Received: 7 March 2022 | Accepted: 30 March 2022

DOI: 10.1111/2041-210X.13868

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

bdc: A toolkit for standardizing, integrating and cleaning biodiversity data

Bruno R. Ribeiro[1] | **Santiago José Elías Velazco**[2,3,4] | **Karlo Guidoni-Martins**[1] | **Geiziane Tessarolo**[5] | **Lucas Jardim**[6] | **Steven P. Bachman**[7] | **Rafael Loyola**[8,9]

[1]Programa de Pós-graduação em Ecologia e Evolução, Universidade Federal de Goiás, Goiânia, Brazil; [2]Department of Botany and Plant Sciences, University of California—Riverside, Riverside, CA, USA; [3]Instituto de Biología Subtropical, Universidad Nacional de Misiones – CONICET, Puerto Iguazú, Argentina; [4]Programa de Pós-Graduação em Biodiversidade Neotropical, Universidade Federal da Integração Latino-Americana (UNILA), Foz do Iguaçu, Brazil; [5]Universidade Estadual de Goiás, UEG, Campus de Ciências Exatas e Tecnológicas – CCET, Anápolis, Brazil; [6]Instituto Nacional de Ciência e Tecnologia em Ecologia, Evolução e Conservação da Biodiversidade, Universidade Federal de Goiás, Goiânia, Brazil; [7]Royal Botanic Gardens, Kew, Richmond, UK; [8]Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Brazil and [9]International Institute for Sustainability, Rio de Janeiro, Brazil

Correspondence
Bruno R. Ribeiro
Email: ribeiro.brr@gmail.com

Funding information
Argentine National Council of Scientific and Technological Research; Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 465610/2014-5, 201810267000023 and 306694/2018-2; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 001; National Science Foundation, Grant/Award Number: 1853697

Handling Editor: Samantha Price

Abstract

1. The increase in online and openly accessible biodiversity databases provides a vast and invaluable resource to support research and policy. However, without scrutiny, errors in primary species occurrence data can lead to erroneous results and misleading information.
2. Here, we introduce the Biodiversity Data Cleaning (*bdc*), an R package to address quality issues and improve the fitness-for-use of biodiversity datasets. The *bdc* package brings together several aspects of biodiversity data cleaning in one place. It is organized in thematic modules related to different biodiversity dimensions, including (a) Merge datasets: standardization and integration of different datasets; (b) Pre-filter: flagging and removal of invalid or non-interpretable information, followed by data amendments; (c) Taxonomy: cleaning, parsing and harmonization of scientific names from several taxonomic groups against taxonomic databases locally stored through the application of exact and partial matching algorithms; (d) Space: flagging of erroneous, suspect and low-precision geographic coordinates; and (e) Time: flagging and, whenever possible, correction of inconsistent collection date. In addition, the package contains features to visualize, document and report data quality—which is essential for making data quality assessment transparent and reproducible. The modules illustrated, and functions within, were linked to form a proposed reproducible workflow that can also integrate functions from other R packages.
3. We demonstrated the *bdc* package's applicability in cleaning more than 30 million occurrence records for terrestrial plant species in Brazil. We found that around one-fifth of the original datasets hold the standard quality requirements.
4. Compared to other available R packages, the main strengths of the *bdc* package are that it brings together available tools—and a series of new ones—to assess

the quality of different dimensions of biodiversity data into a single and flexible toolkit. The functions can be applied to many taxonomic groups, datasets (including regional or local repositories), countries, or world-wide. We hope the *bdc* package can facilitate the data cleaning process and catalyse improvements to allow the wise and efficient use of primary biodiversity data.

KEYWORDS
big data, biodiversity, data cleaning, data quality, fitness-for-use, GBIF, plants, taxonomy

1 | INTRODUCTION

The development of biodiversity informatic tools and new computational platforms in recent decades has led to a significant increase in the online availability of primary species occurrence data retrieved from natural history collections and citizen science observations (Bisby, 2000; Graham et al., 2004; Soberón & Peterson, 2004). Such openly accessible biodiversity databases provide a vast and invaluable resource to document species distributions through time and space for research, education and environmental policy support (Ball-Damerow et al., 2019; Canhos et al., 2015; Chapman, 2005c).

