Distribution of endemic angiosperm species in Brazil on a municipality level

Janaína Gomes-da-Silva‡, João Lanna‡, Rafaela Campostrini Forzza‡

‡ Instituto de Pesquisas do Jardim Botânico do Rio de Janeiro, Rio de Janeiro, Brazil

Corresponding author: Janaína Gomes-da-Silva (jgomes_da_silva@yahoo.com.br)

Academic editor: Quentin Groom

Abstract

Background

Herbarium collections and the data they hold are the main sources of plant biodiversity information. These collections contain taxonomical and spatial data on living and extinct species; consequently, they are the fundamental basis for temporal and spatial biogeographical studies of plants. Mega projects focused on providing digital and free access to accurate biodiversity data have transformed plant science research, mainly in the past two decades. In this sense, researchers today are overwhelmed by the many different datasets in online repositories. There are also several challenges involved in using these data for biogeographical analyses. Analyses performed on the data available in the repositories show that 70-75% of the total amount of data have spatial deficiencies and a high number of records lack coordinates. This shortage of reliable primary biogeographical information creates serious impediments for biogeographical analyses and conservation assessments and taxonomic revisions consequently produces obstacles for evaluations of threats to biodiversity at global, regional and local levels. With the aim of contributing to botanical and biogeographical research, this paper provides georeferenced spatial data for angiosperm species endemic to Brazil. The information from two reliable online databases, i.e. the Flora do Brasil 2020 floristic database (BFG) and Plantas do Brasil: Resgate Histórico e Herbário Virtual para o Conhecimento e Conservação da Flora Brasileira (REFLORA), which are both based on records collected over the course of the last two centuries, is used to create this spatial dataset.

New information

We provide three taxonomically-edited and georeferenced datasets for basal angiosperms, monocots and eudicots, covering a total of 14,992 endemic species from Brazil. Producing this consolidated dataset involved several months of detailed revision of coordinates and nomenclaturally updating of the names in these datasets. The information provided in this

geo-referenced dataset, covering two centuries of specimen collections, will contribute to several botanical and mainly biogeographical studies.

Keywords

Endemic species, data re-use, flowering plants, occurrence records, primary biodiversity data, South America

Introduction

Herbarium collections and the data they hold have been one of the main sources of plant biodiversity information through time (Gasper et al. 2020). They include taxonomical and spatial data on living and extinct species and, therefore, provide a fundamental basis for both temporal and spatial studies of plants (Funk 2003, Hortal et al. 2015, James et al. 2018). Mega projects, focusing on providing accurate, digital open-access data on biodiversity, including digitised specimens and species occurrence data, have transformed biodiversity analysis in the past two decades (Graham et al. 2004, La Salle et al. 2016). In this sense, today's large amount of different datasets in online repositories can be overwhelming, for example, the Global Biodiversity Information Facility (https://www.gbif.org/pt/,GBFI 2021), the Flora of Brasil 2020 floristic database (BFG 2020) and the Plantas do Brasil: Resgate Histórico e Herbário Virtual para o Conhecimento e Conservação da Flora Brasileira (http://reflora.jbrj.gov.br/, REFLORA 2021) virtual herbarium.

Widespread access to taxonomic and distributional data is producing great advances in botanical and biogeographical research, as well as supporting more accurate evaluations of extinction risks (Gomes-da-Silva and Forzza 2020, Robiansyah and Wardani 2020). In spite of this substantial progress, there are several challenges and limitations when applying open-access repository data (see discussion in James et al. 2018). Unfortunately, the quality of the species occurrence records available in most collections is low (Robertson et al. 2016). Evaluations of the data available in repositories show that ca. 70-75% of these data have spatial deficiencies, mainly with regard to the georeferencing quality (Colli-Silva et al. 2020, Jin and Yang 2020, Marcer et al. 2020). Jin and Yang (2020) assessed 30,242,556 occurrence records from different repositories and demonstrated that only 28% of the records had high-quality taxonomic and spatial data. In addition, analyses have shown that erroneous records, containing geographic inaccuracies, affect the spatial patterns for species more significantly than taxonomic uncertainties (Maldonado et al. 2015). The spatial accuracy of the data available in the GBIF database for flowering plants in the Brazilian Atlantic Forest was evaluated recently and the analysis revealed that only 25% of the records contained precise spatial information (Colli-Silva et al. 2020). Similarly, an analysis of the REFLORA 2021 database for the present work showed that the georeferenced data have repetitive errors, of which the most common are missing coordinates (lat/long), zero values entered for the latitude and longitude, points in the oceans (for terrestrial species) or the Antarctic, only the latitude or longitude coordinates

