METHODS AND TECHNIQUES

Evaluating shortfalls and spatial accuracy of biodiversity documentation in the Atlantic Forest, the most diverse and threatened Brazilian phytogeographic domain

Matheus Colli-Silva,[1] Marcelo Reginato,[2] Andressa Cabral,[1] Rafaela Campostrini Forzza,[1,3] José Rubens Pirani[1] & Thais N. da C. Vasconcelos[1,4]

1 *Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, Rua do Matão 277, 05508-090, São Paulo, São Paulo, Brazil*

2 *Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, 91501-970, Porto Alegre, Rio Grande do Sul, Brazil*

3 *Herbário do Jardim Botânico do Rio de Janeiro, Rua Pacheco Leão 915, 22460-030, Rio de Janeiro, Rio de Janeiro, Brazil*

4 *Biological Sciences Department, University of Arkansas, Harmon Ave 227 N, Fayetteville 72701, Arkansas, U.S.A.*

Address for correspondence: *Matheus Colli-Silva, matheus.colli.silva@alumni.usp.br*

DOI https://doi.org/10.1002/tax.12239

Abstract Digital accessible knowledge of biodiversity data is an increasingly important source of information in studies of biogeography and conservation. These databases can also reveal temporal, spatial and taxonomical gaps in biodiversity documentation, even in areas that have been intensively studied and from where accurate species lists are available. Therefore, revealing these gaps may help allocating collecting efforts, conservation priorities and strategies for improving database curation. Here, we evaluate potential shortfalls for flowering plants in a tropical hotspot, the Brazilian Atlantic Forest, by cross-referencing two online repositories of biodiversity data (the Global Biodiversity Information Facility – GBIF – and the Brazilian Flora 2020 floristic database – BFG). We aimed to evaluate the congruence between those repositories, highlighting tendencies in current documentation for this area. We found that from the 7220 reported flowering plant species endemics to the Atlantic Forest, 1573 (22%) have no valid spatial data in GBIF, and 75% of all of the 605,951 records do not present valid spatial information. Most of the missing information is related to species known only from few and old collections with absent or inaccurately georeferenced data. This lack of information may cause a large impact in spatial studies, especially for rare and threatened species. Nevertheless, our analysis also shows that spatial information for the filtered data is highly congruent between GBIF and BFG data, indicating relatively high availability of quality data in large repositories after standard and automatized cleaning procedures. Still, good practices to decrease the impact of losing data are recommended, including more investment in field collections, targeting poorly known species and returning cleaned spatial datasets to online repositories after taxonomic revisions.

Keywords angiosperms; Atlantic Forest; big-data; biodiversity shortfalls; data collection; endemic species; herbarium collections

Supporting Information may be found online in the Supporting Information section at the end of the article.

■ INTRODUCTION

Biodiversity documentation in the "era of big-data". — Biological collections are at the front line of biodiversity research, providing data for accurate documentation (Buerki & Baker, 2016). The maintenance of well-curated databases of species distribution have a major role in spatial analyses and biodiversity studies, which is one service of herbaria (Funk, 2003). These datasets also provide basic data for outreach, increasing awareness and educating society about biodiversity and conservation (Wen & al., 2015). In the past few decades, information regarding biodiversity distribution has been massively compiled in free and online repositories such as GBIF (the Global Biodiversity Information Facility, www.gbif.org). Biodiversity studies have entered in the "era of big-data" (Maldonado & al., 2015), in which it is relatively easy, but sometimes tricky, to retrieve valid information on the spatial distribution of biodiversity (Yesson & al., 2007).

In spite of this great progress, it is widely known that Digital Accessible Knowledge (DAK; see Sousa-Baena & al., 2013) is accompanied by several downsides, especially regarding the quality of georeferencing (Beck & al., 2013; Zizka & al., 2018), taxonomic accuracy (Goodwin & al., 2015; Oliveira & al., 2017, Troudet & al., 2017) and relative collecting effort among different areas (Giaretta & al., 2015; Williams & Crouch, 2017). For instance, it is expected that GBIF contains several of these caveats (Robertson & al., 2014; Yesson & al., 2007), so at least some procedures of data curation (i.e., data cleaning) are mandatory before further use. It is against that background that recent tools and methods have facilitated such procedures by automatizing correcting steps that can deal at least with some of these spatial issues (e.g., Robertson & al., 2016; Zizka & al., 2019).

Article history: Received: 6 Jun 2019 | returned for (first) revision: 18 Oct 2019 | (last) revision received: 17 Jan 2020 | accepted: 23 Jan 2020 | published online: 17 Jul 2020 | **Associate Editor:** Levent Can |

Concurrently, the concept of "biodiversity hotspots"—areas where high levels of species endemism and habitat loss coincide to produce high extinction risk (Myers & al., 2000)—is grounded mainly by the known numbers of exclusive plant species in an area (Joppa & al., 2011). In this sense, proper documentation of spatial distribution of species is critical, and initiatives such as the Systematics Agenda 2020 and the Biodiversity CyberBank (Wen & al., 2015) have increased the application of big-data in biological research and conservation initiatives.

The Brazilian Atlantic Forest as a case-study. — Brazil, the most species-rich country for vascular plants in the world (Forzza & al., 2012; Ulloa-Ulloa & al., 2017; Willis, 2017), is very heterogeneous regarding collecting effort throughout its enormous area. Some regions, such as the Amazon rainforest, are long known to suffer from extensive biasing gaps of largely under-collected areas (Daly & Prance, 1989; Hopkins, 2007; Oliveira & al., 2017). On the other hand, regions such as the Atlantic Forest have gone through more intensive collecting during the last centuries (Mori, 1989; Galindo-Leal & Câmara, 2003).

The Atlantic Forest covers most of the eastern coast of South America and harbors one of the highest levels of species richness and endemism in the Neotropics (Mori, 1989; Murray-Smith & al., 2009). European colonization in Brazil started from the coast, and historical collecting expeditions followed this route. Furthermore, these areas also include the highest concentration of cities and inhabitants today (Galindo-Leal & Câmara, 2003; Ribeiro & al., 2009). As a result, this is one of the most intensively collected phytogeographic domains in Brazil (Oliveira & al., 2019), with a rich collection of preserved plant specimens in herbaria (Morellato & Haddad, 2000). Conversely, the Atlantic Forest is also the phytogeographic domain that has been harshly diminished since the 19th century, when many native and endemic species were described (see Pires-O'Brien, 1993).

Until recently, it was difficult to list exactly how many species of angiosperms were native or endemic to the Brazilian Atlantic Forest. Currently, thanks to a continuous team effort undertaken since 2008, taxonomists are compiling the "Brazilian Flora 2020 Project" (henceforward BFG), a massive collaboration initiative that intends to monograph all plant, algae and fungi species in the country by 2020 (BFG, 2018a). One output of this project is a taxonomically verified list of all species occurring in Brazil and in each of its particular phytogeographic domains—including the Atlantic Forest.

The existence of this list of names, along with the fact that the Brazilian Atlantic Forest has been one of the most vastly collected areas in the Neotropics, makes this phytogeographic domain a good model to explore how cross-checking DAK repositories can help allocating collecting and re-collecting efforts. In this study, we cross-checked two DAK repositories—GBIF in light of the BFG floristic database—to explore patterns related to the accuracy of taxonomical, temporal and spatial scales of biodiversity documentation, using the Brazilian Atlantic Forest as a case-study. We aim to clarify the following statements: (1) there is a temporal pattern in data precision, where more recent collections usually bear more accurate georeferenced data; (2) there is congruence between spatial distribution data from GBIF and the BFG database; and (3) data cleaning procedures discard a significative amount of species distribution points, and this may represent a bias in inferences of spatial patterns of biodiversity. These statements will be discussed in the context of possible solutions to improve biodiversity documentation in Brazil and worldwide.

