Diversity and Distributions, (Diversity Distrib.) (2016) 22, 1232–1244

BIODIVERSITY RESEARCH

The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity

Ubirajara Oliveira[1,2*], Adriano Pereira Paglia[3], Antonio D. Brescovit[4], Claudio J. B. de Carvalho[5], Daniel Paiva Silva[6], Daniella T. Rezende[7], Felipe Sá Fortes Leite[8], João Aguiar Nogueira Batista[9], João Paulo Peixoto Pena Barbosa[4], João Renato Stehmann[9], John S. Ascher[10], Marcelo Ferreira de Vasconcelos[11,12], Paulo De Marco Jr[13], Peter Löwenberg-Neto[14], Priscila Guimarães Dias[15], Viviane Gianluppi Ferro[13] and Adalberto J. Santos[2]

[1]*Centro de Sensoriamento Remoto, Instituto de Geociências, Universidade Federal de Minas Gerais – UFMG, Av. Antonio Carlos 6627, CEP 31270-901, Belo Horizonte, MG, Brazil,* [2]*Departamento de Zoologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais – UFMG, Av. Antonio Carlos 6627, CEP 31270-901, Belo Horizonte, MG, Brazil,* [3]*Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, MG, Brazil,* [4]*Laboratório Especial de Coleções Zoológicas, Instituto Butantan, São Paulo, SP, Brazil,* [5]*Departamento de Zoologia, Universidade Federal do Paraná, Curitiba, Paraná, Brazil,* [6]*Departamento de Biologia, Instituto Federal Goiano – IFGoiano, Urutaí, Goiás, Brazil,* [7]*Independent Researcher, Belo Horizonte, MG, Brazil,* [8]*Laboratório Sagarana, Instituto de Ciências Biológicas e da Saúde, Universidade Federal de Viçosa – UFV, Campus Florestal, Florestal, MG, Brazil,* [9]*Departamento de Botânica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, MG, Brazil,* [10]*Department of Biological Sciences, National University of Singapore, Singapore City, Singapore,* [11]*Coleção Ornitológica, Museu de Ciências Naturais, Pontifícia Universidade Católica de Minas Gerais, Avenida Dom José Gaspar 290, CEP 30535-901, Belo Horizonte, MG, Brazil,* [12]*Instituto Prístino, Rua Santa Maria Goretti, 86, Barreiro, CEP 30642-020, Belo Horizonte, MG, Brazil,* [13]*Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Goiás,*

ABSTRACT

Aim The knowledge of biodiversity facets such as species composition, distribution and ecological niche is fundamental for the construction of biogeographic hypotheses and conservation strategies. However, the knowledge on these facets is affected by major shortfalls, which are even more pronounced in the tropics. This study aims to evaluate the effect of sampling bias and variation in collection effort on Linnean, Wallacean and Hutchinsonian shortfalls and diversity measures as species richness, endemism and beta-diversity.

Location Brazil.

Methods We have built a database with over 1.5 million records of arthropods, vertebrates and angiosperms of Brazil, based on specimens deposited in scientific collections and on the taxonomic literature. We used null models to test the collection bias regarding the proximity to access routes. We also tested the influence of sampling effort on diversity measures by regression models. To investigate the Wallacean shortfall, we modelled the geographic distribution of over 4000 species and compared their observed distribution with models. To quantify the Hutchinsonian shortfall, we used environmental Euclidean distance of the records to identify regions with poorly sampled environmental conditions. To estimate the Linnean shortfall, we measured the similarity of species composition between regions close to and far from access routes.

Results We demonstrated that despite the differences in sampling effort, the strong collection bias affects all taxonomic groups equally, generating a pattern of spatially biased sampling effort. This collection pattern contributes greatly to the biodiversity knowledge shortfalls, which directly affects the knowledge on the distribution patterns of diversity.

Main conclusions The knowledge on species richness, species composition and endemism in the Brazilian biodiversity is strongly biased spatially. Despite differences in sampling effort for each taxonomic group, roadside bias affected them equally. Species composition similarity decreased with the distance from access routes, suggesting collection surveys at sites far from roads could increase the probability of sampling new geographic records or new species.

Keywords

beta-diversity, endemism, Hutchinsonian shortfall, Linnean shortfall, species distribution models, species richness, Wallacean shortfall.

DOI: 10.1111/ddi.12489
http://wileyonlinelibrary.com/journal/ddi

Brazil, [14]*Universidade Federal da Integração Latino-Americana, Foz do Iguaçu, PR, Brazil,* [15]*Departamento de Biologia, Universidade Federal de Lavras – UFLA, Lavras, MG, Brazil*

*Correspondence: Ubirajara Oliveira, Centro de Sensoriamento Remoto, Instituto de Geociências, Universidade Federal de Minas Gerais – UFMG, Av. Antonio Carlos 6627, CEP 31270-901, Belo Horizonte, MG, Brazil. E-mail: ubiologia@yahoo.com.br

INTRODUCTION

The knowledge on species composition, distribution and ecological niche is fundamental for the construction of biogeographic and macroecological hypotheses and to support effective conservation actions (Hortal *et al.*, 2015). The main sources of biodiversity information are specimens deposited in scientific collections and taxonomic publications (Meyer *et al.*, 2004; Sousa-Baena *et al.*, 2014), from which we can have the distribution of each taxa based on the localities where they were collected. However, these data are incomplete and may not represent adequately the actual biodiversity (Dennis *et al.*, 1999; Moerman & Estabrook, 2006; Vale & Jenkins, 2012; Yang *et al.*, 2013). Furthermore, these data do not represent an adequate sampling of biodiversity because they are often biased in different ways (Moerman & Estabrook, 2006; Vale & Jenkins, 2012; Sousa-Baena *et al.*, 2014). These problems are even more pronounced in the tropics (Kier *et al.*, 2005; Collen *et al.*, 2008), which holds most of the planet's biodiversity. This is further aggravated by the lack of resources for taxonomic studies and accelerated rate of destruction of natural ecosystems in these regions.

The intensity and the spatial variation of sampling may affect the biodiversity knowledge (Moerman & Estabrook, 2006; Grand *et al.*, 2007; Yang *et al.*, 2013), generating shortfalls of biological knowledge such as an excess of undescribed species (Linnaean shortfall), poorly known species geographic distribution (Wallacean shortfall; Lomolino, 2004; Whittaker *et al.*, 2005) and the lack of knowledge about the responses and tolerances of species to abiotic conditions (Hutchinsonian shortfall; Hortal *et al.*, 2015). These shortfalls can directly affect the parameters used to quantify biodiversity, as species richness, endemism and beta-diversity (Yang *et al.*, 2013), compromising studies on biogeography, macroecology and conservation (Moerman & Estabrook, 2006; Grand *et al.*, 2007; Yang *et al.*, 2013). However, few studies quantify and evaluate the effects of those biases in biological knowledge using empirical data.