entered and lack of coordinate precision. These analyses (Colli-Silva et al. 2020, Jin and Yang 2020, present work) reveal the value in cleaning data in biodiversity studies and the need to georeference these databases.

Manipulating millions of records is an extremely complicated task. In recent years, workflows, tools and methods have been developed for dealing with taxonomic and geographic errors, simplifying the process (Chamberlain 2016, Robertson et al. 2016, Zizka et al. 2019, Jin and Yang 2020) by identifying potential geographical and temporal errors in databases and converting the coordinates to various text formats (e.g. Chamberlain 2016, Robertson et al. 2016, Zizka et al. 2019, Jin and Yang 2020). In addition, BDcleaner can be used to remove taxonomic errors (Jin and Yang 2020). However, there is no effective tool for correcting geographical errors in lieu of discarding them.

As manual data cleaning is laborious (Marcer et al. 2020), many studies choose to reduce datasets by discarding occurrence data with correctable geographic errors. This incomplete data sampling introduces uncertainties to analyses and compromises the results, particularly in terms of regional analyses (Hortal et al. 2015; Casagranda and Goloboff 2019). To employ the IUCN Categories and Criteria used to create the (IUCN) Red List for species at risk of extinction, mainly criteria B [severely fragmented] and D2 [very restricted area of occupancy], it is necessary to identify the geographical ranges of species accurately and reliably (IUCN 2012). Up until the last decade, ca. 1% (61,914; IUCN 2012) of species have been evaluated using the Red List to define their conservation status (Bachman et al. 2011). Although the number of species assessed has doubled in the last 10 years (120, 372; IUCN 2020), this number is still far from the IUCN target of 160,000 for 2020 (IUCN 2020).

Brazil has the highest biodiversity of vascular plants on the planet (BFG: Filardi et al. 2018). According to the updated version of the BFG database, there are currently 32,696 species of angiosperms on record in Brazil, of which ca. 18,000 species are endemic to the country (BFG 2020). Despite the errors in spatial data, the high number of records lacking coordinates and the gaps in its database, which are common in all databases (Beck et al. 2013, Jin and Yang 2020), the REFLORA repository, used in conjunction with the Flora of Brasil (BFG 2020), provides reliable data. These two repositories represent massive collaborations of taxonomists from various institutions, including experts on every flora family in the country. Using the filters for the BFG database, it is possible to generate a verified taxonomical list of endemic Brazilian species, carefully prepared by several taxonomists. This task, which seems simple nowadays, was exceedingly difficult or impossible prior to the creation of the BFG database.

The geographical range of a species forms the basis for biogeographical studies. Repositories, such as REFLORA 2021, make distributional records accessible, mitigating the poor geographic data. With millions of high-resolution images, the REFLORA project minimises the deficiencies of primary data (Canteiro et al. 2019). However, most of the data provided by the repository lack georeferencing. Thus, with the aim of contributing to botanical and mainly with biogeographical research in Angiospermae, this paper

provides georeferences for 14,992 endemic Brazilian species from 173 families, based on reliable taxonomic data from the REFLORA and BFG datasets.

Project description

Title: Geo-referenced spatial data for angiosperm species endemic to Brazil

Design description: The REFLORA 2021 and BFG (BFG 2020) databases are fed new data daily and edited for changes in nomenclature. The georeferencing work carried out here was developed between August 2018 and December 2019. Thus, the difference between the number of endemic species recorded in 2020 (i.e. about 18,000 species) and the number of georeferenced species provided here (i.e. 14,992 species) is supported by the following factors:

1. In August 2018, 1,393 species had no vouchers in REFLORA.
2. In order to obtain the highest possible accuracy in species occurrence data, we established editing procedures for the use of geographical distributions from the collection records (outlined below). These procedures made it impossible to include 1,615 species with inconsistencies in the collection records.

