bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Defining endemism levels for biodiversity conservation: tree species in the Atlantic Forest hotspot

Renato A. F. Lima[1], Vinícius C. Souza[2], Marinez F. Siqueira[3] & Hans ter Steege[1,4]

1 Biodiversity Dynamics group, Naturalis Biodiversity Center, Leiden, The Netherlands. Email: renato.lima@naturalis.nl

2 Departamento de Ciências Biológicas, Escola Superior de Agricultura 'Luiz de Queiroz' – Universidade de São Paulo, Piracicaba, Brazil. Email: vcsouza@usp.br

3 Diretoria de Pesquisas Científicas, Jardim Botânico do Rio de Janeiro, Rio de Janeiro, Brazil. Email: marinez@jbrj.gov.br

4 Systems Ecology, Free University, De Boelelaan 1087, Amsterdam, 1081 HV, Netherlands. Email: hans.tersteege@naturalis.nl

Running title: Endemism levels for conservation

Number of words in the abstract: 150

Number of words in the manuscript: 2849

Number of references: 39

Number of figures and tables: 3

Name and mailing address: Renato A. F. Lima, Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, The Netherlands. Telephone: +31 (0)71 751 9272. Email: renato.lima@naturalis.nl

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Endemic species are essential for setting conservation priorities. Yet, quantifying endemism remains challenging because endemism concepts can be too strict (*i.e.* pure endemism) or too subjective (*i.e.* near endemism). We use a data-driven approach to objectively estimate the proportion of records outside a given the target area (*i.e.* endemism level) that optimizes the separation of near-endemics from non-endemic species. Based on millions of herbarium records for the Atlantic Forest tree flora, we report an updated checklist containing 5062 species and compare how species-specific endemism levels match species-specific endemism accepted by taxonomists. We show that an endemism level of 90% delimits well near endemism, which in the Atlantic Forest revealed an overall tree endemism ratio of 45%. The diversity of pure and near endemics and of endemics and overall species was congruent in space, reinforcing that pure and near endemic species can be combined to quantify endemism to set conservation priorities.

Key-words: biodiversity hotspot, endemism centres, endemism ratio, near endemism, occasional species, plant conservation, species richness

1 INTRODUCTION

In times of increasing threats to biodiversity and limited resources for its conservation, prioritizing actions is essential. One common practice in biodiversity conservation is to target areas with many flagship species, such as species threatened with extinction (*i.e.* threatened species) or those exclusive to a given region or habitat (*i.e.* endemic species). These flagship species are important for conservation because they have a greater extinction risk than other species (Brooks et al., 2006; Myers et al., 2000; Peterson & Watson, 1998). In addition, patterns of total and endemic species richness can be

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

congruent (Kier et al., 2009; Letters et al., 2002; Storch, Keil, & Jetz, 2012), so the protection of high-endemism areas could also safeguard the remaining biodiversity. However, there have been more efforts to delimit threatened species than endemic ones. Threatened species are grouped by clearly-defined categories enclosed by objective criteria (IUCN, 2018), while species often are classified simply as being endemic or not.

There are proposals to divide endemics species based on spatial scale (*e.g*. narrow, regional and continental endemics), evolutionary history (*e.g*. neo and paleo endemics) or habitat specificity (e.g. edaphic endemics; Ferreira & Boldrini, 2011; Kruckeberg & Rabinowitz, 1985; Peterson & Watson, 1998). These proposals, however, implicitly assume that all individuals of a species are confined to a given region or habitat, also known as true or pure endemism (Tyler, 1996). If one record is found outside the target region, the species is to be (re)classified as non-endemic. Since pure endemism is rather strict, the term near-endemism has been used to describe species with few records outside the target region (Carbutt & Edwards, 2006; Matthews, van Wyk, & Bredenkamp, 1993; Noroozi et al., 2018; Platts et al., 2011). Near-endemics are the result rare dispersal events, temporary establishment in different habitats or the existence small satellite populations (Matthews, van Wyk, & Bredenkamp, 1993; Perera, Ratnayake-Perera, & Procheş, 2011).

The differentiation between pure and near endemics is challenging, because it may not be stable in time: near endemics can become pure endemics if habitat loss is higher outside than inside the target region (Carbutt & Edwards, 2006). Conversely, pure endemics may become near endemics with the accumulation of knowledge on their geographical distribution (Werneck et al., 2011). This is particularly true for geographically-restricted species, which often have scarce occurrence data. Furthermore, pure endemics may be classified as near endemics due to species

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

misidentifications (Carbutt & Edwards, 2006) or by a questionable delimitation of the target region (Platts et al., 2011). In practice, conservation aims at protecting as many individuals as possible for a given species (IUCN, 2018). So, the differentiation between pure and near endemism may have little impact to plan conservation actions. Therefore, the practical question is: how to distinguish both groups of endemic species from non-endemic species? Defining pure endemism is straightforward, but separating near-endemics from non-endemic species can be quite subjective.

Here we establish objective limits to separate near-endemic from non-endemic species for conservation purposes. We also separate widespread species from occasional species, *i.e.*, species common in other regions but sporadic in the target region (Barlow et al., 2010). We use a data-driven approach to estimate what ratio of occurrences inside a target region could be used to separate species into pure-endemics, near-endemics, widespread and occasional species. We perform this evaluation for over 5000 tree species from the Atlantic Forest, a biodiversity hotspot with abundant knowledge on the taxonomy and distribution of its flora. Using millions of carefully curated occurrences from over 500 collections around the world, we evaluate which ratio of occurrences inside the Atlantic Forest match species-specific endemism accepted by taxonomic experts. Finally, we delimit centres of diversity for pure- endemics, near-endemics and occasional species and discuss the implications for the conservation of this biodiversity hotspot.