In addition to the problems mentioned above, certain geographic locations and taxonomic groups have been neglected in biodiversity studies, leading to higher shortfalls of knowledge in these groups and areas (Whittaker *et al.*, 2005; Diniz-Filho *et al.*, 2010). Arthropods, for example, have been historically neglected (Gaston & May, 1992; Diniz-Filho *et al.*, 2010), while it is assumed that vertebrates are better known. This assumption is widely alleged to justify the choice of groups in macroecology and biogeography studies, even without actual empirical support (Cardoso *et al.*, 2011). Likewise, the spatial distribution of sampling effort is widely recognized as strongly biased, resulting in whole areas or ecosystems poorly represented in scientific collections and biodiversity databases. For instance, the Brazilian seasonally dry forests (the Caatinga) is considered historically neglected by taxonomists and biogeographers (Santos *et al.*, 2011). However, no comparative study has demonstrated the effect of this neglect on the knowledge on this biome, compared to others. Moreover, possibly the most pervasive documented sampling bias in biodiversity is the concentration of specimen records near access routes, such as roads and navigable rivers and major cities (Kadmon *et al.*, 2004; Moerman & Estabrook, 2006; Boakes *et al.*, 2010; Vale & Jenkins, 2012; Sousa-Baena *et al.*, 2014; Tessarolo *et al.*, 2014). This sampling bias is mostly a result of financial resource limitation and access difficulty on remote places, particularly in extensive tropical forests (Moerman & Estabrook, 2006; Boakes *et al.*, 2010). Although it is widely accepted that this bias can affect biodiversity assessment in large spatial scales (Nelson *et al.*, 1990; Freitag *et al.*, 1998; Moerman & Estabrook, 2006; Boakes *et al.*, 2010), its effects have never been quantitatively evaluated. Thus, although it is known that some places and taxa have been neglected, it has not been determined how lesser known they actually are.

In this study, we evaluate sampling bias and collection effort of different groups of arthropods, vertebrates and angiosperms in the Brazilian biomes and quantify the influence of those problems on biological knowledge. We quantified the relationship between sampling effort and parameters that are commonly used to quantify biological diversity, as species richness, endemism and beta-diversity. Furthermore, we check the effect of collection bias on Linnean, Wallacean and Hutchinsonian shortfalls.

METHODS

Dataset

We compiled occurrence data of terrestrial angiosperm, arthropods and vertebrates in Brazil. To avoid biasing our

analyses by angiosperm groups that occur predominantly in specific biomes, we included only the most species rich and widely distributed families in Brazil: Asteraceae, Bromeliaceae, Fabaceae, Melastomataceae, Myrtaceae, Orchidaceae, Poaceae and Rubiaceae. For the arthropods, we compiled data for bees, spiders, polydesmid millipedes, flies, tiger moths, dragonflies and Orthoptera. Vertebrates were represented by data on birds, mammals and amphibians.

We compiled occurrence data from the following online databases: GBIF (http://www.gbif.org), SpeciesLink (http://www.splink.org.br), Birdlife International (http://www.birdlife.org), Herpnet (http://www.herpnet.org), Nature Serve (http://www.natureserve.org) and Orthoptera Species File (orthoptera.speciesfile.org). We also compiled data contained in the taxonomic literature and biodiversity inventories. All data were checked for accuracy and validity of the geographic coordinates through data crossing with databases of political units of Brazil in a GIS, totalling 1,144,629 georeferenced records (see details in Appendix S1 and S2 in Supporting information). Records that lacked geographic coordinates or presented georeferencing errors were georeferenced based on databases of localities and municipalities of IBGE (http://mapas.ibge.gov.br). All data were reviewed for the validity of taxonomic names through specific catalogues (Appendix S1) and direct review of data by experts on each group, resulting in 882,468 records with valid names (Appendix S1). The data without taxonomic identification or valid names were used only in the analyses of sampling effort and collection bias. Comparative analyses were performed for the whole dataset and separately for each data partition (angiosperms, arthropods and vertebrates). All statistical analyses were performed using R software (http://www.r-project.org).

Collection bias and sampling effort

To check whether distribution records are biased regarding the proximity of access routes (streets, roads and navigable rivers), we tested whether the density of records was higher near access routes than expected from a random sampling of data. Access route maps were assembled from online databases (Brazilian streets, roads and navigable rivers – http://www.openstreetmap.org/; http://www.dnit.gov.br; http://www.ibge.gov.br/). We created a null model of sampling by: (1) randomly plotting on the map the same number of points present in each database and (2) calculating the distance of these points to the nearest access route. We then compared the distance of null model points to access routes to the distances observed from empirical data. The distance from the nearest access route to each of the random sampled points and the empirical points was quantified in ArcGIS 10.1 (ESRI, Redlands, California). We compared the distribution of the empirical data with the null model through a Mann–Whitney test. This procedure was repeated 1000 times to quantify the percentage of tests in which the difference was significant ($\alpha = 0.001$).

To test for differences in sampling effort intensity in different groups, we used the number of records in 0.5° grid cells as a surrogate of sampling effort and compared the different groups by Kruskal–Wallis test. To test whether the better sampled sites correspond to sites with greater availability of access routes, we performed a Pearson correlation with corrected degrees of freedom (see Clifford *et al.*, 1989) between the density of records and the density of access routes. The density of records and access routes were estimated through kernel interpolation in ArcGIS 10.1. To establish the search radius of kernel estimation, we used the average distance between points of occurrence. To compare different biomes regarding their collection intensity, we classified the density maps in five categories according to the kernel index.

Relationship between diversity, endemism and sampling effort

We checked the influence of spatial distribution of the collection effort on different biodiversity parameters through Spearman correlation with corrected degrees of freedom between the number of records and those parameters in 0.5° grid cells. The effect of collection effort on known species richness was analysed through the correlation between the number of species and the number of records on each cell. To compare the beta-diversity with sampling effort, we use the mean Sørensen index between each grid cell and its eight nearest neighbours. To analyse the relationship between endemism and sampling effort, we used the index of weighted endemism (WE; Williams & Humphries, 1994). The endemism was expressed both through the sum of WE index for all species within each cell, obtaining the general pattern of endemism, and by the average index in the cell, indicating the predominant distribution pattern of the species.

Linnean/Wallacean shortfall and sampling bias

Estimating directly the Linnaean shortfall is hampered by the need of reliable estimates of total species richness. As it is difficult to directly estimate how many species are not yet described, we used an indirect approach to estimate which locations should probably hold more unknown species. We follow other studies (Bini *et al.*, 2006; Reeder *et al.*, 2007; Brito, 2010) and consider that less sampled sites that have greater difference in species composition, compared to better sampled sites, should harbour more undescribed species. We understand that this assumption is not always true. However, considering our results on the collecting bias and the large composition differences observed among the better sampled areas (borders of access roads) and the least sampled sites, it is plausible to consider that this may be an acceptable, albeit crude first approach to this complex problem. As the same poorly sampled sites could harbour both undescribed species and new records of already known species, we treat our index as a measure of a mixture of Linnean and Wallacean shortfalls.

As our findings indicate a strong sampling bias (see 'Results' below), we also tested whether the species composition varies more among samples distant from access routes than along the same routes. We established a 500 m buffer as sampled area along the access routes. From this buffer, we sampled species composition in buffers positioned from 1 to over 50 km away from the route. This procedure was repeated for every one degree grid cell. For each sampling buffer, we quantified the percentage of species also recorded on the 500 m buffer. For comparison, we performed the same test with samples taken along the routes, at the same distance categories from the intersection of the route at the cell border (Fig. 2 in Appendix S2). The potential influence of the number of records in each sampling buffer and its percentage of similarity in species composition with the route was analysed through Pearson correlation with corrected degrees of freedom (Clifford *et al.*, 1989). As the correlation values obtained were always low (Table 3 in Appendix S2), we felt safe to disregard these factor in the analysis. The effect of the distance from the access route, and the distance from the focal point along the route, on the species percentage of similarity was analysed by Kruskal–Wallis test.

Wallacean shortfall and sampling bias

As a way to estimate the Wallacean shortfall, we compared the known distribution of species with the prediction of species distribution models (SDM). As the results of the SDMs can be affected by Hutchinsonian shortfall, SDMs of poorly known species tend to be less accurate, predicting smaller areas of distribution (Loiselle *et al.*, 2007). Thus, this analysis was relatively conservative, because new data for the poorly known species would lead to an even greater difference between the known and the modelled distributions.

