Title:
Limitations in global information on species occurrences

Journal Issue:
Frontiers of Biogeography, 8(2)

Author:
Meyer, Carsten, University of Göttingen

Publication Date:
2016

Permalink:
http://escholarship.org/uc/item/1bm1d0hs

Acknowledgements:
I thank my supervisor, Holger Kreft, for constant support with fund-raising, study design, implementation, writing, and publishing. I thank my other co-authors, Susanne Fritz, Rob Guralnick, Holger Kreft, Walter Jetz and Patrick Weigelt for fruitful collaborations. I thank Yael Kisel and Patrick Weigelt for comments on this manuscript. I am grateful to the Deutsche Bundesstiftung Umwelt (DBU) and the German Academic Exchange Service (DAAD) for PhD scholarships, and the Unibund Göttingen for additional funding of a research stay in the Jetz lab.

Author Bio:
PhDBiodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences, University of Göttingen

Keywords:
Wallacean shortfall, species distributions, occurrence records, data mobilization, data bias, knowledge gaps, data uncertainty

Local Identifier:
fb_28195

Abstract:
Detailed information on species distributions is crucial for answering central questions in biogeography, ecology, evolutionary biology and conservation. Millions of species occurrence records have been mobilized via international data-sharing networks, but inherent biases, gaps and uncertainties hamper broader application. In my PhD thesis, I presented the first comprehensive analyses of global patterns and drivers of these limitations across different taxonomic groups and spatial scales. Integrating 300 million occurrence records for terrestrial vertebrates and plants with comprehensive taxonomic databases, expert range maps and regional checklists, I demonstrated extensive taxonomic, geographical and temporal biases, gaps and uncertainties. I identified key socio-economic drivers of data bias across different taxonomic

groups and spatial scales. The results of my dissertation provide an empirical baseline for effectively accounting for data limitations in distribution models, as well as for prioritizing and monitoring efforts to collate additional occurrence information.

opinions, perspectives & reviews frontiers of biogeography 8.2, e28195, 2016

thesis abstract

Limitations in global information on species occurrences

Carsten Meyer

Biodiversity, Macroecology & Conservation Biogeography Group, University of Göttingen, Göttingen, Germany; carsten.meyer@idiv.de

Abstract. Detailed information on species distributions is crucial for answering central questions in biogeography, ecology, evolutionary biology and conservation. Millions of species occurrence records have been mobilized via international data-sharing networks, but inherent biases, gaps and uncertainties hamper broader application. In my PhD thesis, I presented the first comprehensive analyses of global patterns and drivers of these limitations across different taxonomic groups and spatial scales. Integrating 300 million occurrence records for terrestrial vertebrates and plants with comprehensive taxonomic databases, expert range maps and regional checklists, I demonstrated extensive taxonomic, geographical and temporal biases, gaps and uncertainties. I identified key socio-economic drivers of data bias across different taxonomic groups and spatial scales. The results of my dissertation provide an empirical baseline for effectively accounting for data limitations in distribution models, as well as for prioritizing and monitoring efforts to collate additional occurrence information.

Keywords. Wallacean shortfall, species distributions, occurrence records, data mobilization, data bias, knowledge gaps, data uncertainty

Introduction

Detailed information on species' distributions is crucial for answering central questions in biogeography (Lomolino 2004), ecology (Brown et al. 1996) and evolutionary biology (Holt 2003). Such information is also necessary for the effective allocation of conservation resources (Boitani et al. 2011). In particular, there are many questions that require distribution information over broad spatial extents and at fine spatial grains – for instance, to inform conservation prioritization at scales that match land-use changes and management options (Boitani et al. 2011). Similarly, high temporal coverage of distribution datasets is required to study species' responses to environmental change (Boakes et al. 2010), and for policy-relevant indices of biodiversity change (Butchart et al. 2010). Such detail may come directly from field data, or from modelling approaches such as species distribution modelling (Guisan and Thuiller 2005) or downscaling (Keil et al. 2013).