■ MATERIALS AND METHODS

Study area and data. — The Atlantic Forest is a biodiversity hotspot (Myers & al., 2000), mostly (i.e., over 90% of its area) part of the Brazilian territory (Ribeiro & al., 2009). Recent estimates show that this phytogeographic domain has ca. 15,000 native species of angiosperms only in Brazil, and nearly half of them are endemic to this area (BFG, 2018a).

We considered two DAK repositories of biodiversity information for this area: (1) the BFG floristic database (BFG, 2018b); and (2) GBIF. First, we downloaded a list of all Brazilian angiosperm species retrieved from the BFG floristic database in March 2019, which is indexed in GBIF repository. From that list, we selected all reported species endemic to the Brazilian Atlantic Forest, according to BFG. Only accepted names were selected (i.e., synonyms were not considered). By choosing endemic species only, we limited the sampling area and minimized the already huge amount of data under analysis, making the study operationally feasible.

This initial list of names was used as a search string to download occurrence data from GBIF. Records for all angiosperms based on preserved specimens collected in Brazil were downloaded from GBIF portal (GBIF.org, 2019), as GBIF also incorporates most of the regional and local online repositories, such as SpeciesLink (www.splink.cria.org.br), Reflora (www.reflora.jbrj.gov.br) and the Rio de Janeiro Botanical Garden database (www.jabot.jbrj.gov.br; Silva & al., 2017). In order to standardize taxonomy across datasets, synonyms were updated using the R package "flora" v.0.3.0 (Carvalho, 2017; R Core Team, 2019).

Data cleaning procedures and evaluation. — We performed an automatized data cleaning round using the R package "CoordinateCleaner" v.2.0-11 (Zizka & al., 2019). Functions in this package deal efficiently with technical errors in spatial data by flagging and dropping points located in country or state centroids (which are very imprecise), points in the sea (which do not make sense for terrestrial species) and "lacking coordinates" (i.e., points with latitude and longitude fields marked as "0" or "NA").

Finally, we performed a second round of data cleaning by selecting only occurrence points that are within the borders of

the Brazilian Atlantic Forest domain, using the IBGE (the Brazilian Institute of Geography and Statistics) shapefile of the Atlantic Forest. The same shapefile is used by the BFG project when classifying the phytogeographic domains of each taxon (BFG, 2018a). This round of cleaning was performed using the software QGIS v.3.8 (www.qgis.org). After this procedure, we extracted the final "cleaned" database, which contained what we named as "valid" records. The amount of information lost at each of these cleaning steps were critically evaluated according to their temporal, taxonomic and spatial patterns.

Comparing GBIF and BFG databases of species distribution. — The BFG database also informs which Brazilian first-level administrative divisions (henceforth "states") have confirmed occurrences for a particular species, based on the expertise of the taxonomist responsible for the species to be monographed. In order to get a proxy for the congruence between this information and the distribution records in GBIF, we generated and compared two presence-absence matrices across the states: one for the cleaned occurrence dataset (GBIF) and another for the information retrieved from the BFG database.

For our cleaned database, we applied two different thresholds to consider if a species occurred in a particular state. The first threshold considered that the species should have at least three valid records within a particular state to be coded as an occurrence in it ("strict threshold"). The second threshold considered that only one valid record in GBIF was enough evidence for occurrence in a state ("relaxed threshold"). Conversely, the BFG matrix was built based on distribution information that is already provided by the BFG project for each species. From those matrices, we calculated a Spatial Congruence Index (SCI), as described in the following equation (Equation 1):

$$\mathrm{SCI} = \frac{1}{r} * \left(\sum_{s=1}^{r} [\mathrm{GBIF}_s - \mathrm{BFG}_s] \right) \quad (1)$$

with r being the number of states where a species s can be either present (1) or absent (0) in GBIF (given the two sets of thresholds) and in the BFG database. This gives us the following possible results for each species s in a state r (Equation 2):

$$\mathrm{SCI}_{s,r} = \begin{cases} -1, \text{if } \mathrm{GBIF}_{s,r} = 0 \text{ and } \mathrm{BFG}_{s,r} = 1 \\ 0, \text{if } \mathrm{GBIF}_{s,r} = \mathrm{BFG}_{s,r} = 1 \\ +1, \text{if } \mathrm{GBIF}_{s,r} = 1 \text{ and } \mathrm{BFG}_{s,r} = 0 \end{cases} \quad (2)$$

Hence, the closest the SCI is from 0, the more "congruent" the information on geographical distribution of a species s is between the two databases (i.e., the species s occurs in all the indicated states in both databases). Values closer to −1 indicate that the geographical range is wider in the BFG database (i.e., the occurrence of a species s in a particular state is recorded in the BFG database, but not in GBIF). Conversely, values closer to +1 indicate that the geographical range is wider in GBIF (i.e., the occurrence of a species s in a particular state is recorded in GBIF, but not in the BFG database). The SCI distribution frequencies for each database were statistically compared by applying a Kruskal-Wallis test (Hollander & Wolfe, 1973) at a significance level of 0.05 by using the R package "stats" v.3.5.3 (R Core Team, 2019).

Assessment of temporal patterns. — To untangle temporal patterns in the records, we evaluated how many species had their last collection recorded over 50 years ago in the raw database (i.e., prior to data cleaning), standardizing the "present" to 31 March 2019. This is an arbitrary proxy adopted by the IUCN (International Union for Conservation of Nature) to estimate whether a species is possibly extinct in nature (Magin & al., 1994).

Finally, we also sorted the number of records per year and critically evaluated the impact of data cleaning, considering particularly the year 1995—when GPS reached its full operational capability (Kaplan, 2006)—as a threshold for accuracy in georeferenced records. We built maps of number of records, species richness and a standard weighted endemism (see Guerin & al., 2015) to contrast quality of distribution points prior to 1995 and from 1995 onwards, using the R package "monographaR" v.1.2.0 (Reginato, 2016).

■ RESULTS

The BFG database reported 7220 names of species endemic to the Brazilian Atlantic Forest in 1000 genera. Of this total, 244 species (ca. 3%) lack any information in GBIF (even considering records with missing coordinates), characterizing the first data shortfall. The reasons for their absence in GBIF were assessed by manually searching the problematic names in GBIF portal, checking known records and the protologue of the problematic name. Three major categories within this shortfall were uncovered (Table 1). A full list of these

Table 1. Description of the main causes for null searches in 244 species with zero records in the GBIF database.

Category	Description	Species	Percent
No data	Information is not available at all, because no collections have been digitized nor made online in any repository.	111	45%
Out of date	Records of that species are present in the GBIF database, but under another name (determination "out of date").	80	33%
Misspell	Records of that species are present in the GBIF database, but the name is misspelled.	53	22%
Total		**244**	**100%**

missing species and related additional information is provided in supplementary Table S1.

After discarding these 244 problematic species from the initial BFG list of names, the raw database of distribution points (including the problematic species) had 6976 species in 987 genera, with 605,951 records (see full list of records in suppl. Table S2). After deleting records without valid or lacking coordinates, the total number of records dropped to 153,854, representing 5647 species in 908 genera, meaning that a total of 1573 species had either no valid records or, in case of the 244 problematic species, no information at all in GBIF (Fig. 1A). This represents the second and major shortfall in our database: in total, the number of species dropped ca. 22%, and the number of records ca. 75% when generating a "cleaned" database.

Temporal patterns in shortfalls of biodiversity documentation. — Most of the information lost is related to taxa known from only a few (Fig. 1B) and old collections (Fig. 1C). The proportion of deleted records after the cleaning decreases continuously towards the present, especially after the 1990s, meaning that less recent records are discarded after cleaning procedures. A summary of the information regarding each shortfall and each species is provided in supplementary Table S3.