The known distribution of each species was estimated through three procedures, the sum of 10 and 50 km buffers around distribution records and minimum convex polygons (MCP) constructed through the points of occurrence of the species. To access the sensitivity of the results to differences in SDM methods, the analysis was implemented with species distribution predicted through seven algorithms: Bioclim, Domain, Mahalanobis distance, Maxent, Generalized Linear Model (GLM), Generalized Boosting Model (GBM or Boosted Regression Trees) and Support Vector Machine (SVM; Franklin & Miller, 2009).

The prediction of SDMs can be influenced by the number of records and the quality of georeferencing (Loiselle *et al.*, 2007). Thus, we analysed only 4344 species that presented more than 15 occurrence points accurately georeferenced by locality (Appendix S3). In addition, we partitioned the data into three groups based on number of records: more than 15, 40 and 80 records to determine the effect of sample size on the test conducted with SVM. SDMs were based on the first four axes of a correlation matrix PCA performed on 19 bioclimatic (http://www.worldclim.org/) and two topographic variables, altitude (from Worldclim) and slope, derived from altitude data in ArcGIS 10.1. We used the lowest value of suitability in training points as a threshold. The Principal Component Analysis (PCA) were implemented in ArcGIS with 5 km pixels. To exclude models that presented random predictions, we used the area under the curve (AUC) through pseudo absences. We compared estimated species distribution only to areas predicted by models with AUC over 0.7, using Kruskal–Wallis test. To assess the direct effect of Hutchinsonian shortfall on distribution models, we tested the correlation of the sum of the models with the access route density and the collection density.

Hutchinsonian shortfall and sampling bias

To estimate the Hutchinsonian shortfall, we started from the assumption that large groups, like bees, birds, vertebrates and arthropods, occur throughout the Brazilian territory. Thus, these groups should occupy all the existing terrestrial environmental conditions in Brazil. In this way, we compared the environmental similarity of sampling sites for these large groups with all conditions within the raster map resolution of climatic conditions in Brazil. We used the shortest Euclidean distance between each 5 km pixel in Brazil and the most environmentally similar sample point for each group (Appendix S2). The environmental conditions were estimated through PCA axes of topographic and climatic variables, as explained above. The resulting similarity maps were classified into seven classes, which were used to assess how well sampled are each Brazilian biome in environmental conditions, for each group. We also analysed whether the environmental representability are spatially correlated with sampling effort through Pearson correlation with corrected degrees of freedom with the density of access routes and distribution records, estimated through the kernel procedure described above.

RESULTS

Collection bias and sampling effort

The highest peaks of record density, for all groups and biomes, are located at < 1 km from the access routes (Fig. 1). The distribution of record density along the distance from access routes was significantly different from the null models in all groups and biomes in 99% of the analyses. Additionally, the distribution of record density in relation to distance from access routes showed similar results between groups and biomes (Fig. 1). Sampling effort intensity was significantly different among groups, as there is a significant difference between the number of records per grid cells in different groups ($K = 1592.804$; $P = 0.00001$). The correlation between the density of access routes and the density of records was strong in the analysis with all groups and moderate for the analysis of angiosperms, arthropods and vertebrates separately (Fig. 2). The density of records showed

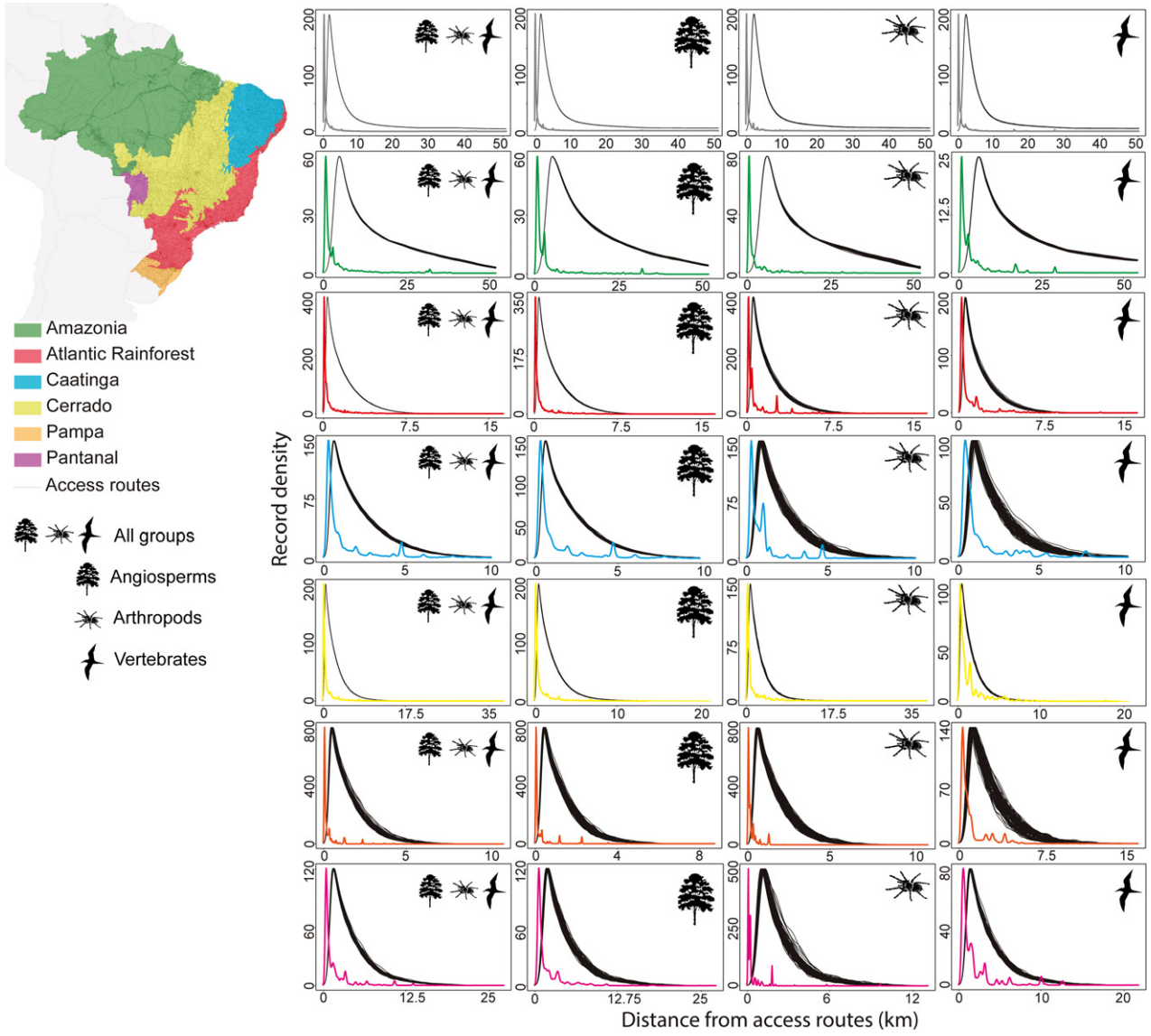

Figure 1 Observed and expected frequency of distribution records at differing distances from access routes for angiosperm, vertebrates and arthropods in Brazil. Colours indicate observed data from each biome, and the black lines show the results from null models. Colour figure can be viewed at wileyonlinelibrary.com

similar proportions between the different groups, although the angiosperms showed larger areas with higher density of records than arthropods and vertebrates in all biomes (Fig. 2). The highest record density for all groups was observed in the Atlantic Rainforest and the lowest in the Amazon (Fig. 2). The Pantanal showed higher record density of vertebrates, while the same group presented lower density in the Pampa (Fig. 2). Arthropods and vertebrates have low record density in the Caatinga, when compared to angiosperms (Fig. 2).