2 METHODS

2.1 Retrieval and validation of occurrence data

The Atlantic Forest covers Argentina, Brazil and Paraguay, so we searched for occurrences using a list of tree names for South America (see Supporting Information

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for more details), compiled from different sources (Grandtner & Chevrette, 2013; Lima et al., 2015; Oliveira-Filho, 2010; ter Steege et al., 2016; Zappi et al., 2015; Zuloaga, Morrone, & Belgrano, 2008). We carefully inspected the list of names to avoid the inclusion of exotic and non-arborescent species. Arborescent species, hereafter 'trees', were defined as species with free-standing stems that often exceed 5 cm of diameter at breast height (1.3 m) or 4 m in total height, including arborescent palms, cactus, tree ferns, and woody bamboos.

The list of South American tree names was used to download occurrence data from *species*Link (www.splink.org.br), JABOT (http://jabot.jbrj.gov.br, Silva et al., 2017), 'Portal de Datos de Biodiversidad Argentina' (https://datos.sndb.mincyt.gob.ar) and the Global Biodiversity Information Facility (GBIF.org, 2019). We excluded all occurrences described as being cultivated or exotic. We checked names for typos, orthographical variants and synonyms in the Brazilian Flora 2020 (BF-2020) project (Filardi et al., 2018; Zappi et al., 2015). Decisions for unresolved names were made by consulting Tropicos (www.tropicos.org) or the World Checklist of Selected Plant Families (http://wcsp.science.kew.org).

Using the 3.11 million records retrieved (Appendix S1), we conducted a detailed data cleaning and validation procedure (see Supporting Information for details). We standardized the notation of different fields (*e.g*. locality description, collector and identifier names, collection and identification dates), which were then used to (i) search for duplicate specimens among herbaria; (ii) validate the geographical coordinates at country, state and/or county levels and (iii) to assess the confidence level of the identification of each specimen (*i.e*. 'validated' and 'probably validated' - Appendix S2). Moreover, (iv) we cross-validated information of duplicate specimens across herbaria to obtain missing or more precise coordinates and/or valid identifications.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, (iv) we removed specimens too distant from their core distributions (i.e. spatial outliers).

2.2 Endemism levels

We obtained the endemism classification of each species from the BF-2020 (Filardi et al., 2018), the best reference currently available for the Atlantic Forest flora. A species was classified as 'endemic' if the BF-2020 field 'phytogeographic domain' contained only the term 'Atlantic Rainforest'. Correspondingly, a species was classified as 'occasional' if this field did not include this term. Species with no information on the 'phytogeographic domain' were omitted from this analysis.

We calculated an empirical level of endemism based on the position of species records in respect to the Atlantic Forest limits (IBGE, 2012; Olson & Dinerstein, 2002). Each record was assigned as being inside, outside or in the transition of the Atlantic Forest to other domains. Records in the transition were those falling inside the Atlantic Forest limits, but in counties with less than 90% of its area inside the Atlantic Forest or vice-versa. Because of the variable precision of the specimen's coordinates and of the arbitrariness of the domain delimitation at the scale of our reference map (1:5,000,000), records in the transition received half the weight other records to calculated species endemism levels:

$$100 \times \left(O_{in} + {}^{O_{ti}}/_{2}\right) / \left(O_{in} + {}^{O_{ti}}/_{2} + O_{out} + {}^{O_{to}}/_{2}\right),$$

where, O_{in}, O_{ti}, O_{out} and O_{to} are the number of specimens inside, inside in the transition, outside and outside in the transition to the Atlantic Forest, respectively. This endemism level is actually a weighted proportion of occurrences inside the Atlantic Forest by the total of valid occurrences found, varying from 0 (no occurrences) to 100% (all occurrences inside the Atlantic Forest).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The comparison between the reference BF-2020 classification and the empirical classification of species endemism was based on thresholds values varying from 0 to 100%, in intervals of 1% (*i.e.* 0, 1, …, 99, 100%). If a given species had an observed endemism level equal or higher than the threshold, it was classified as 'endemic'. For each threshold value, we calculated the number of mismatches between the two classifications (*i.e.* species classified as 'endemic' in the BF-2020 and 'not endemic' from the observed endemism level or vice-versa). The same procedure was used to calculate the number of mismatches for occasional species. We then plotted the number of mismatches against all thresholds and estimated the optimum threshold that minimizes the number of mismatches between classifications. Optimum thresholds were estimated using piecewise regression, allowing up to five segments (*i.e.* four breaking points). Thus, we provided the breaking point of each curve (and its 95% confidence interval). We compared the results using only taxonomically 'validated' and using both taxonomically 'validated' and 'probably validated' records.