The year 1995 appears to be a milestone in universal georeferenced data availability. Records collected before that year often consisted of non-valid records, because the GPS only reached its full operationality from that year on (Kaplan, 2006) (Fig. 1C,D). Still, before 1995, a few peaks were observed in particular years when the proportion of remaining georeferenced records were oddly higher than expected (Fig. 1C, red lines, corresponding to more than 20% of all records for the particular year remained).

An additional temporal pattern recovered from our data analyses is related to collecting efforts. From the raw database, 1843 species (26% of all species with records on GBIF and 33% of all species after the cleaning) have at least one gap of 50 years between consecutive collections, and 425 (6% from the raw list, i.e., before cleaning) have not been collected again in the past 50 years.

Spatial patterns in shortfalls of biodiversity documentation. — In terms of species richness, most taxa are centered in specific portions of the Atlantic Forest, especially in Southern Bahia and in the Serra do Mar region, with the richest areas in Espírito Santo and Rio de Janeiro States (Fig. 2). The overall patterns of species richness and endemism are not different considering records prior to 1995 or from 1995 onwards, and all maps are highly correlated (standard correlation indexes all greater than 0.8 pairing records prior to 1995, 1995 onwards and all records—see full list in supplementary Table S2).

The increase in collections with valid and accurate georeferences from 1995 onwards causes a big impact in the number of records that remain after data cleaning (as observed in Fig. 1A). In fact, the great majority of species have most or all of their distribution points with valid coordinates from 1995 onwards (Fig. 3A). Interestingly, however, quality and quantity of remaining records prior to 1995 and from 1995 onwards do not significantly alter the overall patterns of species distribution. Congruence between databases (Fig. 3B) and general patterns of species richness or endemism per area (Fig. 2) are both similar when considering records from only one of these two time frames.

Also, for the remaining records, the average SCI per species was close to 0 for the two thresholds. This indicates that spatial information of species converges between the present in GBIF and the reported in the BFG database. Visually, the distribution frequency pattern of the SCI per species appears the same for both time frames (Fig. 3B); but this is not corroborated statistically at a 0.05 significance level ($p < 0.001$ according to the Kruskal-Wallis test). See Supporting Information for all values (suppl. Table S4 for the acquired distribution and suppl. and Table S5 for the SCI calculations).

■ DISCUSSION

The possibility of cross-checking species information carefully prepared by taxonomists (BFG database) with a large repository of occurrence records of GBIF provides novel perspectives for biodiversity studies, as well as putative new issues associated with this information. DAK repositories are infamous for their inaccurate taxonomic and spatial data (Graham & al., 2004; Yesson & al., 2007; Beck & al., 2013; Robertson & al., 2016), but some biases recovered in here were not anticipated when we first started our study. Our cross-validation found that ca. 3% of all species listed as endemic to the Atlantic Forest (244 species) are absent in GBIF, while other ca. 18% have no records with valid coordinates. These alarming results indicate that over one-fifth of all angiosperms endemic to the Atlantic Forest have some sort of documentation issue in their spatial data, hampering conservation policies and biogeographical and macroecological analyses.

Reasons for missing species in GBIF. — Our results suggest some potential reasons for these caveats in our data. Our manual inspection of the list of missing species revealed some of the following: (1) misspelled names (i.e., typos), (2) lag in updating determinations in the repository, and (3) truly missing information, as summarized in Table 1 (but see also suppl. Table S1). Misspelled names are somewhat common and expected in any huge database, given that the list of names is primarily provided by humans. However, they hamper communication between databases, resulting in loss of information when they are cross-checked. In our study, typos are present in the reference database of the BFG and would have to be manually corrected by the taxonomists working on this project.

The lag in updating taxonomic determinations represents another important bias. Several species that were not found in GBIF are actually in there, but identified under another name. This lag probably reflects the fact that some herbaria take some time to process new information as they update taxonomical determinations (Robertson & al., 2014). But physical collections are affected as well: in Brazil, processing changes can take from two months (e.g., at the RB Herbarium; R.C. Forzza, pers.

