B. E. (2020). Extinction risk and threats to plants and fungi. *Plants People Planet*, *2*, 389–408.

Nic Lughadha, E., Walker, B. E., Canterio, C., Chadburn, H., Davis, A. P., Hargreaves, S., … Rivers, M. C. (2018). The use and misuse of herbarium specimens in evaluting plant extinction risks. *Philosophical Transactions of the Royal Society B*, *374*, e20170402.

Oliveira, U., Paglia, A. P., Brescovit, A. D., de Carvalho, C. J. B., Silva, D. P., Rezende, D. T., … Santos, A. J. (2016). The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions*, *22*, 1232–1244.

Oliveira-Filho, A. T. (2017). *NeoTropTree, Flora arbórea da Região Neotropical: Um banco de dados envolvendo biogeografia, diversidade e conservação*, Brazil: Universidade Federal de Minas Gerais. Available from. http://www.neotroptree.info

Pender, J. E., Hipp, A. L., Hahn, M., Kartesz, J., Nishino, M., & Starr, J. R. (2019). How sensitive are climatic niche inferences to distribution data sampling? A comparison of biota of North America program (BONAP) and global biodiversity information facility (GBIF) data sets. *Ecological Informatics*, *54*, 100991.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria. Available from: R Foundation for Statistical Computing. https://www.R-project.org/

Rivers, M. C., Bachman, S. P., Meagher, T. R., Nic Lughadha, E., & Brummitt, N. A. (2010). Subpopulations, locations and fragmentation: Applying IUCN red list criteria to herbarium specimen data. *Biodiversity and Conservation*, *19*, 2071–2085.

Rivers, M. C., Taylor, L., Brummitt, N. A., Meagher, T. R., Roberts, D. L., & Nic Lughadha, E. (2011). How many herbarium specimens are needed to detect threatened species? *Biological Conservation*, *144*, 2541–2547.

Roberts, D. L., Taylor, L., & Joppa, L. N. (2016). Threatened or data deficient: Assessing the conservation status of poorly known species. *Diversity and Distributions*, *22*, 558–565.

Robiansyah, I., & Wardani, W. (2020). Increasing accuracy: The advantage of using open access species occurrence database in the red list assessment. *Biodiversitas*, *21*, 3658–3664.

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 in Ecology & Evolution*, *21*, 71–76.

Taylor, B. L. (1995). The reliability of using population viability analysis for risk classification of species. *Conservation Biology*, *9*, 551–558.

Thomson, S. A., Pyle, R. L., Ahyong, S. T., Alonso-Zarazaga, M., Ammirati, J., Araya, J. F., … Zhou, H. Z. (2018). Taxonomy based on science is necessary for global conservation. *PLoS One*, *16*, e2005075.

Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessner, A., & Theirs, B. M. (2012). Increasing the efficiency of digitization workflows for herbarium specimens. *Zookeys*, *209*, 103–113.

Westwood, M., Cavender, N., Meyer, A., & Smith, P. (2020). Botanic garden solutions to the plant extinction crisis. *Plants, People, Planet*, *00*, 1–11. https://doi.org/10.1002/ppp3.10134

Wieczorek, J., Guo, Q., & Hijmans, R. J. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, *18*, 745–767.

Willis, C. G., Ellwood, E. R., Primack, R. B., Davis, C. C., Pearson, K. D., Gallinat, A. S., … Soltis, P. S. (2017). Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology & Evolution*, *32*, 531–546.

Willis, F., Moat, J., & Paton, A. (2003). Defining a role for herbarium data in red list assessments: A case study of *Plectranthus* from eastern and southern tropical Africa. *Biodiversity and Conservation*, *7*, 1537–1552.

Wood, C., Sullivan, B., Lliff, M., Fink, D., & Kelling, S. (2011). eBird: Engaging birders in science and conservation. *PLoS Biology*, *9*, e1001220.

Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J. F. R., … Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. *PeerJ*, *8*, e9916.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C. D., Edler, D., … Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*, 744–751.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Panter CT, Clegg RL, Moat J, Bachman SP, Klitgård BB, White RL. To clean or not to clean: Cleaning open-source data improves extinction risk assessments for threatened plant species. *Conservation Science and Practice*. 2020;2:e311. https://doi.org/10.1111/csp2.311