2.3 Centres of diversity

We used the optimum threshold values obtained above to classify species into pure endemics, near endemics and occasional species and to delimit their centres of diversity (Laffan & Crisp, 2003). We plotted the valid occurrences of each group of species against a 50×50 km grid covering the Atlantic Forest and surrounding domains. Next, we obtained different diversity metrics for each group of species per grid cell. We selected two metrics with best performance to describe our data (Figures S1 and S2): corrected weighted endemism (WE) and rarefied/extrapolated richness (S_{RE}). The WE is the species richness weighted by the inverse of the number of cells where the species is present, divided by cell richness (Crisp et al., 2001). The S_{RE} is the rarefied/extrapolated

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

richness (depending on the observed number of occurrences per cell) for a common number of 100 occurrences, calculated based on the species frequencies per cell (Chao et al., 2014). We also obtained the sample coverage estimate (Chao & Jost, 2012), used here as a proxy of sample completeness. We evaluated the relationship of the diversity of endemic and occasional species with overall species diversity using spatial regression models (*i.e*. linear regression with spatially correlated errors - Pinheiro & Bates, 2000). Centres of diversity were delimited using ordinary kriging and only the grid cells meeting some minimum criteria of sampling coverage (see Supporting Information). We used the 80% quantile of predicted diversity distributions to delimit the centres of endemism.

3 RESULTS

3.1 Number of species found for the Atlantic Forest

After the removal of duplicates, spatial outliers and the geographical and taxonomic validation, we retained 593,920 valid records (disregarding records with 'probably validated' taxonomy). We found 252,911 valid records being collected inside the Atlantic Forest limits, which contained a total of 5062 arborescent species (4057 species excluding tall shrubs; Appendix S3). Most species-rich families were Myrtaceae (681), Fabaceae (658 species), Rubiaceae (328), Melastomataceae (290) and Lauraceae (222). If we consider the valid occurrences in the transitions of the Atlantic Forest to other domains, we could add 294 species as probably occurring in the Atlantic Forest (Appendix S4). Another 3148 names were retrieved but were finally excluded from the list for different reasons (*e.g*. synonyms, typos, orthographical variants, etc; Appendix S5).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2 Endemism levels

We found evidence of pure endemism (i.e. endemism level= 100%) for 1548 tree species (31%; Appendix S6). We found that 90.2% of records inside the Atlantic Forest (95% Confidence Interval, CI: 89.3–91.2%) was the threshold that best matched the endemism accepted by taxonomy experts (Figure 1a). The curve of mismatches between observed and reference classifications decreases until it reaches a minimum and then it increases again, meaning that more or less restrictive thresholds increase the number of mismatches. The 90.2% threshold in the Atlantic Forest added 733 near endemic species (15%). Together, pure and near endemics lead to an overall endemism ratio of 45.1% for the Atlantic Forest arborescent flora (Figure 1b) and 1.01 endemic arborescent species per 100 km^2 of remaining forest (*i.e*. 2261.2 km^2; Fundación Vida Silvestre Argentina & WWF, 2017). Some families had high average endemism level, namely Monimiaceae (94%), Symplocaceae (85%), Poaceae: Bambusoideae (84.2%), Araliaceae (84%), Myrtaceae (80%), Cyatheaceae (80%), Proteaceae (79%), Aquifoliaceae (77%) and Lauraceae (76%). Conversely, we found that 8.7% (95% CI: 8.2–9.3%) was the best threshold for separating occasional from more common Atlantic Forest species (Figure 1a), leading to a total of 646 occasional species (13%). The remaining 42% of the species were classified as widespread (Appendix S6). Results using only occurrences with taxonomy flagged as ‘validated’ were similar (pure endemism: 32%; near endemism: 15%; occasional species: 14%, widespread species: 39% - Figure S3, Appendix S6).

3.3 Centres of diversity

The diversity of endemic species was strongly correlated with the overall species diversity in the Atlantic Forest (Figure 2). There was also a strong and positive

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlation between the number of pure and near endemic species (Figure S4), meaning that the centres of diversity of pure and near endemics are highly congruent in space. The combination of pure and near endemics resulted in the same pattern of high diversity in the rainforests along the coast (Figure 3), corresponding to the Serra do Mar and Bahia Coastal Forests ecoregions (Olson & Dinerstein, 2002). Occasional species in the Atlantic Forest were really rare in the colder Araucaria region. Most of the distribution of occasional species was concentrated in the Brazilian Cerrado, but also in the Amazon and slightly less in the Caatinga domain. General patterns were fairly similar when using other diversity measures (Figures S5-S7).

4 DISCUSSION

Near endemism has been used to assess endemism levels of regional floras and faunas. However, such assessments often use loose (Carbutt & Edwards, 2006; Platts et al., 2011) or arbitrary definitions (Noroozi et al., 2018; Perera et al., 2011) of near endemics. Here, we used a data-driven approach to find that 90% of the occurrences inside a target region can be used to tell apart endemics from non-endemic trees. This is the same limit used to detect plant endemism in the Mediterranean Basin hotspot (Médail & Baumel, 2018), suggesting that a 90% limit could be used in assessment of plant endemism of other species-rich regions. This limit has another important implication: the average endemism concept adopted by taxonomic experts for Atlantic Forest trees implicitly includes the concept of near endemism. Indeed, the overall endemism ratio found here for pure and near endemics combined (45%) is within the range of 40-50% endemism level previously reported for the Atlantic Forest flora (Myers et al., 2000; Stehmann et al., 2009; Zappi et al., 2015). Thus, we propose that pure and near endemics can be used together to objectively delimit endemism or as two

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

categories of endemism, similarly to what already exists for the categories of species threat (IUCN, 2018).