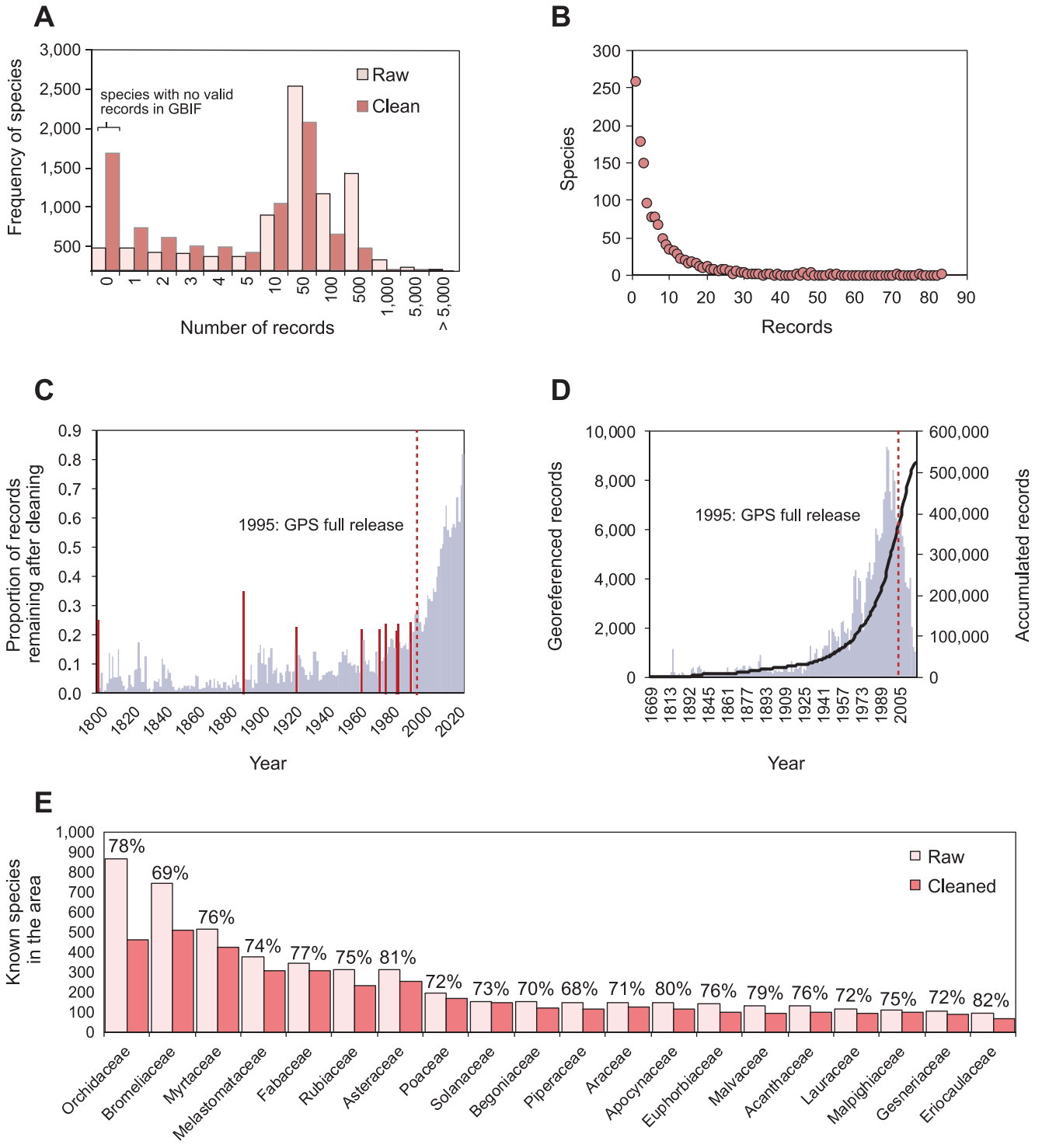


Fig. 1. Evaluation of temporal and taxonomic patterns in big-data shortfalls for documentation of endemic angiosperm species in the Brazilian Atlantic Forest. Raw: raw database, before data cleaning; Clean: cleaned database, after the cleaning steps, with only valid records. **A**, Frequency histogram of records per species, highlighting species lacking any valid records before and after the cleaning. **B**, Number of records for species discarded during cleaning procedures, showing that most discarded species are represented by less than 10 collections. **C**, Proportion of records remaining after cleaning. Note that some years have oddly high proportions of georeferenced records for the period preceding the GPS full release; those with values above 0.2 are highlighted (red bars) and represent older *a priori* georeferenced collections. Also note the relative increase of georeferenced (valid) records from 1995 onwards. **D**, Georeferenced records per year (gray bars) from the raw database and accumulated frequency (thick black curve) of all records, i.e., georeferenced or not. **E**, Frequency of species before and after data cleaning for the twenty species-richest families in the Atlantic Forest according to the BFG database. Numbers above each bar indicate the percent of discarded records after the cleaning.

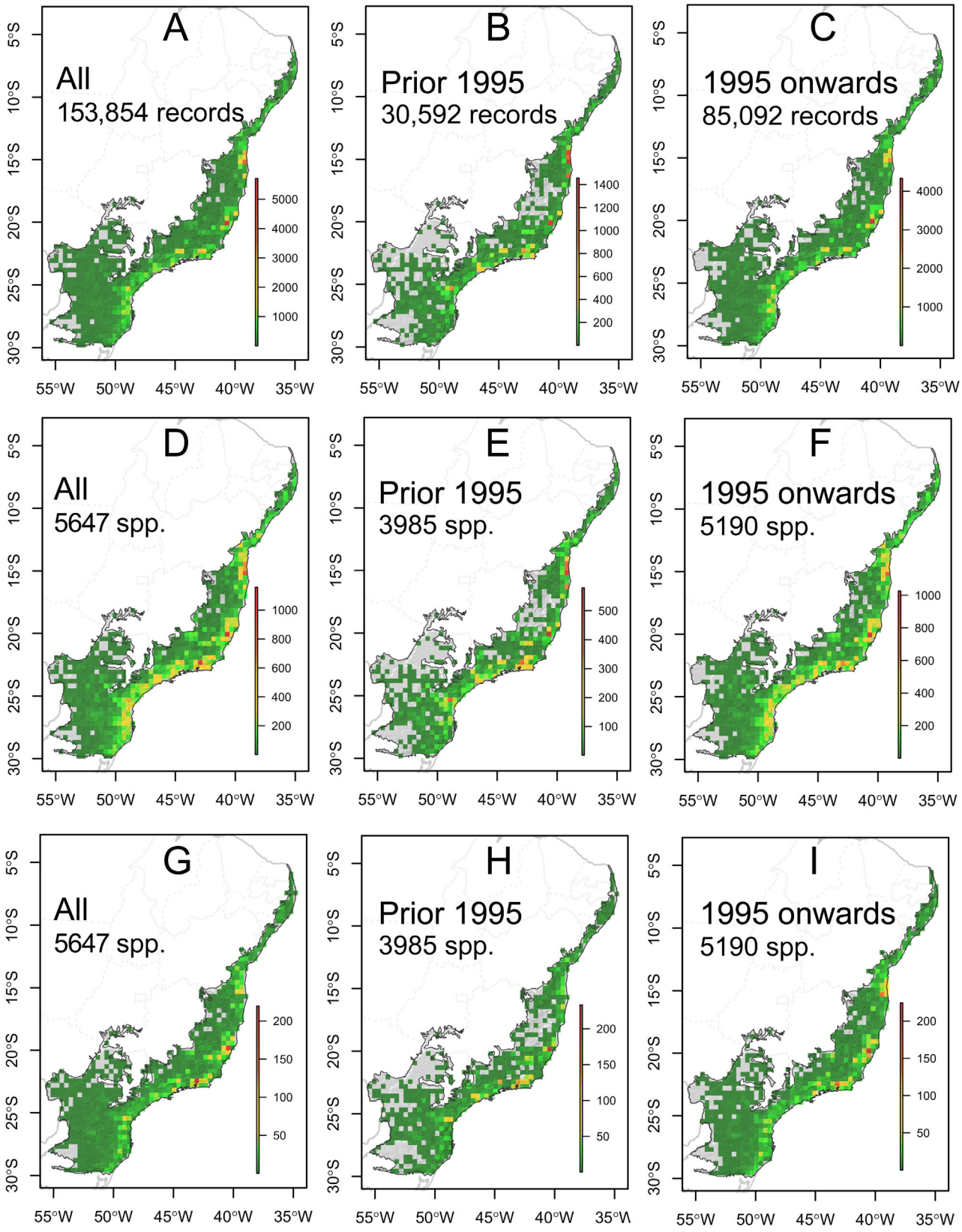


Fig. 2. Number of records (**A** to **C**), species richness (**D** to **F**), and weighted endemism (**G** to **I**) in the Brazilian Atlantic Forest, considering all records (**A**, **D**, **G**), only records prior 1995 (**B**, **E**, **H**) and records from 1995 onwards (**C**, **F**, **I**). Grids of 0.375°.

obs., 2019) up to two years (e.g., at the SPF Herbarium; V.Y. Jono, pers. comm., 2019). Conversely, similar "lags" may be even longer in smaller herbaria, where funding and infrastructure to process and update these changes tend to be minimal (Mann, 1997). This is worrying, as keeping a well-curated and up-to-date collection is mandatory in the era of DAK repositories, and the herbaria continue to be a fundamental source of information for studies at different scales.

Finally, "missing data" is the most common category of the 244 missing species. We identified two main reasons for their absence in GBIF. First, these represent names with nomenclatural issues, e.g., names that require new combinations or are cases for lectotypifications. Some call particular attention: for instance, several are names that were described by Frei José Mariano da Conceição Vellozo or by João Barbosa Rodrigues in *Flora Fluminensis* (Vellozo, 1825), from which types are mostly illustrations and, therefore, cannot be found in GBIF.