Huge numbers of occurrence records, especially from preserved specimens and field observations, have been mobilized via international data-sharing networks, most importantly that of the Global Biodiversity Information Facility (GBIF). Such records provide the primary information on the taxonomic, geographical and temporal dimensions of species' distributions, because they provide direct evidence that a particular species occurred at a particular location at a particular point in time (Soberón and Peterson 2004). GBIF-facilitated records represent by far the largest share of species occurrence information that is both digital and easily accessible in a standard format (hereafter referred to as DAI – *digital accessible information*; originally referred to as *digital accessible knowledge* in Sousa-Baena et al. 2014).

Notwithstanding the increasing accessibility of occurrence information, global knowledge of species' distributions remains extremely limited, a situation termed the 'Wallacean shortfall' (Lomolino 2004). This shortfall is a necessary result of humans' limited and spatio-temporally uneven capacity to collect, organize and process species occurrence information (Hortal et al. 2015). As a result, most taxa and regions lack large-extent, fine-grain datasets, and existing information is furthermore often scattered across multiple sources (Jetz et al. 2012). Moreover, even

available information is prone to many uncertainties, for example from ambiguous scientific names (Jansen and Dengler 2010), imprecisely georeferenced sampling locations (Rocchini et al. 2011) and old age of many records (Boakes et al. 2010). Finally, because most occurrence records were collected opportunistically (ter Steege et al. 2011), they inherit taxonomic, geographical and temporal biases (Dennis and Thomas 2000, Boakes et al. 2010). These biases hamper many important applications, including species distribution modelling (Guisan and Thuiller 2005), macroecological analyses (Yang et al. 2013) and conservation prioritization (Boitani et al. 2011).

Geographical biases may be driven by biased field work, which may result from regional differences in accessibility (Dennis and Thomas 2000), safety concerns (Amano and Sutherland 2013), lack of funding (Ahrends et al. 2011) or preferential interest in endemism-rich, mountainous or protected areas (Soria-Auza and Kessler 2008). However, biases in DAI may also be caused by biased provision of existing information, related to regional differences in financial or institutional resources for digitization (Vollmar et al. 2010), or poor scientific (Amano and Sutherland 2013) or political (Yesson et al. 2007) cooperation that inhibits mobilization into data-sharing networks. Biases towards certain species might reflect such site-specific socio-economic factors, but may also reflect species-specific factors such as lower detectability of nocturnal (Burton 2012) or arboreal species (Chutipong et al. 2014), or deliberate withholding of occurrence records for threatened species (Whitlock et al. 2010). Finally, the geometry of distributional ranges may affect the likelihood that a given researcher's study region intersects with a given species' range, which in turn affects the likelihood that this particular species is recorded.

It is increasingly urgent to address these multiple limitations in DAI, given its many applications in ecology and conservation. The need for better baseline information on species distributions has been frequently emphasized by the scientific community (Lomolino 2004, Boitani et al. 2011). Improving such information is also closely linked to international targets under the framework of the United Nations Convention on Biological Diversity, and plays a central role in current discussions in the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. However, limited funding and the sheer magnitude of the Wallacean shortfall imply that severe limitation in occurrence information will always persist. This makes it imperative to prioritize future data collection and mobilization (Hobern et al. 2013, Sousa-Baena et al. 2014), and also to develop tools to more effectively account for limitations in available information. Improving species' distribution information requires a thorough understanding of global patterns in data limitations and of the underlying causes. Understanding which factors cause biases can help account for these key factors in ecological models by explicitly incorporating them as variables (Dorazio 2014, Fithian et al. 2015). Previous studies of patterns and drivers of distribution information were limited in geographical (Ballesteros-Mejia et al. 2013) or taxonomic (Yesson et al. 2007) scope, by the limited number of tested hypotheses, or by simplistic treatment of distribution information. Before my thesis research, no study had tested the generality of the various information-limiting factors globally across different taxonomic and spatial scales. The main goals of my PhD thesis (Meyer 2015) were therefore to provide:

a) the first global, detailed analyses of limitations in mobilized occurrence information for a large section of biodiversity;
b) a better understanding of global taxonomic, geographical and temporal variation in different aspects of occurrence information;
c) a better understanding of global drivers of this variation across different taxonomic groups and spatial scales;
d) an empirical baseline for prioritizing data collection and mobilization, for monitoring these activities, and for effectively accounting for data limitations in ecological models.