The importance of primary species occurrence data for many biodiversity applications is evident, yet they have limitations, and their quality can vary substantially (Meyer et al., 2016). Without scrutiny, issues related to difficulty standardizing data from different sources (Kissling et al., 2018), discrepancies and errors in taxonomic and nomenclatural data (e.g. Mesibov, 2013; Nic Lughadha et al., 2019), and errors and inaccuracies in geographical and temporal information of primary species occurrence data (e.g. Meyer et al., 2016) can lead to erroneous results and misleading information (Maldonado et al., 2015; Nic Lughadha et al., 2019; Zizka et al., 2020). Although several efforts have already been made to develop tools for cleaning biodiversity data and improve fitness-for-use (e.g. functionalities found in GBIF (www.gbif.org), the Atlas of Living Australia (www.ala.org.au), SpeciesLink (splink.cria.org.br) and many R packages (details in Appendix S1), significant challenges remain, especially when assembling large and heterogeneous databases from online aggregators (Chapman, 2005b; Kissling et al., 2018).

Here, we present the Biodiversity Data Cleaning (*bdc*) package to address quality issues and improve the fitness-for-use of a dataset. In the *bdc* package, we sought to encompass a series of tests regarding the taxonomic, spatial and temporal dimensions of data to resolve the most common data quality issues. Compared to other available R packages, the main strengths of the *bdc* package are that it brings together available tools—and a series of new ones—to assess the quality of different dimensions of biodiversity data into a single and flexible framework. The tools can be applied to many taxonomic groups, datasets (including regional or local repositories), countries, or world-wide. The package builds upon the integration and enhancement of cutting-edge functionalities (Carvalho, 2017; Norman et al., 2020; Zizka et al., 2019) and on a series of new tests and tools developed for validating, documenting and reporting data quality.

2 | DESCRIPTION

2.1 | Overview

The *bdc* package is organized in thematic modules related to different biodiversity dimensions (Figure 1, Table 1; Meyer et al., 2016), including (a) Merge datasets: Standardization and integration of different datasets; (b) Pre-filter: flagging and removal of invalid or non-interpretable information, followed by data amendments; (c) Taxonomy: cleaning, parsing, harmonization of scientific names against multiple taxonomic references; (d) Space: flagging of erroneous, suspect and low-precision geographic coordinates; and (e) Time: flagging and, whenever possible, correction of inconsistent collection date (Figure 1; Table 1). In addition, the package contains functions for documenting the results of the data cleaning tests, including functions for saving (a) records needing further inspection, (b) figures and iii) data quality reports. These files facilitate the interpretation and visualization of the results by users and are automatically saved in a folder named 'Outputs' (Figure 1; Table 1). The modules illustrated, and functions within, can be linked to form a reproducible workflow as illustrated in the empirical example using records of the Brazilian flora, but can also be executed independently depending on user needs.

The *bdc* package is implemented in R (R Core Team, 2020), and it is based on standard tools for data quality assessments as well as data handling and visualization, including *taxadb* (Norman et al., 2020), *flora* (Carvalho, 2017) and *ggplot2* (Wickham, 2016). We provide extensive documentation and tutorials on functions on the package website (brunobrr.github.io/bdc).

2.2 | Standardization and integration of heterogeneous datasets

The lack of terminology standardization makes the integration of large and heterogeneous datasets a challenge. To remedy this, the