Relationship between diversity, endemism and sampling effort

Species richness showed a strong relationship with the number of records in all groups (Fig. 3). Vertebrates showed the strongest relationship between the number of records and species richness ($R^2 = 0.94$), while arthropods showed a weaker relationship ($R^2 = 0.70$). The beta-diversity showed virtually no relationship with the number of records for all groups and biomes (Fig. 3). On the other hand, the sum of the endemism index was moderately correlated with the number of records ($R^2 = 0.79$–0.88). However, the mean of the same index showed no correlation with the number of records for all groups (Fig. 3).

Linnean/Wallacean shortfall and sampling bias

The similarity in species composition of all groups decayed linearly with the increasing distance from access routes (Fig. 4a). The similarity with the buffer around the routes were remarkably low (~20%) even considering the smallest

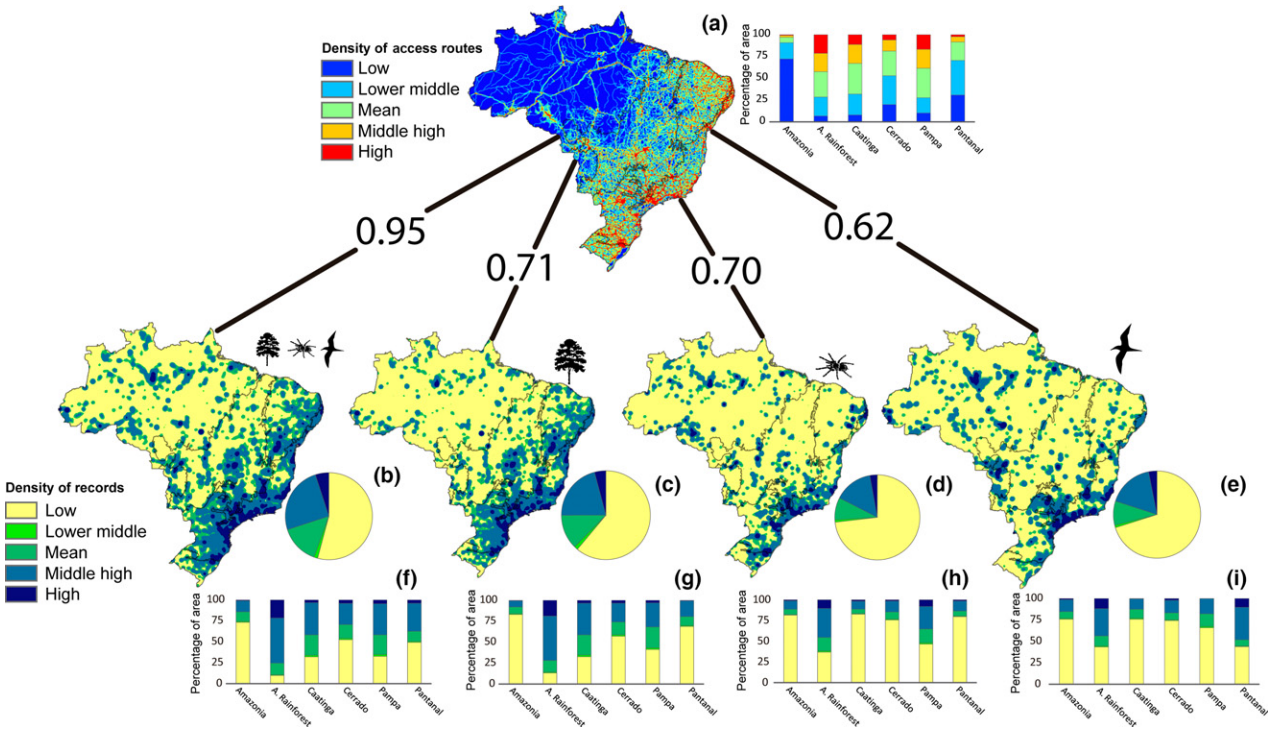

Figure 2 Relationship between density of access routes and density of species distribution records. (a) Kernel density of access routes in Brazil. The graph shows the distribution of density categories among Brazilian biomes. (b–e) Kernel density of distribution records for all groups, angiosperms, arthropods and vertebrates. The numbers over the lines indicate R for Pearson correlation between the density of records and the density of access routes for each data partition. (f–i) Distribution of density categories of each data partition among Brazilian biomes. Colour figure can be viewed at wileyonlinelibrary.com

distance from the routes (1–5 km). However, the decay slope varied between biomes, with lower values for the Atlantic Rainforest and the highest for Amazonia, Caatinga and Cerrado (Fig. 4a,c,e,g in Appendix S2). The Pampa and Pantanal showed no statistically significant effect of the distance from the routes on the species composition similarity (Fig. 4i,k in Appendix S2). The compositional similarity along the routes showed no statistically significant variation (Fig. 4b). This pattern was also observed for each biome separately, and the similarity between samples along the routes was always higher than the highest observed similarity of the ones collected off the road (Table 5 in Appendix S2).

Wallacean shortfall and sampling bias

The SDM algorithms showed significant differences between the median AUC's, except Domain and Maxent, which showed no significant differences between their medians (Fig. 6 in Appendix S2). The SVM showed the highest AUC values, while the Mahalanobis distance showed the lowest. The predicted areas by Domain, Mahalanobis distance, Maxent and SVM were significantly higher (Table 7 in Appendix S2) than the known distribution obtained for each of the three methods (10 and 50 km Buffer and MCP). Only Bioclim, GLM and GBM showed a predicted area smaller than the MCP (Fig. 5). Additionally, predicted areas of distribution were significantly different between algorithms (Table 7 in Appendix S2). The correlation of the sum of the models with the access routes density and the collection density was relatively low ($r > 0.28$). The results of the analysis with data partitions (above 40 and 80 records) showed no differences in the results obtained with the models generated with over 15 records (Fig. 5).

Hutchinsonian shortfall and sampling bias

The similarity of environmental analysis indicated that the majority of Brazilian territory presents more than 0.95 of similarity with the points of occurrence of angiosperms and the points in combined analysis of all groups (Fig. 6a,b). The analysis of arthropods and vertebrates indicated that most of the territory has similarity lower than 0.90 (Fig. 6c,d). The major areas with low environmental similarity are found in Amazonia and Caatinga for all groups, with between 20% and 55% of the area of these biomes showing similarity below 0.90 (Fig. 6). Additionally, the Pantanal showed the smallest areas with low environmental similarity (Fig. 6). The maps of environmental distance showed low correlation with the density of access routes and records (r between 0.07 and 0.15).

