Mapping and understanding the digital biodiversity knowledge about vertebrates in the Atlantic Rainforest

Gracielle T. Higino1, 2 \*, Pablo Henrique da Silva3, Marcos V. C. Vital4

1 Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada.

2 Programme de formation en sciences et services de la biodiversité computationnelle, Canada.

3 Instituto de Matemática e Estatística, Universidade Federal de Goiás, Goiânia, Brazil.

4 Instituto de Ciências Biológicas e da Saúde, Universidade Federal de Alagoas, Maceió, AL, Brazil

\*Corresponding author

\*[**graciellehigino@gmail.com**](mailto:graciellehigino@gmail.com)

# Abstract

Biases and gaps in biodiversity data can have serious consequences for ecological and conservation research and actions, as it can lead to incorrect conclusions. Although there is still digital knowledge to be found and organized, recent efforts on gathering global biodiversity data, such as data papers, have revealed long hidden information. Nevertheless, it is of major importance to map and describe the biases on the data we have available. Here we assessed terrestrial vertebrates’ digital inventory incompleteness at the Atlantic Rainforest and investigated if environmental variables are correlated to biodiversity knowledge. At a resolution of 0.5 degrees and considering the final slope of a rarefaction curve as a measure of inventory completeness, none of the grids can be considered well sampled units. However, the completeness of each group seems to have a non-uniform correlation to different environmental variables. By highlighting undersampled areas in the Atlantic Rainforest, these results provide important information for the biodiversity assessment of this highly disturbed ecoregion. Combining them with the information about the magnitude of these impacts can help shape the agenda and priorities for conservation, but, also, the results alone can help us rethink what we understand about geographically structured biodiversity distribution.  
Keywords: inventory completeness, rarefaction curves, data bias, Atlantic Rainforest.

# Introduction

Information about life diversity and distribution is a fundamental tool for understanding evolutionary and ecological processes (Graham *et al.* 2004; Rocchini *et al.* 2011; Jetz *et al.* 2012; Ladle & Hortal 2013; Meyer *et al.* 2015). After a long history of biodiversity information collection by naturalists, taxonomists and, more recently, citizen scientists (Von Humboldt *et al.* 1850; Hawkins 2001; Willig *et al.* 2003; Chase 2012; Callaghan *et al.* 2020), researchers have been storing these data in electronic cataloges at slow pace since the 1970’s, connecting them, more recently, through web-based initiatives (Graham *et al.* 2004). As a result, we now have accessible and extensive information about biodiversity on big online databases such as the Global Biodiversity Information Facility (GBIF; <http://www.gbif.org/>) and the Map of Life (<https://mol.org/>), which compile museum, survey and observation data (Graham *et al.* 2004; Jetz *et al.* 2012; Beck *et al.* 2013). However, despite these recent efforts, our knowledge on species diversity and distribution is still biased and full of gaps due to the complex nature of this information (Brown & Lomolino 1998; Whittaker *et al.* 2005). Different aspects of shortfalls related to this knowledge have been recently revised (Hortal *et al.* 2015) and there is growing evidence that they can compromise ecological, evolutionary and conservation analyses (IUCN 2012; Ladle & Hortal 2013; Ficetola *et al.* 2014; Hortal *et al.* 2015).

The Wallacean shortfall (the lack of information about species’ real distribution) is present in every spatial and temporal scales (Whittaker *et al.* 2005; Hortal 2008; Hortal *et al.* 2015) and is a consequence of a myriad of biological, environmental and social factors. Characteristics of the species (such as crypsis, its natural history and behaviour), political borders and topography, for example, can lead to biases in biodiversity surveys and form gaps in information. On the other hand, clustered information also can lead to biased surveys, since researchers may prefer to assess places knowingly species-rich or that are undergoing a process of ecological change (Boakes *et al.* 2010; Ahrends *et al.* 2011; Rocchini *et al.* 2011; Yang *et al.* 2014). Information gaps may also be a consequence of data quality decay in space (e.g., when we extrapolate the distribution of a species based on polygons or species distribution models) and time (due to taxonomic reviews, climate change, land use, habitat loss, extinction and migration) (Ladle & Hortal 2013; Tessarolo *et al.* 2017). Therefore, the measurement of geographical variation of biodiversity on the planet (represented by distribution maps) has an error associated that must be assessed (Hortal 2008; Rocchini *et al.* 2011; Ladle & Hortal 2013; Yang *et al.* 2013).