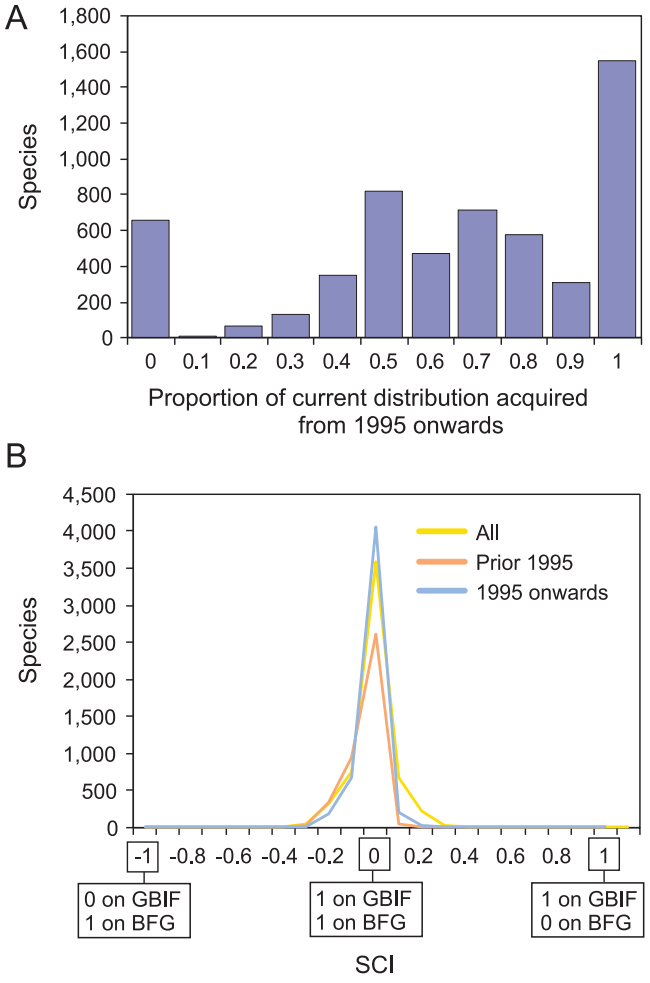


Fig. 3. Spatial congruence analyses of record data from endemic angiosperm species in the Brazilian Atlantic Forest. **A**, Frequency histogram of the proportion of current distribution inferred per species from collections from 1995 onwards. **B**, Frequency histograms of the average Spatial Congruence Index (SCI) for all species of the study area, considering the "strict threshold". The graph for the "lax threshold"' was not represented because the pattern is similar to the "strict threshold".

Second, these can be associated to a few cases where the GBIF taxonomical backbone somehow fails to properly associate data retrieved from regional repositories to a name in GBIF. Some examples in our list include *Guatteria capixabae* Lobão & J.C.Lopes and *Tradescantia atlantica* M.Pell.—none of these have records in GBIF but should have, as there are records in the Rio de Janeiro Botanical Garden database (respectively, 17 and 4 records for *G. capixabae* and *T. atlantica*).

Reasons for non-valid records. — All of the reasons above can explain why species names reported by taxonomists in the BFG database are not in GBIF. However, the major shortfall in GBIF, responsible for dropping 22% of species and 75% of records during the cleaning steps, is related to the high number of records with "lacking coordinates" (i.e., those marked as "0" or "NA" in latitude and longitude). The large amount of data lost in this step can be due to two main reasons: (1) problems in how the data was captured, because the coordinates may be available on the specimen label, but have not been digitized during databasing; or, more commonly, (2) the specimen label has no coordinates at all.

This acute shortfall during data cleaning was particularly severe for Orchidaceae and Bromeliaceae species (Fig. 1E and suppl. Table S3). Possible reasons for that are either (1) biological, because of the high proportion of micro-endemics and rare species (Meirelles & al., 1999; Verola & al., 2007; Menini-Neto & Forzza, 2012); or (2) related to documentation, since their ornamental value make them relatively more abundant in collections of small and private herbaria that have not been fully digitized or that have not been submitted to GBIF; and possibly (3) a combination of both.

This emphasizes the importance in databasing collections of smaller, regional herbaria. As highlighted by Williams & Crouch (2017), rarer species are often found only in local herbaria and not in larger collections. On the other hand, there are also some large Brazilian museums, with rich collections of Bromeliaceae and Orchidaceae from the Atlantic Forest, that still have not been fully digitized. These include the National Museum Herbarium, Rio de Janeiro (R) and the Herbarium Bradeanum, Rio de Janeiro (HB)—all holding historical collections of naturalists such as Elton Leme and Guido F.J. Pabst, who collected hundreds of specimens of bromeliads and orchids in the Atlantic Forest during the 20th century.

In summary, standard steps of data cleaning can remove up to three-quarters of all records, leading to an underestimation of over one-fifth of the species-richness of endemic angiosperms in the Brazilian Atlantic Forest. These are alarming results, as they indicate that somehow we are not managing to fully document spatial information of critically important groups in DAK repositories. Thus, care must be taken particularly regarding rare species, as those may be easily left aside during standard data cleaning procedures.

Even though species only known from a few old collections can possibly represent taxa that have become extinct from nature, there are frequent cases of "re-discoveries" (e.g., Pellegrini & Almeida, 2016; Bochorny & al., 2017; Lírio & al., 2018).

Thus, these missing species should be targeted in future collecting expeditions to truly assess their conservation status and also to improve spatial data in DAK repositories.

Spatial patterns and data quality of the remaining records. — Even though a large number of records and species was discarded during data cleaning, plenty of data still remained afterwards. Remaining records are widely spread over the Atlantic Forest (Fig. 2). The overall distribution pattern of species richness and endemism at different portions of the Atlantic Forest is corroborated by further evidence from particular groups of plants and animals (e.g., Cracraft, 1985; Murray-Smith & al., 2009; DaSilva & al., 2017; Colli-Silva & Pirani, 2019). Moreover, when we split and analyze distribution data in two distinct temporal frames—prior to 1995 and from 1995 onwards—, the overall pattern of species richness is unchanged. In other words, even if we considered only old records with approximated spatial georeferences, we would still have a fair approximation of the overall pattern of richness in the Atlantic Forest, at least at this macro-scale.

Furthermore, old records alone can already provide a good estimation of major patterns of biodiversity, which emphasizes their relevance. We showed that data acquired from 1995 onwards significantly increased information on species distribution (Fig. 3A), and 1509 species have only records with valid coordinates collected after this year. These results are fortunate, because they show that accurate georeferencing and continuous field expeditions have enabled us to expand our understanding on the distribution of many taxa. Thus, even though the overall distribution pattern is unchanged between time frames, a continuous collecting effort is still important to fully understand species distribution, improving diversity patterns at finer scales.

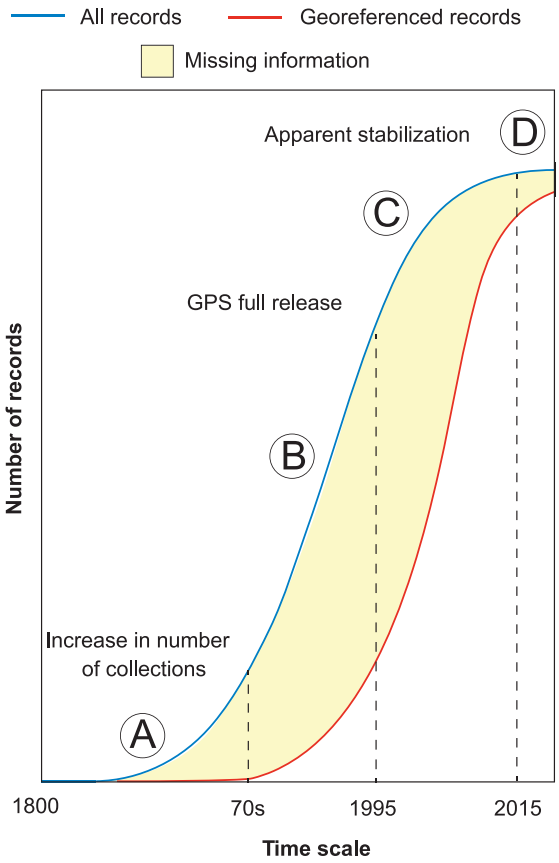
However, imprecisions related to these old collections can also bias estimates of richness and endemism even after data cleaning, and may have negative effects in finer-scale surveys. Examples in our dataset are observed in the year 1900, in which we found only records from the JPB Herbarium and all of them had the same coordinates. These coordinates were not country or state centroids, so they were not discarded during automatized cleaning, but have erroneously included 57 more species in the corresponding cell.

In summary, it is common to refer to GBIF as "dubious" or "non reliable" (Yesson & al., 2007; Beck & al., 2013; Robertson & al., 2014), but it seems that, at least in terms of spatial data, a major issue with this and other DAK repositories is not the data itself, but the enormous amount of information that is discarded after data cleaning. Such lack of information, as we detail in the next section, may have huge implications in theoretical studies of spatial analyses (e.g., Schmidt & al., 2005; Beaman & Cellinese, 2012; Maldonado & al., 2015) and in applied studies aiming to deliver full conservation assessments for threatened species (e.g., Rodrigues & al., 2006; Bachman & al., 2011).