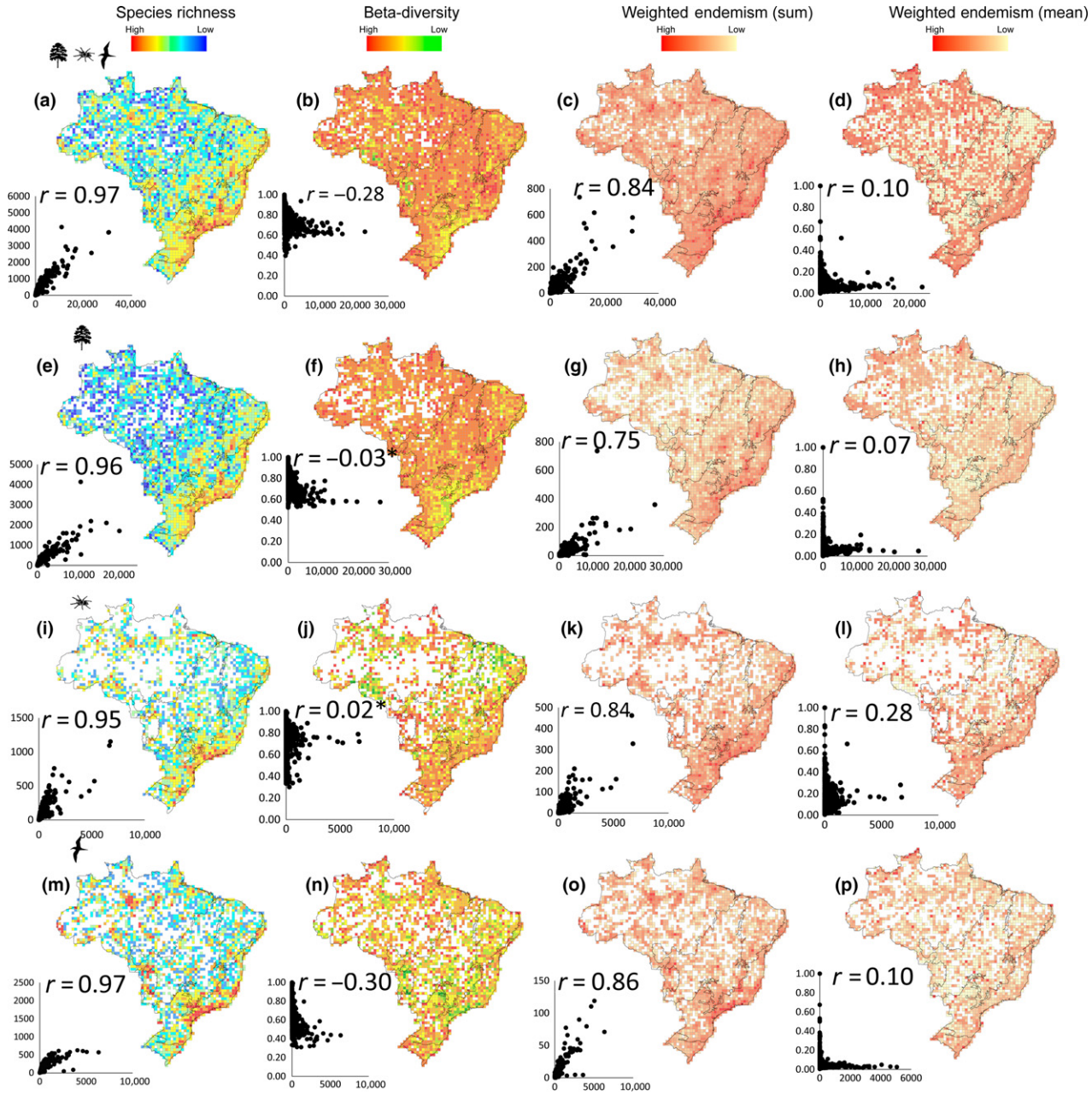

Figure 3 Relationship between diversity measures (species richness, beta-diversity, sum and average weighted endemism) and sampling effort (number of records per grid cell) for each data partition in Brazil. Colour figure can be viewed at wileyonlinelibrary.com

DISCUSSION

In this study, we show large differences in sampling effort among taxonomic groups throughout Brazil. However, independent on the sampling coverage, all groups are equally affected by a collection bias towards access routes. This pattern generates similar biodiversity knowledge deficits, affecting the perception of spatial patterns of diversity. Thus, despite the preferential focus on certain taxa, based on the assumption that they are 'well known' (e.g. Orme *et al.*, 2005; Brooks *et al.*, 2006; Lamoreux *et al.*, 2006), all groups showed similar sampling problems.

All groups showed the same pattern of collection bias despite large differences in sampling intensity between them. Although angiosperms have a density of records higher than arthropods and vertebrates, the best sampled sites are coincident between groups. This can be explained by the fact that research centres, around which most biodiversity sampling is concentrated, are usually located in the same urban centres, leading to common patterns of distribution of records among different taxa (Nelson *et al.*, 1990; Moerman & Estabrook, 2006; Hopkins, 2007).

The sampling bias towards access routes mentioned above was detected for all groups, in all Brazilian biomes. However,

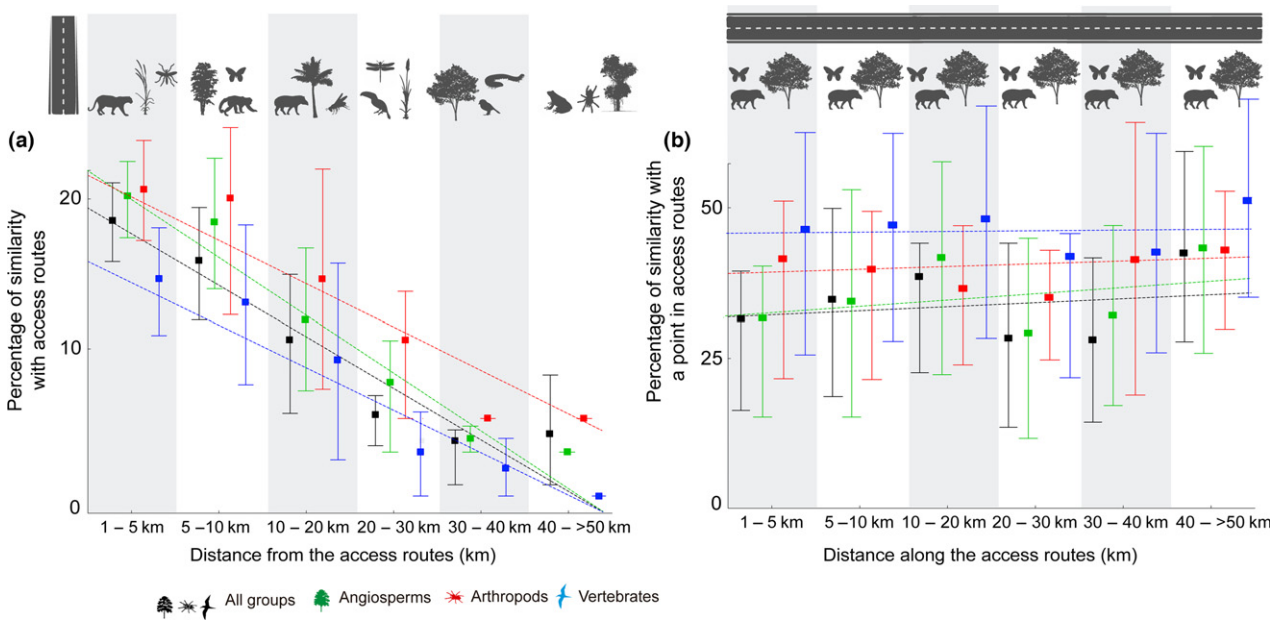

Figure 4 Species composition similarity (% common species) between sampling points near and at increasing distances from access routes, compared to the same parameter measured in points at increasing distance along access routes. Colour figure can be viewed at wileyonlinelibrary.com

the sampling intensity was different between biomes. In the Amazon and the Atlantic Forest, the density of access routes appears to be related to the sampling intensity (Fig. 2). On the other hand, the variation in sampling intensity is not fully explained by the density of routes in others biomes. For instance, the Cerrado, Caatinga and Pampa have a low proportion of their area with a high sampling effort, despite having a considerable density of access routes. Thus, this indicates that the scarcity of biodiversity data for these biomes is not caused by access difficulties, but by a negative bias, compared to humid forest biomes, as suggested by Santos *et al.* (2011) and Werneck (2011). The density of access routes in the Pantanal is relatively high; however, only vertebrates have higher collection intensity in this biome. This is possibly explained by an inherent positive bias from taxonomists, who tend to sample more intensely the places they expect to be more diverse (Sastre & Lobo, 2009). As Pantanal is moderately rich in vertebrate species, most of them relatively common and easier to sample in this biome (Mittermeier *et al.*, 2003), vertebrate taxonomists should tend to sample this biome more intensely. Thus, we can conclude, considering the assumptions of our analysis, that regions further from access routes are those with the lowest biological knowledge in general. However, some of the most neglected regions have little sampling, even in the surroundings of access routes. This pattern of sampling bias can reduce the effectiveness of conservation actions (Grand *et al.*, 2007). As these locations farthest from the access routes have the lowest human impacts in the world (Laurance, 2009) and present distinct species composition, they should be considered as priority areas for inventories.