The Atlantic Forest is arguably the tropical forest with the largest botanical knowledge available, with ca. 680 thousand unique specimens of tree species or 42 per 100 km^2 – average collection density in the Amazon forest is below 10 per 100 km^2 (ter Steege et al., 2016). Nevertheless, here we added 714 new valid occurrences of tree species for this biodiversity hotspot, an increase of 21% to the 3343 trees previously reported by the Brazilian Flora (Zappi et al., 2015). As expected, about 47% of these new records corresponded to occasional species. These species corresponded to 13% of the total richness of the Atlantic Forest tree flora, confirming that occasional species, despite of their rarity, make an important contribution to overall biodiversity patterns (Barlow et al., 2010; ter Steege et al., 2019). More importantly, 53% of the new records corresponded to widespread and endemic species. An increase of 16% in the total richness was also observed for the Espírito Santo state flora compared to the reported in the Brazilian Flora (Dutra, Alves-Araújo, & Carrijo, 2015). These results highlight how data-driven approaches combined with careful validation steps can help to refine the knowledge of local and regional floras. The Brazilian Flora is permanently being improved and it already is of utmost importance for the understanding of the Brazilian flora (Zappi et al., 2015), the richest in the world (Ulloa et al., 2017). Here, we provide products that can be readily integrated into the Brazilian Flora project (*e.g.* more refined endemism filters) and a workflow to perform similar assessments in other regions or for other groups of organisms.

The delimitation of centres of endemic diversity provided here (Figure 3, Appendix S7) has direct implications for conservation planning. For instance, they can assist the identification of Important Plant Areas (IPA - www.plantlifeipa.org), provided

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

by the Target 5 of the Global Strategy for Plant Conservation (www.cbd.int/gspc). Although the delimitation of IPAs predicts the use of endemic species, their definition is mainly based on the presence of threatened species. Moreover, since many of the endemic species identified here are probably also threatened, the Brazilian Alliance for Extinction Zero (www.biodiversitas.org.br/baze) could incorporate the concept of endemism proposed here to select plant trigger-species and to design conservation. Considering that defining threatened and endemic species have the same constraints related to data availability and to the time and spatial scale considered (Ferreira & Boldrini, 2011), the detection of endemics is more straightforward than threatened species, which could speed up the decision-making process for conservation. Therefore, the objective detection of endemic species proposed here (Appendix S6) could help to bridge the scarcity of conservation policies based on endemic species.

ACKNOWLEDGMENTS AND DATA

We thank Sidnei Souza and Renato Giovanni for their help with data compilation from *species*Link network. We also thank Lucie Zinger for helping with GBIF data management and for her suggestions on this manuscript. This study was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 795114. All data providers and their citations are given in Appendix S1.

REFERENCES

Barlow, J., Gardner, T. A., Louzada, J., & Peres, C. A. (2010). Measuring the conservation value of tropical primary forests: The effect of occasional species on estimates of biodiversity uniqueness. *PLoS ONE*, *5*(3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

https://doi.org/10.1371/journal.pone.0009609

Brooks, T. M., Mittermeier, R. A., Da Fonseca, G. A. B., Gerlach, J., Hoffmann, M., Lamoreux, J. F., … Rodrigues, A. S. L. (2006). Global biodiversity conservation priorities. *Science*, *313*(5783), 58–61. https://doi.org/10.1126/science.1127609

Carbutt, C., & Edwards, T. J. (2006). The endemic and near-endemic angiosperms of the Drakensberg Alpine Centre. *South African Journal of Botany*, *72*(1), 105–132. https://doi.org/10.1016/j.sajb.2005.06.001

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67. https://doi.org/10.1890/13-0133.1

Chao, A., & Jost, L. (2012). Coverage-based rarefaction and extrapolation : standardizing samples by completeness rather than size Author (s): Anne Chao and Lou Jost Published by : Wiley Stable URL : http://www.jstor.org/stable/41739612 REFERENCES Linked references are available on. *Ecology*, *93*(12), 2533–2547. https://doi.org/10.2307/41739612

Crisp, Laffan, Linder, & Monro. (2001). Endemism in the Australian flora. *Journal of Biogeography*, *28*(2), 183–198. https://doi.org/10.1046/j.1365-2699.2001.00524.x

Dutra, V. F., Alves-Araújo, A., & Carrijo, T. T. (2015). Angiosperm Checklist of Espírito Santo: Using electronic tools to improve the knowledge of an Atlantic Forest biodiversity hotspot. *Rodriguesia*, *66*(4), 1145–1152. https://doi.org/10.1590/2175-7860201566414

Ferreira, P. M. A., & Boldrini, I. I. (2011). Potential Reflection of Distinct Ecological Units in Plant Endemism Categories. *Conservation Biology*, *25*(4), 672–679. https://doi.org/10.1111/j.1523-1739.2011.01675.x

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Filardi, F. L. R., Barros, F., Baumgratz, J. F. A., Bicudo, C. E. M., Cavalcanti, T. B., Nadruz Coelho, M. A., … Zuntini, A. R. (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

Fundación Vida Silvestre Argentina, & WWF. (2017). *State of the Atlantic Forest: Three countries, 148 million people, one of the richest forests on Earth*. Puerto Iguazú, Argentina

GBIF.org (11 December 2019) GBIF Occurrence Download. https://doi.org/10.15468/dl.mzmat2

Grandtner, M. M., & Chevrette, J. (2013). *Dictionary of Trees, Volume 2: South America: Nomenclature, Taxonomy and Ecology*. Amsterdam: Academic Press.

