ZENTRUM
FÜR BIODIVERSITÄT UND NACHHALTIGE LANDNUTZUNG
SEKTION
BIODIVERSITÄT, ÖKOLOGIE UND NATURSCHUTZ

– CENTRE OF BIODIVERSITY AND SUSTAINABLE LAND USE –
SECTION: BIODIVERSITY, ECOLOGY AND NATURE CONSERVATION

Limitations in Global Information on Species Occurrences

Dissertation zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultäten der
Georg-August-Universität Göttingen

vorgelegt von

Dipl. Biol.

Carsten Meyer

aus
Braunschweig

Göttingen, April, 2015

Referentin/Referent: Prof. Dr. Holger Kreft

Korreferentin/Korreferent: Prof. Dr. Kerstin Wiegand

Tag der mündlichen Prüfung:

It is for such inquiries that the modern naturalist collects his materials; it is for this that he still wants to add to the apparently boundless treasures of our national museums, and will never rest satisfied as long as the native country, the geographical distribution, and the amount of variation of any living thing remains imperfectly known.

Alfred Russel Wallace
Journal of the Royal Geographical Society, 1863

Table of contents

Author contributions

Chapter 1
Multidimensional biases, gaps and uncertainties in global plant occurrence information

Carsten Meyer, Patrick Weigelt and Holger Kreft

All authors designed the research; C.M. and P.W. compiled the data; C.M. analyzed the data; C.M. led the writing with substantial contributions from all authors.

Updated version re-submitted to *Ecology Letters* (after invitation for re-submission). Preprint archived in *PeerJ PrePrints* **3**:e1635. DOI: 10.7287/peerj.preprints.1218v2.

Chapter 2
Global priorities for an effective information basis of biodiversity distributions

Carsten Meyer, Holger Kreft, Robert P. Guralnick and Walter Jetz

H.K. and W.J. led this study; all authors designed this study; C.M. compiled the data; C.M. analyzed the data; C.M.led the writing with major contribution from all authors.

Updated version published: *Nature Communications* **6**:8221. DOI: 10.1038/ncomms9221.

Chapter 3
Global drivers of species-level variation in mobilized occurrence information

Carsten Meyer, Walter Jetz, Robert P. Guralnick, Susanne A. Fritz and Holger Kreft

All authors designed the research; C.M. collected the data; C.M. analyzed the data; C.M. led the writing with substantial contributions from all authors.

Updated version re-submitted to Global Ecology and Biogeography (after invitation for re-submission). Preprint archived in PeerJ PrePrints 3:e1493. DOI: 10.7287/peerj.preprints.1326v2.

Summary

Detailed information on species distributions is crucial to answering central questions in ecology, evolutionary biology and biogeography and for effectively allocating conservation resources among regions. Huge numbers of species occurrence records, the basic data underlying our knowledge of species distributions, have been mobilized via international data-sharing networks, most notably that of the Global Biodiversity Information Facility (GBIF). While these networks have greatly increased accessibility of information, severe knowledge gaps remain, a situation termed the 'Wallacean shortfall'. Moreover, the available information is rife with uncertainties, gaps and biases caused by site-specific factors like accessibility or species-specific factors like detectability. If we are to effectively prioritize future data collection and mobilization, we must understand the gaps, biases and uncertainties in current distribution information and what causes them. So far, patterns and drivers of the different information limitations have never been analyzed in detail at the global scale. In this thesis, I provide the first global analyses of limitations in digital accessible occurrence information for land plant and terrestrial vertebrates.

I retrieved >300 million occurrence records for land plants and three vertebrate groups (amphibians, bird and mammals) from GBIF, and integrated these with taxonomic databases and independent range map and checklist information. I then used these datasets to analyze different types of limitations in occurrence information for different taxonomic groups and spatial scales. In chapter 1, I analyzed taxonomic, geographical and temporal data coverage and uncertainty for land plants. I measured taxonomic, geographical and temporal variation in these aspects of occurrence information and quantified their relationships using pairwise correlations and principal component analysis. In chapter 2, I used terrestrial vertebrates to analyze two aspects of occurrence information at the level of geographical assemblages: i) record density and ii) inventory completeness. I used multi-model inference to compare effects of twelve potential socio-economic drivers across the three vertebrate groups and across four spatial grains. In chapter 3, I focused on terrestrial mammals to analyze three aspects of occurrence information at the species level: 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 parts of their ranges. I used multi-model inference and variation partitioning to test effects of different species attributes, size and shape of their ranges, and socio-economic factors at the global scale and for individual zoogeographical regions.

In my thesis, I found severe biases in all examined aspects of occurrence information. Record counts varied by several orders of magnitude across species and regions. Different coverage and uncertainty measures showed clear taxonomic, geographical and temporal patterns. For instance, taxonomic coverage peaked in Western industrialized countries, but also in several tropical regions. In contrast, information was either antiquated or entirely lacking for many Asian and African regions. As taxonomic, geographical and temporal coverage are all numerically constrained by the number of records, these metrics showed moderate to strong positive correlations. Metrics of data uncertainty generally showed low pairwise correlations with one another and with coverage metrics.

In Chapter 2, I found that only four of my twelve hypothesized drivers of assemblage-level record density and inventory completeness received strong support across vertebrate taxa and spatial grains. These were endemism richness, proximity of grid cells to record-contributing institutions, political participation in GBIF, and locally available research funding. Other factors often assumed to strongly constrain information, like transportation infrastructure or size and funding of Western data-contributing institutions, received surprisingly little support. In Chapter 3, I found that the four key socio-economic factors identified in Chapter 2 also had a strong influence on occurrence information at the species-level, but their relative importance differed depending on the geographical focus of the analysis. Interspecific variation in occurrence information was also strongly determined by range size and shape. This supports our hypothesis that while large ranges are bound to overlap with more sampling locations, large, irregular-shaped ranges constrain the detail with which a given number of records can cover a range. Against expectation, species attributes related to detection or collection probabilities had little impact on species-level differences in occurrence information.

The results of my thesis have important implications for the improvement and effective use of mobilized occurrence information. First, my results prove that digital accessible occurrence information is severely limited, particularly for regions and species of conservation concern. Second, success in refining distribution knowledge for these species will depend on distribution modeling techniques that can deal with low record numbers, data biases and data uncertainties. One promising way to account for biases is explicitly incorporating bias-causing factors into models, and my results can help identify meaningful predictor variables. Third, my results create an empirical baseline for monitoring progress in improving the state of global species occurrence data. Finally, my identification of the main factors limiting occurrence information, and the distinction between different information aspects, will help in identifying activities that will remedy data limitations most effectively. I suggest that key activities include supporting mobilization efforts in institutions near data scarce regions, fostering cooperation of large emerging economies with data-sharing networks, conducting

novel surveys for Central Africa and Southern Asia as local data are often outdated, and generally increasing the focus of collection and mobilization activities on Asia and on range-restricted species.

Zusammenfassung

Detaillierte Informationen über die Verbreitungsareale von Arten sind essentiell für die Beantwortung zentraler Fragen der Ökologie, Evolutionsbiologie und Biogeographie. Solche Informationen sind auch notwendig, um Naturschutzressourcen kostenwirksam zwischen verschiedenen Regionen und Maßnahmen zu verteilen. Unser Wissen über Artverbreitungen beruht vor allem auf Punktdaten, die das Vorkommen einer bestimmten Art an einem bestimmten Ort zu einem bestimmten Zeitpunkt belegen (nachstehend „Records"). Riesige Mengen solcher Records wurden über internationale Data-Sharing-Netzwerke mobilisiert, allen voran durch die Global Biodiversity Information Facility (GBIF). Auch wenn diese Netzwerke die Zugänglichkeit zu solchen Informationen enorm verbessert haben, ist unser Wissen über globale Artverbreitungen immer noch äußerst lückenhaft und von grober räumlicher Auflösung – der sogenannte Wallace'sche Wissensrückstand. Vorhandene Informationen enthalten zudem zahlreiche Unsicherheiten, Fehler und Daten-'Biases'. Diese könnten durch Ort-spezifische Faktoren wie Zugänglichkeit oder durch artspezifische Faktoren, wie Entdeckungswahrscheinlichkeit, verursacht werden. Zukünftiges Sammeln und Mobilisieren von Informationen sollte so gestaltet werden, dass der erreichte Nutzen der Records für Forschung und Naturschutz maximiert wird. Hierfür ist ein tiefgehendes Verständnis der Lücken, Unsicherheiten und Biases in den Informationen sowie der sie verursachenden Faktoren notwendig. Bisher wurden diese Mängel in globalen Artverbreitungsinformationen niemals quantitativ untersucht. Mit meiner Dissertation liefere ich die ersten globalen Analysen zu Mängeln von digital verfügbaren Verbreitungsinformationen für terrestrische Wirbeltiere und Landpflanzen.

Ich habe >300 Millionen Records für Landpflanzen und drei Gruppen terrestrischer Wirbeltiere (Amphibien, Säugetiere, Vögel) über GBIF abgerufen. Diese Informationen habe ich mit taxonomischen Datenbanken sowie unabhängigen Verbreitungskarten und Checklisten verbunden. Auf Grundlage der erstellten Datensätze habe ich unterschiedliche Formen von Informations-Mängeln für verschiedene taxonomische Gruppen und auf mehreren räumlichen Maßstäben untersucht. In Kapitel I habe Daten-Abdeckung sowie Daten-Unsicherheiten in Informationen zu Pflanzenvorkommen jeweils in Bezug auf Taxonomie, Raum und Zeit quantifiziert. Für diese insgesamt 6 Maße habe in anschließend Variation in den drei Dimensionen (Taxonomie, Raum, Zeit) gemessen. Zudem habe ich mithilfe von paarweisen Spearman-Rang-Korrelationen und Hauptkomponentenanalysen die Zusammenhänge

zwischen diesen verschiedenen Formen von Informationsmängeln analysiert. In Kapitel II habe ich anhand von terrestrischen Wirbeltieren zwei spezielle Aspekte von Datenabdeckung zwischen geographischen Regionen verglichen: i) die Datendichte und ii) die Vollständigkeit der abgedeckten Arten. Durch Multi-Modell-Analysen habe ich die Effekte von zwölf potentiellen sozioökonomischen Einflussfaktoren auf Informationsmängel verglichen, und zwar einzeln für jede der drei Wirbeltiergruppen auf jeder von vier verschiedenen räumlichen Auflösungen. In Kapitel III habe ich anhand von Säugetieren drei Aspekte von Datenabdeckung zwischen einzelnen Arten verglichen: i) die Anzahl von Records pro Art, ii) die räumliche Abdeckung der Verbreitungsareale durch Records, und iii) den räumlichen Bias in der Abdeckung verschiedener Teile der Verbreitungsareale. Durch Multi-Modell-Analysen und Variations-Partitionierung habe ich die Effekte von verschiedenen Artmerkmalen, Größe und Form der Verbreitungsareale sowie von sozioökonomischen Faktoren untersucht. Diese Analysen habe ich auf globalem Maßstab sowie einzeln für sechs zoogeographische Gebiete durchgeführt.

In meiner Dissertation habe ich in allen untersuchten Aspekten von Artverbreitungsinformationen starke Biases gefunden. Die Anzahl von Records variierte um mehrere Größenordnungen zwischen Arten und zwischen geographischen Gebieten. Verschiedene Maße von Datenabdeckung und Datenunsicherheiten zeigten klare taxonomische, geographische und zeitliche Muster. Ich fand beispielsweise Höchstwerte von taxonomischer Abdeckung in industrialisierten westlichen Ländern, aber auch in einigen tropischen Gebieten wie Mexiko. Im Gegensatz dazu gab es in weiten Teilen Afrikas und Asiens entweder gar keine oder nur sehr veraltete Informationen. Da taxonomische, räumliche und zeitliche Abdeckung jeweils durch die Anzahl der Records numerisch eingeschränkt sind, fand ich zwischen diesen Maßen gemäßigte bis starke positive Korrelationen. Maße von Datenunsicherheiten hingegen korrelierten kaum untereinander oder mit Datenabdeckungsmaßen.

In Kapitel II habe ich den Einfluss von zwölf potentiellen sozioökonomischen Einflussfaktoren auf Datendichte und Datenvollständigkeit von geographischen Artgemeinschaften untersucht. Nur vier hatten einen durchweg für alle untersuchten Wirbeltiergruppen und räumlichen Auflösungen starken Einfluss. Dies waren der Endemitenreichtum, die räumliche Nähe zu Daten-beisteuernden Institutionen, politische Mitgliedschaft im GBIF-Netzwerk, sowie lokal verfügbare Forschungsgelder. Andere Faktoren, von denen man oft annimmt, dass sie eine große Rolle spielen würden, hatten einen erstaunlich geringen Einfluss, wie z.B. Verkehrsinfrastruktur oder Größe und Finanzausstattungen westlicher Daten-beisteuernder Institutionen. Meine Analysen in Kapitel III ergaben, dass die vier in Kapitel II identifizierten sozioökonomischen Schlüsselfaktoren

ebenfalls einen starken Einfluss auf Artverbreitungsinformationen auf der Ebene von einzelnen Arten hatten. Jedoch unterschied sich ihre relative Wichtigkeit deutlich zwischen geographischen Gebieten. Zwischenartliche Unterschiede in Verbreitungsinformationen waren zudem sehr stark durch Größe und Form der Verbreitungsareale beeinflusst. Dies unterstützt meine Hypothese, dass diese geometrischen Faktoren die Wahrscheinlichkeit beeinflussen, dass sich Verbreitungsgebiete bestimmter Arten mit Untersuchungsgebieten von Feldforschern überschneiden, was wiederum Aufswirkungen auf die Wahrscheinlichkeiten hat, mit denen diese Arten besammelt werden. Entgegen unserer Annahmen hatten Artmerkmale wie etwa Nachtaktivität, die das Entdecken oder Sammeln bestimmter Arten wahrscheinlich machen sollten, kaum einen Einfluss auf zwischenartliche Unterschiede in Verbreitungsinformationen.

Die Ergebnisse meiner Dissertation lassen wichtige Schlussfolgerungen darüber zu, wie mobilisierte Artverbreitungsinformationen effizient genutzt und verbessert werden können. Erstens belegen meine Ergebnisse schwerwiegende Mängel in digital verfügbaren Artverbreitungsinformationen, insbesondere für Gebiete und Arten von besonderer Wichtigkeit für den Naturschutz. Zweitens zeigen sie, dass für die allermeisten Arten feiner aufgelöste Informationen nur durch Artverbreitungsmodelle erreicht werden können, die mit geringen Datenmengen auskommen, die starke Datenunsicherheiten und Biases innehaben. Eine vielversprechende Methode, um in solchen Modellen mit Biases umzugeben, ist das explizite Einbeziehen der Bias-verursachenden Faktoren in die Modelle, und meine Ergebnisse bieten hilfreiche Anhaltspunkte für die Auswahl relevanter Faktoren. Drittens schaffen meine Ergebnisse eine empirische Grundlage zur Überwachung von Fortschritten in der Verbesserung weltweiter Artverbreitungsinformationen. Schließlich schafft mein Identifizieren der global wichtigsten Informations-limitierenden Faktoren sowie das Unterscheiden verschiedener Informationsaspekte eine Grundlage dafür, um Aktivitäten zu identifizieren, die Datenmängel effektiv beheben können. Als wichtigste Aktivitäten empfehle ich unter anderem i) das Unterstützen von Bemühungen zur Datenmobilisierung in Institutionen, die in geographischer Nähe zu datenarmen Gebieten liegen, ii) das Fördern von Kooperation zwischen großen Schwellenländern und Data-Sharing-Netzwerken, iii) die Durchführung von neuen Biodiversitäts-Surveys im zentralen Afrika und südlichen Asien, um weitgehend veraltete Informationen zu aktualisieren, und iv) das Verschieben des Fokus von Datensammel- und Datenmobilisierungsbemühungen auf Asien sowie Arten mit begrenzten Verbreitungsarealen.

Part I

Introduction

Research background

Information on species distributions

The distribution of species in space is central to ecology (Brown *et al.*, 1996), evolutionary biology (Holt, 2003) and biogeography (Lomolino, 2004). Some of the most influential theories in those disciplines (Darwin & Wallace, 1858; Grinnell, 1917; MacArthur & Wilson, 1967) were directly inspired by observations that some species are present in some areas and absent in others. Species' distributions also influence their vulnerabilities to anthropogenic pressures (Fritz *et al.*, 2009; Hof *et al.*, 2011) and underlie schemes for effectively distributing conservation resources among regions (Margules & Pressey, 2000; Brooks *et al.*, 2006).

The diverse research and conservation applications that concern species distributions require solid information about where and when species occur. Many questions require distribution datasets of broad coverage yet high detail. For instance, datasets covering many species over large spatial extents at fine spatial grains would enable the study of the imprint of fine-scale processes like species interactions on larger-scale biodiversity patterns (Beck *et al.*, 2012). Such datasets are also necessary for informing conservation prioritization at scales that match land-use changes and management options (Boitani *et al.*, 2011). Similarly, high temporal coverage of distribution datasets is needed to study species responses to environmental change (Boakes *et al.*, 2010), and to inform policy-relevant indices of biodiversity change (Butchart *et al.*, 2010). Improving baseline information on species distributions is closely linked to international targets in the framework of the United Nations Convention on Biological Diversity (Pereira *et al.*, 2013) and plays a central role in current discussions in the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES, 2015).

Information on species distributions can be derived from different data types, each with different strengths and weaknesses (Jetz *et al.*, 2012a). The most abundant data are point occurrence records, derived from specimens in natural history collections, field surveys, vegetation plots, amateur observations, and other sources. Global natural history collections alone contain an estimated 1 to 3 billion specimens (Vollmar *et al.*, 2010), most of which are associated with data on where they were collected. Such records represent the primary information on the taxonomic, geographical and temporal dimensions of species distributions, as they provide direct evidence that particular species occurred at particular locations at particular points in time (Soberón & Peterson, 2004). Most other information types are

ultimately derived from such occurrence records. For instance, range maps delimit the maximum extent over which species can be expected to occur, and are usually created by experts by drawing polygons around recorded occurrences while incorporating their knowledge of the species' preferred environmental conditions (Graham & Hijmans, 2006). Regional checklists attempt to list all species of a given taxon that occur within a particular region, and are also usually created by consulting primary records of species occurrences. For some species, distribution atlases have been carefully compiled, showing truly occupied and unoccupied areas at relatively fine spatial grains (Robertson *et al.*, 2010). For most species, however, occurrence records are the only type of distribution information available (Jetz *et al.*, 2012a), and researchers have to rely on modeling techniques to estimate fine-scale occupancy. The technique most commonly used is correlative species distribution modeling, where a species' presence is extrapolated beyond the immediate areas where it was recorded based on statistical relationships between occurrences and environmental variables (Guisan & Thuiller, 2005). However, these models are very sensitive to different data limitations (Phillips *et al.*, 2009; Feeley & Silman, 2011a).

Limitations in distribution information

Even when combining all available information, global knowledge of species distributions remains extremely limited, a situation termed the 'Wallacean shortfall' (Lomolino, 2004). Notwithstanding the obvious absence of information for species that are yet to be described (the 'Linnean shortfall'), the majority of species known to science are only known from their type locality, and few species have detailed distribution data across their entire ranges. Occurrence records provide no information on the presence or absence of species beyond the surveyed areas (Rocchini *et al.*, 2011), while range maps and checklists provide no fine-grained information on occupancy within the respective region. The much-needed large-extent, fine-grain atlas datasets exist only for few taxa and regions (Robertson *et al.*, 2010). Furthermore, existing information is often scattered across multiple sources (and thus, difficult to compile; (Jetz *et al.*, 2012a)) and prone to many uncertainties arising from ambiguous scientific names (Jansen & Dengler, 2010), imprecisely geo-referenced sampling locations (Rocchini *et al.*, 2011) and old age of many records (Ladle & Hortal, 2013). Finally, most occurrence records were collected opportunistically, often with the prime aim of maximizing taxonomic diversity in collections in order to support taxonomic, rather than biogeographical or ecological studies (Pyke & Ehrlich, 2010; ter Steege *et al.*, 2011). This created a series of biases (Nelson *et al.*, 1990; Boakes *et al.*, 2010) that hamper many important applications, including species distribution modeling (Phillips *et al.*, 2009),

macroecological analyses (Yang *et al.*, 2013) and conservation prioritization (Boitani *et al.*, 2011).