Patterns of species richness and endemism are strongly correlated, in all groups, with the sample effort, which in turn is related to the collecting bias to access roads. Interestingly, vertebrate species richness data has the same degree of correlation as the arthropod data (Fig. 3). This contradicts the general expectation that vertebrate data are more complete, and thus more adequate for biogeographic and conservation studies, if compared to arthropods (Cardoso *et al.*, 2011). It is possible that this conclusion applies mostly for the tropics, where even vertebrates are poorly known. On the other hand, the vertebrate fauna of North America and Europe are probably much better known, due to the presence of fewer species and a longer history of study. Anyway, our results show that, at least in Brazil, vertebrates should not be considered as better known than arthropods.

Our results show that the least sampled areas within Brazil are those with higher Linnaean/Wallacean shortfall. This means that these shortfalls are directly affected by the collection bias, which, as we demonstrated, equally affects all groups. Clearly, some groups, such as arthropods, should be more affected by this shortfall (Cardoso *et al.*, 2011) because, as invertebrate taxa are much more speciose than vertebrates, they should have more species to be described. Furthermore, we demonstrate that the variation in species composition along the access routes is lower than the observed in distant locations. This may be related to the impacts of access routes on the local biota, affecting the natural patterns of species distribution (Laurance, 2009; Clark *et al.*, 2010; Caro *et al.*, 2014). This negative effect of human presence can be related both to vehicle traffic as to direct impacts such as pollution, edge effect and microclimate variation (Caro *et al.*, 2014). However, in the case of navigable rivers, this effect can be due both to natural differences of these habitats, compared to terrestrial ones, and due to the impact of human activities on these sites. Most biomes showed a tendency to increase the

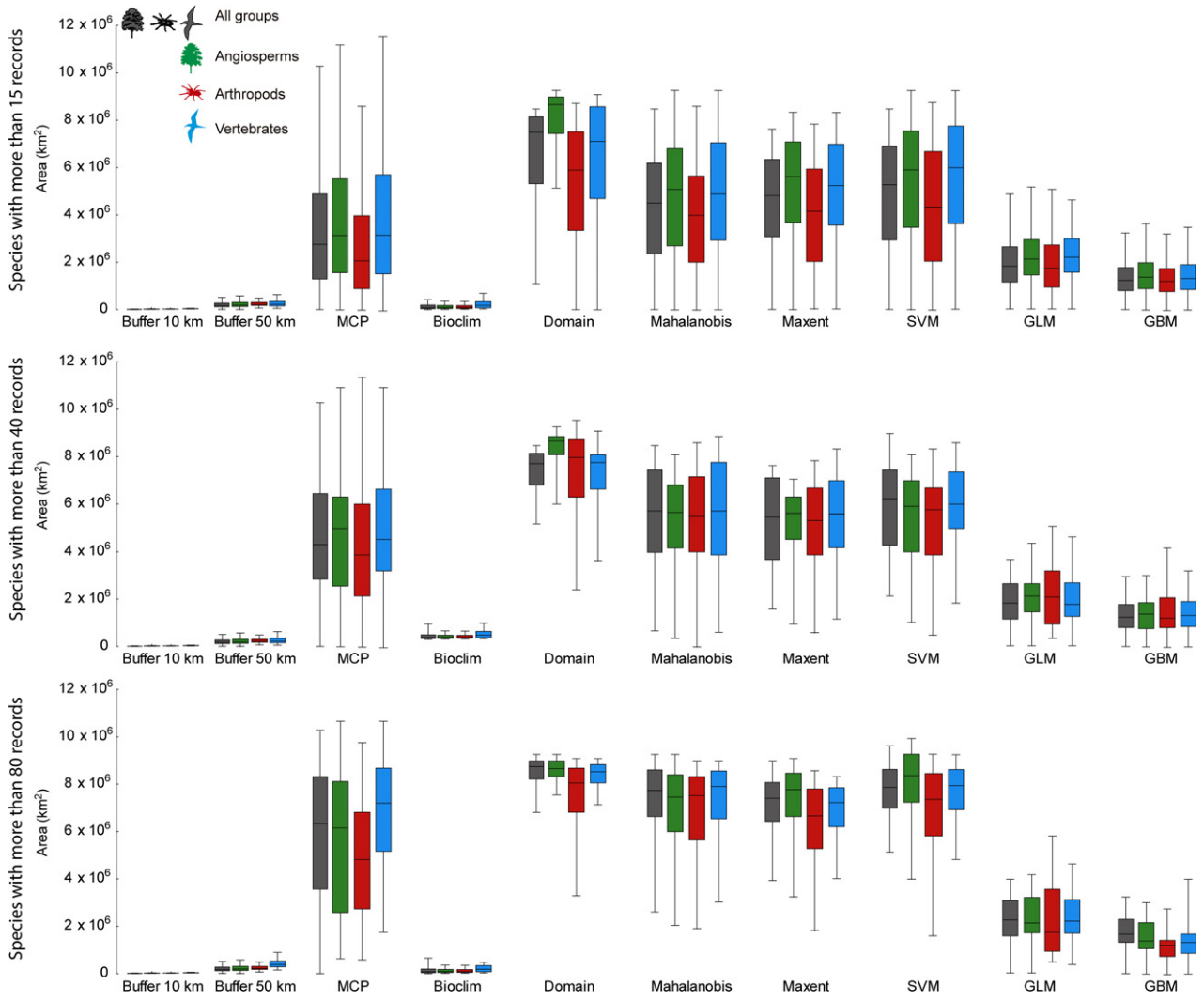

Figure 5 Species distribution areas obtained from empirical data (10 and 50 km buffer and MCP) and SDM's for each data partition. Horizontal line indicates median, box indicates percentiles and line indicates maximum and minimum values. Analyses were repeated with species with at least 15, 40 and 80 reliable distribution records. MCP, minimum convex polygons; SDM, species distribution models. Colour figure can be viewed at wileyonlinelibrary.com

species dissimilarity in relation to the distance from the access routes. However, this pattern was not observed in the Atlantic Rainforest, Pampa and Pantanal. In the Atlantic Rainforest and Pampa, this may be due to the high density of access routes and major impacts suffered over the past few centuries (Tabarelli *et al.*, 2005). In the Pantanal, where the access route density is much lower, this pattern can be caused by lower actual variation in species composition. The differences in sampling intensity observed between the groups did not affect the general pattern of dissimilarity in species composition observed in all groups in relation to the distance to access routes. This indicates that the implementation of inventories in regions distant from access routes should be more effective in increasing knowledge about the species, as had already been suggested from simulations (Sastre & Lobo, 2009).