IBGE. (2012). Mapa da Área de Aplicação da Lei no 11.428 de 2006. Retrieved from http://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/mapas/lei11428_mata_atlantica.pdf

IUCN. (2018). The IUCN Red List of Threatened Species. Retrieved from http://www.iucnredlist.org

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., … Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences*, *106*(23), 9322–9327. https://doi.org/10.1073/pnas.0810306106

Kruckeberg, A. R., & Rabinowitz, D. (1985). Biological aspects of endemism in higher plants. *Annual Review of Ecology and Systematics*, *16*, 447–479.

Laffan, S. W., & Crisp, M. D. (2003). Assessing endemism at multiple spatial scales, with an example from the Australian vascular flora. *Journal of Biogeography*,

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

30(4), 511–520. https://doi.org/10.1046/j.1365-2699.2003.00875.x

Letters, E., Bonn, A., Rodrigues, A. S. L., & Gaston, K. J. (2002). Threatened and endemic species: Are they good indicators of patterns of biodiversity on a national scale? *Ecology Letters*, *5*(6), 733–741. https://doi.org/10.1046/j.1461-0248.2002.00376.x

Lima, R. A. F., Mori, D. P., Pitta, G., Melito, M. O., Bello, C., Magnago, L. F., … Prado, P. I. (2015). How much do we know about the endangered Atlantic Forest? Reviewing nearly 70 years of information on tree community surveys. *Biodiversity and Conservation*, *24*(9), 2135–2148. https://doi.org/10.1007/s10531-015-0953-1

Matthews, W. S., van Wyk, A. E., & Bredenkamp, G. J. (1993). Endemic flora of the north-eastern Transvaal Escarpment, South Africa. *Biological Conservation*, *63*(1), 83–94. https://doi.org/10.1016/0006-3207(93)90077-E

Médail, F., & Baumel, A. (2018). Using phylogeography to define conservation priorities: The case of narrow endemic plants in the Mediterranean Basin hotspot. *Biological Conservation*, *224*, 258–266. https://doi.org/10.1016/j.biocon.2018.05.028

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*, 858–863.

Noroozi, J., Talebi, A., Doostmohammadi, M., Rumpf, S. B., Linder, H. P., & Schneeweiss, G. M. (2018). Hotspots within a global biodiversity hotspot-areas of endemism are associated with high mountain ranges. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-28504-9

Oliveira-Filho, A. T. (2010). TreeAtlan 2.0, Flora arbórea da América do Sul cisandina tropical e subtropical: Um banco de dados envolvendo biogeografia, diversidade e conservação. Universidade Federal de Minas Gerais. Retrieved from www.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cb.ufmg.br/treeatlan

Olson, D. M., & Dinerstein, E. (2002). The Global 200: Priority ecoregions for global conservation. *Annals of the Missouri Botanical Garden*, *89*(2), 199–224.

Perera, S. J., Ratnayake-Perera, D., & Procheş, Ş. (2011). Vertebrate distributions indicate a greater Maputaland-Pondoland-Albany region of endemism. *South African Journal of Science*, *107*(7/8), 68–71. https://doi.org/10.4102/sajs.v107i7/8.462

Peterson, A. T., & Watson, D. M. (1998). Problems with areal definitions of endemism: the effects of spatial scaling. *Diversity and Distributions*, *4*(4), 189–194. https://doi.org/10.1046/j.1472-4642.1998.00021.x

Pinheiro, J., & Bates, D. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-Effects Models in S and S-Plus*. Retrieved from https://link.springer.com/content/pdf/10.1007/0-387-22747-4_1.pdf

Platts, P. J., Burgess, N. D., Gereau, R. E., Lovett, J. C., Marshall, A. R., McClean, C. J., … Marchant, R. (2011). Delimiting tropical mountain ecoregions for conservation. *Environmental Conservation*, *38*(3), 312–324. https://doi.org/10.1017/S0376892911000191

Silva, L. A. E., Fraga, C. N., Almeida, T. M. H., Gonzalez, M., Lima, R. O., Rocha, M. S., … Forzza, R. C. (2017). Jabot - Sistema de Gerenciamento de Coleções Botânicas: a experiência de uma década de desenvolvimento e avanços. *Rodriguésia*, *68*(2), 391–410. https://doi.org/10.1590/2175-7860201768208

Stehmann, J. R., Forzza, R. C., Salino, A., Sobral, M., Pinheiro, D., & Kamino, H. Y. (2009). *Plantas da Floresta Atlântica*. Rio de Janeiro: Jardim Botânico do Rio de Janeiro.