Improving species distribution information has traditionally mainly involved field surveys and data collection, along with taxonomic and curatorial work in museums and herbaria. In recent decades, applications of species occurrence data have been transformed by the new field of biodiversity informatics (Soberón & Peterson, 2004). Key developments included adoption of information technology to manage and analyze data, large-scale digitization of natural history collections and the development of data standards and technological tools, e.g. for the automated capture of collections data (Graham *et al.*, 2004). A quantum leap in the accessibility of distribution information was the creation of distributed online data-sharing networks, most notably that of the Global Biodiversity Information Facility (GBIF; www.gbif.org). These networks allow data providers like museums, government agencies and amateur naturalist communities to publish their occurrence records online, and allow data users to access records from multiple providers with a single query. 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 *digital accessible information* (DAI); originally referred to as DAK in (Sousa-Baena *et al.*, 2014a)). GBIF also plays a key role in disseminating skills, software, tools, and best practices for biodiversity data mobilization. Other approaches concentrate more on drawing best-possible inference on species distributions from accessible data sources. The Map of Life project, for example, is developing tools for integrating different types of distribution information (Jetz *et al.*, 2012a). Such tools are made possible by Bayesian modeling approaches, which can integrate different information types (Keil *et al.*, 2013; Manceur & Kühn, 2014) and incorporate information on factors causing bias (Dorazio, 2014), addressing many of the biases and uncertainties that limit classical species distribution modeling (Phillips *et al.*, 2009).

Despite these encouraging developments, the scale of the Wallacean shortfall means that distribution information will likely remain insufficient for answering many biodiversity research and conservation questions for the foreseeable future. Therefore, it is important to prioritize future data collection and mobilization efforts (Hobern *et al.*, 2013; Sousa-Baena *et al.*, 2014a). If we are to effectively prioritize activities, we must understand the gaps, biases and uncertainties in current occurrence information and what causes them.

Previous research on limitations in point occurrence information

Patterns of limitations in occurrence information have been mainly investigated with respect to geographical data bias. One of the most commonly-quoted data limitations is a broad-scale data gap in tropical countries (Prance, 1977; Collen *et al.*, 2008). Studies have also found finer-scale geographical bias in the completeness with which occurrence records cover the species in different geographical units (Soberón *et al.*, 2007; Ballesteros-Mejia *et al.*, 2013; Yang *et al.*, 2013). Fewer authors have considered biases towards certain species (Schmidt-Lebuhn *et al.*, 2013) or time periods (Boakes *et al.*, 2010). Large-scale patterns in data uncertainties have rarely been studied (Yesson *et al.*, 2007), although their sources and potential impacts received some attention (Feeley & Silman, 2010; Rocchini *et al.*, 2011).

Limitations in distribution information are most commonly attributed to geographically biased field surveys. These may be driven by regional differences in accessibility (Freitag *et al.*, 1998; Dennis & Thomas, 2000), safety concerns (Brito *et al.*, 2013), lack of funding (Ahrends *et al.*, 2011) or preferential interest in endemism-rich, mountainous or protected areas (Soria-Auza & Kessler, 2008; Yang *et al.*, 2014). However, regional gaps in DAI do not necessarily reflect a lack of field work, as often assumed; they can also be caused by biased financial or institutional resources for digitization (Vollmar *et al.*, 2010), or poor scientific (Amano & Sutherland, 2013) or political (Yesson *et al.*, 2007) cooperation constraining international dissemination of information. Similarly, biases may reflect not only site-specific socio-economic constraints, but possibly also species-specific factors like lower detectability of nocturnal (Burton, 2012) and arboreal species (Chutipong *et al.*, 2014) or deliberate withholding of occurrence information for threatened species (Whitlock *et al.*, 2010). Finally, the geometry of distributional ranges may affect the likelihood that the study region of a given researcher intersects with a given range, which in turn affects the likelihood that this particular species is recorded. Understanding which factors limit occurrence information can help prioritize activities for improving information, and account for these known biases in ecological models by explicitly incorporating them as variables (Dorazio, 2014; Fithian *et al.*, 2014).

Despite the urgent need to address limitations in occurrence information, they have never been quantified in detail at the global scale. Previous studies of patterns and drivers of occurrence information were limited in geographical (Ballesteros-Mejia *et al.*, 2013; Yang *et al.*, 2014) or taxonomic (Yesson *et al.*, 2007) scope, by the limited number of tested hypotheses, or by simplistic treatment of distribution information (Amano & Sutherland, 2013). No study has tested the generality of the various information-limiting factors globally across different taxonomic and spatial scales.

The main goals of my dissertation were to

a) provide the first global, detailed analyses of limitations in mobilized occurrence information for a large section of biodiversity,

b) better understand global taxonomic, geographical and temporal variation and biases in different aspects of occurrence information,

c) better understand global drivers of this variation across different taxonomic groups and spatial scales, and to

d) create an empirical baseline for prioritizing data collection and mobilization efforts, for monitoring these activities, and for effectively accounting for data limitations in ecological models.

Study outline

In chapter 1, I focus on land plants, a hyper-diverse group of organisms, to analyze two main aspects of occurrence information, and how they are spread in the three basic dimensions that characterize species distributions – taxonomy, space and time (Fig. I.2.1). The first, mostly quantitative, aspect of occurrence information is i) the coverage of the three dimensions with information. The second, more qualitative, aspect is ii) uncertainty regarding the interpretation of information. I study how these different aspects of occurrence information are related to each other, and identify biases in the three dimensions.

In chapter 2, I focus on terrestrial vertebrates to analyze two aspects of occurrence information at the level of geographical assemblages, i) record density and ii) inventory completeness (Fig. I.2.1). I test the roles of twelve socio-economic drivers of global variation in these information aspects for different vertebrate groups and at different spatial grains.

In chapter 3, I focus on terrestrial mammals to study species-level variation in three aspects of distribution information: i) record count per species, ii) how these records cover individual species' ranges, and iii) the level of geographical bias in how records represent different parts of the ranges (Fig. I.2.1). I tested how species attributes, the size and shape of species ranges, and socio-economic factors drive species-level variation in these information aspects, globally and for individual zoogeographical regions.

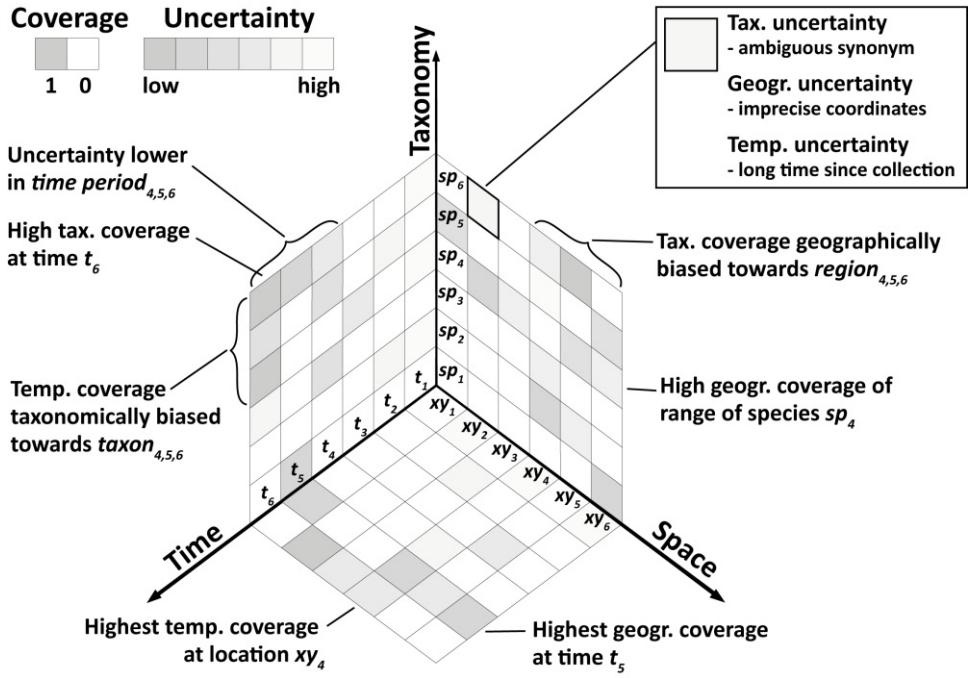


Figure I.2.1. Framework for analyzing limitations in occurrence information. Species distributions are characterized by three main dimensions: taxonomy, space and time. Occurrence records provide direct evidence that particular species (sp_1, sp_2, …) occurred at particular locations (xy_1, xy_2, …) 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 covered). In chapter 1, I study 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 (see inset box for examples). Both coverage and uncertainty may vary in each of the three dimensions, potentially leading to biases (see curly brackets for examples; e.g. center left: temporal coverage taxonomically biased because species of $taxon_{4,5,6}$ have systematically higher coverage, compared to $taxon_{1,2,3}$). In chapter 2 and 3, I focus on specific aspects of coverage. In chapter 2, I compare record density and inventory completeness across geographical assemblages; in chapter 3, I compare record count, range coverage and within-range geographical bias across species.

Part II

Research chapters

Chapter 1

Multidimensional biases, gaps and uncertainties in global plant occurrence information

Carsten Meyer, Patrick Weigelt and Holger Kreft

Updated version re-submitted to *Ecology Letters* (after invitation for re-submission). Preprint archived in *PeerJ PrePrints* **3**:e1635. DOI: 10.7287/peerj.preprints.1218v2.

Abstract

Aim: Detailed information on species distributions is fundamental to ecology, evolution and conservation. Most distribution information is ultimately based on point occurrence records, millions of which have been mobilized via international data-sharing networks. However, biases, gaps and uncertainties hamper broader application. We provide the first global, systematic assessment of taxonomic, geographical and temporal variation in coverage and uncertainty of mobilized occurrence information for all land plants. We assess implications for research, conservation and monitoring possibilities.

Location: Global.

Methods: We integrated 120 million occurrence records available via the Global Biodiversity Information Facility (GBIF) with comprehensive taxonomic databases, checklists for selected plant families and the IUCN Red List of Threatened Species. We calculated different metrics to quantify how mobilized occurrence information covers the taxonomic, geographical and temporal dimensions, and the uncertainty of information with regard to these dimensions. We then assessed taxonomic, spatial and temporal variation in these different aspects of occurrence information, and used pairwise Spearman rank correlations and principal component analysis to investigate relationships between them.

Results: We documented extensive data gaps and uncertainties. For instance, only 5.4% of 110 km x 110 km grid cells had ≥ 80% of species covered, only 28% of species were represented by ≥ 10 unique sampling locations and mobilized information was severely outdated over vast regions. Data limitations varied in all three dimensions, leading to specific combinations of biases. Information metrics were largely uncorrelated in space; different data limitations were predominant in different regions. Filtering could reduce data uncertainties, but caused substantial trade-offs for coverage and additional biases.

Main conclusions: Multidimensional limitations in occurrence information hamper prospects of establishing plants as a model group for global research, and for achieving international conservation targets. Either goal would require both scaling up and prioritizing efforts to collect and mobilize information. Available information should be used effectively, by explicitly accounting for biases and uncertainties.

Introduction

Land plants (Embryophyta, hereafter 'plants') are a hyperdiverse group of organisms and the principal providers of habitat structure and biochemical energy in most terrestrial ecosystems. Geographical distributions of plant species thus determine the spatio-temporal setting for evolutionary and ecological processes (Wright & Samways, 1998; Kissling *et al.*, 2008), and of the ecosystem functions and services upon which most other species, including humans, rely (Isbell *et al.*, 2011; Gamfeldt *et al.*, 2013). Detailed information on plant distributions is necessary for mapping and explaining basic plant diversity patterns (Kreft & Jetz, 2007; Morueta-Holme *et al.*, 2013) and for effectively allocating resources to their conservation (Ferrier, 2002). Improving such information is intrinsically linked to international targets in the framework of the Convention on Biological Diversity's Global Strategy for Plant Conservation (GSPC; www.cbd.int/gspc/targets.shtml). To date however, detailed distribution datasets typically required in research and conservation only exist for few plant groups and geographical regions (Lomolino, 2004).

Most available datasets on plant distributions, including checklists and range maps, are ultimately based on point occurrence records. Such records represent the primary information on the three basic dimensions that characterize species distributions – taxonomy, space and time – as they provide direct evidence that particular plant species occurred at particular locations at particular points in time (Soberón & Peterson, 2004). Vast quantities of such records, from digitized herbarium specimen labels, field observations, literature, and other sources have been mobilized via international data-sharing networks, most notably that of the Global Biodiversity Information Facility (GBIF; Edwards, 2000). Unlike un-mobilized datasets or expert knowledge, GBIF-facilitated records represent by far the largest share of plant occurrence information that is both digital and easily accessible in a standard format (hereafter referred to as *digital accessible information* (DAI); originally referred to as DAK by Sousa-Baena *et al.*, (2014a)). Potential uses are manifold, spanning research on diversity patterns, range dynamics, plant invasions, or phenological changes (Lavoie, 2013), as well as threat assessments, monitoring (Brummitt *et al.*, 2015) and conservation planning (Ferrier, 2002). However, broader application is hampered by severe gaps and biases in each of the three basic dimensions (Nelson *et al.*, 1990; Boakes *et al.*, 2010; Schmidt-Lebuhn *et al.*, 2013).

Apart from mere quantity of mobilized records, at least two further aspects of occurrence information directly influence applicability in research and conservation (Fig. II.1.1). One aspect closely connected to quantity is i) the '*coverage*' of the three dimensions with information. For instance, *taxonomic coverage* of assemblages determines how reliably

biodiversity can be compared across sites in conservation prioritization (Funk *et al.*, 1999), high *geographical coverage* of species' ranges with records may facilitate species distribution modeling (Feeley & Silman, 2011a), and high *temporal coverage* is essential for monitoring species' responses to environmental change (Brummitt *et al.*, 2015). A second, more qualitative aspect of occurrence information is ii) '*uncertainty*' regarding the interpretation of information on the three dimensions. For instance, ambiguous scientific names entail *uncertainty* regarding taxonomic identities (Jansen & Dengler, 2010), imprecisely geo-referenced sampling locations regarding the environmental context in which species were found (Rocchini *et al.*, 2011), early sampling dates regarding their continued presence near those locations (Boitani *et al.*, 2011).

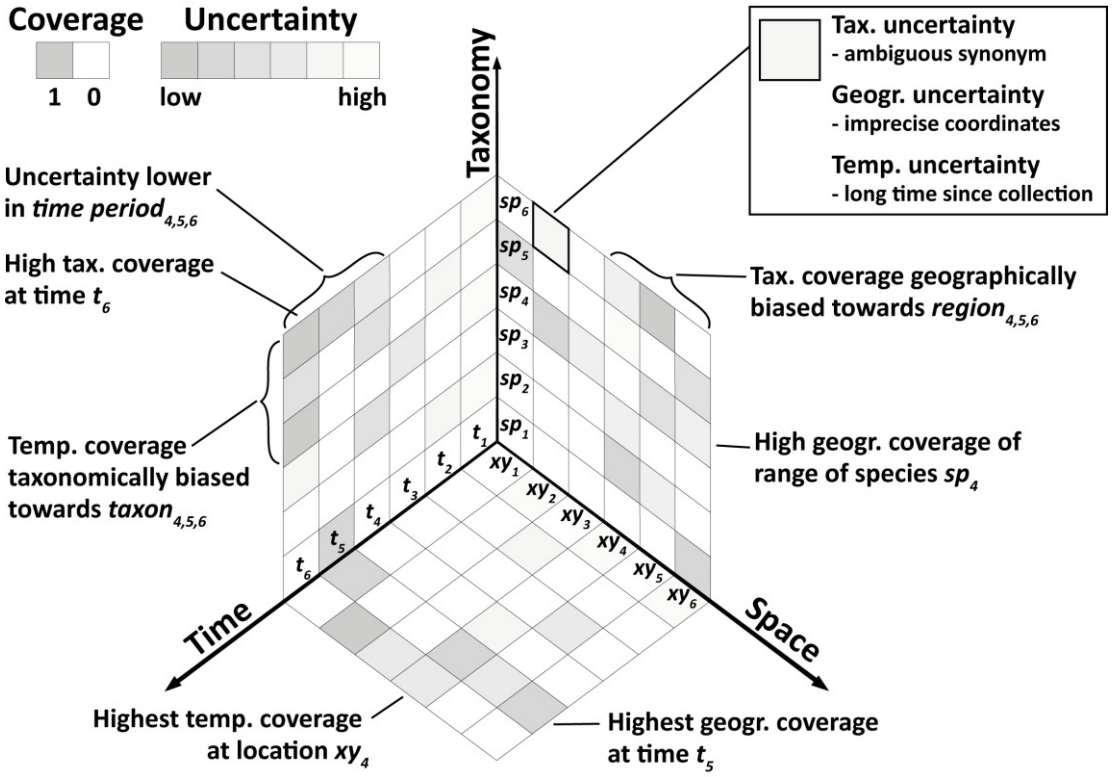


Figure II.1.1. Framework for analyzing limitations in occurrence information. Species distributions are characterized by three main dimensions: taxonomy, space and time. Occurrence records provide direct evidence that particular species (sp_1, sp_2, …) occurred at particular locations (xy_1, xy_2, …) 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. At least two aspects of occurrence information 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 (see inset box for examples). 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 covered). Both *coverage* and *uncertainty* may vary in each of the three dimensions, potentially leading to biases (see curly brackets for examples; e.g. center left: temporal coverage taxonomically biased because species of $taxon_{4,5,6}$ have systematically higher coverage, compared to $taxon_{1,2,3}$).

Both *coverage* and *uncertainty* may be biased in the taxonomic, geographical and temporal dimension (Fig. II.1.1), potentially leading to biased ecological inferences and inefficient conservation. For instance, *taxonomic coverage* of plant assemblages may be geographically biased to certain regions (Yang *et al.*, 2013; Sousa-Baena *et al.*, 2014a), and *geographical uncertainty* may be greater in records from earlier time periods (Murphey *et al.*, 2004). Understanding magnitude and biases in different aspects of occurrence information with regard to the three dimensions is necessary for evaluating the potential for research and conservation applications, and for prioritizing and monitoring activities to improve information. Identifying botanical information gaps has long been of scientific interest (Jäger, 1976; Prance, 1977; Kier *et al.*, 2005), while most recent analyses emphasized effects on specific applications (Feeley & Silman, 2011a; Yang *et al.*, 2013; Sousa-Baena *et al.*, 2014b). Despite the need to comprehensively evaluate global DAI, a quantitative assessment for the World's plants has never been attempted.

Here, we provide such an assessment for all land plants, by integrating 120 million occurrence records facilitated via GBIF with comprehensive taxonomic databases, distribution checklists for selected plant families, and the global red list (see *Methods*). We analyze taxonomic, geographical and temporal variation in information *coverage* and *uncertainty*, and their relationships with one another. We further assess implications of global *coverage* and *uncertainty* patterns for research, conservation and monitoring possibilities, with particular emphasis on GSPC targets. Our work provides the first quantitative global overview of strengths and weaknesses in DAI for a hyperdiverse taxonomic group, and a baseline for more effective information usage and mobilization.