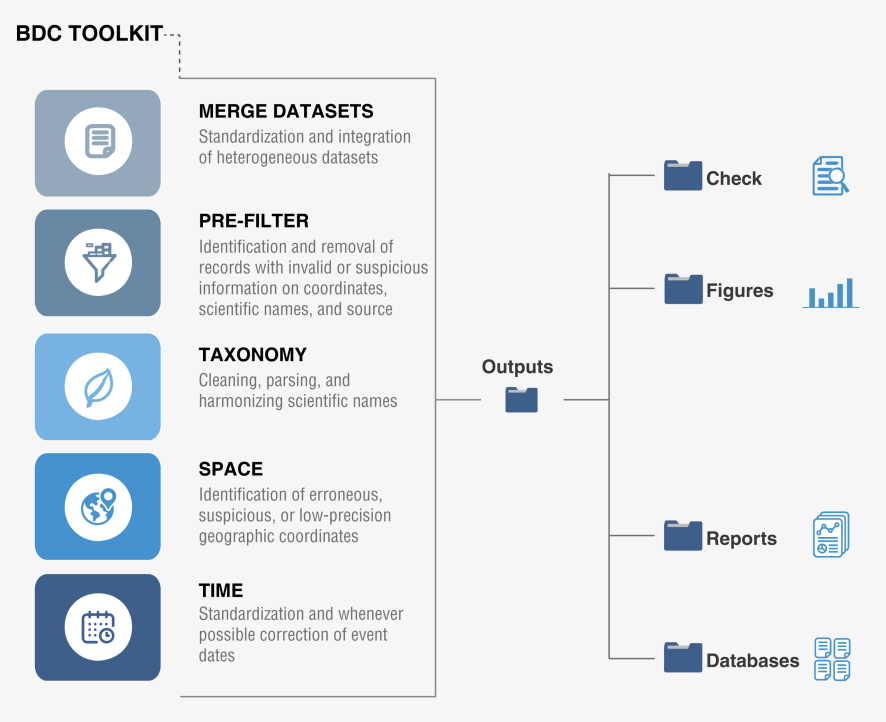

FIGURE 1 The Biodiversity Data Cleaning (*bdc*) package contains functionalities for standardizing and integrating data from different sources and implements several tests to flag, document, clean and correct biodiversity data. The *bdc* package is organized in thematic modules (merge datasets, pre-filter, taxonomy, space and time). Several outputs documenting the data cleaning process can be saved, including files needing further inspections, figures and reports

function *bdc_standardize_datasets* specifically handles the standardization of heterogeneous datasets. To do so, users must fill in a configuration table (see example in Appendix S2) to indicate which field names (i.e. column headers) of each original dataset match a list of Darwin Core standard terms (Wieczorek et al., 2012). Once standardized, datasets are then integrated into a formatted database having a minimum set of terms required to share biodiversity data and metadata across a wide variety of biodiversity applications (Table S1; see also Simple Darwin Core standards at dwc.tdwg.org/simple).

2.3 | Pre-filter

Large and heterogeneous datasets may contain thousands of records missing spatial or taxonomic information (partially or entirely) and several records outside a region of interest (Jin & Yang, 2020; Peterson et al., 2018). The pre-filter module contains functions to flag and remove (a) records missing species names, (b) records missing partial or complete information on geographic coordinates, (c) out-of-range coordinates (latitude >90 or −90; longitude >180 or −180), (d) records from doubtful sources (e.g. from drawings, photographs or multimedia objects, among others) and (e) records outside a region of interest, that is, records in other countries or at an informed distance from the coast (e.g. in the ocean). This last step avoids falsely flagging records close to country limits as invalid (e.g. records of coast or marshland species; see Table S2 for additional details about this test). The pre-filter module also includes functions for data enhancement, such as deriving country names from valid geographic coordinates, standardizing country names, identifying records with potentially transposed geographic coordinates and saving a table containing records with missing coordinates but with potentially useful locality information (Table 1, Table S2; see Supporting Information for more details).

2.4 | Taxonomic harmonization

Combining large datasets from several sources requires careful harmonization of potentially thousands of taxonomic names. The *bdc* package includes functions to help the taxonomic name harmonization by comparing scientific names against one of 10 taxonomic databases. The taxonomic harmonization uses *taxadb* package (Norman et al. (2020), which contains functions that allow querying millions of taxonomic names quickly, efficiently and consistently using high-quality, locally stored taxonomic databases. Querying names against these databases avoids significant drawbacks inherent to tools that implement queries using web APIs (application programming interfaces), such as the need for internet access to perform queries, and limitations on number of names that can be queried at once (Norman et al., 2020).