We showed that the lack of knowledge about the distribution of species in Brazil is very large and equally affects all groups. This similarity in knowledge deficits in all groups contradicts the expected from Cardoso *et al.* (2011), which postulates that arthropods are most affected by the knowledge shortfalls. The SDM's indicated a strong lack of knowledge about the distribution of taxa. Considering that models are affected by incomplete distribution data (Wisz *et al.*, 2008), as we know it is the case of our database, this lack of knowledge about the distribution may be even higher than we observed. Therefore, although our results indicated strong Wallacean shortfall for all groups, that deficit could be even higher. However, the models were not directly affected by Hutchinsonian shortfall, as there is little relationship between the models and the access routes or collection density. The Wallacean shortfall is more worrying because, even the less conservative estimate of the distribution of known species (MCP) was still lower than the distribution predicted by the models (Fig. 5), even in the supposedly

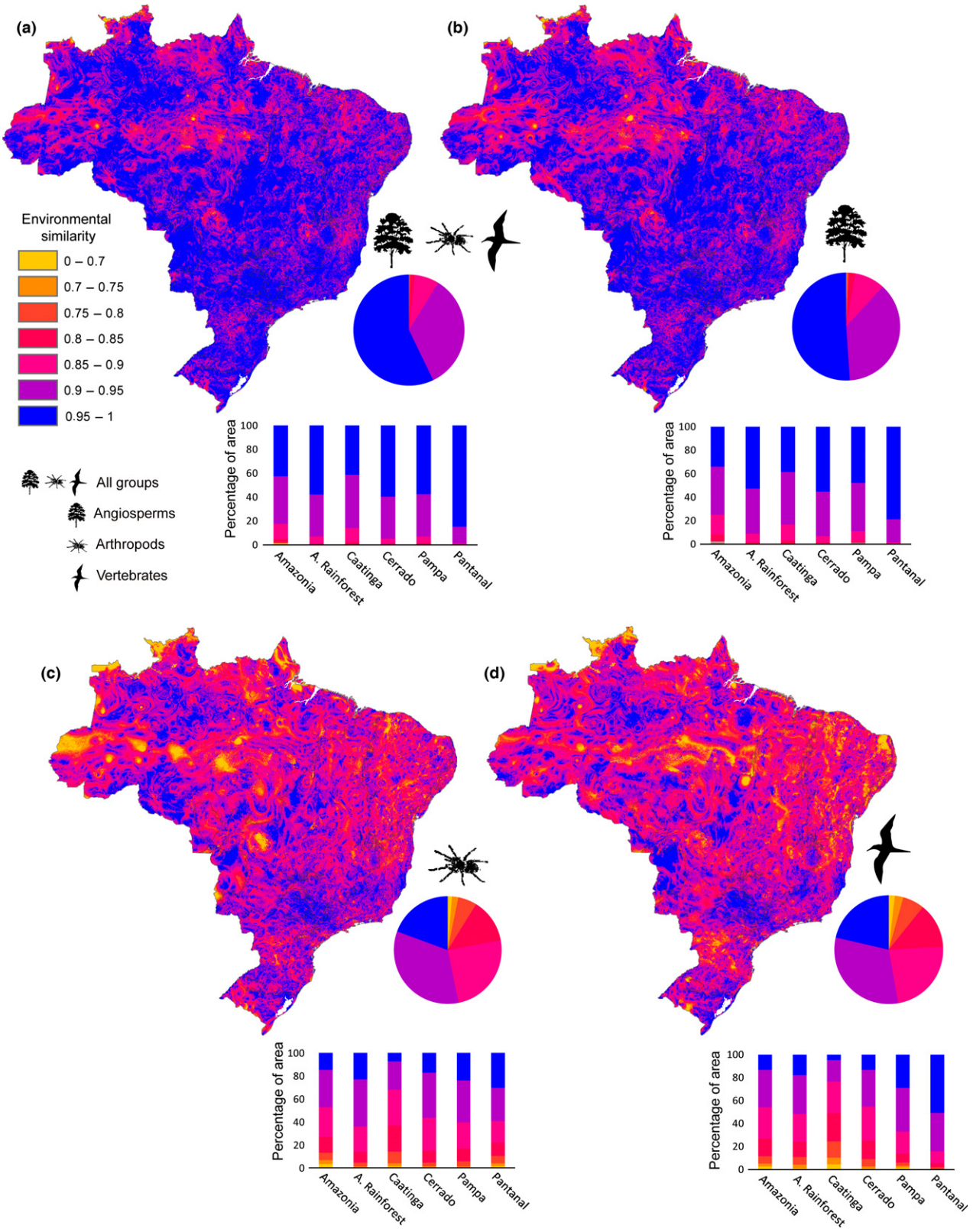

Figure 6 Environmental similarity between distribution points of each taxonomic group and each pixel of a raster map of environmental conditions (climatic PCA) of Brazilian territory. The value of similarity indicates how similar is each site to its most similar distribution point. Graphics show the distribution of similarity categories among the Brazilian biomes. Colour figure can be viewed at wileyonlinelibrary.com

well-known vertebrate groups. Thus, biogeographic studies must take into account the uncertainty about the distribution of taxa, and SDM's can be an alternative for estimating this uncertainty.

Although our results show a weak correlation between beta-diversity and average endemism, these parameters may be affected indirectly by the Wallacean shortfall. Thus, the collecting bias affects the spatial distribution of sampling effort, affecting the different parameters that quantify the diversity, directly or indirectly. This is worrying as this sampling pattern does not allow a good representation of the actual distribution of diversity (Sastre & Lobo, 2009). The direct use of these data can affect the conclusions from macroecology and conservation studies (Moerman & Estabrook, 2006; Grand *et al.*, 2007; Yang *et al.*, 2013). Thus, it is also important to quantify and control this kind of problem in macroecological analysis and conservation programs through uncertainty mapping.

Unlike observed in other forms of biological knowledge shortfall, the Hutchinsonian shortfall was the least affected, directly, by collecting bias. Additionally, the intensity of the Hutchinsonian shortfall differs between taxonomic groups, suggesting that this shortfall is more influenced by the sample intensity, as suggested by studies of SDM (Wisz *et al.*, 2008). This is even more evident in arthropods and vertebrates as the sample gaps in environmental conditions are higher in these groups. The largest number of angiosperm records seems to have a strong effect on the completeness of knowledge about the ecological niche of members of this group, despite the fact that angiosperm records are as biased as those of vertebrates and arthropods. Thus, samples near roads, but in somewhat different environmental conditions of the sampled elsewhere, may have significantly contributed to the lowest Hutchinsonian shortfall in plants. Considering these results, SDM's should provide a better overall prediction for plants than for animals in Brazil. However, the Hutchinsonian shortfall, in all groups, should vary greatly between species of each group analysed, as well as between species within the same taxonomic group.

In this study, we show that, despite the differences in intensity of sampling effort, strong collecting bias affects equally angiosperm, vertebrate and arthropod distribution knowledge in Brazil. Our results suggest that macroecological and biogeographic studies intended to identify processes that determine the biogeographic patterns should consider the direct consequences of biodiversity knowledge shortfalls. Future studies should focus on understanding the differences of Hutchinsonian shortfall between species and quantify other biodiversity knowledge shortfalls not yet explored, as Darwinian and Prestonian shortfalls. It is also worth mentioning that poorly sampled sites may present higher combined Linnaean, Wallacean and Hutchinsonian shortfalls, as unknown species may occur at these sites (Cardoso *et al.*, 2011). These sites can be used to guide biological inventories, optimizing resource use and significantly expanding the knowledge about biodiversity. Additionally to solutions already proposed for these shortfalls (e.g. Cardoso *et al.*, 2011), we suggest that sites distant from access routes, particularly those with poorly sampled environmental conditions, should be preferential targets for future biodiversity surveys.

ACKNOWLEDGEMENTS

We would like to thank Ary Teixeira de Oliveira Filho, Cristiano de Campos Nogueira, Eduardo Andrade Botelho de Almeida and Mario Alberto Cozzuol for discussions, sharing of ideas and critical readings of early versions of the manuscript. This study is part of U. Oliveira's PhD dissertation at 'Pós-graduação em Zoologia da UFMG'. Authors of this study received financial support from: U. Oliveira (CAPES graduate fellowship), CJB de Carvalho (CNPq 304713/2011-2), AD Brescovit (CNPq 303028/2014-9; FAPESP 2011/50689-0), AJSantos (CNPq 407288/2013-9, 306222/2015-9; FAPEMIG PPM 00651-15; and Instituto Nacional de Ciência e Tecnologia dos Hymenoptera Parasitoides da Região Sudeste Brasileira –http://www.hympar.ufscar.br/). Data acquisition on stingless bee specimens at the American Museum of Natural History by JSA occurred with help the from HH Go, A Pfister, M Tuell and ES Wyman. It was supported by RG Goelet and by a NSF-DBI grant (#0956388 with JS Ascher as the P.I., and JG Rozen Jr. and D Yanega as co-P.I.s).