Methods

Point occurrence information

We downloaded all data for land plants available via GBIF in Jan 2014 (c. 120 M). GBIF-facilitated records represent by far the largest source of digital accessible information, and a substantial part of the digitized portion of the c. 350 M records that exist in the World's herbaria (New York Botanical Garden, 2014). Gaps in global coverage in these data may represent genuinely under-sampled regions or regions whose digital information is not yet integrated into international data-sharing networks, such as Brazil or China (see Sousa-Baena *et al.* (2014a) and Yang *et al.* (2013) for regional bias assessments). We cleaned,

taxonomically standardized and validated verbatim scientific names, using comprehensive taxonomic information provided via The Plant List (TPL, 2014) and iPlant's Taxonomic Name Resolution Service (TNRS, 2014). We applied basic taxonomic and geographical filtering (see below) and excluded duplicated combinations of accepted species, sampling location and year-month combination (see Fig. V.1.S1 for an overview of our workflow, see *Supplementary Information (SI) V.1.1* for details). These steps led to a reduction of 119,058,280 raw records with 2,206,831 verbatim name strings to 55,930,12 unique records for 229,218 accepted species from 3,947,969 sampling locations and 3,172 year-month combinations (*SI V.1.1*).

Coverage

We used a suite of simple metrics to estimate the extent to which available records cover the taxonomic, geographical and temporal dimensions (Fig. II.1.1). We estimated taxonomic coverage of grid cells as the ratio between recorded vascular plant richness and an environment-richness model for that group (Kreft & Jetz, 2007). Spatial patterns of taxonomic coverage may be affected by non-native species' records, However, independent information on species' native ranges to geographically validate records (e.g. chapter 2) does not exist for most plants. In a side analysis, we validated 16.8M records for 105,031 (34% of all) species of seed plants against checklists for 'botanical countries' (level-3 regions of Biodiversity Data Standards, formerly International Working Group on Taxonomic Databases – TDWG; www.tdwg.org/standards/109/), derived from the World Checklist of Selected Plant Families (WCSP, 2013) and compared the ratio between recorded and checklist-based richness among 'botanical countries'. To estimate geographical coverage of species' ranges and grid cells, respectively, we used the quantity of unique sampling locations per species and per land area in grid cells. To measure temporal coverage of species and cells, we calculated the negative mean minimum Euclidean distance (in years) of all months between 1750 and 2010 to the sampling dates of their respective temporally closest record. This metric has large negative values if that time span contains large temporal gaps with no records. We analyzed temporal patterns of taxonomic and geographical coverage by comparing percentages of species and grid cells covered within 5-year periods, and cumulatively up to those periods.

Uncertainty

To investigate uncertainty regarding the interpretation of information (Fig. II.1.1), we created three potential uncertainty filters ('basic', 'moderate', 'strict') that a user of GBIF-facilitated data might consider. We defined three taxonomic uncertainty filters based on criteria and decisions taken during taxonomic validation (see *SI V.1.1*):

- TaxStrict: Recorded name matches a name treated by TPL as an accepted species with high expert confidence (three 'stars'; www.theplantlist.org/about), with no more than 5% orthographic distance (see *SI V.1.1*), either directly or through an unambiguous synonym (i.e., one that only links to one accepted name);
- TaxModerate: Recorded name matches a name treated by TPL as an accepted species with high or medium expert confidence (two or three 'stars') with no more than 15% orthographic distance, either directly or through an unambiguous or ambiguous synonym;
- TaxBasic: Recorded name matches a name treated by TPL or TNRS as an accepted name (no criteria for expert confidence in TPL) with no more than 25% orthographic distance, either directly or through an unambiguous or ambiguous synonym. This basic filter was always applied before other analyses.

We defined three geographical uncertainty filters, based on x, y coordinates and indicated country:

- GeoStrict: Coordinates reported with an precision of at least 1/1000 of a degree (~100m at the equator) and falling within the indicated country;
- GeoModerate: Records reported with an precision of at least 1/100 of a degree and falling within the indicated country;
- GeoBasic: Records reported with a precision of at least 1/10 of a degree and falling within the indicated country. This filter was always applied before other analyses.

We defined three temporal uncertainty filters:

- TempStrict: Records reportedly collected after 1990;
- TempModerate: Records reportedly collected after 1970;
- TempBasic: Records reportedly collected after 1950.

If not otherwise stated we hereafter refer to a dataset to which basic taxonomic and geographical filters, but no temporal filter, were applied. The necessity for a specific filter depends on the research question at hand. Some analyses might make good use of data that

was excluded by our basic filtering, such as higher-level taxonomic information, or locations geo-referenced to full degrees.

We investigated patterns in taxonomic and geographical uncertainty by comparing across species and grid cells the percentages of records that would be additionally excluded when applying moderate or strict taxonomic and geographical uncertainty filters, respectively, compared to the basic filter. We investigated patterns in temporal uncertainty by comparing percentages of species additionally excluded by moderate or strict temporal uncertainty filters. Similarly, we investigated patterns in combined uncertainty by comparing percentages of additionally lost species if all three filters were applied.

Analyses of variation in occurrence information

To quantify and visualize taxonomic, geographical and temporal variation in coverage and uncertainty of information, we compared the respective metrics among major plant groups (bryophytes, pteridophytes, gymnosperms and angiosperms), geographical units (110km grid cells and TDWG level-3 'botanical countries'), and 5-year periods.

We investigated relationships between geographical patterns of nine different information metrics, including the three dimensions of coverage and uncertainty and combined uncertainty (see above; uncertainty measured as information loss under moderate filtering). We also included two further aspects of limitations in occurrence information: the number of vascular plant species that were not recorded but expected to occur in an area based on the environment-richness model (Kreft & Jetz, 2007), and the time (in years) since the last record was recorded. We analyzed their relationships with pairwise Spearman rank. We used principal component analysis (PCA) to reduce co-linear metrics to orthogonal principal components. Grid cells were located in the multidimensional PCA-space according to their scores in the different information metrics. We assigned red, green and blue components of the RGB color space to the grid cells according to their positions in the three-dimensional space formed by the first three PCA (Weigelt *et al*., 2013). We then mapped these colored grid cells on a world map to visualize which regions are predominantly characterized by the different aspects and dimensions of occurrence information.

We assessed the potential for selected research and conservation applications that depend on distribution information, by counting species that would meet minimum data requirements of different distribution estimation methods (e.g., 10 locations for simple extent-of-occurrence polygons (Rivers *et al.*, 2011), 25 to 200 for species distribution modeling (Feeley & Silman, 2011a), if all three basic, moderate or strict uncertainty filters were applied.

All analyses were carried out in R version 3.x (R Core Team, 2014).

Results and Discussion

The high number of globally mobilized plant records (119 M; Fig. V.1.S1A) greatly overestimates the actual available information on species occurrences. Basic validation and filtering steps (see Methods) excluded 38.2 M records, including 27.9 M in the sea (Fig. V.1.S1C) and 12.5 M with non-validatable verbatim name strings (Fig. V.1.S1G, *SI V.1.1*). Removing duplicated species-location-month combinations excluded a further 25 M, leaving 56 M records for analyses (47% of all). Record numbers varied by five orders of magnitude across species, and by six orders of magnitude across 110 km grid cells (Fig. V.1.S1B). For instance, a single cell in the Netherlands had 2.8 M records, whereas 21.2% of all cells lacked any records. In the following we provide a detailed assessment of information coverage and uncertainty in the taxonomic, geographical and temporal dimension.

Coverage of the different dimensions

Taxonomic coverage

Globally, 229,218 plant species (65.4% of all) were represented with ≥ 1 record that passed basic filtering (see Methods). *Taxonomic coverage* was itself taxonomically biased, with only 28.3% of bryophytes but 82.9% of pteridophyte species represented (Fig. II.1.2A).

Recorded species richness of grid cells was mainly a function of record number (r_S=0.94, P_{Dut}=0; Fig. V.1.S1B/F/K), demonstrating that perceived centers of plant diversity may simply reflect better documentation (Nelson *et al.*, 1990). Accordingly, *taxonomic coverage* of plant assemblages was extremely heterogeneous in space (Fig. II.1.2B). Only 5.4% of cells had a ratio between recorded and modeled species richness >0.8 and could thus be considered taxonomically well-*covered*, notably in Europe, parts of Australia, North America, South Africa, Ecuador, Costa Rica, and scattered parts in the rest of the World (Fig. II.1.2B). Conversely, 78.6% of the world was severely under-inventoried with ratios below 0.25 (Fig. II.1.2B). Large numbers of 'missing' species, i.e. that portion of modeled richness that was not confirmed by records, were typical for Eastern Amazonia and Borneo (Fig. V.1.S2A). Surprisingly, we did not find significant differences in *coverage* between tropical and non-tropical (Collen *et al.*, 2008; P_{Dut}=0.37), nor between neo- and palaeotropical grid cells (Prance, 1977; P_{Dut}= 0.64). The overall low *taxonomic coverage* over vast extents seriously impairs estimations of plant diversity and site-based plant conservation prioritization (GSPC target 5).

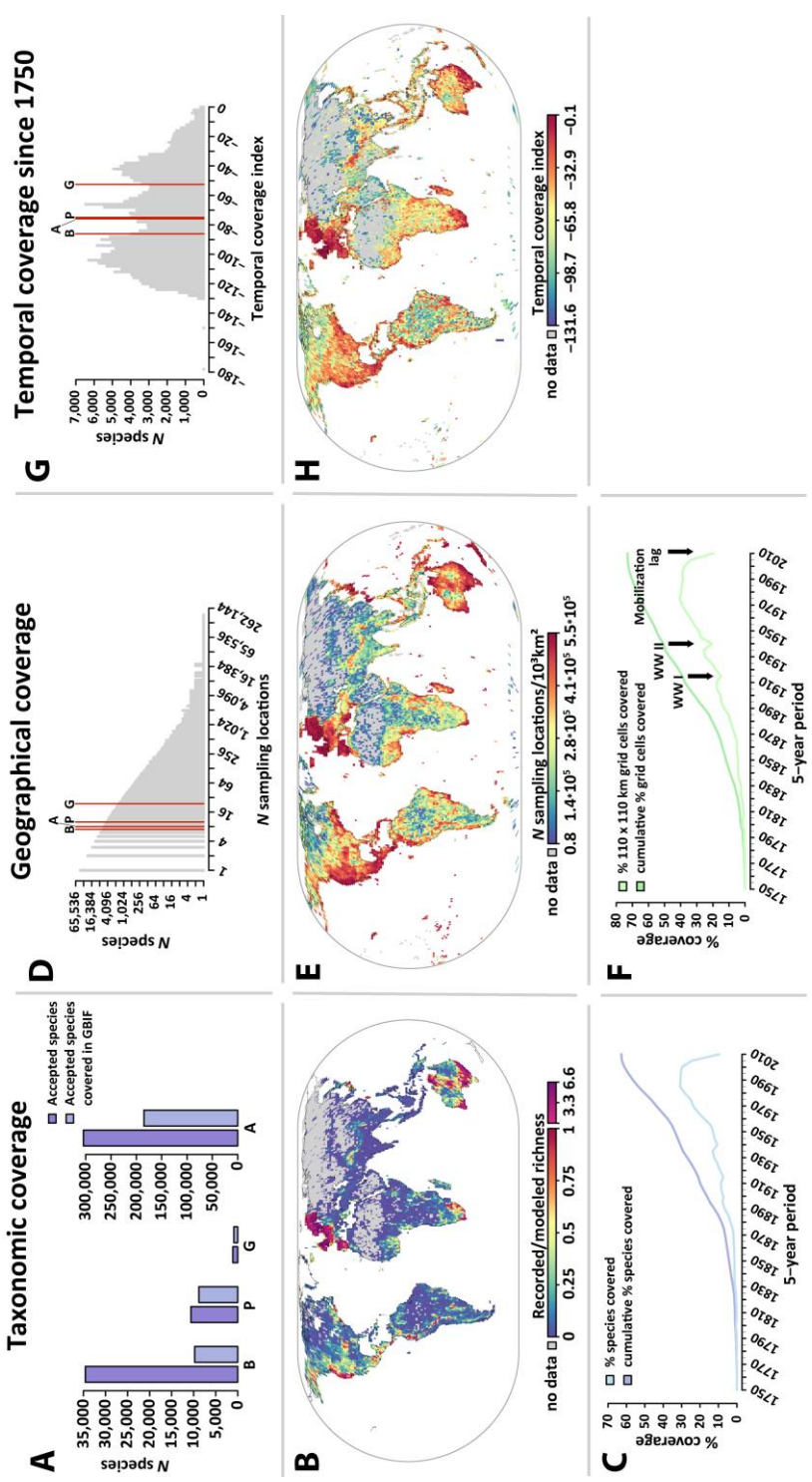


Figure II.1.2. Global variation in occurrence information *coverage*. A) *Taxonomic coverage* of major plant groups (species accepted in The Plant List vs. accepted recorded species; B – bryophytes, P – pteridophytes, G – gymnosperms, A - angiosperms); B) Geographical variation across 110 x 110 km grid cells of *taxonomic coverage* for vascular plants (recorded richness (GBIF) / modeled richness; Kreft & Jetz, (2007)); values >1 mean larger recorded than modeled richness; note that mobilized records include non-native species whereas the model predicts native species richness; C) Percentages of species *covered* within 5-year periods between 1750 and 2010, and cumulatively until each period. D) *Geographical coverage* of plant species, estimated from number of sampling locations (red bars: medians for major plant groups); E) *Geographical coverage* of grid cells (sampling locations / 10^4 km² land area); F) Percentages of grid cells *covered* within and up to 5-year periods since 1750. G) *Temporal coverage* of G) plant species and H) grid cells since 1750, calculated as negative mean minimum Euclidean distance between all possible months and their respective closest month with records (red bars in G: medians for major plant groups). In C/F early dips correspond to two World Wars; the latest drop likely represents a time lag between field collection and mobilization of information.

Taxonomic coverage scores exceeded 1 in 3.6% of cells (Fig. II.1.2B). Scores >1 may stem from an underestimation of richness by the environment-richness model, records of non-native species, or inaccurate information on sampling locations. For instance, the score of 6.6 around Stockholm was mainly due to undated records for non-European species provided by the Bergius Herbarium, possibly from collections assembled in the 19th century by the East India Company. Such factors could influence recorded/modeled richness ratios anywhere in the world, therefore *taxonomic coverage* cannot directly be interpreted as completeness of native plant inventories. Anyone using mobilized records to study native biodiversity should carefully consider these potential sources of error.

A more robust measure of *taxonomic coverage* can be attained for 'botanical countries' and 105,031 species of native spermatophytes (seed plants), based on records that were geographically validated against WCSP checklists (Fig. V.1.S3A). However, these coarser patterns only moderately correlated with mean grid-level *coverage* (r_P=0.68, P_{Dut}=0), and underestimated local data gaps in regions where *coverage* was achieved by combining scattered species records over vast areas, such as in Argentina or the Democratic Republic of the Congo (Fig. II.1.2B, Fig. V.1.S3A). Due to their higher spatial resolution, grid-level metrics therefore better indicate global data gaps, and provide an important first step in identifying priority regions for improving botanical baseline information (GSCP target 3).

We found that 10.1% (1.7 M) of records for WCSP-listed species were collected outside the species' indicated native ranges, but even these records could play an important role for progress towards GSPC targets. 45% of these were collected immediately adjacent to indicated native ranges, and potentially represent valid additions to those regions' native floras, notably in the Neotropics (Fig. V.1.S3B), which highlights the importance of occurrence records for target 1, the completion of an online flora of all plants (Paton, 2013). The 55% collected well beyond native ranges (Fig. V.1.S3C) could support target 10 by facilitating study and effective management of plant invasions (Broennimann & Guisan, 2008).

Geographical coverage

If a species has been recorded at sufficient sampling locations, available records may be used to estimate its extent of occurrence (Gaston & Fuller, 2009) or to model occurrences at finer scales (Feeley & Silman, 2011a). However, mobilized records for a given species were typically collected from only 7 unique sampling locations (median; Fig. II.1.2D), making meaningful estimations unlikely in most cases.

Estimates of *geographical coverage* of regions may aid in pinpointing under-collected areas where new species might be found (Bebber *et al.*, 2010), and in controlling for uneven survey

effort in biodiversity analyses (Schulman *et al.*, 2007; Lobo, 2008). As expected, *geographical coverage* was generally high in traditionally well-studied North America, Western Europe and Australia, with peaks in Australia and Scandinavia (Fig. II.1.2E). Outside those regions, high *geographical coverage* often appeared associated with specific botanical interest and major research and data mobilization programs. For instance, Madagascar is renowned among botanists for its exceptional plant diversity (>11,000 species, 82% endemic, (Callmander, 2011)). The Missouri Botanical Garden has long focused on Madagascar (Raven & Axelrod, 1974), was one of the first institutions to engage in data mobilization (Crosby & Magill, 1988), and now contributes 66% of Madagascan records.

Temporal variation in coverage

Globally, percentages of *covered* species and grid cells mostly increased through time, apart from dips during the World Wars (Fig. II.1.2C/F). *Geographical coverage* leveled off since the 1970s and *taxonomic coverage* since the 1980s, while cumulative *coverage* continued to increase, but at shallower slopes (Fig. II.1.3E-F). The most recent drops in *coverage* since the mid-1990s likely reflect time lags between field collection and mobilization of records, but may also in part be due to decreasing survey effort (Lavoie, 2013).

While *covered* species and areas mostly increased globally, there was strong spatio-temporal variation in certain regions. Going from the 1950s/60s via the 1970s/80s to the 1990s/2000s, sampling activity decreased in the Afrotropics and Middle East, while it increased in the Neotropics and circum-Tibetan mountain ranges (Fig. V.1.S4C-E). Accordingly, regional percentages of *covered* species also changed over recent decades. In many part of the world, *taxonomic coverage* during a given time period was always well below cumulative *coverage* (Fig. V.1.S4I-L), demonstrating that regionally high *coverage* is often reached only by aggregating information over long time periods. This in turn suggests that most species are not continuously *covered* with records.

Temporal coverage

Continuous *temporal coverage* of species and regions is important to reveal and monitor changes in status and distribution of biodiversity (Boakes *et al.*, 2010) and to provide historical baselines for evaluating present-day observations (Willis *et al.*, 2007). *Temporal coverage* since 1750 was extremely low for most species, with a given point in time typically decades away from the nearest record (median: 77.3 years; Fig. II.1.2G). *Temporal coverage* of grid cells was very high across non-eastern Europe while large temporally well-*covered* areas also spanned North America, Central America, the Caribbean, the northern Andes, south-eastern Brazil, South Africa, Madagascar, the Kashmir region, south-western Australia,

and New Zealand (Fig. II.1.2H). In contrast, most of the Amazon and Asia showed extremely poor *temporal coverage*.

For many applied questions in global change research and monitoring, *temporal coverage* specifically of recent decades may be more relevant, and *coverage* since 1950 was indeed higher (Fig. V.1.S2B-C). Worryingly however, several tropical and high arctic regions thought to undergo rapid environmental change were characterized both by poor *temporal coverage* and aging records, notably in Canada, central Africa and Asia (Fig. V.1.S2C-D). For instance, in grid cells in the Democratic Republic of the Congo, the last record was typically collected 28 years ago (median, measured from 2010).

Uncertainty regarding the interpretation of information

Compared to *coverage*–related aspects of species occurrence information (Yang *et al.*, 2013; Sousa-Baena *et al.*, 2014a; Meyer *et al.*, 2015), patterns in more qualitative aspects like information *uncertainty* have received little attention.

Taxonomic uncertainty

Globally *Taxonomic uncertainty* regarding interpretations of scientific names can arise from missing clarity on whether names are accepted or synonyms, from ambiguous synonyms linked to several accepted names, or from orthographic variations and spelling mistakes (Jansen & Dengler, 2010; see *SI V.1.1*). We found that applying our moderate filter to reduce *taxonomic uncertainty* would mean losing 8.4% of records globally (see *Methods*). If however, a very strict taxonomic filter was applied, most information would be lost (66.5% of records; 62.7% of species). Pteridophytes would disproportionately lose records under moderate filtering, compared to other groups (Fig. II.1.3I). All non-spermatophyte records would be excluded by our strict taxonomic filter (Fig. II.1.3I), because *The Plant List* only assigns highest confidence levels to names sourced from taxonomically comprehensive and peer-reviewed databases, which do not exist for these groups (www.theplantlist.org/1.1/about). Depending on the rigor of taxonomic filtering, geographical peaks in lost information appeared either in insular South-East Asia (moderate filter, Fig. II.1.3A) or in the North American Midwest and the Caribbean (strict filter, Fig. II.1.3B). Contrasting these strong taxonomic and geographical patterns, *taxonomic uncertainty* varied very little through time, with usually around 10% and 70% of records in a given 5-year period falling above our moderate and strict *taxonomic uncertainty* threshold, respectively (Fig. II.1.3J).