A major limitation of *taxadb* package is that misspelled scientific names—commonly found in biodiversity databases—cannot be resolved by an exact matching algorithm, which may result in many unresolved names. To troubleshoot this, we developed additional functions (*bdc_clean_names* and *bdc_query_names_taxadb* functions) for (a) cleaning and parsing scientific names; (b) resolving misspelled names or variant spellings using a fuzzy matching application, (c) converting nomenclatural synonyms to the currently accepted name and (d) flagging ambiguous results. The *bdc_clean_names* comprises several name-checking routines that optimize the taxonomic queries by unifying writing style and thus increasing the probability of finding matching names (Tables

TABLE 1 List and description of the main functions implemented in the *bdc* package (more details are presented in Table S2). Functions are grouped in thematic modules, namely merge datasets, pre-filter, taxonomy, space and time. The function *clean_coordinates* (Zizka et al., 2019) is part of the proposed data cleaning workflow but not part of the *bdc* package

Modules	Label	Description	Source
Merge datasets	*bdc_standardize_datasets*	Harmonization and integration of different datasets into a standard database	*bdc*
Pre-filter	*bdc_scientificName_empty*	Identification of records lacking names or with names not interpretable	*bdc*
	bdc_coordinates_empty	Identification of records lacking information on latitude or longitude	*bdc*
	bdc_coordinates_outOfRange	Identification of records with out-of-range coordinates (latitude >90 or −90; longitude >180 or −180)	*bdc*
	bdc_basisOfRecords_notStandard	Identification of records from doubtful sources (e.g. fossil or machine observation)	*bdc*
	bdc_country_from_coordinates	Deriving country name from valid geographic coordinates	*bdc*
	bdc_country_standardized	Standardization of country names and retrieving country code	*bdc*
	bdc_coordinates_transposed	Identification of records with potentially transposed latitude and longitude	*bdc*
	bdc_coordinates_country_inconsistent	Identification of coordinates in other countries or far from a specified distance from the coast of a reference country (i.e. in the ocean)	*bdc*
	bdc_coordinates_from_locality	Identification of records lacking coordinates but with a detailed description of the locality from which coordinates can be derived	*bdc*
Taxonomy	*bdc_clean_names*	Name-checking routines to clean and split a taxonomic name into its binomial and authority components	*bdc*; rgnparser
	bdc_query_names_taxadb	Harmonization of scientific names by correcting spelling errors and converting nomenclatural synonyms to currently accepted names	bdc; taxadb
	bdc_filter_out_names	This tool is used to filter out records according to their taxonomic status present in the column 'notes'. For example, to filter only valid accepted names categorized as 'accepted'	*bdc*
Space	*bdc_coordinates_precision*	Identification of records with a coordinate precision below a specified number of decimal places	*bdc*
	clean_coordinates	Identification of potentially problematic geographic coordinates based on geographic gazetteers and metadata	CoordinateCleaner v2.0-18
Time	*bdc_eventDate_empty*	Identification of records lacking information on event date (i.e. when a record was collected or observed)	*bdc*
	bdc_year_outOfRange	Identification of records with illegitimate or potentially imprecise collecting year	*bdc*
	bdc_year_from_eventDate	This function extracts four-digit year from unambiguously interpretable collecting dates	*bdc*
All modules	*bdc_create_report*	Creation of data quality reports documenting the results of data quality tests and the taxonomic harmonization process	*bdc*
	bdc_create_figures	Creation of figures (i.e. bar plots and maps) reporting the results of data quality tests	*bdc*
	bdc_filter_out_flags	Removal of columns containing the results of data quality tests (i.e. column starting with '.') or other columns specified	*bdc*
	bdc_quickmap	Creation of a map of points using *ggplot2*. Helpful in inspecting the results of data cleaning tests	*bdc*
	bdc_summary_col	This function creates or updates the column summarizing the results of data quality tests (i.e. the column '*.summary*')	*bdc*

S2–S4). After cleaning and parsing, names are then standardized based on one out of 10 taxonomic databases (docs.ropensci.org/taxadb/articles/data-sources.html) available in the *taxadb* package using an exact matching algorithm. Even after running name-checking routines, a scientific name can remain unresolved because of typos or spelling variants. In such cases, a fuzzy matching algorithm processes name-matching queries to find a potential matching candidate from the specified taxonomic database based

on a match distance defined by the user. A detailed explanation of the taxonomic harmonization process can be found in the Supporting Information and Table S2.