This georeferenced occurrence dataset for endemic species provides the basis for a wide range of biodiversity studies, for example, spatial studies conducted at various hierarchical levels, i.e. family, genus, species; effects of global change; changes in distributions of species; conservation; and systematics.

Funding: Conselho Nacional de Desenvolvimento Cientıfico e Tecnológico (CNPq) and FA PERJ - Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro for the postdoctoral fellowship granted to JGS. RCF received a Research Productivity Fellowship from CNPq (proc.303420/2016-2) and FAPERJ (processes n° E-26/202.778/2018) through Programa Cientista do Nosso Estado.

Sampling methods

Description: Brazilian angiosperms dataset

Species list compilation:

The list of species was established in two phases. First, the initial list of names of all endemic species of Angiospermae was generated through the BFG in the Brazilian Flora (BFG 2020) website (http://floradobrasil.jbrj.gov.br), edited by several taxonomic experts in each sampled family, using the following search filters: (1) Group: Angiospermae; (2) Occurs in Brazil: yes; (3) Occurrence: only occurs in Brazil; (4) Endemism: only endemic to Brazil; and (5) Origin: native (Fig. 1).

Based on this list of all endemic Brazilian angiosperm species retrieved from the BFG floristic database between August 2018 and October 2019, all occurrence records were

downloaded from the REFLORA 2021 virtual herbarium (www.reflora.jbrj.gov.br) from 73 herbaria (ALCB, ASE, B, BRBA, CEN, CEPEC, CESJ, CGMS, COR, CRI, DVPR, E, EAC, ECT, ESA, EVB, FIG, FLOR, FURB, GH, HACAM, HBR, HCF, HDCF, HEPH, HRCB, HSTM, HTO, HUCO, HUCP, HUEFS, HUEM, HUEMG, HUENF, HUFU, HUNEB, HUNI, HUPG, HVASF, IAN, IBGE, ICN, K, LUSC, MAC, MBM, MBML, MG, MO, MUFAL, NY, P, PEL, PMSP, R, RB, RBR, REAL, RFA, UERJ, RFFP, RON, S, SJRP, SPF, UB, UFRN, UNIP, UNOP, UPCB, US, VIES and W, the herbaria acronyms following Thiers (2020), continuously updated). After these two phases, 18,000 endemic species to Brazil were identified, corresponding to the raw database. Producing this consolidated dataset involved 1 1/2 years of detailed revision of coordinates and nomenclaturally updating of the names in these datasets, as follows:

We created a protocol to clean the datasets (Fig. 1), the data were processed carefully by checking nomenclatural status and excluding records with erroneous occurrence data. The accuracy of species identification follows the list of endemic species of BFG 2020. Four steps were conducted for cleaning the taxonomic data. In the first step, we checked and cleaned the data taxonomically and nomenclaturally; only vouchers identified to species level about which we were uncertain were removed, including 'cf.,' 'aff.,' 'sp.,' and 'spp.'. In the second step, we corrected the spelling of taxon names, which, for some species, had multiple entries with different spellings. In the third step, varieties and subspecies were grouped at the species level. In the fourth step, hybrids were excluded, synonyms were checked and accepted names were adopted according to the BGF. We performed the first four steps using the "filter" tool in Microsoft Excel v. 14.5 (Microsoft Office 2010 Proofing Tools).

Subsequently, we conducted manual cleaning procedures on the records. For cleaning the records, three steps were performed on the geographic data. In the first step, records of specimens with imprecise or vague descriptions of locations (e.g. Negro River, north coast, south coast) and incomplete (e.g. Amazonia, Bahia, Brazil) or incongruent information concerning locations (e.g. with no administrative unit, location in the ocean) were excluded. In the second step, we removed the taxonomic duplicates and records of duplicate samples with the same species name and place of occurrence and voucher information. In the final dataset, each record corresponds to a single herbarium specimen for which the geographical location has been checked and is unique to that locality. Duplicates were removed from the list, based on locality, collector name, collector number and the year in which the sample was collected. After data cleaning, the total number of records dropped from 827,016 to 183,201 occurrence records with complete voucher information.