Methods

In chapter 1 (Meyer et al. 2016b), I focused on land plants. I obtained ca. 120M records from

GBIF, standardized taxonomic information against comprehensive taxonomic databases and carried out plausibility checks of the recorded sampling locations. I used the resulting vetted dataset to calculate metrics describing two main aspects of occurrence information, each with regard to the three basic dimensions that characterize species distributions: taxonomy, space and time (Fig. 1). The first set of metrics quantified aspects of coverage of each dimension with information and the second set of metrics quantified uncertainty regarding the interpretation of information. I measured taxonomic, geographical and temporal variation in these information aspects and assessed their relationships using pairwise correlations and principal components analysis.

In chapter 2 (Meyer et al. 2015), I focused on terrestrial vertebrates and analyzed two aspects of occurrence information at the level of geographical assemblages (Fig. 1) based on ca. 183M records from GBIF (1.7M for amphibians, 177M for terrestrial birds, 4.7M for terrestrial mammals). I standardized species' names and used expert range maps to validate records geographically (details in chapters 2–3). I calculated two measures of coverage, i) the density of rec-

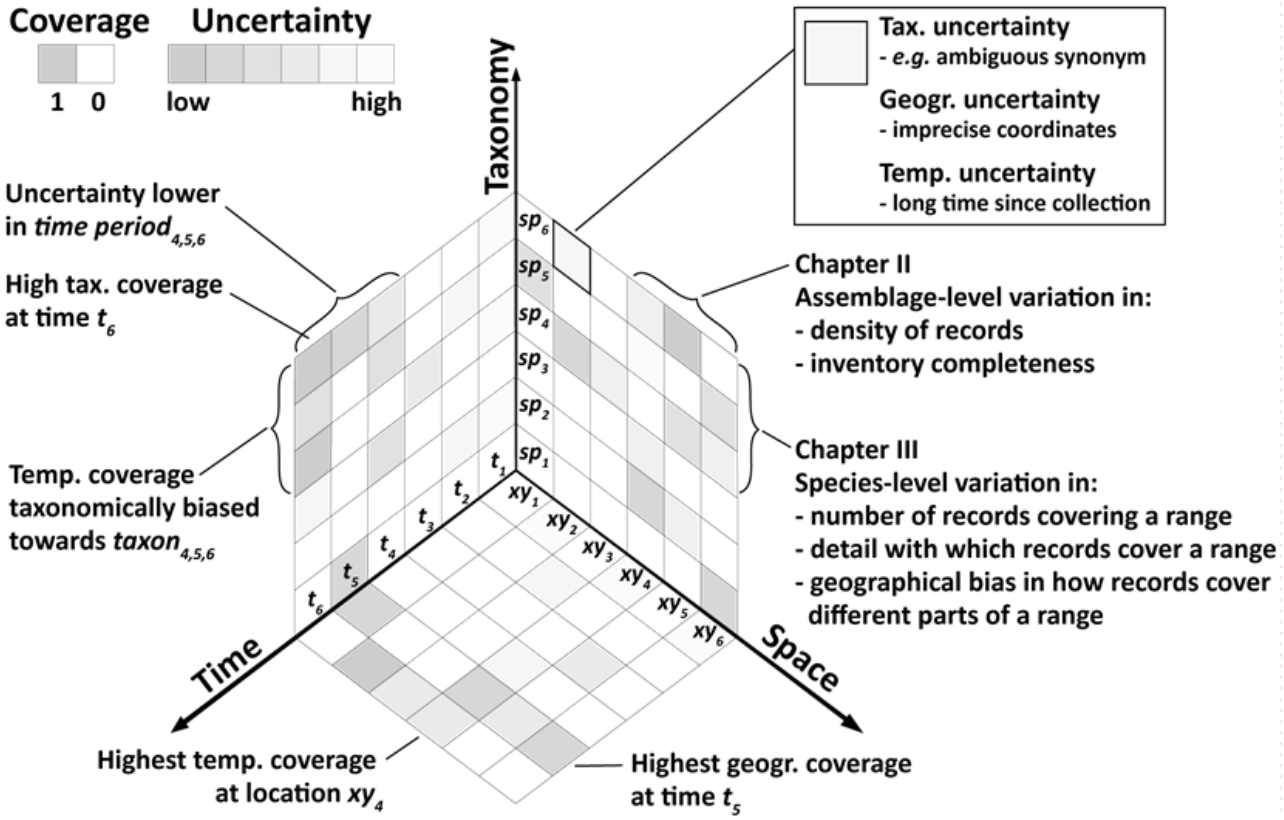

Figure 1. Framework for analyzing limitations in occurrence information (figure and caption text adapted from Meyer et al. 2016b). Species' distributions are characterized by three main dimensions: taxonomy, space and time. Occurrence records provide direct evidence that particular species (sp_1, sp_2, …) occurred at particular locations (xy_1, xy_2, …) at particular points in time (t_1, t_2, …). Planes of cells illustrate spread of information between pairs of dimensions; occurrence information from anywhere along the third dimension is vertically projected onto the plane. Integrating across cells in one dimension summarizes information per unit of the other dimension (e.g. bottom right: highest geographical coverage at time t_5 because four out of six *xy* locations have occurrences