2.5 | Identification of errors in geographic coordinates

We used the *CoordinateCleaner*, an R package based on geographic gazetteers, to flag potential erroneous coordinates (Zizka et al., 2019), which include records with (a) zero coordinates in a radius around the point at zero latitudes and longitude; (b) equal latitude and longitude; (c) possible duplicate records with equal longitude, latitude and accepted species name. Likewise, the package identifies records assigned to (d) country capitals; (e) province centroids; (f) urban areas; (g) biodiversity institutions; and (h) geographic outliers (see Figure 1, Table 1; Table S2). Finally, we also developed a tool for identifying records with low-precision coordinates (Robertson et al., 2016). More details on each test can be found in Table S2 and Zizka et al. (2019). We stressed that *CoordinateCleaner* (Zizka et al., 2019) is part of the proposed data cleaning workflow but not part of the *bdc* package.

2.6 | Standardization and validation of temporal information

To standardize and validate temporal data, *bdc* contain a function (*bdc_year_from_eventDate*) to extract the collection year whenever possible from complete and legitimate date information (Figure 1; Table S2). Records with dubious collection year (e.g. 10/10/12) as well as with illegitimate (e.g. 1450, 2050) or no collection date supplied (e.g. 0 and NA) are flagged and can be subsequently removed (*bdc_year_outOfRange* function).

2.7 | Empirical example: The Brazilian flora

We demonstrated the *bdc*'s applicability in cleaning >30 million occurrence records for terrestrial plant species in Brazil. All package functions were used to assess the quality of Brazilian flora data (the workflow used can also be checked on the package website). The R scripts used in the analyses are available in Appendix S3 (Ribeiro et al., 2022a). More specifically, we aimed to assess the impact of data cleaning on species richness (i.e. number of species before and after both the taxonomic harmonization and the application of spatial and temporal filters) and on the spatial pattern of species richness. Brazil harbours ~38,680 plant species (angiosperm, gymnosperm, ferns and lycophytes, and bryophyte, Flora do Brasil, 2020 under construction), whose records are distributed in several heterogeneous online databases, making it an ideal case study. We assembled records for terrestrial plant species in Brazil that could be accessed via nine public, freely and openly available online databases (Table S5).

3 | RESULTS

From ~31 million records included in the original databases, only ~2.7 million records were considered high-quality data after the data cleaning and validation processes, representing a reduction of nearly 91% of the initial dataset (Figure 2; Figure S1; Table S6). Without removing records lacking information on collecting date, 13% of the original database were considered fit-for-use (Figure 2; Figure S1; Table S5). The number of records flagged in each data cleaning test can be found in Table S5.

Overall, 59% of the initial records, most corresponding to records lacking georeferencing (43%) or occurring outside Brazil (13%; Table S5; Figure 2), were excluded after applying filters of the pre-filter module. Data amendments were also performed by using functions available in the pre-filter module. Country names of 11% of records were derived from valid coordinates and others 9% had country names standardized (Table S7). Finally, 0.14% of records with transposed coordinates were corrected (Table S7; Figure S2). The taxonomic step flagged 1.9% of records with names linked to non-accepted scientific names (Table S6). The most common spatial issues flagged corresponded to duplicate records (26%) followed by records within urban areas (2.2%; Table S6; Figure S3). Around 19% of records lacked information on collecting date and 56 records with an out-of-range year (e.g. record collected before the year 1600 or in the future, e.g. 2030; Table S6). Records with valid date information had collecting years spanning from 1600 to 2020; most of them were recorded between 1980 and 2016 (Figure S4).

The data-cleaning altered the number of occurrence records available to build species richness maps (Figure 2). In the pre-filter module, 213,303 specimens were recognized. This number was reduced to 38,783 species after taxonomic harmonization and to 38,207 and 36,540 species after applying space and temporal filters respectively. The application of space and temporal filters led to a loss of 576 and 1,667 species, respectively, which had all records flagged as suspect or erroneous. While most records flagged in the space filters occurred within urban areas, around country or province centroids, or presented imprecise coordinates, most records removed in the time module were recorded before 1970.