The underestimation of species distribution can have consequences in conservation planning, such as the classification of species in risk of extinction following the range restriction criterion (IUCN 2012; Ladle & Hortal 2013; Ficetola *et al.* 2014; Hortal *et al.* 2015). Furthermore, bias can influence and even reverse ecogeographical patterns, leading us to associate certain factors to species richness when they are only proxies for sampling quality (Ficetola *et al.* 2014). These are examples of why the acknowledgment of error in biodiversity information is of major importance. For that reason, it has been recommended to include maps of ignorance in the results or to map data quality and use only well sampled locations on analyses (Hortal 2008; Ladle & Hortal 2013; Ficetola *et al.* 2014; Yang *et al.* 2014). These maps could help detecting biasing variables, and the researcher could therefore chose the most apropriate procedure to deal with them - by removing them or adding weights, for example (Stockwell & Peterson 2002; Boakes *et al.* 2010) -, in order to perform better analyses, but they are only possible once researchers are aware of the error in their data sets. This practice, in addition to guide future research, produces more reliable results, since the exact measure of uncertainty clarifies how explicative an inference can be.

There is a growing interest in biodiversity data biases in the literature (see Boakes *et al*. (2010); Yang *et al*. (2013); Sousa-Baena *et al*. (2014)). Nevertheless, studies mapping South American under-sampled sites are relatively few. This is worrying especially for the Atlantic Forest because this ecoregion is an important biodiversity and socio-climatic hotspot (Scarano & Ceotto 2015). Human activities and the growth of urban centres have significantly reduced its original, and estimations suggest the remaining native vegetation area to be of only 28% (Rezende *et al.* 2018), while others suggest it could be as little as 8% (Galindo-Leal & Câmara 2003; Scarano & Ceotto 2015), resulting in substantial loss of habitat. Despite that, it still hosts 1-8% of the world’s total species (Silva & Casteleti 2003). The Atlantic Rainforest is also a good model for ecological and evolutionary research because of its large latitudinal and altitudinal range, high endemicity, variation in temperature and precipitation, and historical connexion with other biomes (Silva & Casteleti 2003; Ribeiro *et al.* 2009; Batalha-Filho *et al.* 2013). Given that the Atlantic Rainforest is a highly impacted biodiversity hotspot, it becomes urgent to describe and map the information about species occurrence in this ecoregion that is available online.

Here we aim to map and quantify the gaps on digital occurrence data of terrestrial vertebrates available on GBIF, Integrated Digitized Biocollections (iDigBio; www.idigbio.org) and those published by the ATLANTIC project data papers (Bovendorp *et al.* 2017; Lima *et al.* 2017; Muylaert *et al.* 2017; Gonçalves *et al.* 2018; Hasui *et al.* 2018; Vancine *et al.* 2018; Culot *et al.* 2019; Souza *et al.* 2019). Although there is valuable and important data stored on other data repositories such as Dryad and Zenodo, they do not follow a standard structure, and very often miss quality metadata (Rousidis *et al.* 2014), which often makes their use in macroecological research more challenging. Moreover, the most frequent use of the GBIF data is in species distribution research (Heberling *et al.* 2021), given the large-resolution, integrated datasets design (König *et al.* 2019). The GBIF database also encompasses other more specific and local databases, such as eBird (a specialized database for birds; Sullivan *et al.* (2009)) and SiBBr, the Brazilian node of GBIF focused on local occurrence records (Sistema de Informação sobre a Biodiversidade Brasileira, www.sibbr.gov.br).