then). In chapter 1, I studied two main aspects of occurrence information that determine applicability in research and conservation: i) coverage of the three dimensions with information (grey cells), and ii) uncertainty regarding the interpretation of information (shade of grey cells). Uncertainty may consist of different components. Both coverage and uncertainty may vary in each of the three dimensions, potentially leading to biases (see curly brackets for examples; e.g. centre left: temporal coverage is taxonomically biased because species 4, 5 and 6 have systematically higher coverage, compared to species 1, 2 and 3). In chapters 2 and 3, I focused on specific aspects of coverage. In chapter 2, I compared record density and inventory completeness across geographical assemblages; in chapter 3, I compared record count, range coverage and within-range geographical bias across species.

ords and ii) inventory completeness, calculated as the percentage of expert-opinion species richness (inferred from range maps) that is documented by records. I tested twelve hypotheses on the geographical and socio-economic drivers of global variation in these information aspects, separately for each vertebrate group at each of four spatial grain sizes between 110 and 880 km. I used multi-model inference to quantify the relative importance of predictor variables.

In chapter 3 (Meyer et al. 2016a), I used the same records for terrestrial mammals and combined them with range maps to analyze aspects of occurrence information at the species level (Fig. 1). These aspects were i) record count per species, ii) how these records cover individual species' ranges, and iii) the level of geographical bias in their representation of different range parts. I calculated metrics of range coverage and geographical bias by relating the positions of records to those of randomly placed points across the range maps. I used multi-model inference and variation partitioning to test how different species attributes, size and shape of their ranges, and socio-economic factors drive species-level variation in these information aspects globally and for individual zoogeographical regions.