4 | DISCUSSION

The *bdc* package is a toolkit that offers the means to convert raw data into high-quality information through a suite of core functions used to flag, clean, document and enrich data quality. Such tools allow an automated and faster quality assessment of large datasets containing millions of records and thousands of species. The package comprises flexible functions that can be applied to evaluate data quality. Nevertheless, there is no silver bullet to assess the quality of biodiversity data suitable for all purposes (Chapman, 2005a; Zizka et al., 2020). Biodiversity data have inherent limitations, and no hard rule exists to judge data quality needs to make data fit-for-use in all biodiversity applications; essentially, the data's adequacy depends

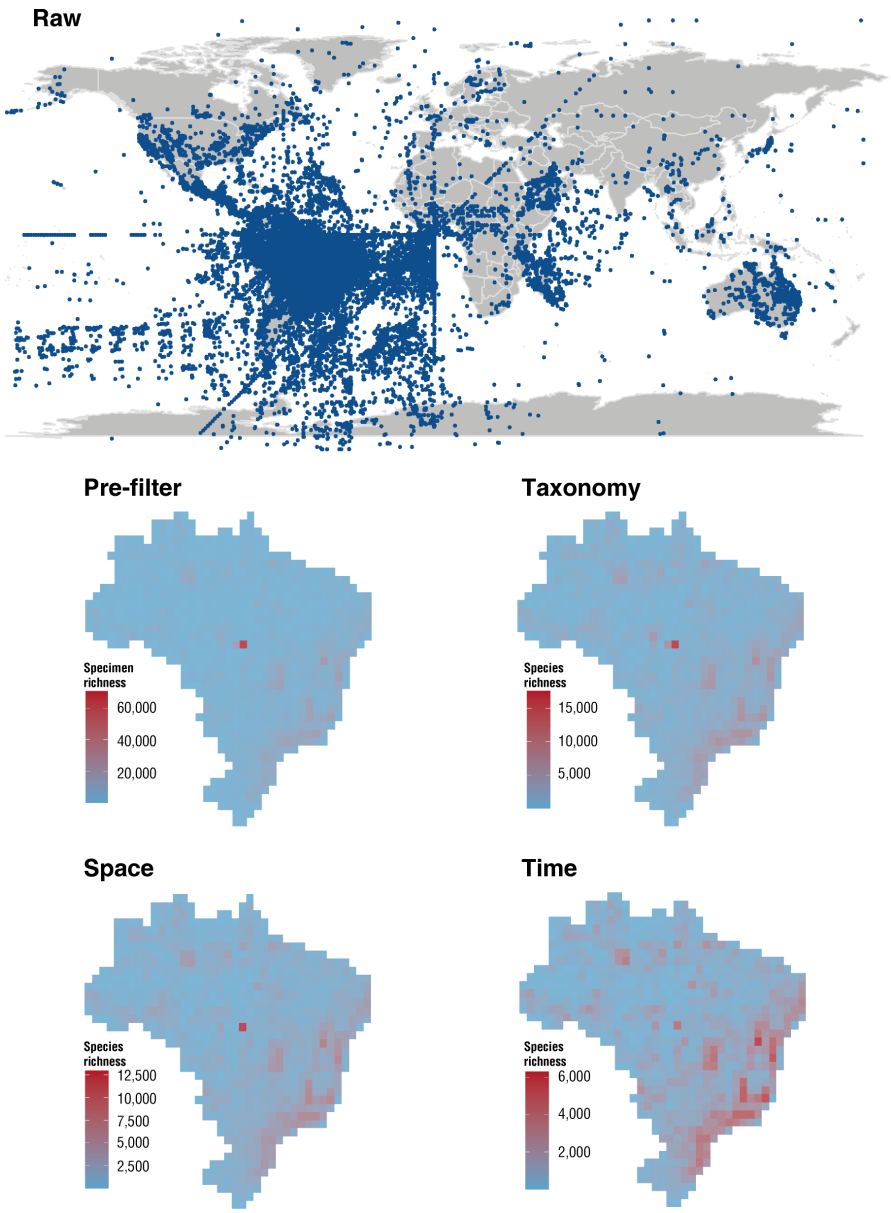

FIGURE 2 Records of terrestrial plants occurring in Brazil in the raw database. Richness maps after applying data cleaning tests grouped in thematic modules named pre-filter, taxonomy, space and time. Note legend has different scales