As a further investigation, we try to identify environmental variables that may be related to these shortfalls. As beforementioned, the Wallacean shortfall can be the result of many factors, including environmental Because of the extensive use of large biodiversity databases to investigate species richness distribution and its relationship with environmental gradients, we hypothesize that these relationships could influence the amount of samples taken at each location. Therefore we chose four environmental variables that are known to have positive relationships with species richness at large scale (mean annual temperature - characterizing warmer regions -, altitudinal range variation - characterizing regions with high speciation probability - and potential evapotranspiration - characterizing highly productive environments). Additionally, we investigated weather the inventory completeness is related to distance to conservation units, since preserved areas could be highly attractive to naturalists.

# Methods

### Data collection and cleaning

The inventory completeness of Atlantic Rainforest fauna was analysed for amphibians, birds, mammals and reptiles. We used occurrence data from the Global Biodiversity Information Facility, which was downloaded on March 16th 2021 by classes’ names, using geometric filtering and excluding fossil records (GBIF.org 2021). We also used all datasets from the ATLANTIC project related to these groups of vertebrates (complete list of data sources available in Supplementary Table 1). Since not all databases had the same metadata, we filtered occurrences based only on common variables, such as scientific name, class and geographic coordinates.

We checked scientific names for validity using the ‘taxize’ R package (Chamberlain & Szöcs 2013; Chamberlain *et al.* 2014), based on the National Center for Biotechnology Information Taxonomy Database (<http://www.ncbi.nlm.nih.gov/taxonomy>), Mammal Species of the World (3rd edition, <http://vertebrates.si.edu/msw/mswcfapp/msw/index.cfm>) and The Reptile Database (<http://reptile-database.reptarium.cz/>), BirdLife International (<https://www.birdlife.org/>), Integrated Taxonomic Information System (<https://www.itis.gov/>) and GBIF. We filtered the remaining occurrence points by the Atlantic Rainforest domain *sensu* Olson *et al*. (2001) (Fig. 01) and then assessed species richness and number of occurrences by grid cells of 30 arc-minutes (~55km at Equator), which we used for calculating inventory completeness. This resolution was chosen because it captures macroecological trends while revealing important regional variations. It was not in the scope of this study to investigate the inventory completeness and its relationship with environmental variables at local scale, nor should we use coarser resolutions given the limited extension of the study area.

### Inventory completeness

Two approaches were used to evaluate inventory completeness in the Atlantic Forest: the species accumulation curve for the whole region followed by the analysis of its final 10% slope (Yang *et al.* 2013), and a rarefaction method for each sampling unit (SU), also assessing the sample slope. The species accumulation curve is a sample-based method for assessing sampling effort and estimate species richness (Colwell & Coddington 1994; Gotelli & Colwell 2001). This approach was performed with the method ‘exact’ of the function‘specaccum’ and the final 10% slopes were extracted with function ‘specslope’ in the R package ‘vegan’ (Oksanen *et al.* 2015). The output of this function was analysed according to Yang *et al*. (2013) and slopes higher than 0.05 were considered as indicators of inventory incompleteness. The rarefaction curve is an individual-based method that represents the sampling effort needed to reach total estimated richness within an area (Gotelli & Colwell 2001). <!—>review the definition of rarrefaction curve and explicit what was used as samples in both rarefy and rareslope functions<—>The ‘rarefy’ and ‘rareslope’ functions in R calculate the rarefaction and slopes of each SU, both operating in the same package abovementioned.