Storch, D., Keil, P., & Jetz, W. (2012). Universal species–area and endemics–area

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

relationships at continental scales. *Nature*, *488*(7409), 78–81. https://doi.org/10.1038/nature11226

ter Steege, H., Vaessen, R. W., Cárdenas-López, D., Sabatier, D., Antonelli, A., Oliveira, S. M., … Salomão, R. P. (2016). The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. *Scientific Reports*, *6*(1), 29549. https://doi.org/10.1038/srep29549

ter Steege, H., Mota de Oliveira, S., Pitman, N. C. A., Sabatier, D., Antonelli, A., Guevara Andino, J. E., … Salomão, R. P. (2019). Towards a dynamic list of Amazonian tree species. *Scientific Reports*, *9*(1), 1–5. https://doi.org/10.1038/s41598-019-40101-y

Tyler, P. A. (1996). Endemism in freshwater algae. In *Biogeography of Freshwater Algae* (pp. 127–135). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-0908-8_12

Ulloa, C. U., Acevedo-Rodríguez, P., Beck, S., Belgrano, M. J., Bernal, R., Berry, P. E., … Jørgensen, P. M. (2017). An integrated assessment of the vascular plant species of the Americas. *Science*, *358*(6370), 1614–1617. https://doi.org/10.1126/science.aao0398

Werneck, M. S., Sobral, M. E. G., Rocha, C. T. V., Landau, E. C., & Stehmann, J. R. (2011). Distribution and endemism of angiosperms in the atlantic forest. *Natureza a Conservacao*, *9*(2), 188–193. https://doi.org/10.4322/natcon.2011.024

Zappi, D. C., Ranzato Filardi, F. L., Leitman, P., Souza, V. C., Walter, B. M. T., Pirani, J. R., … Forzza, R. C. (2015). Growing knowledge: An overview of Seed Plant diversity in Brazil. *Rodriguésia*, *66*(4), 1085–1113. https://doi.org/10.1590/2175-7860201566411

Zuloaga, F., Morrone, O., & Belgrano, M. (2008). *Catálogo de las plantas vasculares*

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

del cono sur (Argentina, southern Brazil, Chile, Paraguay y Uruguay). Monographs in Systematic Botany from the Missouri Botanical Garden 107.

AUTHOR CONTRIBUTIONS SUPPORTING INFORMATION

RAFL developed the idea of this study, conducted data compilation, analysis and drafted the paper. HTS assisted with herbarium data editing and study design. RAFL Geographical and taxonomical validation of the records were supported by MFS and VCS, respectively. All authors contributed to the interpretation and discussion of the results and writing the final manuscript.

SUPPORTING INFORMATION

Supporting Methods and References

Supplementary Figures

LIST OF APPENDICES

(Note: Full Appendices will be provided after the publication of the manuscript)

Appendix S1: List of collections and data providers used for data compilation. The numbers of records retrieved per collection correspond to overall sum of records before data validation, thus including both valid and invalid records.

Appendix S2: List of names of taxonomists per family used for taxonomical validation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The 'tdwg.name' represents the taxonomist name following the standard notation of the Biodiversity Information Standards (https://www.tdwg.org), which includes different variants of notation found for the same taxonomist name.

Appendix S3: Updated, taxonomically vetted checklist of the Atlantic Forest tree flora. For each name included in the checklist we provide the life form, the status of the name in respect to the Brazilian Flora 2020 project, the number of records found inside the Atlantic Forest (both 'validated' and 'probably validated' taxonomy) and a list of up to 30 vouchers (only specimens with 'validated' taxonomy), giving priority to type specimens. We also indicate which species were regarded as being taxa of low taxonomic complexity (TBC) or taxa commonly cultivated outside its original range.

Appendix S4: List of species with probable occurrence in the Atlantic Forest. We present all names with valid records found only in the transition of the Atlantic Forest to other domains and those names cited in the Brazilian Flora 2020 project as being an Atlantic Forest species, but for which we did not find any valid records. Again, we present for each name the life form, the number of records found and a list of up to 30 vouchers.

Appendix S5: List of names excluded from the final Atlantic Forest checklist. For each name on the list we provide the life form and the reason why the name was excluded. For synonyms, orthographical variants, common typos we also provide the corresponding valid name used in this study.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix S6: Endemism levels for the Atlantic Forest tree flora and the corresponding classification into pure endemic, near endemic, widespread and occasional species. For each species name, we provide the number of valid records outside the Atlantic Forest, outside but in the transition to the Atlantic Forest, inside the Atlantic Forest but in the transition to other domains, and inside the Atlantic Forest. We present the endemism levels and species classifications using only records with validated taxonomy and using records with validated and probably validated taxonomy. Finally, we present the endemism classification currently accepted in the Brazilian Flora 2020 in respect to the Atlantic Forest.

Appendix S7: Shapefiles delimiting the centres of the endemic and occasional species diversity in the Atlantic Forest for pure endemics, near endemics, pure + near endemics and occasional species.

Each shapefile contains the isoclines corresponding to the 75%, 80%, 85%, 90% and 95% quantiles of the distribution of rarefied/extrapolated richness for 100 specimens, predicted using ordinary kriging.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