mainly on the user's needs (Chapman, 2005b; Veiga et al., 2017). The uncritical use of overly strict filters can result in loss of valid information; otherwise, errors can persist if no data cleaning is applied. In this sense, some researcher judgement will always be required to choose appropriate tools and criteria to evaluate data quality and make data adequate for specific purposes (Zizka et al., 2019).

Perhaps the main novelty of the package is that it brings together several aspects of biodiversity data cleaning in one package and workflow. Further, *bdc* contains hands-on functions to harmonize scientific names of several taxonomic groups using taxonomic databases locally stored and through the application of exact and partial matching algorithms. Finally, *bdc* processes are documented and auditable, making biodiversity data management more transparent and reproducible (a detailed comparison to available R packages can be found in Appendix S1). We hope the *bdc* package can facilitate and scale the data cleaning process and catalyse improvements to allow the wise and efficient use of primary biodiversity data. We plan to add new functionalities in the future versions of the package. In this sense, we encourage and welcome users' support to fix issues and suggest new features.

ACKNOWLEDGEMENTS

We are grateful to Eimear Nic Lughadha and Barnaby Walker (Review 1) for their valuable comments and suggestions on this piece. We also thank researchers and citizens all over the world working to make knowledge on plants openly available online. B.R.R. and K.G.M. were supported by CAPES scholarships. G.T. was supported by PNPD/CAPES postdoctoral fellowship. S.J.E.V. thanks the postdoctoral fellowships supported by the National Science Foundation (Award 1853697) and the Argentine National Council of Scientific and Technological Research received during this project. R.L. research is funded by CNPq (grant #306694/2018-2). L.J. thank the postdoctoral fellowship supported by CNPq (165615/2020-6). 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).

CONFLICT OF INTEREST

All authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

B.R.R., S.J.E.V., K.G.-M., G.T., L.J., S.P.B. and R.L. conceived the ideas and designed the methodology; B.R.R., S.J.E.V., K.G.-M., G.T. and L.J. collected the data; B.R.R. analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13868.

DATA AVAILABILITY STATEMENT

The package is available as R package from the CRAN repository (Ribeiro et al., 2022b). The code of *bdc* is open and available on GitHub (brunobrr.github.io/bdc) and Zenodo's webpage (doi.org/10.5281/zenodo.6450390, Ribeiro et al., 2022c). The scripts used in the analyses of the Brazilian flora are available in Appendix S3 (Ribeiro et al., 2022a). Extensive documentation and tutorials on *bdc* package can be found at brunobrr.github.io/bdc. All data on the Brazilian flora were downloaded from nine public, freely and openly available data sources (Table S4).

ORCID

Bruno R. Ribeiro https://orcid.org/0000-0002-7755-6715
Santiago José Elías Velazco https://orcid.org/0000-0002-7527-0967
Karlo Guidoni-Martins https://orcid.org/0000-0002-8458-8467
Geiziane Tessarolo https://orcid.org/0000-0003-1361-0062
Lucas Jardim https://orcid.org/0000-0003-2602-5575
Steven P. Bachman https://orcid.org/0000-0003-1085-6075
Rafael Loyola https://orcid.org/0000-0001-5323-2735

REFERENCES

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*(9), 1–26. https://doi.org/10.1371/journal.pone.0215794

Bisby, F. A. (2000). The quiet revolution: Biodiversity informatics and the internet. *Science*, *289*(5488), 2309–2312. https://doi.org/10.1126/science.289.5488.2309

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 Biology*, *13*(7), e1002204. https://doi.org/10.1371/journal.pbio.1002204

Carvalho, G. (2017). flora: Tools for interacting with the Brazilian Flora 2020. Retrieved from http://www.github.com/gustavobio/flora

Chapman, A. D. (2005a). *Principles and methods of data cleaning*. Report for the Global Biodiversity Information Facility, 1–72.