### Geographic analyses

Additionally, we investigated if four environmental variables were correlated with inventory completeness. These variables were chosen based on previous studies indicating their influence on data bias or if they are commonly mentioned as proxies for species richness (Currie 1991; Sánchez-Fernández *et al.* 2008; Toranza & Arim 2010; Martin *et al.* 2012; Ficetola *et al.* 2014; Vasconcelos *et al.* 2014; Yang *et al.* 2014). We used annual mean temperature and altitudinal range downloaded from WorldClim (resolution 30“), through the package raster (Hijmans 2016) and potential evapotranspiration (PET) data from the Consortium for Spatial Information of the Consultative Group for International Agricultural Research (resolution 30”) (Trabucco & Zomer 2009). Temperature and PET represent ecosystems’ energy income, while altitudinal range represents topographical and, therefore, habitat homogeneity. All variables were also rescaled to 30 arc-seconds resolution grids.

TODO #11 Análise descritiva das variáveis ambientais e rodar um stepwise regression para selecionar as variáveis que entram no modelo. Entender para cada grupo quais variáveis que são mais interessantes para o gwr.

We did a Geographically Weighted Regression (GWR) (Brunsdon *et al.* 1998), that considers the spatial structure of the data, to investigate if these correlations are consistent through space. This analysis was made with the ‘spgwr’ package (Bivand & Yu 2017). Finally, we assessed taxonomic bias by investigating collinearity through a Principal Component Analysis (PCA) of all the slopes, using the ‘prcomp’ function of the package ‘vegan.’ Statistical analyses were performed using the computing environment R 3.5.1 (R Core Team 2018).

# References

Ahrends, A., Burgess, N.D., Gereau, R.E., Marchant, R., Bulling, M.T., Lovett, J.C., *et al.* (2011). Funding begets biodiversity. *Diversity and Distributions*, 17, 191–200.

Batalha-Filho, H., Fjeldså, J., Fabre, P.H. & Miyaki, C.Y. (2013). Connections between the Atlantic and the Amazonian forest avifaunas represent distinct historical events. *Journal of Ornithology*, 154, 41–50.

Beck, J., Ballesteros-Mejia, L., Nagel, P. & Kitching, I.J. (2013). Online solutions and the ’Wallacean shortfall’: What does GBIF contribute to our knowledge of species’ ranges? *Diversity and Distributions*, 19, 1043–1050.

Bivand, R. & Yu, D. (2017). *Spgwr: Geographically weighted regression*. Manual.

Boakes, E.H., McGowan, P.J., Fuller, R.A., Chang-qing, D., Clark, N.E., O’Connor, K., *et al.* (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biol*, 8, e1000385.

Bovendorp, R.S., Villar, N., de Abreu-Junior, E.F., Bello, C., Regolin, A.L., Percequillo, A.R., *et al.* (2017). Atlantic small-mammal: A dataset of communities of rodents and marsupials of the Atlantic forests of South America. *Ecology*, 98, 2226.

Brown, J.H. & Lomolino, M.V. (1998). *Biogeography*. Second. Sinauer Associates, Inc. Publishers, Massachusetts.

Brunsdon, C., Fotheringham, S. & Charlton, M. (1998). Geographically Weighted Regression-Modelling Spatial Non-Stationarity. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47, 431–443.

Callaghan, C.T., Roberts, J.D., Poore, A.G.B., Alford, R.A., Cogger, H. & Rowley, J.J.L. (2020). Citizen science data accurately predicts expert-derived species richness at a continental scale when sampling thresholds are met. *Biodiversity and Conservation*.

Chamberlain, S. a & Szöcs, E. (2013). Taxize: Taxonomic search and retrieval in R. *F1000Research*, 2, 191.

Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I. & Baumgartner, J. (2014). Taxize: Taxonomic information from around the web.

Chase, J. (2012). Historical and Contemporary Factors Govern Global Biodiversity Patterns. *PLoS Biol*, 10, e1001294.

Colwell, R.K. & Coddington, J.A. (1994). Estimating Terrestrial Biodiversity through Extrapolation. *Philosophical Transactions: Biological Sciences*, 345, 101–118.