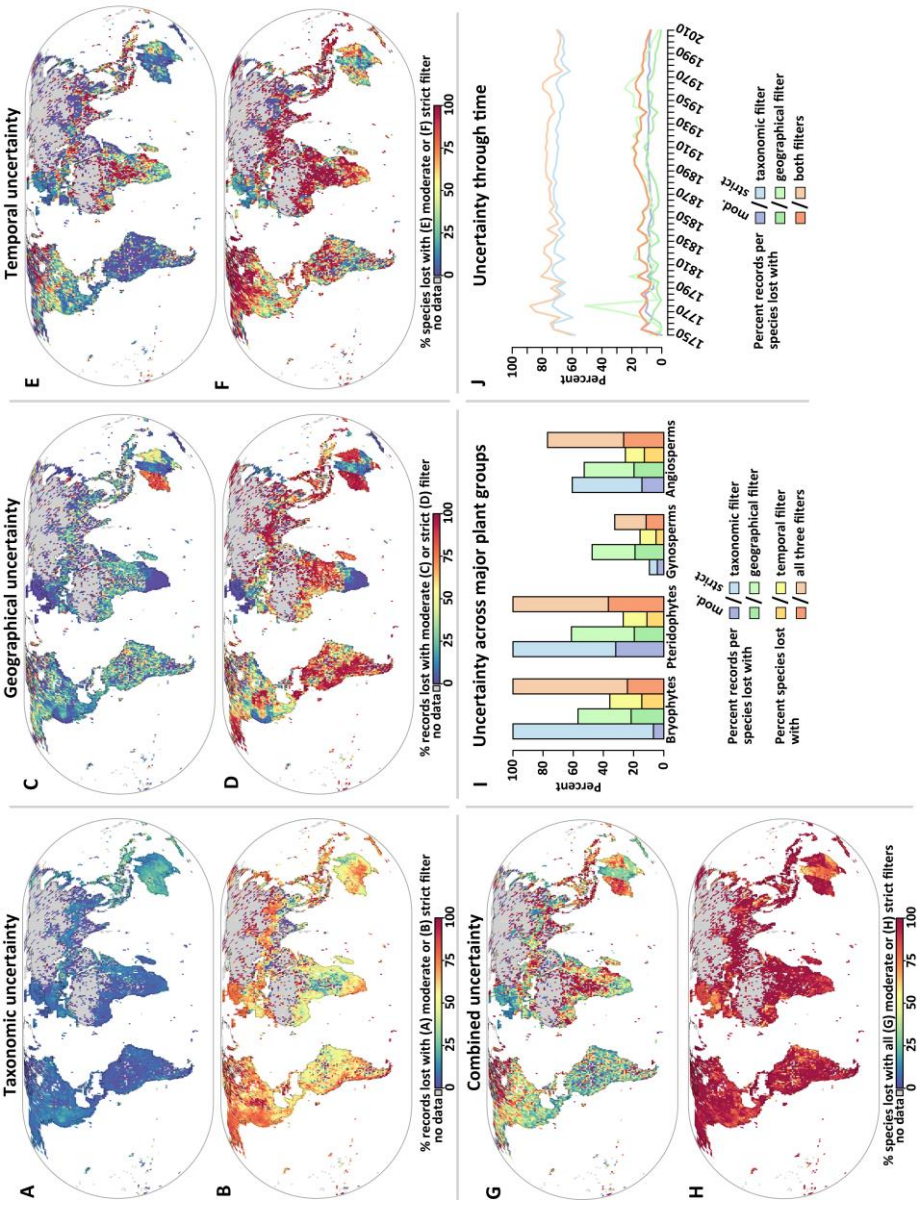


Figure II.1.3. Global variation in occurrence information *uncertainty*. Geographical patterns across 110 km x 110 km grid cells of percentages of records excluded by A) moderate and B) strict *taxonomic uncertainty* filtering, by C) moderate and D) strict *geographical uncertainty* filtering; geographical patterns of percentages of species excluded by E) moderate and F) strict *temporal uncertainty* filtering; by applying all three G) moderate and H) strict filters. I) Taxonomic patterns across major plant groups of mean percentages of records per species excluded by taxonomic and geographical filters, and of species entirely excluded by temporal and combined filters; J) Temporal patterns across 5-year periods between 1750 and 2010 of percentages of records excluded by taxonomic, geographical and the two combined filters.

Geographical uncertainty

Imprecisely geo-referenced sampling locations lead to *uncertainty* regarding the geographical and environmental context of species' occurrences. This *uncertainty* hampers applications built on linking occurrences with high-resolution environmental data, such as species distribution modeling (Feeley & Silman, 2010; Rocchini *et al.*, 2011). Applying our basic geographical filter already lead to a 38% loss in accepted species from the raw dataset (from 367,703 to 229,218), confirming a strong trade-off between geographical quality and *taxonomic coverage* of occurrence information (Feeley & Silman, 2010). Compared to our pre-filtered dataset, further applying moderate and strict geographical filters would lead to an additional reduction of, respectively, 5.5% and 25.3% in species, and 1.9% and 13.9% in records.

These relatively low percentages of globally excluded records were mainly due to high numbers of precisely geo-referenced records in North-Western Europe (Fig. V.1.S1J). However, such global statistics of data *uncertainty* can tremendously underestimate local *uncertainty*, as demonstrated by substantially higher mean percentages of excluded records across grid cells (moderate filter: 22.3% (*sd*: 26.0); strict filter: 58.6% (*sd*: 35.7); Fig. II.1.3C-D). Large areas of relatively low *geographical uncertainty* were in Europe, the western United States, Southern Africa, Japan, New Zealand and parts of Australia (Fig. II.1.3C-D). Records not fulfilling the strictest *geographical uncertainty* criteria were typical for the tropics, but also for remote non-tropical regions, including Alaska, temperate Asia, and Western Australia (Fig. II.1.3D). Imprecise geo-referencing in those regions may be related to a lack of high-quality maps and more sparsely distributed settlements, which often serve as geographical reference points, particularly in older records. *Geographical uncertainty* may also be created at the time of data mobilization. For instance, in Australia, differences in *geographical uncertainty* closely mirrored administrative boundaries, reflecting different mobilization policies of Australian state departments, which contributed 53.8% of Australian records (Fig. II.1.3C). At the time of downloading our records (Jan 2014), certain Australian datasets were mobilized into the GBIF network via intermediaries that generalized location coordinates. Mobilization pathways have since changed and generalizations are now restricted to lists of sensitive species (e.g. those threatened by illegal collecting), therefore, *geographical uncertainty* in Australia will appear lower in future assessments (N. Klazenga, *pers. comm.*).

Geographical uncertainty of records appeared similar across major plant groups (Fig. II.1.3I), but there were several notable changes through time. Older sampling locations were not generally reported with lower precision (Murphey *et al.*, 2004), although such patterns could be observed in several regions, like Spain or south-eastern Australia (Fig. V.1.S4M-Q).

Instead, there were two major periods during which global *geographical uncertainty* increased, in both cases likely reflecting increased explorations of tropical and remote regions, one between 1860 and 1910, coinciding with the second wave of European colonial expansion, and one between 1940 and 1965 (Fig. II.1.3J; Fig. V.1.S4B-C; Fig. II.1.3C-D). The steady decrease in *geographical uncertainty* since 1965 may reflect increasing availability of high-quality maps, and later of geo-referencing technology.

Temporal uncertainty

Early-collected records inherit vital information about past biota. However, they also lead to greater *temporal uncertainty* regarding species' continued presences near sampling locations, as distributions may change in response to environmental change (Thuiller *et al.*, 2008) and biological processes (Schurr *et al.*, 2012). Therefore, many applications like conservation planning or SDMs that link occurrences with recent habitat data usually require recent occurrence information (Boitani *et al.*, 2011).

86.3% of globally represented species had at least one record collected after 1970 and 72.4% even had records collected after 1990. Using these dates as filters for excluding records would cause an average loss of, respectively, 32.0 and 61.8% of species in a given grid cell (Fig. II.1.3E-F). Regions were most species had records collected after 1990 include continuously well-sampled north-western Europe, but also areas where most species were only recorded during recent surveys, such as Benin, the circum-Tibetan mountain ranges, or Indochina (Fig. II.1.3F, Fig. V.1.S4E). In contrast, much of arctic Canada, central Africa, Iraq, eastern India, Myanmar and Java were characterized by generally outdated information, as most recorded species did not even have records collected after 1970 (Fig. II.1.3E). Local reasons for spatio-temporal changes in sampling activity may include shifting funding priorities (Ahrends *et al.*, 2011), arising security concerns (Brito *et al.*, 2013), or lowered botanical appeal of environmentally degraded regions (Boakes *et al.*, 2010). Whatever the reasons, it is important to detect and account for such spatio-temporal biases and uncertainties. Mean percentages of excluded species also varied three-fold across major plant groups (5.4%-15.10%; moderate filter; Fig. II.1.3I), showcasing potential taxonomic biases introduced by temporal filters.

Combined uncertainty

Combining filters to minimize *uncertainty* in all three dimensions lead to substantial trade-offs for *coverage* (compare Feeley & Silman (2010); Boitani *et al.* (2011)). 78.9% of all species in our dataset had no record that passed all strict filters; 52.2% had no record passing all moderate filters. *Uncertainty* was even more apparent in geographical patterns: North-western Europe was the only larger regions where typically ≥ 80% of species in a grid cell had at least

one record that passed moderate combined filters (Fig. II.1.3G). No region retained much of available information under strict combined filtering; even regions where 20% of species would withstand such filters were confined to parts of Europe, Benin, Indochina, and central and south-eastern Australia (Fig. II.1.3H).

Given such pervasive levels of data *uncertainty*, it is very likely that species identities and their environmental associations are frequently misinterpreted. Furthermore, our documented patterns of *uncertainty* demonstrate that the likelihood of such misinterpretations is biased to particular taxonomic groups, geographical regions, and time periods. Overall, these issues seriously hamper opportunities for ecological inference, and need to be carefully accounted for whenever records of variable or unknown quality are used in biodiversity analyses (Rocchini *et al.*, 2011).

Relationships between different aspects of occurrence information

Pairwise Spearman rank correlations across 9 variables of occurrence information mostly yielded weak to moderate associations (Fig. V.1.S5). Different *coverage* aspects correlated moderately to strongly (r_S: 0.63 to 0.86), which was expected as *coverage* of any dimension is numerically constrained by the number of available records (correlations with record number: 0.65 to 0.92; compare Yang *et al.* (2013)). *Taxonomic* and *geographical coverage* also moderately correlated with time since the last recording activities (r_S: -0.67 to -0.70). In contrast, most *uncertainty* aspects showed no or only weak correlations, the only moderately strong correlation being that between *temporal* and combined *uncertainty* (r_S: 0.75). Most metrics showed no strong correlations with quantities of mobilized raw records (Fig. V.1.S5), suggesting that such simpler indicators cannot reliably inform about different aspects of occurrence information.

To reduce complexity, we included the 9 variables in a PCA (Fig. II.1.4A-C) and mapped PCA site scores on a world map to identify regions that are predominantly characterized by specific aspects of occurrence information (Fig. II.1.4D). The first three axes of the PCA accounted for 69.8% of the variance. The most important axis (38%) mainly separated regions of high *taxonomic* and *geographical coverage*, e.g. in Europe (r_S: 0.86/0.85; Fig. II.1.4A-B/D), from regions where a long time has passed since the last recording activities, e.g. in Central Africa and South Asia (r_S: -0.85; Fig. II.1.4A-B/D). The second axis (20% of variance) mainly correlated with combined and *temporal uncertainty* (r_S: 0.74/0.75; Fig. II.1.4A/C/D), highlighting e.g. arctic Canada. For instance, combined *uncertainty* was characteristic for much of Asia, such as the Altai or the mountain ranges between Eastern

Tibet and Sichuan (Fig. II.1.4D). *Geographical* and *taxonomic uncertainty* varied mainly along the third axis (11.8% of variance; r_S: 0.69/0.47; Fig. II.1.4B-C), characterizing e.g. Borneo. Some metrics were poorly correlated with all three major PCA axes (e.g. for number of missing species r_S: -0.34 to 0.31 for first three axes; r_S: 0.7 for fourth axis).

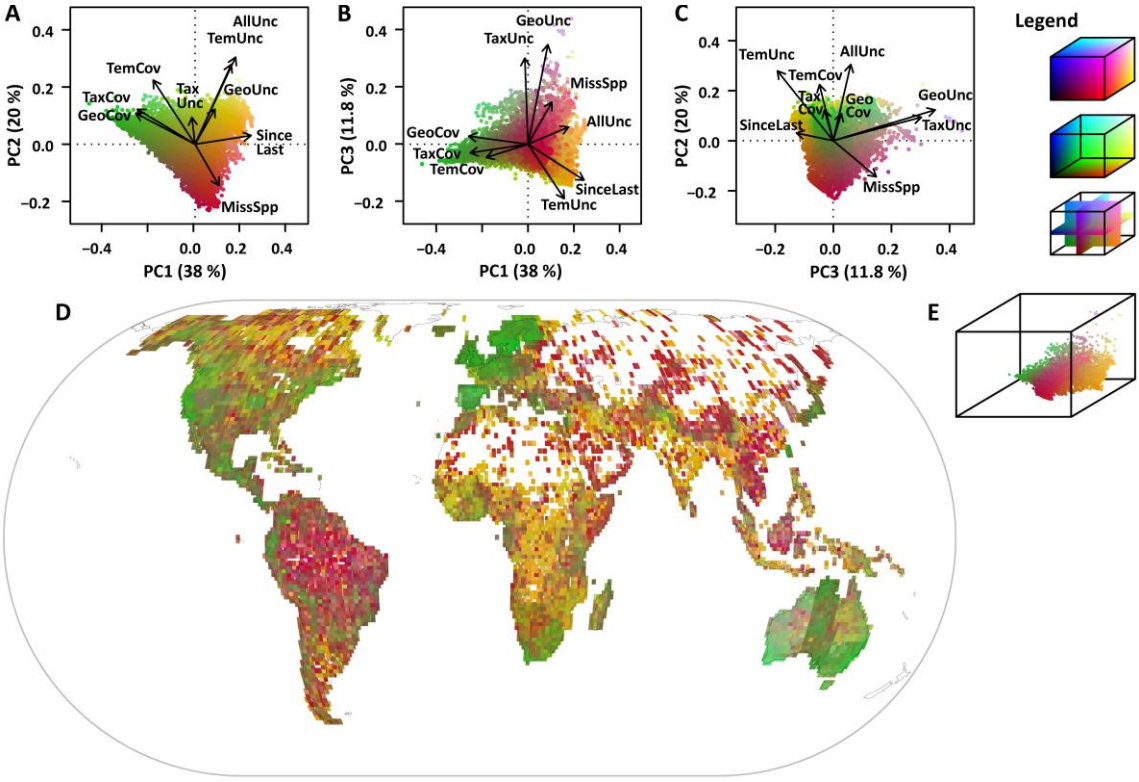


Figure II.1.4. Principal component analyses of 9 metrics of plant occurrence information. PCA for 8,567 110 km x 110 km grid cells with ≥ 1 validated record. A-C) Biplots of the first three PCA axes (numbers in parentheses indicate percentages of variance accounted for by each axis); points represent grid cells. Colors refer to a red–green–blue (RGB) color space (cubes in legend and E) projected onto the 3D PCA space E) following (Weigelt *et al.*, 2013). Each grid cell consistently has the same color across A-E. In A-C, Points are plotted in decreasing order of the respective component not shown to give an impression of three-dimensionality. **TaxCov**: *taxonomic coverage*, calculated as the ratio between recorded richness and richness modeled by (Kreft & Jetz, 2007); **GeoCov**: *geographical coverage*, estimated as the number of sampling locations per 10^4 km² land area; **TempCov**: *temporal coverage* since 1750, estimated as the mean minimum Euclidean distance between all possible months between 1750 and 2010 to their respective closests month with records; **TaxUnc**: percentage of records lost under moderate taxonomic filtering; **GeoUnc**: percentage of records lost under moderate geographical filtering; **TempUnc**: percentage of species lost under moderate temporal filtering; **AllUnc**: percentage of species lost with all three moderate filters applied (*see Methods* for information on filters). **MissSpp**: number of species that are not recorded but expected based on the environment-richness; **SinceLast**: Time (in years) since the last mobilized record was recorded. All correlations based on z-transformed variables. D) Global map of ordination site scores; similar colors denote regions whose occurrence information is mainly characterized by similar information aspects. Cube in D shows PCA results in a 3D space.

The above patterns and analyses highlight the differences, rather than the similarities, between geographical patterns of different aspects and dimensions of occurrence information. Different

limitations predominate in different regions. This multidimensionality of limitations in occurrence information should be considered in research and conservation applications, as well as in future assessments of data limitations. How combinations of different data limitations bias ecological analyses, and how these biases can be effectively controlled for, remains largely un-investigated.

Opportunities for using DAI in plant research, conservation and monitoring

Despite the showcased limitations in DAI, there is an urgent need to apply this information in plant research and conservation. For instance, distribution estimates play a vital role in advancing conservation assessments (GSPC target 2), developing management plans for threatened species (GSPC target 7), and monitoring changes in species' statuses (Brummitt *et al.*, 2015). As we show below, the potential for such applications largely depends on the ability of methods to deal with low record numbers and high data uncertainty.

If distribution estimates could be derived from 10 sampling locations (Rivers *et al.*, 2011) and methods were robust to relatively high data uncertainty, DAI might facilitate preliminary assessments for 85,787 non-red-listed or 'Data-Deficient' species globally (24.5% of all plants; Fig. II.1.5). This represents a potential seven-fold increase compared to the IUCN Red List (as of Aug 2014). However, this number would drop to 1,921 for uncertainty-sensitive methods requiring ≥ 200 locations. Similarly, depending on methods' data requirements, estimates might be feasible for 0.1 to 15.7% of 'Threatened' plants, and for 0.1 to 6.6% of all plants during three 20-year periods since 1950. Furthermore, the potential for such applications is geographical highly biased (Fig. II.1.5). For instance, based on DAI, distributional changes since 1950 might be documented for 386 to 3,682 plant species in Europe, but only 0 to 26 in the Pacific region (Fig. II.1.5).

Most distribution modeling methods are highly sensitive to both number and quality of records (Guisan *et al.*, 2007). Restricting analyses to highest-quality data is often recommended (Feeley & Silman, 2010), but cutoffs are always arbitrary, and strict filters wipe out most available information (Fig. II.1.4H). Moreover, depending on strictness of filters, they introduce different biases to already-biased datasets (Fig. II.1.4). More effective usage of available information would be to explicitly incorporate biases and uncertainties into analyses. Methods for doing so are increasingly available (e.g. McInerny & Purves (2011); Dorazio (2014)) and further developing such methods holds great potential for advancing global plant research and conservation.

Taxonomic standardization and basic geographical plausibility checks, as carried out in this study, are essential when using mobilized occurrence records. However, such measures obviously cannot ensure that information is in fact accurate, given likely taxonomic misidentifications and geo-referencing errors. Sampled taxonomic re-assessments of original material and sampled ground-truthing of occurrences could provide vital information on typical rates of such errors for different taxa, regions and data sources, which could additionally be accounted for in analyses.