Chapman, A. D. (2005b). *Principles of data quality*. Global Biodiversity, 58. Retrieved from http://www2.gbif.org/DataQuality.pdf

Chapman, A. D. (2005c). Uses of primary species-occurrence data. *Australian Biodiversity Information Services*, *GBIF Papers*, *6*, 22–36. https://doi.org/10.1007/s00520-011-1353-z

Flora do Brasil. (2020). *Jardim Botânico do Rio de Janeiro*. Retrieved from http://floradobrasil.jbrj.gov.br/

Graham, C., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, *19*(9), 497–503. https://doi.org/10.1016/j.tree.2004.07.006

Jin, J., & Yang, J. (2020). BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation*, *21*, e00852. https://doi.org/10.1016/j.gecco.2019.e00852

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., … Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews*, *93*(1), 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? *Global Ecology and Biogeography*, *24*(8), 973–984. https://doi.org/10.1111/geb.12326

Mesibov, R. (2013). A specialist's audit of aggregated occurrence records. *ZooKeys*, *293*, 1–18. https://doi.org/10.3897/zookeys.293.5111

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*(8), 992–1006. https://doi.org/10.1111/ele.12624

Nic Lughadha, E. M., Graziele Staggemeier, V., Vasconcelos, T. N. C., Walker, B. E., Canteiro, C., & Lucas, E. J. (2019). Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. *Conservation Biology*, *33*(3), 511–522. https://doi.org/10.1111/cobi.13289

Norman, K. E. A., Chamberlain, S., & Boettiger, C. (2020). taxadb: A high-performance local taxonomic database interface. *Methods in Ecology and Evolution*, *11*(9), 1153–1159. https://doi.org/10.1111/2041-210X.13440

Peterson, A. T., Asase, A., Canhos, D., de Souza, S., & Wieczorek, J. (2018). Data leakage and loss in biodiversity informatics. *Biodiversity Data Journal*, *6*, e26826. https://doi.org/10.3897/BDJ.6.e26826

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ribeiro, B. R., Velazco, S. E. V., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P., & Loyola, R. (2022a). Data from: Appendix S3. Online resource. *figshare*, https://doi.org/10.6084/m9.figshare.19390949

Ribeiro, B. R., Velazco, S. E. V., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P. & Loyola, R. (2022b). Data from: bdc: Biodiversity data cleaning. R package version 1.0.0. Retrieved from https://brunobrr.github.io/bdc/ (website) https://github.com/brunobrr/bdc

Ribeiro, B. R., Velazco, S. E. V., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P., & Loyola, R. (2022c). Data from: *bdc*: A toolkit for standardizing, integrating, and cleaning biodiversity data (v1.0.0). *Zenodo*, https://doi.org/10.5281/zenodo.6450390

Robertson, M. P., Visser, V., & Hui, C. (2016). Biogeo: An R package for assessing and improving data quality of occurrence record datasets. *Ecography*, *39*(4), 394–401. https://doi.org/10.1111/ecog.02118

Soberón, J., & Peterson, A. T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *359*(1444), 689–698. https://doi.org/10.1098/rstb.2003.1439

Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., & Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLoS ONE*, *12*(6), e0178731. https://doi.org/10.1371/journal.pone.0178731

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. Retrieved from http://ggplot2.org

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, *7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

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., Moreira, F. G. L., Santos, N. M. C., Santos, T. A. B., dos Santos-Costa, R. C., Serrano, F. C., da Silva, A. P. A., de Souza Soares, A., de Souza, P. G. C., Tomaz, E. C., … 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 in Ecology and Evolution*, *10*(5), 744–751. https://doi.org/10.1111/2041-210X.13152

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Ribeiro, B. R., Velazco, S. J., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P., Loyola, R. (2022). *bdc*: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution*, 13, 1421–1428. https://doi.org/10.1111/2041-210X.13868