Shortfall patterns in DAK repositories. — There are several biodiversity shortfalls, or limitations in documenting biodiversity. For instance, the "Linnean shortfall" relates to the still undescribed biodiversity, i.e., taxa that occur in an area but are still unknown to science, while the "Wallacean shortfall" refers to the non-complete understanding of the geographical range of a species (Beck & al., 2014; Hortal & al., 2015). Here, perhaps we are not dealing with a shortfall *per se*, but missing coordinates directly strengthens the Wallacean shortfall and underestimates our knowledge for the distribution of the flora.

These gaps in our knowledge of biodiversity resulting from the necessary cleaning of inaccurate data associated with large databases are summarized in Fig. 4. This hampers our full comprehension of species distribution, especially rare taxa with scarce, old collections. In the case of the Atlantic Forest, we have demonstrated that over 1300 species were either absent or removed from the database during simple cleaning procedures.

Such limitation seems to be less acute in recent collections, and two main events may have contributed for that: (1) the development of electronic catalogues and voucher



(A) Few collections in absolute numbers

(B) Number of collections increase, but many present inaccurate georeferences

(C) Number of collections with accurate georeferences increases

(D) Time taken to update collections in big-data repositories

Fig. 4. Four phases in the documentation of biodiversity through time, highlighting the described "missing information" for distribution data.

digitization in the 1970s (B in Fig. 4) (Graham & al., 2004) and (2) the GPS full release and operational capability (C in Fig. 4) (Kaplan, 2006). However, after an outstanding progress in the last 40 years, the accumulation of online records seems to have slowed down in the past few years (D in Fig. 4). This is perhaps due to the "lag" in updating data repositories, or to a potential decrease of field expeditions and, consequently, of new collections and records (see Ríos-Saldaña & al., 2018; Daly & Martinez-Habibe, 2019).

Good practices to move forward. — We believe there are some good practices that, if adopted by collectors and herbaria, could safeguard new data, lessening the effect of accumulating non-valid spatial information. First: even when analyzing only collections from 1995 onwards, ca. 20% of the vouchers still had no valid coordinates at all (Fig. 1C). As GPS devices are now universalized and more accessible, such proportion should be nearing zero; so, it is worrying to know that this issue still persists. Thus, the first good practice is to always try to take precise GPS coordinates during field expeditions; otherwise, the specimen might be less valuable for any subsequent study of geographical distribution or conservation assessment.

Another practice that could lessen the loss of spatial data is georeferencing all the vouchers by a gazetteer or by the locality informed on the label. Unfortunately, to do this in a mass-scale in Brazil is now unfeasible and unpractical given operational and financing issues (Cai & Zhu, 2015; Zamudio & al., 2018). Nevertheless, this is commonly done by taxonomists who are working with particular groups, as they carefully revisit each distributional record. Researchers, however, often keep these revisited distribution data to themselves until a monograph is published.

We argue that researchers should publish their geographical databases as datapaper whenever possible (e.g., Costello & al., 2013), indexing the records and suggesting updated coordinates to those they manually georeferenced once. Each voucher could be properly updated in GBIF for further reuse then. Also, curators could consider adding additional labels informing revisited or estimated georeferencing, as they do with determination labels. Automatized image recognition by machine learning may be another way to produce valuable spatial data in a more efficient way, too (Collins & al., 2018; Lorieul & al., 2019). We believe future investment and production of new georeferencing frameworks using one of the suggestions presented above might help to improve biodiversity big-data, increasing its valuability.

To conclude, we emphasize that increasing funding for collections and museums can still have plenty of applications for biology and society, especially since many people use distribution data from biological collections (Funk, 2003; Suarez & Tsutsui, 2004; Wen & al., 2015). However, biological collections are now at risk across the globe due to decline in funding and shifts in scientific interests (Ríos-Saldaña & al., 2018; Zamudio & al., 2018). In Brazil, increased devaluation of science and funding instability towards maintaining biodiversity collections and accomplish new field expeditions can also intensify this shortfall (Ríos-Saldaña & al., 2018; Zamudio & al., 2018). We advocate that maintaining and funding biodiversity collections and field expeditions contribute to the diminishment of shortfalls in DAK, towards an effective documentation, evaluation and conservation of our threatened flora. We also stress that our study focused mainly on the spatial accuracy of data retrieved from DAK repositories. However, increased funding in biodiversity collections and curation would very likely help solving other relevant issues with DAK repositories, such as taxonomic misidentifications (e.g., Goodwin & al., 2015).

■ AUTHOR CONTRIBUTIONS

MCS and TNCV conceived the original idea of the study; MCS, MR, TNCV and AC performed the analyses; MC and TNCV wrote the first versions of the manuscript and MR, RCF and JRP reviewed and helped to write the discussion and the final version of the manuscript. — MCS, https://orcid.org/0000-0001-7130-3920; MR, https://orcid.org/0000-0002-3511-6586; AC, https://orcid.org/0000-0003-4330-0946; RCF, https://orcid.org/0000-0002-7035-9313; JRP, https://orcid.org/0000-0001-7984-4457; TNCV, https://orcid.org/0000-0001-9991-7924

■ ACKNOWLEDGMENTS

We are grateful to the following Brazilian funding agencies for maintaining the authors' research: FAPESP (the São Paulo Research Foundation, Grant IDs: 17/19295-1, 17/09447-9 and 18/02191-1); CAPES (Coordination of Improvement of Higher Education Personnel), for maintaining MCS & AC on the institution in which they are enrolled; FAPERJ (the Rio de Janeiro Research Foundation, Grant ID E-26/202.778/2018), which provided research grants to RCF through the "Programa Cientistas do Nosso Estado"; and CNPq (National Council for Scientific and Technological Development) for researcher grants to JRP and RCF. We also thank all the Brazilian taxonomists for maintaining the Brazilian Flora 2020 Project active, as well as herbarium curators for their efforts in maintaining the collections and the data properly available. Finally, we also would like to thank V.Y. Jono for her support in relating common technical issues in her work of digitization processing that have improved the discussion of this paper.

■ LITERATURE CITED

Bachman, S., Moat, J., Hill, A.W., 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

Beaman, R.S. & Cellinese, N. 2012. Mass digitalization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys* 209: 7–17. https://doi.org/10.3897/zookeys.209.3313

Beck, J., Ballesteros-Meija, L., Nagel, P. & Kitching, I.J. 2013. Online solutions and the 'Wallacean shortfall': What does GBIF contribute to our knowledge of species' ranges? *Diversity & Distrib.* 19: 1043–1050. https://doi.org/10.1111/ddi.12083

Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. 2014. Spatial bias in the GBIF and its effect on modelling species' geographic distributions. *Ecol. Informatics* 19: 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

BFG (The Brazil Flora Group) 2018a. Brazilian Flora 2020: Innovation and collaboration to meet Target 1 of the Global Strategy for Plant Conservation (GSPC). *Rodriguésia* 69: 1513–1527. https://doi.org/10.1590/2175-7860201869402

BFG (The Brazil Flora Group) 2018b. Brazilian Flora 2020 Project – Projeto Flora do Brasil 2020. Version 393.173. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Checklist database. https://doi.org/10.15468/1mtkaw (accessed via GBIF.org, 3 Jan 2019).