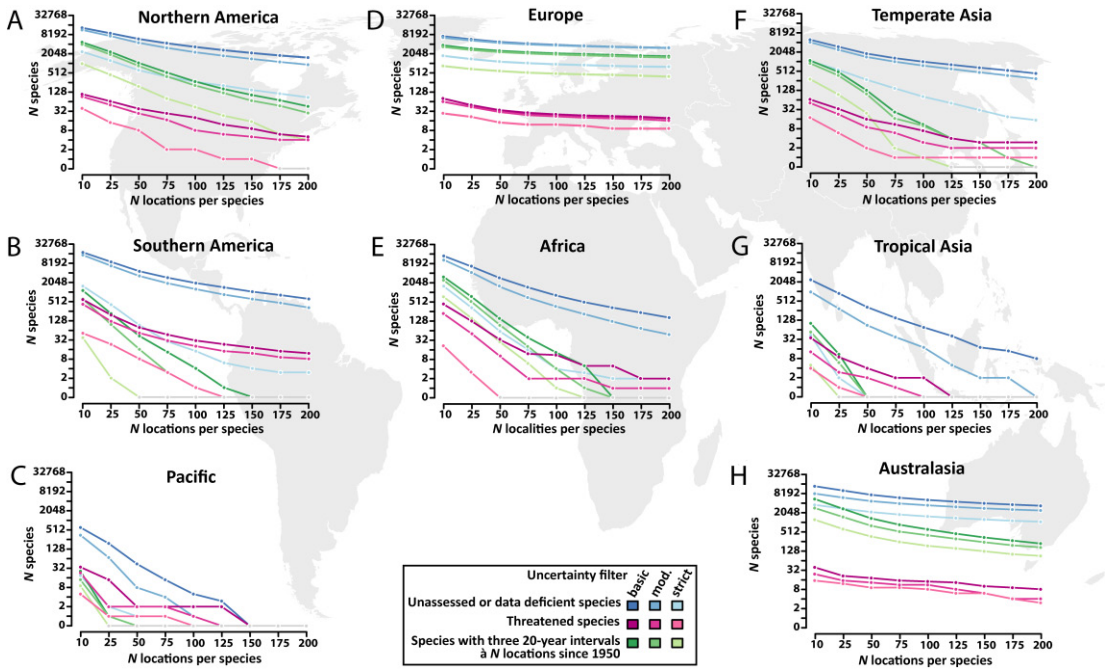


Figure II.1.5. Global trade-offs between plant occurrence information *coverage* and *uncertainty*. Graphs show the number of species for which their distribution could be estimated with different hypothetical methods, in dependence of those methods' minimum requirements (10 to 200 unique sampling locations; Rivers *et al.* (2011); Feeley & Silman (2011a)) and abilities to cope with different levels of data *uncertainty*. A) Northern America, B) Southern America, C) Pacific, D) Europe, E) Africa, F) Temperate Asia, G) Tropical Asia, H) Australasia. Blue colors denote species that are either un-assessed or 'Data Deficient' on the International Union for the Conservation of Nature's Red List. Violet colors denote species with Red List categories 'Vulnerable', 'Endangered' or 'Critically Endangered'. Green colors denote species for which the indicated number of sampling locations exists in each of three 20-year time periods since 1950. Different color shadings denote basic, moderate or strict filtering of datasets (combined taxonomic, geographical and temporal filters). World regions are level-1 regions of Biodiversity Information Standards (TDWG, formerly International Working Group on Taxonomic Databases).

After decades of intensive data mobilization, options for using plants as a model group for global research and conservation are still severely compromised by different data limitations. Even under our most optimistic scenario regarding methods' data requirements and robustness to uncertainty, distribution estimations based on DAI were unfeasible for three quarters of all plants. The multidimensional data limitations also highlight flaws in the accuracy of datasets that are derived from primary biodiversity records, including checklists, range maps and many atlas data. This is exemplified by the many WSCP-listed species recorded in regions adjacent

to their supposedly correct native ranges. Severe data gaps will likely persist for decades to come, as evident in slow progress towards regional floras (Paton, 2013). Meeting GSPC targets on plant conservation seems unlikely without substantial increases in funds and personnel allocated to data collection and mobilization. Given limited resources, efforts to improve occurrence information should be globally coordinated and prioritized (Meyer *et al.*, 2015).

Towards more effective improvement of DAI

One way of efficiently improving DAI could be reducing the level of information duplication (Fig. V.1.S6A). For instance, 0.22 M validated records in south-western Mexico yielded 0.19 M unique records (ratio 0.8), whereas substantially more (0.50 M) validated records in Peru yielded only 0.16 M (ratio 0.3). While these duplicated species-location-month-combinations may benefit certain research applications, such localized, dense sampling entails trade-offs for global *coverage* (Meyer *et al.*, 2015). Broad *coverage* is the main strength of occurrence records (Boakes *et al.*, 2010; Jetz *et al.*, 2012a), compared to more localized, dense recording schemes (e.g. Scholes *et al.* (2008); Dengler *et al.* (2011)). This strength should be fostered by prioritizing poorly-*covered* regions and species in future record collection and mobilization efforts.

Our analyses provide an important first step towards such prioritizations. Detailed distinctions between issues of *coverage* and *uncertainty* in taxonomic, geographical and temporal information allow narrowing down key improvements necessary for different regions and taxa. For instance, peaks in *taxonomic uncertainty* for South-East Asian and pteridophyte floras could be overcome by targeted taxonomic revisions and better integration of available taxonomic resources into The Plant List. New surveys are urgently required for Central Africa, Mozambique, and tropical Asia, as available information is largely outdated. More generally, Asian and bryophyte floras are woefully under-represented. Efforts to digitize respective collections and integrate digital datasets into international data-sharing networks should be a top priority. Such preliminary global priorities could be focused further by accounting for environmental dissimilarity to well-sampled areas (Sousa-Baena *et al.*, 2014a) or by focusing on species and areas of conservation concern (Pyke & Ehrlich, 2010). Identifying relevant collections through metadata digitization (Berendsohn & Seltmann, 2010) and identifying information-limiting socioeconomic factors (Meyer *et al.*, 2015) can help prioritize specific activities likely to have a large impact on improving global DAI.

As we demonstrated, limitations in occurrence information are multidimensional and different information aspects are not strongly correlated. This raises the question of how to effectively monitor progress in data mobilization, and more broadly, progress towards international targets on improving and sharing conservation-relevant knowledge (e.g. Aichi target 19, GSPC target 3). Straightforward, simple indicators like global or per-country quantities of mobilized raw data (e.g. Tittensor *et al.* (2014)) cannot inform about data *uncertainties* or fine-scale biases in *coverage*. To monitor improvements of DAI, rather than mere increases in volume, we recommend developing and routinely evaluating a set of indicators that inform about different information aspects and dimensions at relevant scales.

Conclusions

Severe multidimensional gaps, biases and uncertainties are prevalent in global DAI on plant occurrences. These limitations hamper prospects of establishing plants as a model group for global biodiversity research, and achieving international targets on plant conservation. Either goal would require both substantially scaling up and prioritizing efforts to collect and mobilize additional occurrence information. Success of such efforts should be monitored using meaningful indicators. Furthermore, available information should be used effectively by explicitly accounting for biases and uncertainties in ecological analyses.

Acknowledgements. We thank those active in collecting, curating, and sharing plant distribution data. We thank Tim Robertson for assistance with data retrieval from GBIF, and Janet Scott for facilitating access to IUCN red list data. CM acknowledges funding from the Deutsche Bundesstiftung Umwelt (DBU). HK acknowledges funding by the German Research Council (DFG) in the framework of the German Excellence Initiative within the Free Floater Program at the University of Göttingen. PW acknowledges funding in the scope of the BEFmate project from the Ministry of Science and Culture of Lower Saxony.

Chapter 2

Global priorities for an effective information basis of biodiversity distributions

Carsten Meyer, Holger Kreft, Robert P. Guralnick and Walter Jetz

Updated version published: *Nature Communications* **6**:8221. DOI: 10.1038/ncomms9221.

Abstract

Severe gaps and biases in digital accessible information (DAI) of species distributions hamper prospects of safeguarding biodiversity and ecosystem services and reliably addressing central questions in ecology and evolution. Accordingly, governments have agreed on improving and sharing biodiversity knowledge by 2020 (United Nations Convention on Biological Diversity's Aichi target 19). To achieve this target, gaps in DAI must be identified, and actions prioritized to address their root causes. We take terrestrial vertebrates, an iconic and comparatively well-studied group, as a model and present the first globally comprehensive assessment of patterns and drivers of gaps in DAI, based on an integration of 157 million validated point records with 21,170 expert-based distribution maps. We demonstrate that outside a few well-sampled regions, DAI provides a very limited and spatially highly biased inventory of actual biodiversity. Coarser spatial grains result in more complete inventories, but provide insufficient detail for conservation and resource management. Surprisingly, large emerging economies are particularly under-represented in global DAI, even more so than species-rich, developing countries in the tropics. Multi-model inference reveals that completeness is mainly limited by distance to researchers, locally available research funding, and political participation in data-sharing networks, rather than transportation infrastructure, or size and funding of Western data contributors as often assumed. Our study provides an empirical baseline to advance strategies of enhancing the global information basis of biodiversity. In particular, our results highlight the need for targeted data integration from non-Western data holders and intensified cooperation to more effectively address societal biodiversity information needs.

Introduction

The parties to the Convention on Biological Diversity (CBD) have agreed on 20 targets to improve the state of biodiversity by 2020 (cbd.int/sp/targets/). Aichi target 19 specifically mandates the development of an advanced and shared biodiversity knowledge base. The distribution of species in space is a central aspect of biodiversity knowledge that can enable the effective management of biodiversity and associated ecosystem services in a rapidly changing world (Whittaker *et al.*, 2005; Butchart *et al.*, 2010; Boitani *et al.*, 2011; Pereira *et al.*, 2012; Guisan *et al.*, 2013). Species distributions are critical for informing actions towards

multiple Aichi targets, associated environmental indicators (Pereira *et al.*, 2013), and the recently launched assessment work of the Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services, IPBES (Inouye, 2014).

International efforts to mobilize and aggregate distribution data, most notably the Global Biodiversity Information Facility (GBIF), have facilitated access to large quantities of digital species occurrence records from a variety of data sources, especially museum specimens and field observations (Edwards, 2000; Graham *et al.*, 2004). Such records provide vital, fine-scale information about where and when species occur and are widely used in ecology, evolution and conservation research. In contrast to expert knowledge and un-digitized datasets, which are effectively inaccessible, such mobilized records form the bulk of *de facto* 'digital accessible information' (DAI, originally referred to as DAK in Sousa-Baena *et al.* (2014a)). While in a recent study (Tittensor *et al.*, 2014) the authors saw evidence for progress towards Aichi target 19 in increasing volumes of GBIF-facilitated DAI, they had to acknowledge the critical caveat of unclear "taxonomic coverage (e.g. number of species), record completeness or geographic biases".

Severe gaps and biases usually exist in DAI (Boakes *et al.*, 2010; Feeley & Silman, 2011b; Jetz *et al.*, 2012a; Sousa-Baena *et al.*, 2014a), and require careful consideration in ecological modelling (Guisan & Thuiller, 2005; Phillips *et al.*, 2009; Yang *et al.*, 2013) and conservation research (Boitani *et al.*, 2011). Data limitations may arise from a multitude of socio-economic and geographic factors, including inadequate financial and institutional resources (Vollmar *et al.*, 2010; Ahrends *et al.*, 2011; Amano & Sutherland, 2013), poor international scientific cooperation (Amano & Sutherland, 2013), lack of access or regional safety concerns (Freitag *et al.*, 1998; Moerman & Estabrook, 2006; Amano & Sutherland, 2013; Ballesteros-Mejia *et al.*, 2013), or a focus on regions with certain appeal like endemism-, species-rich or protected areas (Freitag *et al.*, 1998; Boakes *et al.*, 2010; Yang *et al.*, 2014). The amount of data required to completely inventory species assemblages is a function of their richness and the spatial grain (Soberón *et al.*, 2007; Feeley & Silman, 2011b; Jetz *et al.*, 2012a). To be relevant for conservation applications, distribution datasets must inform about species occurrences at fine spatial grains (Smith *et al.*, 2009), either directly or by facilitating derived, fine-grain data products (Jetz *et al.*, 2012a; Guisan *et al.*, 2013). Such fine-grain data products are integral to conservation research, but can also directly influence conservation decision-making. For instance, in Madagascar, occurrence records have facilitated the identification of 'priority areas' (Kremen *et al.*, 2008) where following a legal decree, no mining and forestry activities can be permitted (*Arrêté Interministériel* n18633/2008/MEFT/MEM, renewed in 2014; further examples in Guisan *et al.* (2013)).

Identifying information gaps and factors limiting the dissemination of biodiversity information are recognized as priorities both at the political (Conference of the Parties to the Convention on Biological Diversity, 2010) and scientific (IPBES, 2015) levels of the CBD. To date, magnitude and exact location of gaps in global DAI as well as the relative importance of underlying causes remain unclear, hampering prioritization of future data mobilization efforts (Hobern *et al.*, 2013) and thus cost-effective progress towards Aichi target 19. International efforts to mobilize biodiversity records remain un-assessed for their success and effectiveness in addressing targets to improve and share biodiversity knowledge.

Here, we perform this assessment for 21,170 species of birds, mammals, and amphibians and c. 157 million geographically and taxonomically validated point records that were provided to GBIF by 160 data publishers, including small institutions with a distinct taxonomic and geographic focus as well as large internationally active research museums and citizen science programs (see *Supplementary Information: Table V.2.S7* and *Methods*). We determine the factors currently limiting biodiversity inventory completeness in global DAI and identify priority regions and activities to advance it.

Results and Discussion

At a grain size of 110 km grid cells, the density of terrestrial vertebrate records varies by five orders of magnitude (Fig. II.2.1 A), peaking in parts of Europe, North and Central America, and Australia. Conversely, 48% of Asian, 35% of African and 21% of South American cells have no records mobilized into DAI. At this spatial grain, the finest ensuring sufficient accuracy of species expert-range maps (Hurlbert & Jetz, 2007; Hawkins *et al.*, 2008), species richness derived from point records shows little concordance with expected richness (Fig. II.2.1 B, C). While spatial patterns between the two data sources show at least weak associations (rs=0.28-0.39, see Table V.2.S1 a), only 4.2% of all 12,029 cells reach ≥ 80% completeness (Fig. II.2.1 D).

Completeness, defined as percentage of expected richness documented with point records, is moderately to strongly predicted by record density (binomial GLM, d^2=0.59-0.90, Fig. V.2.S1, Table V.2.S1 b; see *SI V.2 Methods* for details). Whereas high record density results in high levels of completeness in much of the Nearctic and Australasia, this is less the case for the more species-rich Neo-, and Afrotropics (Fig. II.2.1 A-B, D-E, Fig. V.2.S1 D). The Eastern Palaearctic and Indomalayan realms are characterized by particularly low levels of completeness. Average completeness also varies greatly among the World's major biomes and biomes within biogeographical realms (Fig. II.2.1 E, Table V.2.S2 a-c). Specifically, tropical

and subtropical forests, grasslands and savannas, but also boreal forests and tundra biomes remain vastly under-inventoried. Surprisingly, we cannot confirm a pronounced "tropical data gap" (Collen *et al.*, 2008; P_{Dut}=0.27; tropics versus non-tropics). Instead, a severe gap emerges across most of Asia (including temperate regions), non-Southern Africa, and Brazil (P_{Dut}<0.01; when comparing mean completeness in these areas to all others).

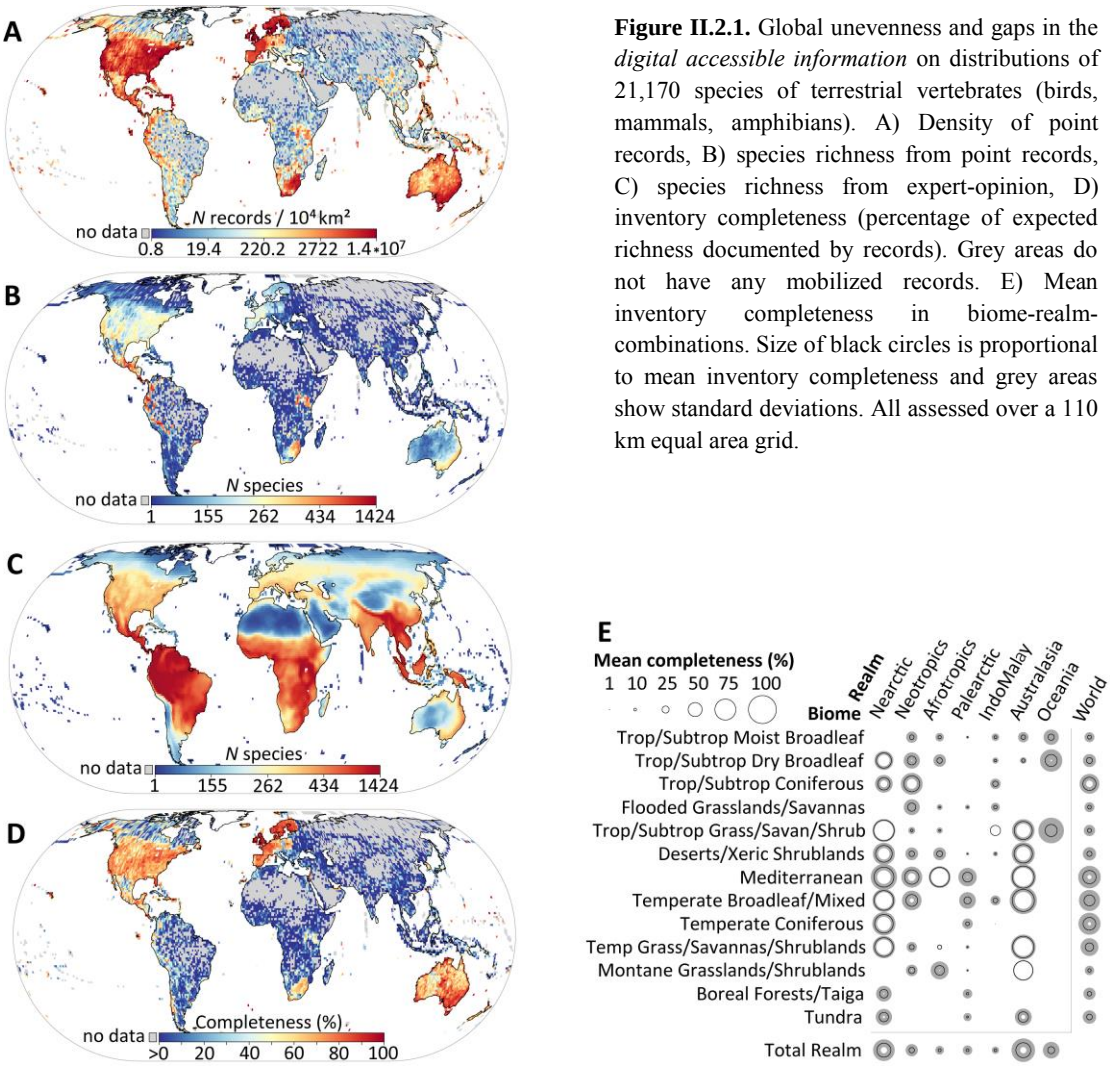


Figure II.2.1. Global unevenness and gaps in the *digital accessible information* on distributions of 21,170 species of terrestrial vertebrates (birds, mammals, amphibians). A) Density of point records, B) species richness from point records, C) species richness from expert-opinion, D) inventory completeness (percentage of expected richness documented by records). Grey areas do not have any mobilized records. E) Mean inventory completeness in biome-realm-combinations. Size of black circles is proportional to mean inventory completeness and grey areas show standard deviations. All assessed over a 110 km equal area grid.

While these strong geographic differences in completeness are broadly repeated among the three vertebrate groups (Fig. II.2.2 A), completeness patterns among the three taxa only show moderately strong positive associations (rs=0.65–0.74 depending on taxon and grain, all P_{Dut}<0.001). This suggests that the completeness pattern of a single-taxon is a poor predictor for un-assessed taxa and highlights the need to identify taxon-specific information gaps (Vale & Jenkins, 2012). As expected from substantially fewer records for mammals and amphibians compared to birds (~3M and ~1M, compared to ~150M, see *SI V.2 Methods*), their overall

level of completeness is significantly lower (Tukey-test, P_{Dut}<0.001 for all spatial grains, when comparing mammal / amphibian completeness with bird completeness).