Results

To my knowledge, this thesis represents the first comprehensive global analyses of different aspects of occurrence information (e.g. coverage, uncertainty). Rather than merely assessing global taxonomic completeness (as in Pelayo-Villamil et al. 2015, for example), I evaluated both data quality and coverage along the taxonomic, geographical and temporal dimensions (Fig. 1), and systematically compared them across different spatial scales and taxonomic groups. As expected, I found extensive gaps and biases in the representation of different taxa, regions and time periods. In all taxonomic groups, record numbers varied across geographical assemblages and individual species by several orders of magnitude (chapters 1–3). Large proportions of records were identified as having high data uncertainty (chapter 1; Feeley & Silman 2010), and many records fell outside species' presumed native ranges (chapters 1–3). I found clear taxonomic bias. For instance, record counts per species tended to be higher in gymnosperms than in other plants (chapter 1), in birds than in other vertebrates (chapter 2), and in Australian marsupials than in other mammals (chapter 3). Patterns of data limitations differed depending on the aspect of occurrence information in focus. For instance, pteridophytes were taxonomically better covered in DAI compared with other plant groups, but pteridophyte records also showed the most severe levels of taxonomic uncertainty (chapter 1). DAI was also geographically biased. For instance, peaks in the coverage of species assemblages emerged in 'Western' industrialized countries, but also in several tropical regions such as Central America or parts of the Andes (chapters 1–2). In contrast, broad regions were without any mobilized occurrence records, particularly in Asia and most of Africa. Surprisingly, there was no pronounced 'tropical data gap' (Collen et al. 2008), neither in plants nor in vertebrates, but this was because several temperate and Arctic regions also emerged as extremely data scarce. I also found strong temporal variation in occurrence information (Boakes et al. 2010). Several areas, notably in parts of Africa and Asia, had peaks in coverage before the 1970s and little recording activity since (chapter 1).

Coarsening grain sizes led to higher coverage of species' assemblages (Soberón et al. 2007), but also to lower opportunities for inference (chapter 2; comparison between 110-km to 880-km cells) and an underestimation of local data gaps (chapter 1; comparison between 110-km cells and countries). The grain size where a given percentage of an assemblage is covered directly relates to the coverage of individual species' ranges. For instance, the few scattered vertebrate records available for much of Asia can only cover few species in any one grid cell (chapter 2), and only provide limited range coverage for the species that occur in the region (chapter 3). Thus, different coverage metrics are naturally constrained by record quantity (chapters 1–3; Yang et al. 2013) and, accordingly, show at least moderate positive pairwise associations (chapter 1). However, the

generally positive relationships between data quantity and aspects of coverage are disturbed by aggregation, duplication and biases in those records (chapters 1–3). In contrast, different metrics of data uncertainty generally showed poor correlations with one another, as well as with coverage metrics (chapter 1).

I also provided the most comprehensive analyses of possible underlying causes of bias in occurrence information to date. Of twelve potential geographical and socio-economic drivers of assemblage-level record density and inventory completeness, only four received strong support across taxa and grain sizes (chapter 2). First, regions with many range-restricted species were generally better inventoried, supporting the hypothesis that researchers preferentially survey regions where they can hope to find such species (Soria-Auza and Kessler 2008). Second, an effect of accessibility was mainly evident in strong positive effects of proximity of grid cells to record-contributing institutions (Moerman and Estabrook 2006), while transportation infrastructure (Ballesteros-Mejia et al. 2013) played a surprisingly minor role. Third, political participation in GBIF (Yesson et al. 2007) was much more important than a region's integration into scientific activities that may lead to peer-reviewed publications. Finally, locally available research funding (Vollmar et al. 2010, Ahrends et al. 2011) limited distribution information much more than size or funding of the Western institutions that contributed the majority of mobilized records. These four key socio-economic variables were also strongly correlated with species-level variation in different aspects of DAI (chapter 3), but their relative importance differed substantially depending on the geographical extent and focus of the analysis (global vs. realm-wide).

Discussion

Together, the results of my research have several important implications for the effective improvement of DAI and its effective use in ecological research, conservation and species distribution modelling. After more than a decade of intensive mobilization, DAI is still – and probably always will be – characterized by severe biases, gaps and uncertainties. Unless carefully accounted for, these limitations seriously impair research and conservation applications (Boitani et al. 2011, Rocchini et al. 2011, Yang et al. 2013). The magnitude of data limitations shows that relying only on highest-quality records (Soberón and Peterson 2004, Feeley and Silman 2010) or data-intensive distribution-modelling techniques (Feeley and Silman 2011) is unrealistic for many species and regions of particular conservation concern (chapters 1–3). Further improving the ability of distribution modelling techniques to draw useful inference from low numbers of records, and to account for data bias and uncertainty (e.g. McInerny & Purves 2011), should therefore be a top priority. One promising way to account for biases is explicitly incorporating bias-causing factors into models (Dorazio 2014, Fithian et al. 2015), and my results can help identify meaningful predictor variables. In such models, accounting for site-specific socio-economic data collection and mobilization constraints appears more promising for addressing these biases than focusing on species-specific detectability.