The use of GPS became more widespread in 1995-1996, but there were still few satellites at that time (Kaplan 2005). Given that the occurrence records for all species endemic to Brazil were collected mainly over the last two centuries, it was not surprising that more than 75% records were not georeferenced. Hence, in the third step, we manually edited and included the coordinates of each voucher, based on databases of localities and municipalities maintained by the Brazilian Institute of Geography and Statistics website (IBGE) (http://mapas.ibge.gov.br), for 161,563 occurrence records of 14,992 endemic angiosperm species. For 21,632 records, it was not possible to perform georeferencing

due to lack of sufficient information on the voucher. In this step, we removed the complete voucher data, since the main objective concerns the use of the dataset for biogeographical analysis. We performed the three steps using the “filter” tool in Microsoft Excel v. 14.5 (Microsoft Office 2010 Proofing Tools).

The final checklist is composed of native and endemic angiosperms and includes only vouchers identified to the species level, based on the Brazilian Flora (BFG 2020) and complete records, based on REFLORA. The complete list of vouchers, including all geographical duplicates (duplicate samples for same location) and photos to check the identity of the species, is available at REFLORA 2021(http://reflora.jbrj.gov.br/).

Geographic coverage

Description: The geographic coverage encompasses the national territory of Brazil, which extends from 5° to -34° Latitude; -34° to -73° Longitude and covers a total area of approximately 8.5 million km² (IBGE). The dataset comprised all species of Angiospermae found exclusively in Brazil and it contains occurrence records in six phytogeographic domains, i.e. Amazonia, Caatinga, Cerrado, the Atlantic Forest, Pampa and Pantanal, in Chacoan, Parana, South Brazilian and South-eastern Amazonian dominions (Fig. 2, sensu Morrone (2014)).

Coordinates: -34 and -5° Latitude; -73° and -34 Longitude.

Taxonomic coverage

Description: To facilitate the search for taxa at different hierarchical levels, the dataset comprises three different worksheets of specimens collected over the past two centuries organised according to APG IV classification (Chase et al. 2016) and these have been organised alphabetically, as follows:

(1st Worksheet) A total of 649 species of basal angiosperms belonging to five orders, i.e. Canellales, Laurales, Magnoliales, Nymphaeales and Piperales from 13 families and 50 genera. Number of records is georeferenced by order in Fig. 3A.

(2nd Worksheet) A total of 3,854 species of monocots belonging to nine orders, i.e. Alismatales, Arecales, Asparagales, Commelinales, Dioscoreales, Liliales, Pandanales, Poales and Zingiberales from 32 families and 370 genera. Number of records is georeferenced by order in Fig. 3B.

(3rd Worksheet) A total of 10,489 eudicots, belonging to 31 orders, i.e. Apiales, Aquifoliales, Asterales, Boraginales, Brassicales, Caryophyllales, Celastrales, Cornales, Cucurbitales, Dilleniales, Dipsacales, Ericales, Escalloniales, Fabales, Gentianales, Geraniales, Gunnerales, Lamiales, Malpighiales, Malvales, Myrtales, Oxalidales, Picramniales, Proteales, Ranunculales, Rosales, Santalales, Sapindales, Solanales,

Vitales and Zygophyllales from 128 families and 1,199 genera. Number of records is georeferenced by order in Fig. 3C.

Usage licence

Usage licence: Creative Commons Public Domain Waiver (CC-Zero)

Data resources

Data package title: Distribution of endemic angiosperm species in Brazil on a municipality level.

Resource link: https://ckan.jbrj.gov.br/dataset/mitigating-the-question-of-the-geographic-distribution

Number of data sets: 3

Data set name: Basal_Angiosperms_Brazil_Gomes_da_Silva_Forzza_Lanna.tsv

Download URL: https://ckan.jbrj.gov.br/dataset/e1eb798c-601a-4d20-bf17-87dc037ed73e/resource/5ceeb350-b071-46bb-86b4-071b1bbf1372/download/basal_angiosperms_brazil_gomes_da_silva_forzza_lanna.tsv

Data format: TSV

Description: Data containing the geographic distribution of 649 species of basal angiosperms from 13 families.