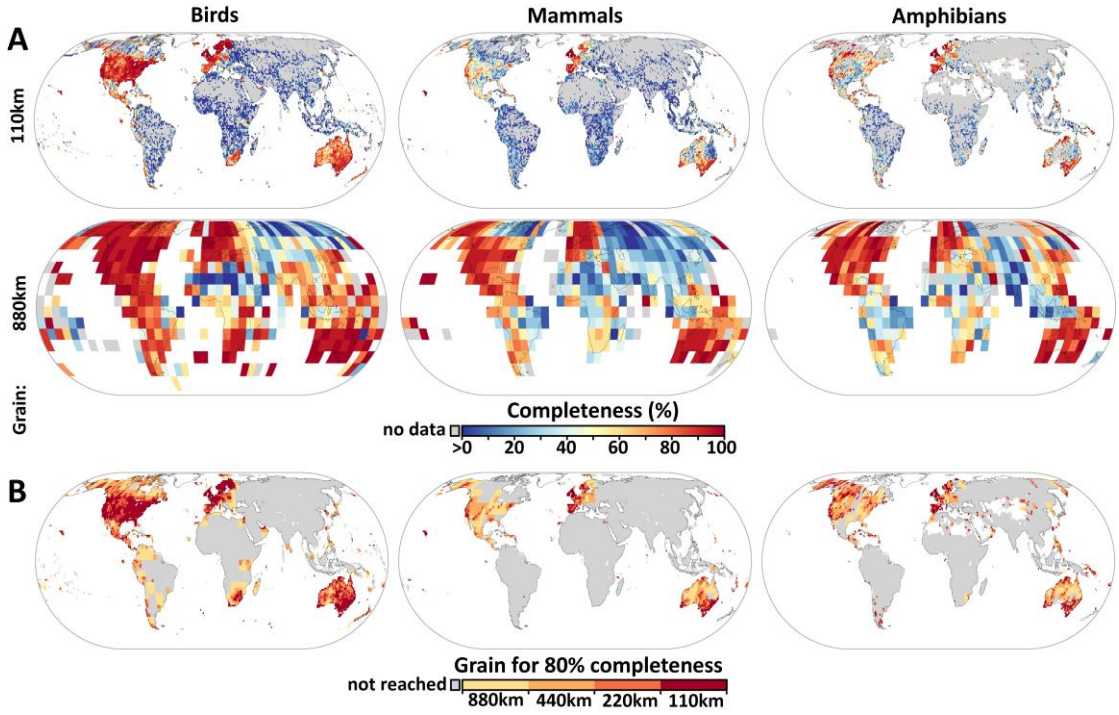


Figure II.2.2. Spatial variation in point record-based inventory completeness for three vertebrate taxa at different spatial grains. A) Inventory completeness at the 110 km and 880 km grain (for 220 km and 440 km grain, see Fig. V.2.S1.). B) Minimum grain size to reach 80% inventory completeness, mapped at 110 km. Grey grid cells in A) show areas within the taxon's global range without mobilized records, in B) areas that do not reach 80% completeness at 880 km.

Completeness levels of ≥ 80% over large extents, even at a relatively coarse grain of 110 km, are only achieved in birds and only in North America, Europe, and Australia (Fig. II.2.2 A). Coarsening grains even further to 440 or 880 km substantially increases completeness in all groups (Kruskal-Wallis-test, all P<0.001, Fig. II.2.2 A-B, Fig. V.2.S2), but necessarily leads to inferior opportunities for inference and application. Such coarse grains are not adequate for most questions in ecology (Beck *et al.*, 2012) and, with land-use and conservation actions typically set at the kilometer scale or finer, are unsuited for effective resource management. Most species distribution models (SDM) connecting records with fine-grained environmental data for extrapolation (Guisan & Thuiller, 2005) are unable to provide a general remedy here, due to their known sensitivity to environmental bias (Menke *et al.*, 2009; Feeley & Silman, 2011b). This pervasive lack of DAI over vast extents (e.g., only <20% completeness at 880 km grain over much of Asia, Fig. II.2.2 A) demonstrates that for many regions with large conservation opportunities (Venter *et al.*, 2014) there are not sufficient mobilized occurrence data to facilitate even the most sophisticated modeling approaches. Global numbers of sampling locations for the majority of species are far below the 50-100 typically

recommmended (Wisz *et al.*, 2008; Boitani *et al.*, 2011; Feeley & Silman, 2011a) as minimum SDM requirements (54.9% of all bird species have <50 records, median=37; mammals: 79.2%, median=6; amphibians: 91.3%, median=2; compare Cayuela *et al.* (2009); Feeley & Silman (2011a)).

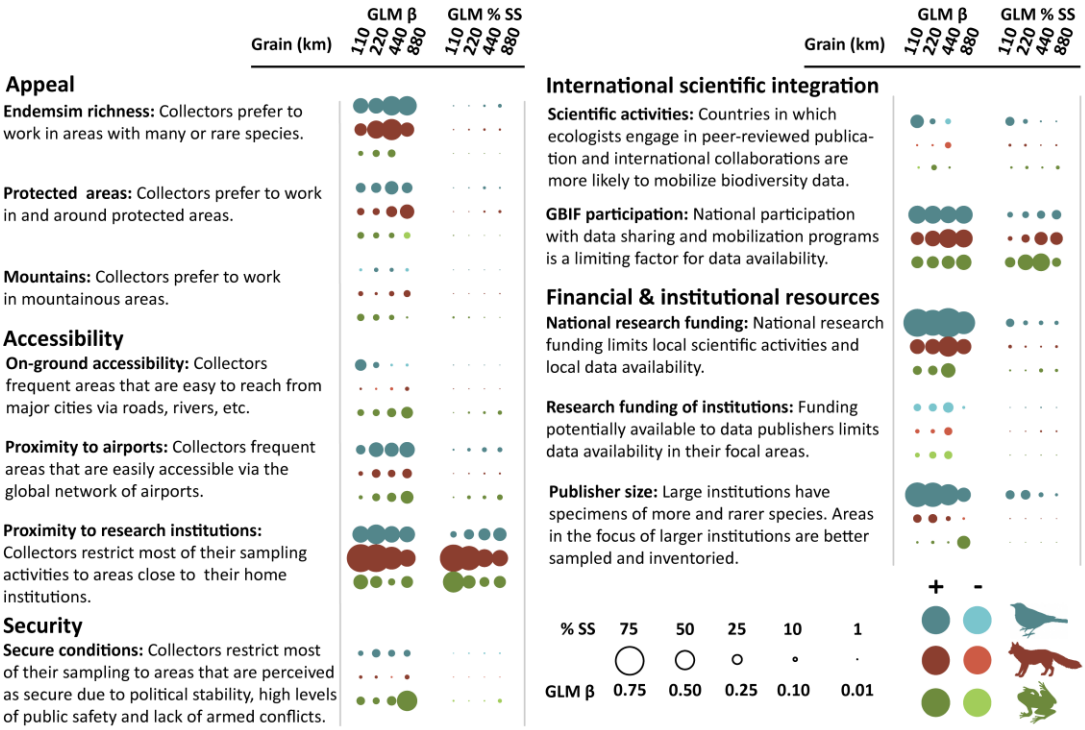


Figure II.2.3. Determinants of inventory completeness in digital accessible information. Effects were tested in multiple generalized linear regression models with a binomial distribution and a logit link (GLM β and GLM % SS). All possible model subsets were ranked based on AIC scores and subsets with ΔAIC < 10 re-run as spatial models to account for spatial autocorrelation in model residuals. Bubble size represents the relative strength of predictor-response relationships. Vertebrate groups are represented by different colors, with shading denoting the direction of the relationship. We show the relative importance of predictors using two different metrics: i) the standardized coefficients of the reduced spatial multiple regression models with the lowest AIC score (blank cells indicate variables that were not included in these models) (GLM β), and ii) the percentage each predictor has in the total sum of squares (GLM % SS) of a type III ANOVA. For results of bivariate models and similar tests on record density, see Fig. V.2.S3. For details on hypotheses, methods, and results, see SI Materials and Methods, Fig. V.2.S3-4, Tables V.2.S3-5).

Such glaring data gaps highlight the need to identify and, where possible, address the root causes of low inventory completeness. Understanding of the key driving factors of bias is important to prioritize activities in data mobilization. Further, drivers of bias can be explicitly incorporated into biodiversity models (Dorazio, 2014; Fithian *et al.*, 2014; Manceur & Kühn, 2014). Gaps in DAI may result from the way data are collected in the field, digitized in museums, or mobilized and aggregated as digital species records into global biodiversity data-sharing networks. To this end, we tested twelve hypotheses falling into five broad categories: appeal, accessibility, security, international scientific integration, and financial and institutional resources (details in Fig. II.2.3 and *SI V.2 Methods*). Most hypotheses receive at least some support in our multi-model inference framework, highlighting the complex

interplay of geographic and socio-economic factors as drivers of inventory completeness (Fig. II.2.3; for record density and bivariate model results see Fig. V.2.S3-4; detailed results in Tables V.2.S3-5). Depending on taxon and grain, minimum adequate models of inventory completeness explain 60-78% of the deviance (Table V.2.S3) and the relative importance of factors varies more strongly among taxonomic groups than among grain sizes (depending on the predictor, percentages of sums of squares explained in an ANOVA are 16.5-72.5% higher for factor "taxon" compared to factor "spatial grain").

A strong role for data collection has been attributed to region or species "appeal", e.g., researchers' preference for reserves, mountains or other areas of high total, rare and range-restricted species richness (Freitag *et al.*, 1998; Soria-Auza & Kessler, 2008; Yang *et al.*, 2014). We find this supported in birds and mammals by strong positive effects on inventory completeness of endemism richness, and weaker effects of protected area coverage. Surprisingly, we find relatively low importance of on-ground accessibility from cities and proximity to airports (Fig. II.2.3), which have previously been suggested to strongly constrain field collections (Freitag *et al.*, 1998; Ballesteros-Mejia *et al.*, 2013). In contrast, spatial distance to data-contributing institutions (Table V.2.S7) consistently emerges as a key predictor of inventory completeness and record density (Fig. II.2.3, Fig. V.2.S3). This highlights the imprint that long-term logistics of maintaining field sampling and specimen transport leave on global biodiversity information (compare Moerman & Estabrook (2006); Yang *et al.* (2014)). Insecure conditions may discourage field sampling (Amano & Sutherland, 2013; Brito *et al.*, 2013), but we find little evidence that security aspects are important in limiting completeness or record density (Fig. II.2.3; Fig. V.2.S3, *SI V.2 Methods*: *1.B*). We expected our index of integration into scientific activities, i.e., country's H-index in ecology multiplied by level of international collaboration, to be strongly correlated with inventory completeness, as it should reflect the routine of making research results public (Collen *et al.*, 2008; Amano & Sutherland, 2013) However, it is neither important for explaining completeness nor record density (Fig. II.2.3; Fig. V.2.S3). Conversely, GBIF participation emerges as a consistently strong factor determining completeness in DAI. Supporting previous suggestions (King, 2002; Ahrends *et al.*, 2011), national research funding (gross expenditure on research and development) is strongly positively correlated with completeness (Fig. II.2.3). Surprisingly, however, research funding of countries where data-publishing institutions are situated does not affect inventory completeness in the regions of their sampling activity (*Supplementary Information: Methods*). Finally, publisher size, estimated from contributed data volume, only weakly predicts inventory completeness for mammals and amphibians, but it has much stronger effects for birds, where the largest data contributors are not museums but aggregators of citizen-science observation data (Table V.2.S7), pointing to the potential of alternative, non-institution-based ways of producing DAI

for certain taxa (see discussions in Hochachka *et al.*, (2012); Jetz *et al.*, (2012a); Beck *et al.*, (2013)).

Most of the strongest limiting factors of completeness affect digitization and mobilization of existing data rather than the actual collection of new records in the field. While adequate national research funding is vital for producing DAI on local biodiversity, our results suggest that funding for university research usually leading to peer-reviewed publications is not improving our ability to close information gaps as greatly as direct support for data mobilization programs (Fig. II.2.3: 'Scientific activities' vs. 'GBIF participation'). A likely reason is that current data-archiving policies (Whitlock, 2011) and academic reward systems (Enke *et al.*, 2012) do not favor data-sharing activities. They further suggest that the largest or best-funded museums alone are unable to guarantee high inventory completeness in distant regions, unless their efforts are backed by supportive local conditions, such as locally available research funding, mobilization efforts in local research institutions and national commitment to data-sharing. The most effective strategy for closing gaps in DAI may therefore lie in supporting mobilization efforts in institutions nearby identified data gaps and supporting participation in international data-sharing programs. Funds and specialized personnel for data mobilization in developed, often low-diversity countries may be better applied to support efforts in countries that lag behind due to lack of expertise or cyber-infrastructure (Ariño *et al.*, 2011), e.g., through direct partnerships or capacity building assistance.

The need to mobilize more data to increase completeness is obvious: 69-95% of the deviance in completeness explained by our minimum adequate models can also be explained by differences in record density (Table V.2.S4 a). However, we find that there is much room for improving the effectiveness of such mobilization. Theoretically, it would take 3.7M evenly sampled records to represent each known species of the three vertebrate groups once in every 110 km cell overlapping its range, and thus achieve 100% inventory completeness globally at that spatial grain. Currently, about forty-two times that many (157M) validated records represent only 21.6 % (0.8M) of these unique species-grid cell combinations, demonstrating a huge level of informational redundancy concentrated in a few places (Fig. II.2.4). Such intensive but localized sampling and data mobilization may benefit local conservation efforts as well as many purely scientific endeavors, but surely trades off against global-scale data needs, such that gaps in DAI are particularly severe in regions where higher-resolution datasets are most needed to support cost-effective progress towards multiple Aichi targets (Pereira *et al.*, 2010; Venter *et al.*, 2014). Strategic mobilization of data sources that likely contain many missing species-grid cell combinations could prove effective in quickly closing gaps and reducing geographical bias in global DAI. This in turn would facilitate robust, fine-

grain distribution data products from SDMs or downscaling models (Keil *et al.*, 2013) for a greater and geographically more representative sample of species than previously possible (Boitani *et al.*, 2011) and could immediately support various Aichi targets (Pereira *et al.*, 2013). Examples include land-use planning to minimize biodiversity loss (target 7), creating species lists for protected areas and improving global reserve networks (target 11), safeguarding threatened species (target 12) and mapping and securing associated ecosystem services (target 14). Targeting sufficiently recent data sources would furthermore create strong synergies with keeping conservation assessments up-to-date (Rondinini *et al.*, 2014). As a concrete example of potential conservation impacts, GBIF-facilitated records were recently used in the legal listing of five species of sawfish (Pristidae) under the US Endangered Species Act(Department of Commerce. National Oceanic and Atmospheric Administration, 2014). Increased access to occurrence information alone cannot ensure sound application nor conservation outcomes, but it can facilitate sound, data-driven decision-making (Guisan *et al.*, 2013), which in many parts of the world is currently impossible. We therefore argue that data mobilization efforts should be coordinated and strive to maximize return-on-investment for global conservation applicability.

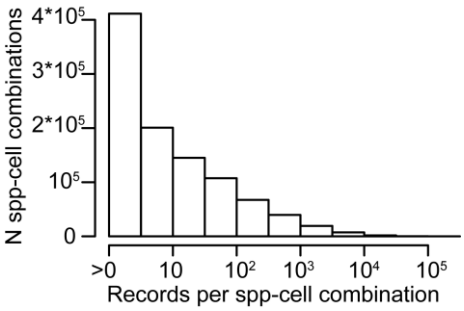


Figure II.2.4. Redundancy of information in 157M globally mobilized point records that constitute 'digital accessible information' of species distributions. The histogram shows the frequency of different degrees of information duplication (duplicated species-grid cell combinations) at the 110 km grain. Theoretically, and under ideal sampling, representing each of 3.7M species-grid cell combinations by one record would achieve 100% inventory completeness at that spatial grain.

Immediate opportunities for addressing gaps in DAI are most apparent at the national level: We find that even after controlling for all investigated factors (which explain 92.1–97.2% of cross-national variation), country identity still explains a significant portion of inventory completeness (2.4–7.1% of D^2; Table V.2.S4 b), pointing to an important role of country-specific political, legal, historical, linguistic or cultural factors (*Supplementary Information: Methods 1.D*). If countries were equally committed to providing access to their biodiversity information, as agreed upon by CBD signatories, completeness should be mainly limited by available financial resources. However, there is only a moderate relationship between country-level completeness and per capita gross domestic product (r^2=0.34, P<0.001; Fig. II.2.5 A, B) or total conservation spending (Waldron *et al.*, 2013; r^2=0.16, P<0.001). Notably, several large emerging economies including Brazil, China, India, Indonesia, Russia, or Turkey lag

behind (Fig. II.2.5 B, C, Table V.2.S6), which is worrying given increasing pressure on their biodiversity from rising global and domestic consumption (Naidoo & Adamowicz, 2001; Lenzen *et al.*, 2012). Success in building an adequate information basis for global biodiversity conservation and thus globally informed policies for environmental sustainability will depend on their support, and may be determined by political rather than economic factors. For example, despite the large mobilization needs due to its megadiverse biota, Mexico has a leading role in biodiversity informatics due to early political support for establishment of a national biodiversity program (CONABIO, 2012). Data-rich institutions in economically powerful countries like Brazil, China and Russia (Boakes *et al.*, 2010; Feeley & Silman, 2011b; Yang *et al.*, 2014), which together account for 31% of missing species-grid cell combinations (Fig. II.2.5 C, Table V.2.S6), seem particularly well-poised to contribute significantly to globally accessible species distribution information.

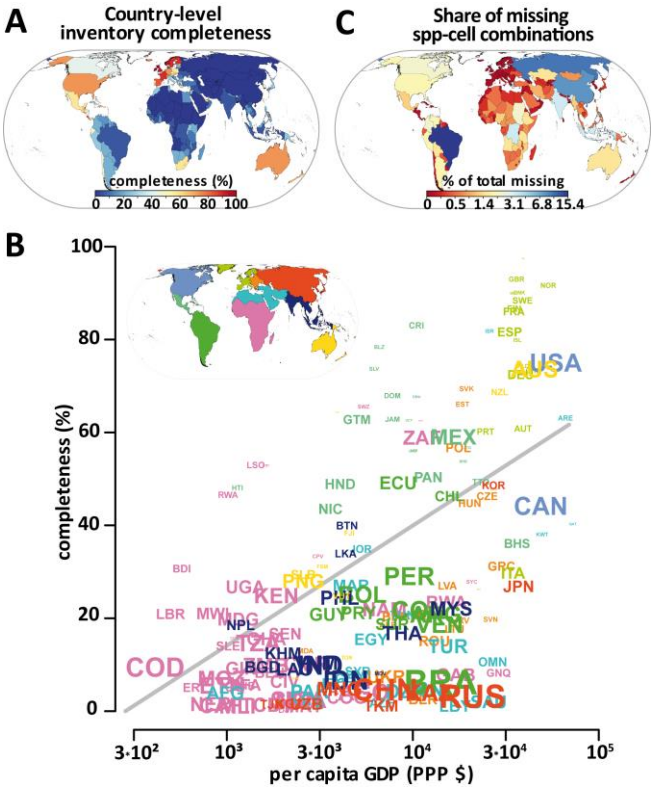


Figure II.2.5. Gaps in *digital accessible information* of biodiversity distributions at the country level. A) Country-level inventory completeness, measured as the percentage of the total unique species-grid cell combinations in each country that are covered by GBIF records. B) Country-level inventory completeness in relation to per capita gross domestic product (in purchase power parity dollars, PPP); r^2=0.34, P<0.001. Font size of country ISO codes is proportional to the total number of unique species-grid cell combinations that need to be recorded in each country to reach 100% inventory completeness at the 110 km grain. Font color is for geographical reference (compare inset map). Countries mentioned in the main text: BRA – Brazil, CHN – China, IDN – Indonesia, IND – India, MEX – Mexico, RUS – Russia, TUR – Turkey. C) Share that each country has in the unique species-grid cell combinations that are missing globally from a complete inventory at the 110 km grain.