My analysis of potential drivers of assemblage-level record density and inventory completeness demonstrates that regional contexts determine which socio-economic factors are important causes of biases in occurrence information. Interspecific variation in occurrence information was additionally strongly determined by range size and shape. This is consistent with my hypothesis that while large ranges are bound to overlap with more sampling locations, large, irregular-shaped ranges constrain the ways in which a given number of records can cover a range. Against expectations, species' attributes that were related to detection or collection probabilities received little support as predictors of species-level variation in occurrence information.

My identification of key factors limiting occurrence information, and the distinction between different information aspects, will help identify priority activities to remedy data limitations most effectively. Priorities include: supporting mobilization efforts in institutions near identified data

gaps; fostering cooperation of large emerging economies with data-sharing networks (chapters 2–3); updating the mostly old information for much of Africa and Southern Asia by carrying out novel surveys (chapter 1); and generally increasing the focus on Asia (chapters 1–2) and on range-restricted species (chapter 3). My results also provide a baseline for monitoring progress in data mobilization, and more generally in efforts towards international targets for improving biodiversity knowledge (e.g. Aichi target 19[1]). They show that simple indicators like the number of GBIF-facilitated records (Tittensor et al. 2014) are unable to reliably reveal changes in coverage of species and areas, and even less so changes in data uncertainties. I therefore recommend that DAI should be monitored by a range of indicators that represent different aspects of occurrence information at grains relevant for biodiversity research and management.

Acknowledgments

I thank my supervisor, Holger Kreft, for constant support with fund-raising, study design, implementation, writing, and publishing. I thank my co-authors, Susanne Fritz, Rob Guralnick, Holger Kreft, Walter Jetz and Patrick Weigelt for fruitful collaborations. I thank Yael Kisel and Patrick Weigelt for comments on this manuscript. I am grateful to the Deutsche Bundesstiftung Umwelt (DBU) and the German Academic Exchange Service (DAAD) for PhD scholarships, and the Universitätsbund Göttingen for additional support of a research stay in the Jetz lab.

References

Ahrends, A., Burgess, N.D., Gereau, R.E. et al. (2011) Funding begets biodiversity. Diversity and Distributions, 17, 191–200.

Amano, T. & Sutherland, W.J. (2013) Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. Proceedings of the Royal Society B Biological Sciences, 280, 20122649.

Ballesteros-Mejia, L., Kitching, I.J., Jetz, W., Nagel, P. & Beck, J. (2013) Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. Global Ecology and Biogeography, 22, 586–595.

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.

Boitani, L., Maiorano, L., Baisero, D., Falcucci, A., Visconti, P. & Rondinini, C. (2011) What spatial data do we need to develop global mammal conservation strategies? Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 366, 2623–2632.

Brown, J.H., Stevens, G.C. & Kaufman, D.M. (1996) The geographic range: size, shape, boundaries, and internal structure. Annual Review of Ecology and Systematics, 27, 597–623.

Burton, A.C. (2012) Critical evaluation of a long-term, locally-based wildlife monitoring program in West Africa. Biodiversity and Conservation, 21, 3079–3094.

Butchart, S.H.M., Walpole, M., Collen, B. et al. (2010) Global biodiversity: indicators of recent declines. Science, 328, 1164–1168.

Chutipong, W., Lynam, A.J., Steinmetz, R., Savini, T. & Gale, G.A. (2014) Sampling mammalian carnivores in western Thailand: Issues of rarity and detectability. Raffles Bulletin of Zoology, 62, 521–535.