Culot, L., Pereira, L.A., Agostini, I., de Almeida, M.A.B., Alves, R.S.C., Aximoff, I., *et al.* (2019). ATLANTIC-PRIMATES: A dataset of communities and occurrences of primates in the Atlantic Forests of South America. *Ecology*, 100, e02525.

Currie, D.J. (1991). Energy and large-scale patterns of animal - and plant - species richness. *The American Naturalist*, 137, 27–49.

Ficetola, G.F., Cagnetta, M., Padoa-Schioppa, E., Quas, A., Razzetti, E., Sindaco, R., *et al.* (2014). Sampling bias inverts ecogeographical relationships in island reptiles. *Global Ecology and Biogeography*, 1303–1313.

Galindo-Leal, C. & Câmara, I.G. (2003). Atlantic Forest hotspots status: An overview. In: *The Atlantic Forest of south america: Biodiversity status, threats, and outlook* (eds. Galind-Leal, C. & Câmara, I.G.). Conservation International, pp. 3–11.

GBIF.org. (2021). GBIF Occurrence Download.

Gonçalves, F., Hannibal, W., Godoi, M.N., Martins, F.I., Oliveira, R.F., Figueiredo, V.V., *et al.* (2018). Non-volant mammals from the Upper Paraná River Basin: A data set from a critical region for conservation in Brazil. *Ecology*, 99, 499.

Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4, 379–391.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, a.T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, 19, 497–503.

Hasui, É., Metzger, J.P., Pimentel, R.G., Silveira, L.F., Bovo, A.A.d.A., Martensen, A.C., *et al.* (2018). ATLANTIC BIRDS: A data set of bird species from the Brazilian Atlantic Forest. *Ecology*, 99, 497.

Hawkins, B.A. (2001). Ecology’s oldest pattern? *Trends in Ecology & Evolution*, 16, 470.

Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B. & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118.

Hijmans, R.J. (2016). *Raster: Geographic Data Analysis and Modeling*.

Hortal, J. (2008). Uncertainty and the measurement of terrestrial biodiversity gradients. *Journal of Biogeography*, 35, 1335–1336.

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.

IUCN. (2012). *IUCN Red List Categories and Criteria Version 3.1 Second Edition*. *Gland*. IUCN, Switzerland and Cambridge, UK.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012). Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology and Evolution*, 27, 151–159.

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J. & Kreft, H. (2019). Biodiversity data integrationthe significance of data resolution and domain. *PLOS Biology*, 17, e3000183.

Ladle, R. & Hortal, J. (2013). Mapping species distributions: Living with uncertainty. *Frontiers of Biogeography*, 5, 4–6.

Lima, F., Beca, G., Muylaert, R.L., Jenkins, C.N., Perilli, M.L.L., Paschoal, A.M.O., *et al.* (2017). ATLANTIC-CAMTRAPS: A dataset of medium and large terrestrial mammal communities in the Atlantic Forest of South America. *Ecology*, 98, 2979.

Martin, L.J., Blossey, B. & Ellis, E. (2012). Mapping where ecologists work: Biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, 10, 195–201.

Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, 6, 8221.

Muylaert, R.d.L., Stevens, R.D., Esbérard, C.E.L., Mello, M.A.R., Garbino, G.S.T., Varzinczak, L.H., *et al.* (2017). ATLANTIC BATS: A data set of bat communities from the Atlantic Forests of South America. *Ecology*, 98, 3227.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O’Hara, R.B., *et al.* (2015). *Vegan: Community Ecology Package*.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., *et al.* (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, 51, 933.

R Core Team. (2018). *R: A language and environment for statistical computing*. Manual. R Foundation for Statistical Computing, Vienna, Austria.

Rezende, C.L., Scarano, F.R., Assad, E.D., Joly, C.A., Metzger, J.P., Strassburg, B.B.N., *et al.* (2018). From hotspot to hopespot: An opportunity for the Brazilian Atlantic Forest. *Perspectives in Ecology and Conservation*, 16, 208–214.