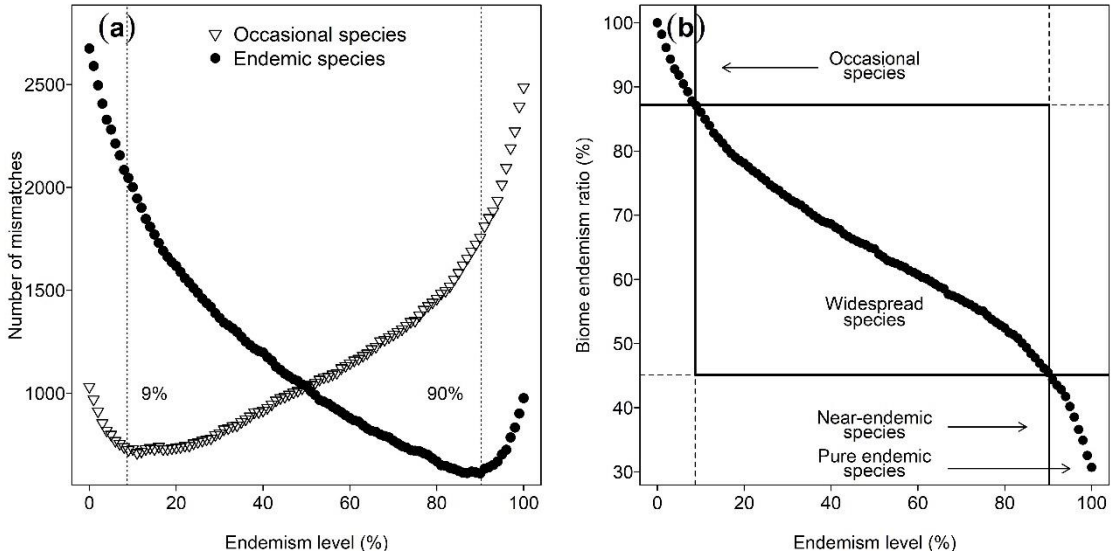

Figure 1. Defining near endemic and occasional tree species using herbarium records for the Atlantic Forest biodiversity hotspot. For both endemic (black circles) and occasional species (triangles), we present (a) the optimum endemism levels (vertical dashed lines) estimated from the distribution of mismatches between the empirical and the Brazilian Flora 2020 classifications and (b) the overall endemism ratio of the Atlantic Forest in intervals of 1% (*x*-axis in both panels).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

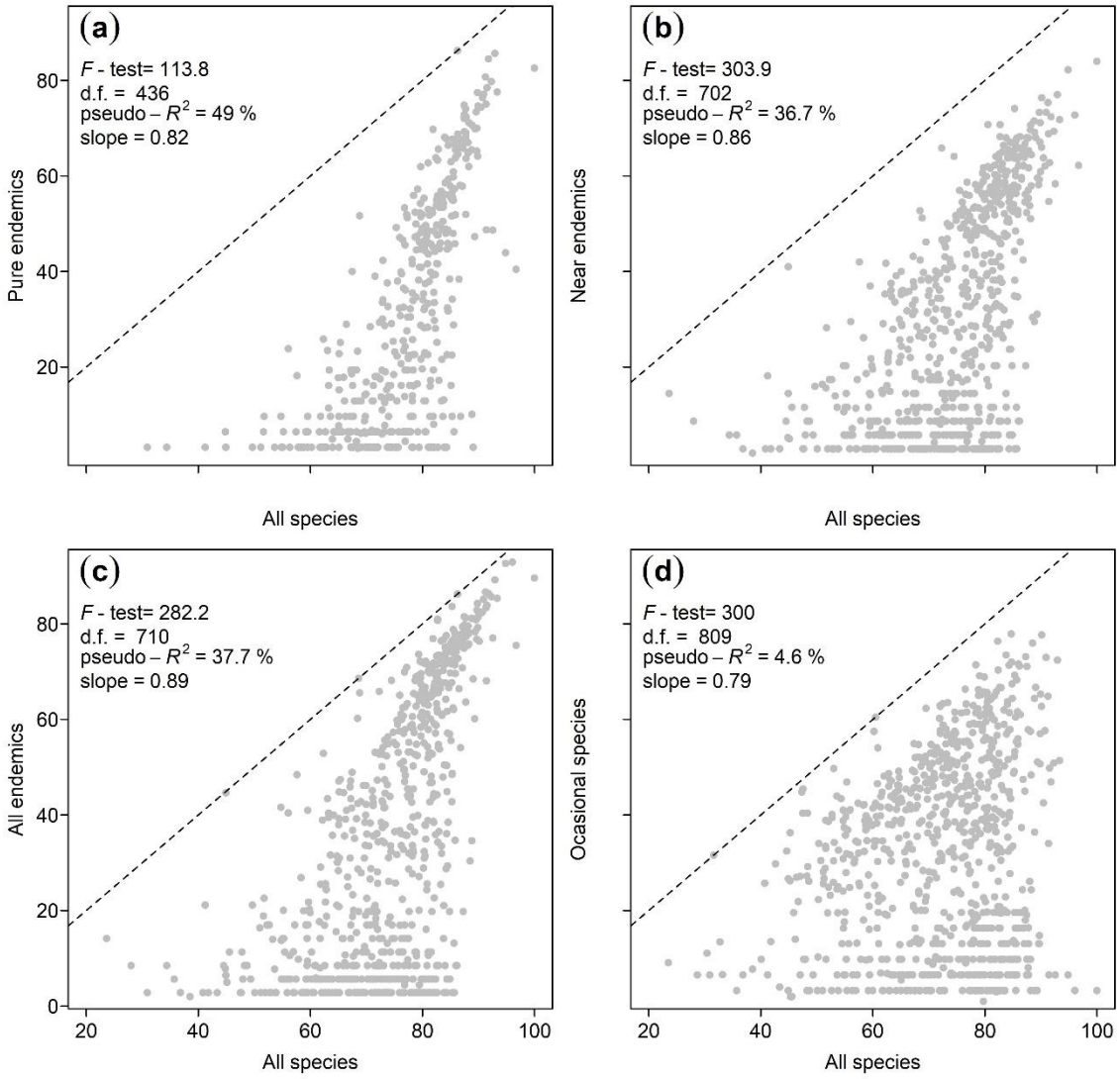

Figure 2. Relationship between the number of rarefied/extrapolated richness per 50×50 km grid cell and the same diversity metric obtained for (a) pure endemics, (b) near endemics, (c) all endemics (pure + near endemics) and (d) occasional species. For each group of species, we present the summary statistics of each spatial regression model (top left; d.f.= degrees of freedom), including the predicted slope of the regression prediction. The spatial regression analysis was performed only for grid cells meeting some minimum criteria of sampling coverage (see Supplementary Methods). The dashed line represents the 1:1 line.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.08.939900; this version posted February 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