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.L.H. & Thomas, C.D. (2000) Bias in butterfly distribution maps : the influence of hot spots and recorder's home range. Journal of Insect Conservation, 4, 73–77.

Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography, 23, 1472–1484.

Feeley, K.J. & Silman, M.R. (2010) Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. Journal of Biogeography, 37, 733–740.

Feeley, K.J. & Silman, M.R. (2011) Keep collecting: accurate species distribution modelling requires more collections than previously thought. Diversity and Distributions, 17, 1132–1140.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution, 6, 424–438.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8, 993–1009.

Hobern, D. et al. (2013) Global biodiversity informatics outlook: delivering biodiversity knowledge in the information age. Availble at: http://www.gbif.org/orc/?doc_id=5353.

Holt, R.D. (2003) On the evolutionary ecology of species' ranges. Evolutionary Ecology Research, 5, 159–178.

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

1 https://www.cbd.int/sp/targets/ last accessed 30 June 2016

Review of Ecology, Evolution, and Systematics, 46, 523–549.

Jansen, F. & Dengler, J. (2010) Plant names in vegetation databases – a neglected source of bias. Journal of Vegetation Science, 21, 1179–1186.

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.

Keil, P., Belmaker, J., Wilson, A.M., Unitt, P. & Jetz, W. (2013) Downscaling of species distribution models: a hierarchical approach. Methods in Ecology and Evolution, 4, 82–94.

Lomolino, M.V. (2004) Conservation biogeography – Introduction. In: Frontiers of biogeography: New directions in the geography of nature (ed. by M.V. Lomolino and L.R. Heaney), pp. 293–296. Sinauer Associates, Sunderland, MA.

McInerny, G.J. & Purves, D.W. (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. Methods in Ecology and Evolution, 2, 248–257.

Meyer, C. (2016) Limitations in Global Information on Species Occurrences. Doctoral thesis. Göttingen, Georg-August-Universität.

Meyer, C., Jetz, W., Guralnick, R.P., Fritz, S.A. & Kreft, H. (2016a) Range geometry and socio-economics dominate species-level biases in occurrence information. Global Ecology and Biogeography (early view). DOI: 10.1111/geb.12483.

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

Meyer, C., Weigelt, P. & Kreft, H. (2016b) Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology Letters (early view). DOI: 10.1111/ele.12624.

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.

Pelayo-Villamil, P., Guisande, C., Vari, R.P.et al. (2015) Global diversity patterns of freshwater fishes – potential victims of their own success. Diversity and Distributions, 21, 345–356.

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

Soberón, J., Jiménez, R., Golubov, J. & Koleff, P. (2007) Assessing completeness of biodiversity databases at different spatial scales. Ecography, 30, 152–160.

Soberón, J.M. & Peterson, A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 359, 689–698.

Soria-Auza, R.W. & Kessler, M. (2008) The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: a case study of ferns from Bolivia. Diversity and Distributions, 14, 123–130.

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.

ter Steege, H., Haripersaud, P.P., Bánki, O.S. & Schieving, F. (2011) A model of botanical collectors' behavior in the field: never the same species twice. American Journal of Botany, 98, 31–7.

Tittensor, D.P. et al. (2014) A mid-term analysis of progress toward international biodiversity targets. Science, 346, 241–244.

Vollmar, A., Macklin, J.A. & Ford, L.S. (2010) Natural history specimen digitization: challenges and concerns. Biodiversity Informatics, 1, 93–112.

Whitlock, M.C., McPeek, M.A., Rausher, M.D., Rieseberg, L. & Moore, A.J. (2010) Data archiving. The American Naturalist, 175, 145–6.

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.

Yesson, C., Brewer, P.W., Sutton, T. et al. (2007) How global is the global biodiversity information facility? PloS ONE, 2, e1124.

Submitted: 21 July 2015
First decision: 15 February 2016
Accepted: 01 March 2016
Edited by Joaquín Hortal and Richard Ladle