As countries like Brazil recently announced intentions to improve and unlock their data store and existing national programs (e.g., speciesLink; http://splink.cria.org.br/) will be integrated into global DAI, information gaps and priorities may rapidly shift. More than current snapshots, tools for ongoing re-evaluation are therefore needed. The Map of Life project (Jetz *et al.*, 2012a) now provides such a tool (see http://patterns.mol.org/completeness) which may

help researchers to assess or account for data bias (Rocchini *et al.*, 2011) and to monitor progress in data mobilization (Tittensor *et al.*, 2014).

This global cross-taxon assessment represents a first in a number of steps required for more effective understanding and confrontation of information gaps on biodiversity distributions. While terrestrial vertebrates represent only c. 1.6% of described species (Costello *et al.*, 2013), addressing the factors that emerged as important across vertebrate taxa may hold the greatest promise for closing gaps for biodiversity in general. Vitally, and confirmed by the strong taxon-dependence of our results, assessments of distribution information need to be extended to more species-rich groups such as fishes, plants and invertebrates (see e.g. (Soberón *et al.*, 2007; Ballesteros-Mejia *et al.*, 2013; Sousa-Baena *et al.*, 2014a) for regional assessments)). Comparing ratios between mobilized record volumes and described species numbers suggests that gaps in DAI may be one to three orders of magnitude more severe in those groups (average records per species: tetrapods (31,032 spp.): 6,909; fishes (31,658 spp.): 347; vascular plants (283,701 spp.): 317; invertebrates (1.38M spp.): 31; numbers of geo-referenced records from GBIF website, June 2014, species numbers from Costello *et al.* (2013)).

The number of Aichi targets connected to species distributions indicates that they are a particularly essential biodiversity variable (Pereira *et al.*, 2013). In fact, accurate species distribution information is a prerequisite for more nuanced conservation strategies targeting critical or declining populations, or associated ecosystem services. However, such datasets require equally systematic assessments and prioritizations in order to effectively proceed towards Aichi target 19 (*Supplementary Information: Methods 1.D*).

Rapid biodiversity loss, limited funding, and potential trade-offs with direct conservation investments (Grantham *et al.*, 2008) require priorities for future collection and mobilization of biodiversity records into DAI. Our assessment highlights potential ways for making institution-based data mobilization more effective, but also the limitations of such efforts. Point records from museum specimens provide vital information but only represent one of a variety of data sources (Jetz *et al.*, 2012a) and their targeted mobilization should be complemented by other ways to address biodiversity information needs. Thorough biodiversity assessments led by trained field biologists will continue to play an important role in the creation of information about as yet un-surveyed, biodiverse areas, while novel biodiversity informatics infrastructure can facilitate more rapid integration of expert knowledge into DAI (Jetz *et al.*, 2012a). Citizen science projects are rapidly growing and poised to provide increasingly valuable records for certain taxa at comparatively low cost (Hochachka *et al.*, 2012). Improved reward systems (Enke *et al.*, 2012) as well as new data publishing mechanisms and journal requirements (Whitlock, 2011) can incentivize individual

data-holders to share biodiversity records. Novel SDM techniques (Dorazio, 2014; Fithian *et al.*, 2014; Manceur & Kühn, 2014) and downscaling approaches (Keil *et al.*, 2013) hold promise to overcome many of the typical data limitations. Further integration of available information and assessments of gaps, along with continued evaluation of effectiveness of DAI for conservation needs, are as vital as increased commitment to biodiversity data-sharing by political stakeholders, institutions, and individual scientists. With time running out to meet CBD targets on biodiversity knowledge, more effective data use and mobilization and a cultural shift about data-sharing are urgently needed.

Methods

Species distribution data

We overlaid expert-based extent-of-occurrence range maps for terrestrial birds (excluding pelagic feeders; N = 9,712), terrestrial mammals (N = 5,270), and amphibians (N = 6,188) with four nested equal-area grids (grain sizes: 110, 220, 440, 880 km) to infer coarse-resolution species richness patterns. As a representation of international efforts to collect, digitize, and share biodiversity records, we compiled a database of nearly 200M records for the three groups, aggregated by the Global Biodiversity Information Facility (GBIF). We focus on GBIF given that it is by far the largest such effort in geographic and taxonomic scope(Edwards, 2000; Graham *et al.*, 2004) and has an intergovernmental mandate to openly make accessible data from a worldwide base of data publishers. Data from GBIF represent the greatest body of existing DAI on species occurrences, based on centuries' worth of museum specimens, citizen science observations, surveys, literature and other sources. GBIF also has a vital role in sharing skills, software, tools, and best practices for biodiversity data use and mobilization. Thus, GBIF-facilitated DAI is currently the best available indicator of "shared biodiversity knowledge, science base and technologies" as referred to by Aichi target 19 (Tittensor *et al.*, 2014). To link GBIF-facilitated records with range maps, extensive taxonomic standardization was necessary (our approach as well as various filtering and validation steps are explained in the *SI V.2 Methods*). We defined inventory completeness as the percentage of expert-opinion species richness documented by mobilized records. We note that other DAI sources play vital and often complementary roles in progressing towards Aichi targets (*Supplementary Information: Methods 1.D*). Yet, other datasets may not be shared but nevertheless influence regional research and conservation. Thus results here should not be interpreted as definite maps of knowledge gaps, but the analyses of drivers are likely indicative of factors limiting biodiversity information in other data sources.

Geographic and socio-economic drivers of gaps in DAI

We analyzed relationships of twelve geographic and socio-economic factors with record density and inventory completeness. We used three variables to describe the appeal of areas to attract collectors: i) Endemism richness (Kier & Barthlott, 2001), i.e., the sum of inverse range sizes of all species present in a grid cell, was calculated from the number of 110 km cells. ii) To model effects of mountains on record collection, we calculated the topographic range in each cell based on a digital elevation model. iii) We modeled the effects of protected areas using proportions of land area in grid cells that fall within protected areas of International Union for Conservation of Nature categories I-IV. We investigated three aspects of accessibility: i) To test for effects of on-ground accessibility, we used a dataset on the time needed to travel to cities with a population >50,000 (Nelson, 2008). ii) To model effects of the proximity to airports, we created an index based on the locations of >9,300 airports and airfields (Partow, 2003). iii) 'Proximity to institutions' was expressed as weighted geographic proximity of a grid cell to those data publishers that contributed records for the area surrounding the cell. Index values are high if the majority of records are contributed by geographically close data publishers. We modeled effects of secure conditions using the Global Peace Index (Institute for Economics and Peace, 2012), which aggregates information on political stability, armed conflicts and levels of public safety. We investigated two aspects of international scientific integration: i) To quantify integration into 'scientific activities', we extracted the H-index for every country based on peer-reviewed papers published in the field 'Ecology, Evolution, Behavior and Systematics' from Elsevier's Scopus database (covering the years 1996-2011), and multiplied it with the proportion of papers resulting from international collaborations (see *Supplementary Information: Methods*). ii) We tested for effects of political cooperation with data-sharing networks using the proportion of the land area within each grid cell that falls within GBIF-participating countries. We used three measures of financial and institutional resources: We estimated financial resources that are potentially available for biodiversity research from per capita gross domestic expenditure on research and development i) within grid cell-overlaying countries ('National research funding') as well as ii) in countries where the publishers of records for a particular cell are situated ('Research funding of institutions'). iii) We used record volumes contributed to GBIF by different data publishers to estimate institution size. Details on calculation and transformation of predictor variables, along with detailed information on the respective hypotheses and the limitations of our data sources are in *SI V.2 Methods*.

Statistical methods

We investigated effects of predictor variables on inventory completeness separately for amphibians, birds and mammals at each of the four spatial grains with simple and multiple regressions. Specifically, we used non-spatial and spatial generalized linear models with a binomial distribution, where completeness enters as a composite variable ('species covered by records', 'species not covered but presumed present') and where differences in species richness are automatically accounted for. Spatial models account for residual spatial autocorrelation by including a 'residuals autocovariate' build from residuals of the non-spatial model and an optimized spatial neighborhood structure (Crase *et al.*, 2012). Because of long computation times for spatial models, we ran all possible non-spatial models, and re-ran those model subsets that would likely be among the minimum adequate spatial models (with ΔAIC <10 to the lowest AIC score) as spatial models. We assessed model fits of minimum adequate spatial models as the % deviance explained (D^2; Table V.2.S3). We investigated interactions among variables as well as non-linear effects, but - although many were significant - accounting for them did not greatly alter model fit or parameter estimates of main effects in preliminary analyses. To maintain as much simplicity as possible given twelve predictor variables and twelve separate sets of models (3 taxa x 4 spatial grains), we decided to focus on the main effects. We used standardized coefficients (β) of minimum adequate spatial models (with the lowest AIC scores) to measure the relative importance of predictor variables. As an alternative measure, we used percentages of the sums of squares attributable to each factor, based on ANOVAs with a response variable consisting of the AIC scores of all possible models and predictor variables coding the presence/absence of each predictor in the respective model. For further details and references see *SI V.2 Methods*.

Acknowledgements. We thank those active in collecting, sharing, curating, digitizing and mobilizing species distribution data. We thank Jeremy Malczyk, Javier Otegui, Tim Robertson, Gaurav Vaidya and Patrick Weigelt for help with data assembly and handlings, and Carsten Dormann for advice on statistical methods.

Chapter 3

Global drivers of species variation in mobilized occurrence information

Carsten Meyer, Walter Jetz, Robert P. Guralnick, Susanne A. Fritz and Holger Kreft

Updated version re-submitted to *Global Ecology and Biogeography* (after invitation for re-submission). Preprint archived in *PeerJ PrePrints* **3**:e1493. DOI: 10.7287/peerj.preprints.1326v2.

Abstract

Despite the central role of species distributions in ecology and conservation, occurrence information remains geographically and taxonomically incomplete and biased. Numerous socio-economic and ecological drivers of uneven record collection and mobilization have been suggested, but the generality of their effects remains untested. We develop scale-independent metrics of range coverage and geographical record bias and apply them to 2.8M point-occurrence records of 3,625 mammal species to test 13 putative constraints on data availability. We find that data limitations can be linked to species attributes related to detection and collection probabilities, such as body size, diurnality, or description date. However, species attributes are much weaker predictors of the amount and range coverage of available records than range size and shape, and the geography of socio-economic conditions. Our results highlight the need to prioritize range-restricted species and to address the key socio-economic drivers of data bias in ecological modeling and data mobilization efforts.

Introduction

Detailed information on species distributions is fundamental to basic and applied ecology (Whittaker *et al.*, 2005; Boitani *et al.*, 2011). Expert range maps have become a key basis for many large-scale analyses, but they incur high errors of commission toward finer spatial scales and their accuracy varies with species-level ecological and range attributes (Jetz *et al.*, 2008). Moreover, range maps exist only for few groups of organisms. This makes point occurrence records a critical resource for developing distribution datasets for more taxonomic groups and at relevant spatial scales (Jetz *et al.*, 2012a). Large amounts of digital occurrence records from field observations, museum specimens and other sources have been mobilized via international data-sharing networks, most notably that of the Global Biodiversity Information Facility (GBIF; Edwards, 2000). While such records represent vital fine-scale information on spatial and temporal occurrences of species, severe gaps and biases hamper broader application (Rocchini *et al.*, 2011). These data limitations have been mostly studied with a focus on geographical assemblages (Soria-Auza & Kessler, 2008; Meyer *et al.*, 2015), whereas differences among species received less attention (Cayuela *et al.*, 2009).

Bias towards species with certain (bio-)geographical, phylogenetic or ecological attributes can lead to biased ecological inference (Garamszegi & Møller, 2011) and inefficient conservation

(Gonzalez-Suarez *et al.*, 2012). For instance, in comparative studies, species-level bias violates the statistical assumptions that missing species occur at random across the entire range of relevant dimensions and that data quality (i.e., occurrence information) is constant across observations (Garamszegi & Møller, 2011). A better understanding of species-level variation in occurrence information is crucial for effectively closing information gaps and for developing robust ecological models that can differentiate between true absences of species and missing information (Iknayan *et al.*, 2013; Dorazio, 2014). While the reliability of range maps in relation to range size and species' ecology has been assessed (Jetz *et al.*, 2008), patterns and drivers of species-level variation in point occurrence information remain largely un-investigated.

Species-level variation and bias in point occurrence information arise from at least three different characteristics of available occurrence records: i) *record count* per species, the most commonly studied and perhaps most intuitive metric (Cayuela *et al.*, 2009; Burton, 2012), ii) *range coverage*, i.e. the degree to which records document a species throughout its entire range, and iii) *geographical bias*, i.e. the non-randomness in records' representation of different range parts. Depending on the research question at hand, bias in these three aspects of occurrence information can have different ramifications. For instance, species distribution models do not necessarily require high range coverage as long as a minimum number of environmentally unbiased records is available (Varela *et al.*, 2014). In contrast, protected area gap analyses require high coverage of species ranges, whereas *geographical bias* is less important.

Many possible drivers of species-level variation in occurrence records have been suggested. An often-cited, but rarely tested cause for species-level variation may be that species attributes affect detection and collection probabilities. For instance, more records might be available for species that are easily detected due to higher abundances (Dorazio, 2007), or because they possess specific traits that make them more conspicuous, such as terrestrial foraging behavior or diurnal activity (Iknayan *et al.*, 2013). Further, more records might have accumulated for early-described species as well as for species that attract more scientific or public interest, or for which records are logistically, legally or ethically easier to collect and share (Amori & Gippoliti, 2000; Whitlock *et al.*, 2010).

Besides species attributes, geographical factors could constrain occurrence information. First, range geometry, i.e. the size and shape of a range, might affect the likelihood that a given range part is close or distant to a given record. Second, socio-economic factors, such as area appeal, proximity to research institutions, cooperation with data-sharing networks, and financial resources may limit occurrence information by affecting the likelihood that records from within a given range are collected, digitized and shared (Meyer *et al.*, 2015). While all

above-mentioned factors might drive species-level variation in *record count* and *range coverage*, within-range *geographical bias* of records should be driven by range size and within-range variation in socio-economic factors (see Box II.3.1).

Here, we provide the first analysis of global patterns and drivers of species-level variation in point occurrence information. We integrated c. 2.8 M geographically and taxonomically validated records mobilized via GBIF for 3,625 terrestrial mammal species (c. 72% of all extant species) with their expert-opinion range maps. To this end, we developed scale-independent metrics for *range coverage* and *geographical bias*. We first explored relationships among the three different aspects of occurrence information – *record count*, *range coverage* and *geographical bias* – while accounting for range geometry. We expected *range coverage* to increase with *record count* and to decrease with *geographical bias*, range size and range shape irregularity. We then tested three major classes of hypotheses about constraints on *record count* and *range coverage*, namely species attributes, range geometry, and socio-economic factors (represented by 13 variables; Box II.3.1). Additionally, we tested whether range size and within-range variations in socio-economic factors drive *geographical bias*. We assessed the relative importance of variables at the global scale and additionally at the scale of zoogeographical realms. Our work provides the first global assessment of species-level variation in different aspects of mammalian occurrence information, and the first comparison of the relative effects of species-specific, geometric and socioeconomic factors.

Methods

Measuring occurrence information

We overlaid 4,524,585 point occurrence records mobilized via GBIF (retrieved Oct 2012) with extent-of-occurrence range maps of 5,057 species of terrestrial mammals (IUCN, 2010). Occurrence records provide direct evidence that a particular species occurred at a particular geographical point at a specific point in time. In contrast, range maps delimit the geographical distribution of known and assumed species occurrences, based on expert interpretation of different distribution data types (Jetz *et al.*, 2012a). Range maps overestimate distributions at fine scales, but typically provide a less biased view of distributions than occurrence records and can serve as geographical reference of likely distributions at coarse scales. We matched taxonomies between records and range maps and used range map overlays to validate records geographically (see *Supporting Information* (*SI*) *V.3.1.1*). The final, rigorously cleaned dataset contained 2,849,058 records for 3,625 species.

Box II.3.1 Putative drivers of species-level variation in occurrence information

Species-level variation in occurrence information (*record count*, *range coverage*, *geographical bias*) may be driven by species attributes, range geometry and socio-economic factors. For each of these groups of hypotheses, we first provide a brief rationale for including individual factors and then summarize their hypothesized effects.

Species attributes:

Certain species attributes may drive *record count* and *range coverage* because they positively affect species' detectability, popularity or sampling logistics.

i) Diurnality: Predominantly diurnal species are more likely to be detected (Burton, 2012).

ii) Body size: Despite the often-cited conspicuousness and appeal of large-bodied species (Knight, 2008; Brooke *et al.*, 2014), their lower abundances (Robinson & Redford, 1986), and greater sensitivity to disturbance (Blumstein, 2006) lead to lower detectability. Furthermore, larger specimens are logistically more difficult to transport and store.

iii) Foraging stratum: Terrestrial species are more easily detected than arboreal species with standard sampling techniques (Chutipong *et al.*, 2014).

iv) Dietary level: Higher dietary levels (i.e., specialization on high-energy but low-abundance resources) are associated with lower abundances (Robinson & Redford, 1986) and larger home ranges (Tucker *et al.*, 2014), resulting in lower detectability.

v) Years since description: Early-described species have had more time to accumulate records.

vi) Public interest: It is more appealing and easier to attract funding for sampling and data mobilization of species for which there is great public interest due to commercial, medicinal, aesthetic, psychological or cultural reasons (Knight, 2008; Perry, 2010; Tyler *et al.*, 2012).

vii) Threat status: Despite higher interest for threatened species (Tyler *et al.*, 2012), their often lower abundances and smaller ranges lead to lower detectability (Dorazio, 2007) and their threat status prohibits specimen collection. Records of threatened species are less often shared to prevent exposing exact occurrences to the public (Whitlock *et al.*, 2010).

We hypothesized *record count* and *range coverage* to be positively affected by diurnality, time since description and public interest, and negatively by body size, foraging stratum, dietary level, and threat status. We did not expect these factors to influence within-range *geographical bias.*

Range geometry:

Under geographically non-random sampling, range geometry is expected to affect the likelihood of ranges intersecting records.

viii) Range Size: We expected clusters of sampling locations interspersed with areas of lower record availability. Unless records are perfectly clumped, large ranges are bound to intersect with more clusters of sampling locations. Under this scenario, species with larger ranges are more likely to have higher *record counts* and, when controlling for *record count*, lower *geographical bias* in the representation of different range parts. Conversely, larger range sizes are increasingly less likely to achieve high *range coverage*.

ix) Range shape irregularity: The same natural constraints that cause non-uniform dispersal and elongated ranges, like rivers, coast lines and mountain ranges (Pigot *et al.*, 2010), have historically determined human transportation routes (Rodrigue *et al.*, 2006). Hence, *record counts* should be higher for elongated ranges, because researchers' study areas and species' ranges are more likely to intersect *Range coverage*, however, should be lower for more elongated or fragmented ranges, as random points in such ranges would be increasingly less likely to be close to a given record.

We hypothesized that both range size and range shape irregularity positively affect *record count*, and negatively affect *range coverage*. We further hypothesize that when controlling for *record count*, *geographical bias* is negatively correlated with range size.

Socio-economic factors:

We considered four socio-economic factors that are particularly important for mammalian assemblage-level occurrence information (Meyer *et al.*, 2015).

x) Area appeal: Biologists prefer to work in areas with many rare or range-restricted species (Soria-Auza & Kessler, 2008).

xi) Proximity to research institutions: Species close to researchers' home institutions are more likely to be well-sampled, due to easier logistics of carrying out multiple field surveys at different sites. Areas remote from research institutions are visited more occasionally, making it likely that rare species evade detection (Dennis & Thomas, 2000).

xii) GBIF participation: Participation in international data-sharing networks enhances data mobilization (Yesson *et al.*, 2007) .

xiii) Financial resources: Financial resources for data collection and mobilization, associated with research or conservation programs, limit record availability for species in a given country (Soberón & Peterson, 2004).