REFERENCES

Bini, L.M., Diniz-filho, J.A.F., Rangel, T.F.L.V.B., Bastos, R.P. & Pinto, M.P. (2006) Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Diversity and Distributions*, **12**, 475–482.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, **8**, e1000385.

Brito, D. (2010) Overcoming the Linnean shortfall: data deficiency and biological survey priorities. *Basic and Applied Ecology*, **11**, 709–713.

Brooks, T.M., Mittermeier, R. a, da Fonseca, G. a B., Gerlach, J., Hoffmann, M., Lamoreux, J.F., Mittermeier, C.G., Pilgrim, J.D. & Rodrigues, a S.L. (2006) Global biodiversity conservation priorities. *Science*, **313**, 58–61.

Cardoso, P., Erwin, T.L., Borges, P.A.V. & New, T.R. (2011) The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, **144**, 2647–2655.

Caro, T., Dobson, A., Marshall, A.J. & Peres, C.A. (2014) Compromise solutions between conservation and road building in the tropics. *Current Biology*, **24**, R722–R725.

Clark, R.W., Brown, W.S., Stechert, R. & Zamudio, K.R. (2010) Roads, interrupted dispersal, and genetic diversity in timber rattlesnakes. *Conservation Biology: The Journal of the Society for Conservation Biology*, **24**, 1059–1069.

Clifford, P., Richardson, S. & Hemon, D. (1989) Assessing the significance of the correlation between two spatial processes. *Biometrics*, **45**, 123–134.

Collen, B., Ram, M., Zamin, T. & Mcrae, L. (2008) The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science*, **1**, 75–88.

Dennis, R., Sparks, T. & Hardy, P. (1999) Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation*, **3**, 33–42.

Diniz-Filho, J.A.F., De Marco, Jr P. & Hawkins, B.A. (2010) Defying the curse of ignorance: perspectives in insect macroecology and conservation biogeography. *Insect Conservation and Diversity*, **3**, 172–179.

Franklin, J. & Miller, J.A. (2009) *Mapping species distributions spatial inference and prediction*. Cambridge University Press, Cambridge.

Freitag, S., Hobson, C., Biggs, H.C. & Van Jaarsveld, A.S. (1998) Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*, **1**, 119–127.

Gaston, K.J. & May, R.M. (1992) Taxonomy of taxonomists. *Nature*, **356**, 281–282.

Grand, J., Cummings, M.P., Rebelo, T.G., Ricketts, T.H., Neel, M.C. & Letters, E. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters*, **10**, 364–374.

Hopkins, M.J.G. (2007) Modelling the known and unknown plant biodiversity of the Amazon Basin. *Journal of Biogeography*, **34**, 1400–1411.

Hortal, J., de 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 Review of Ecology, Evolution, and Systematics*, **46**, 523–549.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Kier, G., Mutke, J., Dinerstein, E., Ricketts, T.H., Küper, W., Kreft, H. & Barthlott, W. (2005) Global patterns of plant diversity and floristic knowledge. *Journal of Biogeography*, **32**, 1107–1116.

Lamoreux, J.F., Morrison, J.C., Ricketts, T.H., Olson, D.M., Dinerstein, E., Mcknight, M.W. & Shugart, H.H. (2006) Global tests of biodiversity concordance and the importance of endemism. *Nature*, **440**, 212–214.

Laurance, W.F. (2009) Roads to ruin: expanding transportation networks imperil global biodiversity. *The multiple faces of globalization* (ed. by C. Gandarias), pp. 198–211. BBVA Group, Spain.

Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jiménez, I., Blake, J.G., Lohmann, L.G. & Montiel, O.M. (2007) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, **35**, 105–116.

Lomolino, M.V. (2004) Conservation biogeography. *Frontiers of biogeography* (ed.by M.V. Lomolino and L.R. Heaney), pp. 293–296. Sinauer, Sunderland, MA.

Meyer, R., Dikow, T. & Meier, R. (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology*, **18**, 478–488.

Mittermeier, R.A., Mittermeier, C.G., Gil, P.R. & Pilgrim, J. (2003) *Wilderness – Earth's last wild places*. Conservation International, Washington, DC.

Moerman, D.E. & Estabrook, G.F. (2006) The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography*, **33**, 1969–1974.

Nelson, B.W., Ferreira, C.A.C., da Silva, M.F. & Kawasaki, M.L. (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature*, **345**, 714–716.

Orme, C.D.L., Davies, R.G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V.A., Webster, A.J., Ding, T.S., Rasmussen, P.C., Ridgely, R.S., Stattersfield, A.J., Bennett, P.M., Blackburn, T.M., Gaston, K.J. & Owens, I.P.F. (2005) Global hotspots of species richness are not congruent with endemism or threat. *Nature*, **436**, 1016–1019.

Reeder, D.M., Helgen, K.M. & Wilson, D.E. (2007) Global trends and biases in new mammal species discoveries. Occasional Papers, 1–35.

Santos, J.C., Leal, I.R., Almeida-Cortez, J.S., Fernandes, G.W. & Tabarelli, M. (2011) Caatinga: the scientific negligence experienced by a dry tropical forest. *Tropical Conservation Science*, **4**, 276–286.

Sastre, P. & Lobo, J.M. (2009) Taxonomist survey biases and the unveiling of biodiversity patterns. *Biological Conservation*, **142**, 462–467.

Sousa-Baena, M.S., Garcia, L.C. & Peterson, A.T. (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions*, **20**, 369–381.

Tabarelli, M., Pinto, L.P., Silva, J.M.C., Hirota, M. & Bede, L. (2005) Challenges and opportunities for biodiversity conservation in the Brazilian Atlantic Forest. *Conservation Biology*, **19**, 695–700.

Tessarolo, G., Rangel, T.F., Araújo, M.B. & Hortal, J. (2014) Uncertainty associated with survey design in Species Distribution Models. *Diversity and Distributions*, **20**, 1258–1269.

Vale, M.M. & Jenkins, C.N. (2012) Across-taxa incongruence in patterns of collecting bias. *Journal of Biogeography*, **39**, 1744–1748.

Werneck, F.P. (2011) The diversification of eastern South American open vegetation biomes: historical biogeography and perspectives. *Quaternary Science Reviews*, **30**, 1630–1648.

Whittaker, R.J., Araújo, M.B., Jepson, P., Ladle, R.J., Watson, J.E.M. & Willis, K.J. (2005) Conservation biogeography: assessment and prospect. *Diversity and Distributions*, **11**, 3–24.

Williams, P.H. & Humphries, C.J. (1994) *Biodiversity, taxonomic relatedness, and endemism in conservation*. Oxford University Press, Oxford.

Wisz, M.S., Hijmans, R.J.J., Li, J., Peterson, A.T., Graham, C.H. & Guisan, A. (2008) Effects of sample size on the

performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Yang, W., Ma, K. & Kreft, H. (2013) Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *Journal of Biogeography*, **40**, 1415–1426.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Detailed information on the database, including data sources, georeferencing criteria, number of records per group, number of species per group and taxonomic catalogues used to access taxonomic identity within each data partition.

Appendix S2 Detailed results and information of database.

Appendix S3 List of species used in the SDM analyses, with the sample size per species.

BIOSKETCH

Ubirajara Oliveira is a PhD in Zoology interested in theoretical, empirical and methodological aspects of biogeography. His current research includes studies on areas of endemism, species distribution models, macroecology, effects of sampling bias on biogeographic methods and on the assessment of species richness, evolutionary biogeography and conservation biogeography.

Editor: Jeremy VanDerWal