Bochorny, T., Bacci, L.F. & Goldenberg, R. 2017. Following Glaziou's footsteps: Rediscovery and updated description of three species of *Behuria* Cham. (Melastomataceae) from the Atlantic Forest (Brazil). *Phytotaxa* 302: 229–240. https://doi.org/10.11646/phytotaxa.302.3.2

Buerki, S. & Baker, W.L. 2016. Collections-based research in the genomic era. *Biol. J. Linn. Soc.* 117: 5–10. https://doi.org/10.1111/bij.12721

Cai, L. & Zhu, Y. 2015. The challenges of data quality and data quality assessment in the big data era. *Data Sci. J.* 14: 2. https://doi.org/10.5334/dsj-2015-002

Carvalho, G. 2017. flora: Tools for interacting with the Brazilian Flora 2020. R package version 0.3.0. https://CRAN.R-project.org/package=flora

Collins, M., Yeole, G., Frandsen, P., Dikow, R., Orli, S. & Figueiredo, R. 2018. A pipeline for deep learning with specimen images in iDigBio – Applying and generalizing an examination of mercury use in preparing herbarium specimens. *Biodivers. Inform. Sci. & Standards* 2: e25699. https://doi.org/10.3897/biss.2.25699

Colli-Silva, M. & Pirani, J.R. 2019. Biogeographic patterns of Galipeinae (Galipeeae, Rutaceae) in Brazil: Species richness and endemism at different latitudes of the Atlantic Forest "hotspot". *Flora* 251: 77–87. https://doi.org/10.1016/j.flora.2019.01.001

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z. & Bourne, P.E. 2013. Biodiversity data should be published, cited, and peer reviewed. *Trends Ecol. Evol.* 28: 454–461. https://doi.org/10.1016/j.tree.2013.05.002

Cracraft, J. 1985. Historical biogeography and patterns of differentiation within the South American avifauna: Areas of endemism. *Ornithol. Monogr.* 36: 49–84.

Daly, D.C. & Martinez-Habibe, M.C. 2019. Ten new species of *Dacryodes* from Amazonia and the Guianas. Studies in neotropical Burseraceae XXIII. *Brittonia* 71: 201–224. https://doi.org/10.1007/s12228-018-09564-7

Daly, D.C. & Prance, G.T. 1989. Brazilian Amazon. Pp. 401–426 in: Campbell, D.G. & Hammond, H.D. (eds.), *Floristic inventory of tropical countries: The status of plant systematics, collections, and vegetation, plus recommendations for the future*. New York: The New York Botanical Garden.

DaSilva, M.B., Pinto-da-Rocha, R. & Morrone, J.J. 2017. Historical relationships of areas of endemism of the Brazilian Atlantic rain forest: A cladistic biogeographic analysis of harvestman taxa (Arachnida: Opiliones). *Curr. Zool.* 63: 525–535. https://doi.org/10.1093/cz/zow092

Forzza, R.C., Baumgratz, J.F.A., Bicudo, C.E.M., Canhos, D.A.L., Carvalho, A.A., Jr., Coelho, M.A.N., Costa, A.F., Costa, D.P., Hopkins, M.G., Leitman, P.M., Lohmann, L.G., Lughadha, E.N., Maia, L.C., Martinelli, G., Menezes, M., Morim, M.P., Peixoto, A.L., Pirani, J.R., Prado, J., Queiroz, L.P., Souza, S., Souza, V.C., Stehmann, J.R., Sylvestre, L.S., Walter, B.M.T. & Zappi, D.C. 2012. New Brazilian floristic list highlights conservation challenges. *BioScience* 62: 39–45. https://doi.org/10.1525/bio.2012.62.1.8

Funk, V.A. 2003. 100 Uses of an Herbarium: Well at least 72. *A. S. P. T. Newslett.* 17: 17–19.

Galindo-Leal, C. & Câmara, I.G. 2003. Atlantic Forest hotspot status: an overview. Pp. 3–11 in: Galindo-Leal, C. & Câmara, I.G. (eds.), *The Atlantic Forest of South America: Biodiversity status, threats, and outlook*. Washington, D.C.: Center for Applied Biodiversity Science and Island Press.

GBIF.org 2019. GBIF Occurrence Download. https://doi.org/10.15468/dl.uonz9q (12 Mar 2019).

Giaretta, A., Menezes, L.F.T. & Peixoto, A.L. 2015. Diversity of Myrtaceae in the southeastern Atlantic Forest of Brazil as a tool for conservation. *Brazil. J. Bot.* 38: 175–185. https://doi.org/10.1007/s40415-014-0121-y

Goodwin, Z.A., Harris, D.J., Filer, D., Wood, J.R.I. & Scotland, R.W. 2015. Widespread mistaken identity in tropical plant collections. *Curr. Biol.* 25: 1066–1067. https://doi.org/10.1016/j.cub.2015.10.002

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

Guerin, G.R., Ruokolainen, L. & Lowe, A.J. 2015. A georeferenced implementation of weighted endemism. *Methods Ecol. Evol.* 6: 845–852. https://doi.org/10.1111/2041-210X.12361

Hollander, M. & Wolfe, D.A. 1973. *Nonparametric statistical methods*. New York: John Wiley & Sons.

Hopkins, M.J.G. 2007. Modelling the known and unknown plant biodiversity of the Amazon Basin. *J. Biogeogr.* 34: 1400–1411. https://doi.org/10.1111/j.1365-2699.2007.01737.x

Hortal, J., Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Rev. Ecol. Evol. Syst.* 46: 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Joppa, L.N., Roberts, D.L., Myers, N. & Pimm, S.L. 2011. Biodiversity hotspots house most undiscovered plant species. *Proc. Natl. Acad. Sci. U.S.A.* 108: 13171–13176. https://doi.org/10.1073/pnas.1109389108

Kaplan, E. 2006. Introduction. Pp. 1–19 in: Kaplan, E. & Hegarty, C. (eds.), *Understanding GPS: Principles and applications*, 2nd ed. Boston & London: Artech House Publishers.

Lírio, E.J., Freitas, J., Negrão, R., Martinelli, G. & Peixoto, A.L. 2018. A hundred years' tale: Rediscovery of *Mollinedia stenophylla* (Monimiaceae) in the Atlantic rainforest, Brazil. *Oryx* 52: 437–441. https://doi.org/10.1017/S0030605316001654

Lorieul, T., Pearson, K.D., Ellwood, E.R., Goëau, H., Molino, J., Sweeney, P.W., Yost, J.M., Sachs, J., Mata-Montero, E., Nelson, G., Soltis, P.S., Bonnet, P. & Joly, A. 2019. Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical and equatorial floras. *Applic. Pl. Sci.* 7: e01233. https://doi.org/10.1002/aps3.1233

Magin, C.D., Johnson, T.H., Groombridge, B., Jenkins, M. & Smith, H. 1994. Species extinctions, endangerment and captive breeding. Pp. 3–32 in: Olney, P.J.S., Mace, G.M. & Feistner, A.T.C. (eds.), *Creative conservation: Interactive management of wild and captive animals*. Salisbury: Springer.

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 Ecol. Biogeogr.* 24: 973–984. https://doi.org/10.1111/geb.12326

Mann, D.G. 1997. The economics of botanical collections. Pp. 68–82 in: Nudds, J.R. & Pettitt, C.W. (eds.), *The value and valuation of natural science collections: Proceedings of the International Conference, Manchester, 1995*. London: Geological Society of London.

Meirelles, S.T., Pivello, V.R. & Joly, C.A. 1999. The vegetation of granite rock outcrops in Rio de Janeiro, Brazil, and the need for its protection. *Environm. Conservation* 26: 10–20.

Menini-Neto, L. & Forzza, R.C. 2012. Biogeography and conservation status assessment of *Pseudolaelia* (Orchidaceae). *Bot. J. Linn. Soc.* 171: 191–200. https://doi.org/10.1111/j.1095-8339.2012.01304.x

Morellato, L.P.C. & Haddad, C.F.B. 2000. Introduction: The Atlantic Forest. *Biotropica* 32: 786–792.