Column label	Column description
family	The scientific name of the family in which the taxon is classified.
genus	The scientific name of the genus in which the taxon is classified.
specificEpithet	Scientific name.
country	The country where the species occur.
stateProvince	State of Brazil where species occur.
municipality	Municipality of Brazil where species occur.
decimalLatitude	The latitude component (N/S) of the coordinates of the municipality where the species occur, in decimal degrees.
decimalLongitude	The longitude component (E/W) of the coordinates of the municipality where the species occur, in decimal degrees.

Data set name: Eudicots_Brazil_Gomes_da_Silva_Forzza_Lanna.tsv

Download URL: https://ckan.jbrj.gov.br/dataset/e1eb798c-601a-4d20-bf17-87dc037ed73e/resource/d2160257-a141-4ff4-89f2-d93edef0e6a6/download/eudicots_brazil_gomes_da_silva_forzza_lanna.tsv

Data format: TSV

Description: Data containing the geographic distribution of 10,489 eudicots from 128 families.

Column label	Column description
family	The scientific name of the family in which the taxon is classified.
genus	The scientific name of the genus in which the taxon is classified.
specificEpithet	Scientific name.
country	The country where the species occur.
stateProvince	State of Brazil where species occur.
municipality	Municipality of Brazil where species occur.
decimalLatitude	The latitude component (N/S) of the coordinates of the municipality where the species occur, in decimal degrees.
decimalLongitude	The longitude component (E/W) of the coordinates of the municipality where the species occur, in decimal degrees.

Data set name: Monocots_Brazil_Gomes_da_Silva_Forzza_Lanna.tsv

Download URL: https://ckan.jbrj.gov.br/dataset/e1eb798c-601a-4d20-bf17-87dc037ed73e/resource/4326f085-dbbe-48ff-812d-aba565f64c8d/download/monocots_brazil_gomes_da_silva_forzza_lanna.tsv

Data format: TSV

Description: Data containing the geographic distribution of 3,854 species of monocots from 32 families.

Column label	Column description
family	The scientific name of the family in which the taxon is classified.
genus	The scientific name of the genus in which the taxon is classified.
specificEpithet	Scientific name.
country	The country where the species occur.
stateProvince	State of Brazil where species occur.
municipality	Municipality of Brazil where species occur.

decimalLatitude	The latitude component (N/S) of the coordinates of the municipality where the species occur, in decimal degrees.
decimalLongitude	The longitude component (E/W) of the coordinates of the municipality where the species occur, in decimal degrees.

Additional information

Despite the digitisation efforts of numerous museums and herbaria, data gaps remain. We strongly encourage and recommend that distributional data be correctly georeferenced in collections in order to increase the quality of the spatial data used in future analyses.

Due to the immeasurable importance of primary occurrence data and the difficulties in georeferencing inaccurate geographical distribution data, we recommend that collectors strive to prioritise and record exact coordinates for their collections (see discussion in Colli-Silva et al. 2020). In addition, the sharing of georeferenced data should become standard procedure, in line with sharing DNA sequences data in GenBank. As well as the georeferenced data in the present work being returned to the REFLORA database, we recommend that small and large datasets of georeferenced data should be returned to the collections database and published in a data paper. Unquestionably this "standard procedure" will improve the quality of primary data and provide greater accuracy in future biogeographical analyses, thus promoting the advancement of science.

Acknowledgements

We would like to thank the scientific and technical teams of the Flora do Brasil and REFLORA. The work, presented in this paper, is part of a postdoctoral study conducted by the first author at the Jardim Botânico do Rio de Janeiro. The authors are grateful to the following Brazilian funding agencies: FAPERJ - Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (2021) and CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (2020) for the postdoctoral fellowship granted to JGS. RCF received a Research Productivity Fellowship from CNPq (proc. 303420/2016-2) and FAPERJ (processes n° E-26/202.778/2018) through Programa Cientista do Nosso Estado. This work was supported by funds from Natura. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Author contributions

Gomes-da-Silva, Janaina (conceived the presented idea, dataset preparation, dataset editing, manuscript writing, manuscript editing).