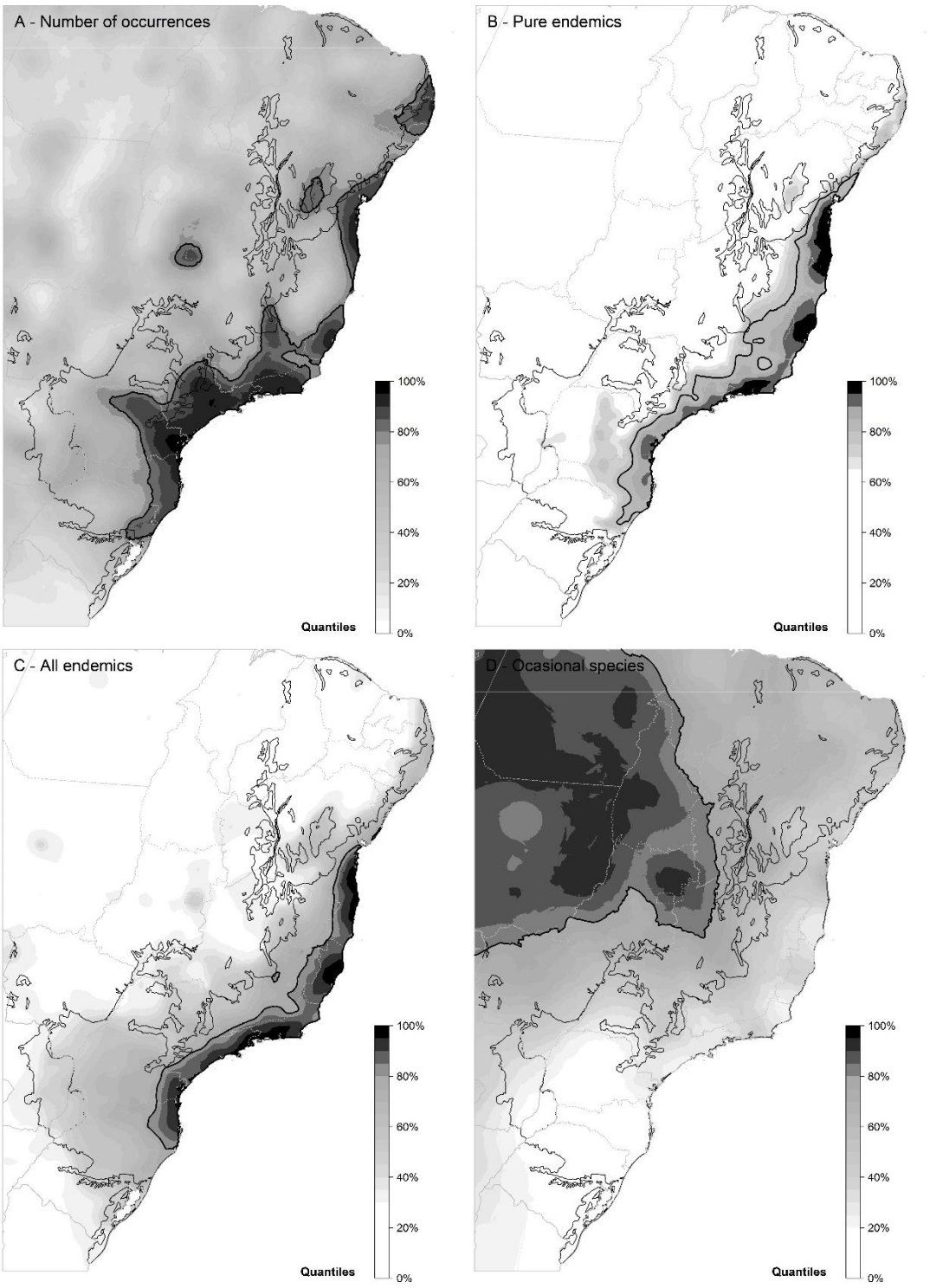

Figure 3. The spatial distribution of (A) the number of occurrences retrieved for the species occurring in the Atlantic Forest, and the centres of diversity of (B) pure endemics, (C) all endemics (pure + near) and (D) occasional species. Maps were produced using ordinary kriging based on rarefied/extrapolated species richness obtained for a common number of 100 records per grid cell. The colour scale represents the 5% quantiles of the metrics distribution, from 0-5% (white) to 95-100% (black). Bold black lines are the area containing the 80% higher richness values. The black line marks the limits of the Atlantic Forest, while the solid and dashed grey lines mark the limits of South American countries and of the Brazilian states, respectively.