Ribeiro, M.C., Metzger, J.P., Martensen, A.C., Ponzoni, F.J. & Hirota, M.M. (2009). The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. *Biological Conservation*, 142, 1141–1153.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C., *et al.* (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, 35, 211–226.

Rousidis, D., Garoufallou, E., Balatsoukas, P. & Sicilia, M.-A. (2014). Metadata for big data: A preliminary investigation of metadata quality issues in research data repositories. *Information Services & Use*, 34, 279–286.

Sánchez-Fernández, D., Lobo, J.M., Abellán, P., Ribera, I. & Millán, A. (2008). Bias in freshwater biodiversity sampling: The case of Iberian water beetles. *Diversity and Distributions*, 14, 754–762.

Scarano, F.R. & Ceotto, P. (2015). Brazilian Atlantic forest: Impact, vulnerability, and adaptation to climate change. *Biodiversity and Conservation*, 24, 2319–2331.

Silva, J.M.C. & Casteleti, C.H.M. (2003). Status of the biodiversity of the Atlantic Forest of Brazil. In: *The Atlantic Forest of South America: Biodiversity, status, trends, and outlook.* (eds. Galindo-Leal, C. & Câmara, I. de G.). Center for Applied Biodiversity Science and Island Press, pp. 43–59.

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.

Souza, Y., Gonçalves, F., Lautenschlager, L., Akkawi, P., Mendes, C., Carvalho, M.M., *et al.* (2019). ATLANTIC MAMMALS: A data set of assemblages of medium- and large-sized mammals of the Atlantic Forest of South America. *Ecology*, 100, e02785.

Stockwell, D. & Peterson, A.T. (2002). Controlling bias in biodiversity data. *Predicting species occurrences: issues of scale and accuracy. Island Press, Washington, DC*, 537–546.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142, 2282–2292.

Tessarolo, G., Ladle, R., Rangel, T. & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution*, 7, 6863–6870.

Toranza, C. & Arim, M. (2010). Cross-taxon congruence and environmental conditions. *BMC ecology*, 10, 18.

Trabucco, A. & Zomer, R.J. (2009). *Global Potential Evapo-Transpiration (Global-PET) and Global Aridity Index (Global-Aridity) Geo-Database*. *CGIAR Consortium for Spatial Information*.

Vancine, M.H., Duarte, K. da S., de Souza, Y.S., Giovanelli, J.G.R., Martins-Sobrinho, P.M., López, A., *et al.* (2018). ATLANTIC AMPHIBIANS: A data set of amphibian communities from the Atlantic Forests of South America. *Ecology*, 99, 1692.

Vasconcelos, T.S., Prado, V.H.M., da Silva, F.R. & Haddad, C.F.B. (2014). Biogeographic distribution patterns and their correlates in the diverse frog fauna of the Atlantic Forest hotspot. *PloS one*, 9, e104130.

Von Humboldt, A., Otté, E.C. & Bohn, H.G. (1850). *Views of Nature: Or Contemplations on the Sublime Phenomena of Creation*. Cambridge University Press.

Whittaker, R.J., Araujo, M.B., Jepson, L., P., Watson, J.E.M. & Willis, K.J. (2005). Conservation biogeography: Assessment and prospect. *Diversity and Distributions*, 11, 3–23.

Willig, M.R., Kaufman, D.M. & Stevens, R.D. (2003). Latitudinal Gradients of Biodiversity: Pattern, Process, Scale, and Synthesis. *Annual Review of Ecology, Evolution, and Systematics*, 34, 273–309.

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.

Yang, W., Ma, K. & Kreft, H. (2014). Environmental and socio-economic factors shaping the geography of floristic collections in China. *Global Ecology and Biogeography*, 23, 1284–1292.