João Lanna (dataset editing).

Forzza, Campostrini Rafaela (conceived the presented idea, supervised the findings of this work and the project, manuscript editing).

References

- Bachman S, Moat J, Hill A, de la Torre J, Scott B (2011) Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. ZooKeys 150: 117-126. https://doi.org/10.3897/zookeys.150.2109
- Beck J, Ballesteros-Mejia L, Nagel P, Kitching I (2013) Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? Diversity and Distributions 19 (8): 1043-1050. https://doi.org/10.1111/ddi.12083
- BFG, et al. (2020) Flora do Brasil 2020. Jardim Botânico do Rio de Janeiro. http://floradobrasil.jbrj.gov.br/. Accessed on: 2021-2-26.
- Canteiro C, Barcelos L, Filardi F, Forzza R, Green L, Lanna J, Leitman P, Milliken W, Pires Morim M, Patmore K, Phillips S, Walker B, Weech M, Nic Lughadha E (2019) Enhancement of conservation knowledge through increased access to botanical information. Conservation Biology 33 (3): 523-533. https://doi.org/10.1111/cobi.13291
- Casagranda MD, Goloboff PA (2019) On stability measures and effects of data structure in the recognition of areas of endemism. Biological Journal of the Linnean Society 127 (1): 143-155. https://doi.org/10.1093/biolinnean/blz019
- Chamberlain S (2016) scrubr: clean biological occurrence records R package version 0.1.1. URL: https://CRAN.R-project.org/package=scrubr
- Chase MW, Christenhusz MJM, Fay MF, et al. (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181 (1): 1-20. https://doi.org/10.1111/boj.12385
- Colli-Silva M, Reginato M, Cabral A, Forzza RC, Pirani JR, Vasconcelos TdC (2020) Evaluating shortfalls and spatial accuracy of biodiversity documentation in the Atlantic Forest, the most diverse and threatened Brazilian phytogeographic domain. Taxon 69 (3): 567-577. https://doi.org/10.1002/tax.12239
- Filardi FLR, Barros FD, Baumgratz JFA, Bicudo C, Cavalcanti TB, Coelho MAN, et al. (2018) Brazilian flora 2020: Innovation and collaboration to meet target 1 of the Global Strategy for Plant Conservation (GSPC). Rodriguésia 69 (4): 1513-1527. https://doi.org/10.1590/2175-7860201869402
- Funk V (2003) The importance of herbaria. Plant Science Bulletin 49: 94-95.
- Gasper ALD, Stehmann JR, Roque N, Bigio NC, et al. (2020) Brazilian herbaria: an overview. Acta Botanica Brasilica 34 (2): 352-359. https://doi.org/10.1590/0102-33062019abb0390
- GBFI (2021) Global Biodiversity Information Facility. https://www.gbif.org/pt/ ,. Accessed on: 2021-5-21.
- Gomes-da-Silva J, Forzza RC (2020) Two centuries of distribution data: detection of areas of endemism for the Brazilian angiosperms. Cladistics https://doi.org/10.1111/cla.12445
- 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