Mori, S.A. 1989. Eastern, extra-Amazonian Brazil. Pp. 427–454 in: Campbell, D.G. & Hammond, H.D. (eds.), *Floristic inventory of tropical countries: The status of plant systematics, collections, and vegetation, plus recommendations for the future*. New York: The New York Botanical Garden.

Murray-Smith, C., Brummitt, N.A., Oliveira-Filho, A.T., Bachman, S., Moat, J., Lughadha, E.M. & Lucas, E.J. 2009. Plant diversity hotspots in the Atlantic Coastal forests of Brazil. *Conservation Biol.* 23: 151–163. https://doi.org/10.1111/j.1523-1739.2008.01075.x

Myers, N., Mittermeier, R.A., Mittermeier, C.G., Fonseca, G.A.B. & Kent, J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403: 853–858. https://doi.org/10.1111/2041-210X.13152

Oliveira, U., Soares-Filho, B.S., Paglia, A.P., Brescovit, A.D., Carvalho, C.J.B., Silva, D.P., Rezende, D.T., Leite, F.S.F., Batista, J.A.N., Barbosa, J.P.P.P., Stehmann, J.R., Ascher, J.S., Vasconcelos, M.F., Marco, P., Löwenberg-Neto, P., Ferro, V.G. & Santos, A.J. 2017. Biodiversity conservation gaps in the Brazilian protected areas. *Sci. Rep.* 7: a9141. https://doi.org/10.1038/s41598-017-08707-2

Oliveira, U., Soares-Filho, B.S., Santos, A.J., Paglia, A.P., Brescovit, A.D., Carvalho, C.J.B., Silva, D.P., Rezende, D.T., Leite, F.S.F., Batista, J.A.N., Barbosa, J.P.P.P., Stehmann, J.R., Ascher, J.S., Vasconcelos, M.F., Marco, P., Löwenberg-Neto, P. & Ferro, V.G. 2019. Modelling highly biodiverse areas in Brazil. *Sci. Rep.* 9: a6355. https://doi.org/10.1038/s41598-019-42881-9

Pellegrini, M.O.O. & Almeida, R.F. 2016. Rediscovery, identity and typification of *Dichorisandra picta* (Commelinaceae) and comments on the short-stemmed *Dichorisandra* species. *Phytotaxa* 245: 107–118. https://doi.org/10.11646/phytotaxa.245.2.2

Pires-O'Brien, M.J. 1993. An essay on the history of natural history in Brazil, 1500–1900. *Arch. Nat. Hist.* 20: 37–48.

R Core Team 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reginato, M. 2016. monographaR: An R package to facilitate the production of plant taxonomic monographs. *Brittonia* 68: 212–216. https://doi.org/10.1007/s12228-015-9407-z

Ribeiro, M.C., Metzger, J.P., Martensen, A.C., Ponzoni, F.J. & Hirota, M.M. 2009. The Brazilian Atlantic Forst: How much is left, and how is the remaining forest distributed? Implications for conservation. *Biol. Conservation* 142: 1141–1153. https://doi.org/10.1016/j.biocon.2009.02.021

Ríos-Saldaña, C.A., Delibes-Mateos, M., Ferreira, C.C. 2018. Are fieldwork studies being relegated to second place in conservation science? *Global Ecol. Conservation* 14: e00389. https://doi.org/10.1016/j.gecco.2018.e00389

Robertson, T., Döring, M., Guralknick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L. & Desmet, P. 2014. The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PLOS ONE* 9: e102623. https://doi.org/10.1371/journal.pone.0102623

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

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

Schmidt, M., Kreft, H., Thiombiano, A. & Zizka, G. 2005. Herbarium collections and field data-based plant diversity maps for Burkina Faso. *Diversity & Distrib.* 11: 509–516. https://doi.org/10.1111/j.1366-9516.2005.00185.x

Silva, L.A.E., Fraga, C.N., Almeida, T.M.H., Gonzalez, M., Lima, R.O., Rocha, M.S., Bellon, E., Ribeiro, R.S., Oliveira, F.A., Clemente, L.S., Magdalena, U.R., Medeiros, E.S. & Forzza, R.C. 2017. Jabot – Sistema de gerenciamento de coleções botânicas: A experiência de uma década de desenvolvimento e avanços. *Rodriguésia* 68: 391–410. https://doi.org/10.1590/2175-786020176820

Sousa-Baena, M.S., Garcia, L.C., Peterson, A.T. 2013. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity & Distrib.* 20: 369–381. https://doi.org/10.1111/ddi.12136

Suarez, A.V. & Tsutsui, N.D. 2004. The value of museum collections for research and society. *BioScience* 54: 66–74. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. 2017. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* 7: a9132. https://doi.org/10.1038/s41598-017-09084-6

Ulloa-Ulloa, C, Acevedo-Rodríguez, P., Beck, S., Belgrano, M.J., Bernal, R., Berry, P.E., Brako, L., Celis, M., Davidse, G., Forzza, R.C., Gradstein, R., Hokche, O., León, B., León-Yánez, S., Magill, R.E., Neill, D.A., Nee, M., Raven, P.H., Stimmel, H., Strong, M.T., Villaseñor, J.L., Zarucchi, J.L., Zuloaga, F.O. & Jørgensen, P.M. 2017. An integrated assessment of the vascular plant species of the Americas. *Science* 358: 1614–1617. http://doi.org/10.1126/science.aao0398

Vellozo, J.M.C. 1825. *Florae Fluminensis*. Flumine Januario [Rio de Janeiro]: ex Typographia Nationali. https://doi.org/10.5962/bhl.title.745

Verola, C.F., Semir, J., Antonelli, A. & Koch, I. 2007. Biosystematic studies in the Brazilian endemic genus *Hoffmannseggella* H.G. Jones (Orchidaceae: Laeliinae): A multiple approach applied to conservation. *Lankersteriana* 7: 419–422. https://doi.org/10.15517/lank.v7i1-2.19651

Wen, J., Ickert-Bond, S.M., Appelhans, M.S., Dorr, L.J. & Funk, V.A. 2015. Collections-based systematics: Opportunities and outlook for 2050. *J. Syst. Evol.* 53: 477–488. https://doi.org/10.1111/jse.12181

Williams, V.L. & Crouch, N.R. 2017. Locating sufficient plant distribution data for accurate estimation of geographic range: The relative value of herbaria and other sources. *S. African J. Bot.* 109: 116–127. https://doi.org/10.1016/j.sajb.2016.12.015

Willis, K. 2017. *The state of world's plants report – 2017*. Richmond: Royal Botanic Gardens, Kew.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A. & Culham, A. 2007. How global is the Global Biodiversity Information Facility? *PLOS ONE* 2: e1124. https://doi.org/10.1371/journal.pone.0001124

Zamudio, K.R., Kellner, A., Serejo, C., Britto, M.R., Castro, C.B., Buckup, P.A., Pires, D.O., Couri, M., Kury, A.B., Cardoso, I.A., Monné, M.L., Pombal, J., Patiu, C.M., Padula, V., Pimenta, A.D., Ventura, C.R.R., Hajdu, E., Zanol, J., Bruna, E.M., Fitzpatrick, J. & Rocha, L.A. 2018. Lack of science support fails Brazil. *Science* 361: 1322–1323. https://doi.org/10.1126/science.aav3296

Zizka, A., Steege, H., Pessoa, M.C. & Antonelli, A. 2018. Finding needles in the haystack: Where to look for rare species in the American tropics. *Ecography* 41: 321–330. https://doi.org/10.1111/ecog.02192

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C.D., 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. *Meth. Ecol. Evol.* 10: 744–751. https://doi.org/10.1111/2041-210X.13152