- Hortal J, de Bello F, Diniz-Filho J, Lewinsohn T, Lobo J, Ladle R (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. Annual Review of Ecology, Evolution, and Systematics 46 (1): 523-549. https://doi.org/10.1146/annurev-ecolsys-112414-054400
- IUCN (2012) IUCN Red List Categories and Criteria:. Version 3.1. Gland,. Switzerland and 558 Cambridge, UK.
- IUCN (2020) The IUCN Red List of Threatened Species. IUCN Red List Categories and Criteria: version 3.1. 2012 Version 2020-2. https://www.iucnredlist.org. Accessed on: 2020-7-09.
- James S, Soltis P, Belbin L, Chapman A, Nelson G, Paul D, Collins M (2018) Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in Plant Sciences 6 (2). https://doi.org/10.1002/aps3.1024
- 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 https://doi.org/10.1016/j.gecco.2019.e00852
- Kaplan E (2005) Chapter 1: Introduction. In: Hegarty C, Kaplan E (Eds) Understanding GPS: Principles and applications. 2. Artech House Publishers, Boston, London, 726 pp.
- La Salle J, Williams K, Moritz C (2016) Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society B: Biological Sciences 371 (1702). https://doi.org/10.1098/rstb.2015.0337
- Maldonado C, Molina C, Zizka A, Persson C, Taylor C, 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
- Marcer A, Haston E, Groom Q, Ariño A, Chapman A, Bakken T, Braun P, Dillen M, Ernst M, Escobar A, Fichtmüller D, Livermore L, Nicolson N, Paragamian K, Paul D, Pettersson L, Phillips S, Plummer J, Rainer H, Rey I, Robertson T, Röpert D, Santos J, Uribe F, Waller J, Wieczorek JR (2020) Quality issues in georeferencing: From physical collections to digital data repositories for ecological research. Diversity and Distributions 27: 564-567. https://doi.org/10.1111/ddi.13208
- Morrone J (2014) Biogeographical regionalisation of the Neotropical region. Zootaxa 3782 (1). https://doi.org/10.11646/zootaxa.3782.1.1
- REFLORA (2021) REFLORA Resgate Histórico e Herbário Virtual para o Conhecimento e Conservação da Flora Brasileira. http://reflora.jbrj.gov.br/. Accessed on: 2021-5-01.
- Robertson M, 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
- Robiansyah I, Wardani W (2020) Increasing accuracy: The advantage of using open access species occurrence database in the Red List assessment. Biodiversitas Journal of Biological Diversity 21 (8). https://doi.org/10.13057/biodiv/d210831
- Thiers (2020) A global directory of public herbaria and associated staff. http://sweetgum.nybg.org/ih/. Accessed on: 2020-10-20.
- 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

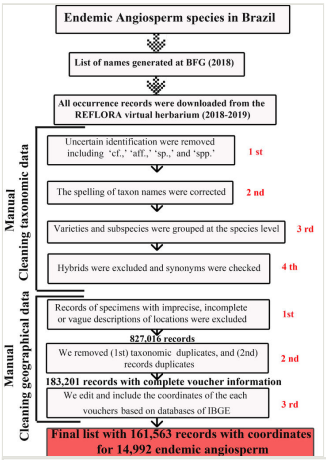

Figure 1.

Taxonomic and geographic data refinement workflow. Steps of data filtering to obtain the endemic angiosperm species list for Brazil, based on the list available from BFG 2018 (Filardi et al. 2018), in the Brazilian Flora 2020 website and records from Reflora Herbarium Virtual (2018-2019).

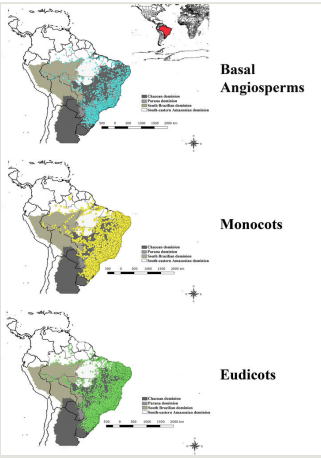

Figure 2.

Spatial distribution of angiosperms for all georeferenced data available at the Reflora Herbarium Virtual after data cleaning.

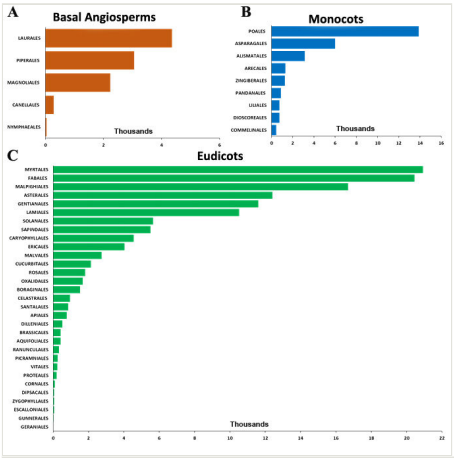

Figure 3.

Number of records georeferenced for endemic angiosperm species in Brazil on a municipality level, by order for: **A.** basal angiosperms; **B.** monocots; and **C.** eudicots.