We hypothesized *record count* and *range coverage* to be positively influenced by favorable socio-economic conditions averaged within ranges, and *geographical bias* to be positively related to within-range variation in these factors.

In addition to simple *record count*, we then used these two data types to develop two response metrics for occurrence information: '*range coverage*' and '*geographical bias*'. *Range coverage* describes the detail with which a species' range is documented by available records. *Geographical bias*, in contrast, describes the level of non-randomness with which records represent different range parts. Both metrics are based on the great-circle distance (in km) of every one of 1000 random points placed across the range map to its geographically closest occurrence record (i.e., the record 'covering' that range part). Parts of ranges with random points close to their nearest records can be considered 'well-covered' (Fig. II.3.1, Fig. V.3.S1).

Range coverage. *Range coverage* is the negative mean minimum distance (MMD) between 1000 random points and *n* available records, such that less negative values corresponded to higher *range coverage*:

$$Range\ coverage = -\text{MMD} = -\frac{1}{1000}\sum_{i=1}^{1000} MinDistRPi,$$

where $MinDistRP_i$ is the minimum distance of the *i*-th random point to its nearest record (Fig. II.3.1 for examples; Fig. V.3.S1).

Geographical bias. To quantify *geographical bias* in records' representation of different range parts, we related the MMD to a null model of the potential MMD under random sampling. We randomly placed *n* (number of actually available records) 'pseudo records' across the range 1000 times, and calculated MMD each time. *Geographical bias* is then the standardized effect size, calculated as the difference between observed MMD and null model mean divided by the null model standard deviation:

$$Geographical\ bias = \frac{\text{MMD}Observed - \text{mean}\ (\text{MMD}NullModel)}{\text{sd}\ (\text{MMD}NullModel)}.$$

Higher *geographical bias* scores result if sampling locations are highly clumped and concentrated in one range part, as well as from high levels of information duplication, e.g. large *record counts* from exactly the same sampling locations (Fig. II.3.1, Fig. V.3.S1). The large number of random points ensures that even large ranges are appropriately represented and that commission errors due to range map inaccuracies do not greatly affect *range coverage* and *geographical bias* metrics.

Predictors of occurrence information

We focus here on predictors of *record count* and *range coverage* but provide further details in the *SI* on models of *geographical bias* (*SI V.3.1.4)*, as well as on models of whether species have any mobilized records (*SI V.3.1.3*). We tested three major classes of hypotheses related

to species attributes, range geometry, and socio-economic factors, which were represented by 13 variables as potential drivers of *record count* and *range coverage* (Box II.3.1; see *SI V.3.1.7* for information on omitted variables):

Species attributes: i) We estimated diurnality by assigning the activity period of each species on an ordinal scale based on data in Wilman *et al.* (2014): 1=nocturnal only; 2=nocturnal and crepuscular; 3=crepuscular only (active only around dusk/dawn); 4=nocturnal, crepuscular and diurnal; 5=crepuscular and diurnal; 6=diurnal only. Data on **ii)** adult body mass (in g) and **iii)** dietary level was also taken from Wilman *et al.* (2014). For the latter, we first grouped ten diet categories into an ordinal scale: 1=low-nutrition/high-abundance plant matter (e.g. leaves, wood); 2=high-nutrition/low-abundance plant matter (e.g. fruits, seeds, nectar); 3=animal matter (e.g. vertebrates, invertebrates and carrion). We then calculated weighted averages of dietary level scores, such that an omnivore with a diet composed of 25% leaves, 25% fruit and 50% invertebrates was assigned a score of 2.25. We assigned categorical data from Wilman *et al.* (2014) on **iv)** main foraging stratum on an ordinal scale: 1=terrestrial (including bats that forage close to the water surface); 2=scansorial (climbing); 3=arboreal; 4=aerial. We calculated **v)** time since description (in years until 2014) from dates in species author information (IUCN, 2010). **vi)** public interest for species was estimated based on the prominence of species names in internet activity, represented by numbers of Google hits for verbatim scientific names (as of November 2013). As an estimate of **vii)** threat status, we assigned threat categories from the International Union for the Conservation of Nature's Red List (IUCN, 2010) on an ordinal scale: 1=LC, 2=NT, 3=VU, 4=EN, 5=CR, 6=EW.

Range geometry: To model effects of **viii)** range size, we used the area of the original expert range map polygons (in km²). Because existing methods to quantify range shape are either grain-size dependent or only focus on specific shape aspects (usually elongation; compare Pigot *et al.* (2010)), we developed a new metric of **ix)** range shape irregularity: the ratio of the mean distance between 1000 random points within the range to the mean distance between 1000 random points within a circle of the same area (see Fig. II.3.1 for examples). Ratios increase from 1 (perfect circle) as shapes become more elongated or fragmented.

Socio-economic factors: To estimate **x)** area appeal to researchers, we calculated the mean mammalian endemism richness score across range map-overlapping 110-km grid cells. Endemism richness is the sum of inverse range sizes of all species present in a cell (Kier & Barthlott, 2001). To calculate the **xi)** proximity of a species' range to research institutions, we first identified institutions that could have potentially contributed records for that species because they have performed surveys in range-overlapping countries (inferred from sampling locations of all their contributed mammal records). Proximity to institutions was then the mean inverse great circle distance of 100 random points placed across that species' range to

those institutions, weighted by the institutions' relative contribution to all mammal records in range-overlapping countries:

$$10^8 * \sum_{i=0}^{n} \left(\frac{\text{RelProp}i}{\text{D}i}\right),$$

where RelProp_i is the relative contribution of the *i*-th publisher to records from the range-overlapping countries and D_i its distance (in km) to the random point. We calculated **xii)** GBIF participation of range-overlapping countries as the proportion of a species' range that falls within GBIF-participating countries (as of 2012). We estimated **xiii)** locally available financial resources from conservation funding data (Waldron *et al.*, 2013). Large, species-rich countries require more resources to attain high *coverage* for all species (Meyer *et al.*, 2015). We therefore first divided country-level conservation funds by the country's total area of mammal ranges to calculate a country's available resources per species range size to-be-covered (in million USD/10,000 km² range size). For each species, we then calculated the mean available resources across all range-overlapping countries, weighted by relative overlap.

Statistical modeling

First, we modeled effects of *record count*, *geographical bias*, and range geometry (size and shape) on *range coverage*. Then, we used species attributes, range geometry and socio-economic factors to model *record count* and *range coverage.* Finally, we modeled effects of range size and within-range variation in socio-economic factors on *geographical bias*. We modeled *record count* using generalized linear models (GLM) with a quasi-Poisson distribution to account for over-dispersion (O'Hara & Kotze, 2010). We modeled *range coverage* and *geographical bias* with ordinary least squares models (OLS).

Preliminary tests for taxonomic bias yielded strong effects of species' order memberships on *record count*, *range coverage* and *geographical bias* (also weaker effects of family memberships; memberships following IUCN (2010); see *SI V.3.1.2*, Table V.3.S2 A). We therefore included 'mammal order' as a covariate in all models. We inspected model residuals for normality and autocorrelation, using global Moran's I for spatial autocorrelation (Dormann *et al.*, 2007), and a phylogenetic adaptation of Moran's I for phylogenetic autocorrelation (Abouheif, 1999) based on the phylogeny in Fritz *et al.* (2009). These tests revealed that further accounting for phylogenetic or spatial non-independence was not necessary (Fig. V.3.S3, for details see *SI V.3.1.6*). We used multi-model inference (Burnham & Anderson., 2002) to assess model support and relative importance of predictor variables by running all possible model subsets and performing model selection based on Akaike's Information Criterion (AIC) for OLS and quasi-AIC for GLM. After assessing the relative support of all predictor variables, we calculated fractions of total explained variation in *record count* and

range coverage attributable uniquely and jointly to the three major hypotheses using variation partitioning based on the respective minimum adequate models (Peres-Neto *et al.*, 2006).

We $\log_{10}$-transformed and z-transformed continuous predictor and response variables to improve linearity and to obtain standardized coefficients. We used negative $\log_{10}$-transformed MMDs to model *range coverage*, such that variables causing high *range coverage* yield positive effects. We limited collinearity by only including variables with generalized variance inflation factors ≤10 (Dormann *et al.*, 2013; Table V.3.S4-5). We modeled *record count*, *range coverage* and *geographical bias* at the global scale and separately for each of six zoogeographical realms (Olson *et al.*, 2001). We assigned species to realm-scale models if their ranges overlapped the realm by >70%.

All analyses were performed in R 2.15.2–3.1.2 (R Core Team, 2014).

Results

Patterns in occurrence information

3,625 or 72% of the 5,057 mammal species considered had at least one validated record (see Fig. V.3.S2, *SI V.3.1.3* for models of whether species have any records). Among these, *record count* varied by five, *range coverage* and *geographical bias* by four orders of magnitude, respectively (Fig. II.3.2 A-C, Table V.3.S1). Globally, the mean *record count* per species was 563 (SD=3,073, median=13, Table V.3.S1). *Range coverage* averaged -205.5 km across species (SD=375.5, median=-199).

For all three aspects of occurrence information, we observed significant variation between higher taxonomic levels (Table V.3.S1, *ANOVA* results in Table V.3.S2 A). Among the more speciose mammal orders, primates stood out for below-average *record counts*, and carnivores for below-average *range coverage* scores. High *record counts* and *range coverage* scores characterized Australasian marsupials (Fig. II.3.2 D-E), which also had above-average *geographical bias* scores (Fig. II.3.2 F, Table V.3.S1). Phylogenetic and spatial autocorrelation analyses attributed this taxonomic bias in occurrence information mainly to a better representation of species living in certain regions, rather than to a strong phylogenetic component (*SI V.3.1.6*, Fig. V.3.S3).

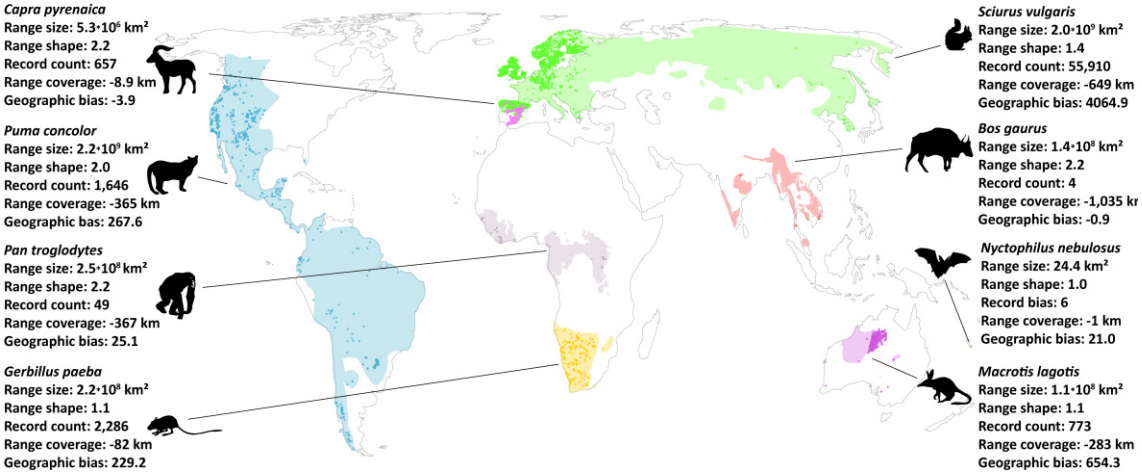


Figure II.3.1. Range geometry and occurrence information for eight selected mammal species. Pale colours denote extent-of-occurrence range maps (IUCN, 2010), brightly colored dots indicate locations of GBIF-facilitated occurrence records. Examples show global variation in *record count*, *range coverage* by those records, and *geographical bias* in how records represent different range parts. Comparing *Puma concolor* with *Sciurus vulgaris* demonstrates how substantially fewer records can *cover* a larger and more irregularly-shaped range better, if less *geographically biased*. The negative *geographical bias* score for *Capra pyrenaica* indicates more even *coverage* than under random sampling. The New Caledonian bat *Nyctophilus nebulosus* is highly range-restricted; therefore six records suffice for extremely high *coverage*. In contrast, random points within the range of *Bos gaurus* are on average 1,035 km from the closest one of just four mobilized records. See *Materials and Methods* for further

explanations.

Accordingly, occurrence information differed more strongly among geographical realms (Fig. II.3.2 G-I) than among mammal orders (Fig. II.3.2 D-F, Table V.3.S1, and *ANOVA* results: *SI V.3.1.2*, Table V.3.S2 B). The Nearctic, northern Neotropical, western and northern Palaearctic and Australasian realms had mostly species with above-average *record counts*, whereas Madagascar and the south-eastern Palaearctic and Indomalayan realms had mostly below-average species (Fig. II.3.2G). High *record counts* often coincided with high *geographical bias* and *range coverage* scores. However, high *record counts* did not coincide with high *range coverage* in the Palaearctic realm, where records were extremely biased towards Europe and therefore covered most species' ranges only poorly (Fig. II.3.2 G-I). Species without GBIF records had highest concentrations in Southern China and South-East Asia (Fig. V.3.S2).

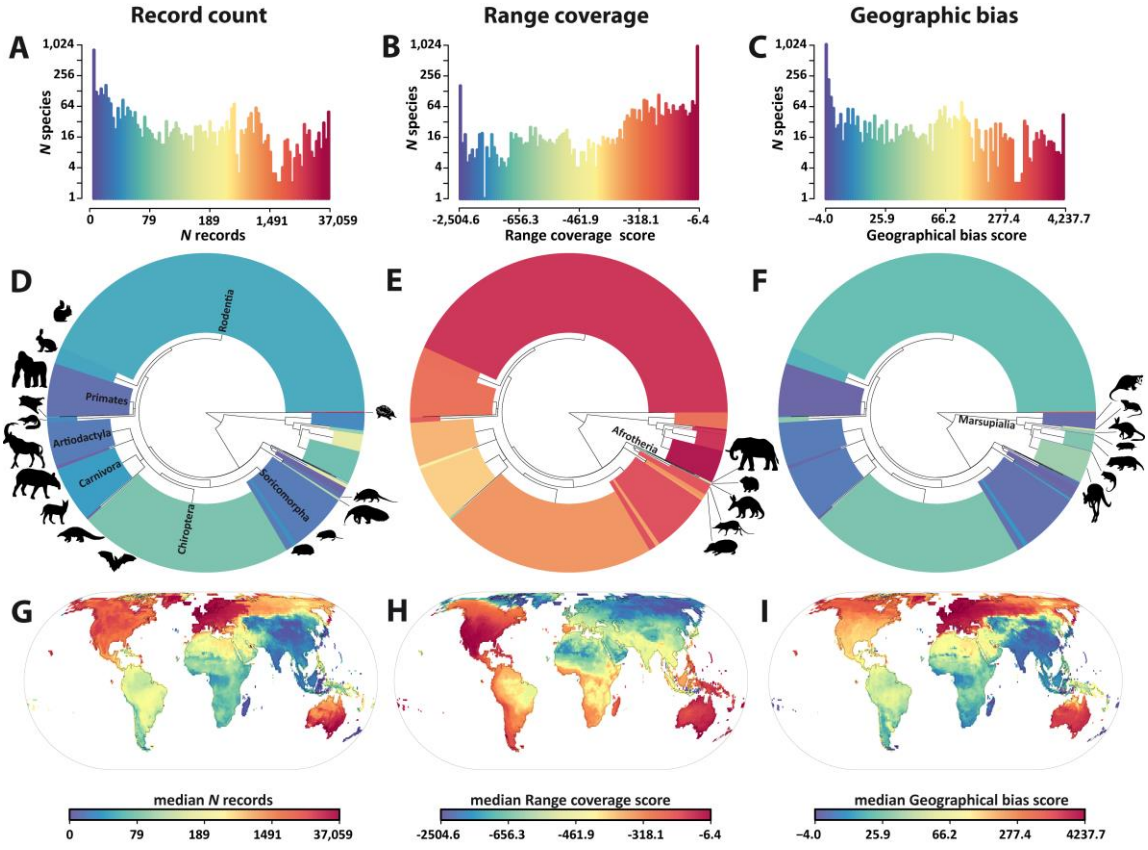


Figure II.3.2. Species-level variation in *record count*, *range coverage* and *geographical bias* for 3,625 mammal species. Shwon are frequencies of scores of A) *record count*, B) *range coverage* and C) *geographical bias* across global mammal species, median scores for each mammal order (D-F) and for each 110 km x 110 km grid cell (G-I). Phylograms in D-F based on Fritz *et al.* (2009). Colored areas of mammal orders have widths proportional to their species number. Labels in D) highlight the six most speciose orders. Silhouettes are for visual orientation, those for Afrotheria and Marsupialia shown in E-F because of limited space. Occurrence information metrics are calculated across the entire range of a species. Consequently, values for a particular grid cell in G-I show what occurrence information is available for the species occurring there, not for the specific region. Color scales are calibrated on grid-cell percentiles and identical in A/D/G, B/E/H, and C/F/I. Most species have few records (mostly cooler colors in D), yet most species also have relatively small ranges which often have higher *range coverage* scores (warmer colors in E).

Range coverage was strongly positively correlated with *record count*, negatively with *geographical bias* and furthermore strongly constrained by range geometry (Fig. II.3.3, Table V.3.S3). These effects appeared general across global and realm-scale models (Fig. II.3.3) and together accounted for 73-89% of inter-specific variation in *range coverage* (Table V.3.S3). Furthermore, *record count* was strongly positively correlated with *geographical bias* (r_S=0.62, P<0.001).

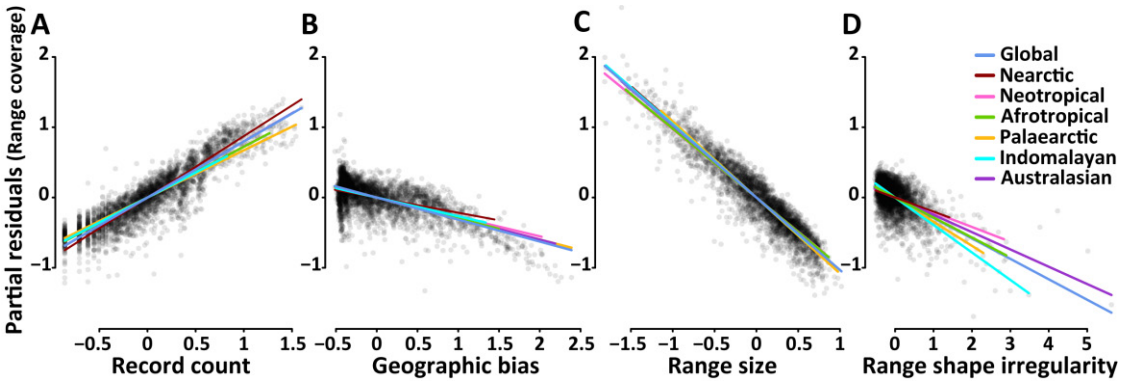


Figure II.3.3. Effects of *record count*, *geographical bias* and range geometry on *range coverage*. Partial residuals show effects of A) *record count*, B) *geographical bias*, C) range size and D) range shape irregularity on *range coverage* while controlling for all other variables in the global model. Partial residuals refer to the relationship at the global scale. Partial fits of global and realm-scale models are indicated by different colors. All variables were $\log_{10}$-transformed and z-transformed (see Table V.3.S3 for details).

Predictors of occurrence information

Record count and *range coverage* were well-predicted by a combination of species attributes, range geometry and socio-economic factors, which explained 44-86% of the variation depending on geographical focus (Fig. II.3.4). All 13 predictor variables showed at least weak effects in some of the models (Fig. II.3.4, Table V.3.S4). Numbers of variables retained in minimum adequate models varied between 5 (*record count* and *range coverage* in the Palaearctic model) and 12 (*range coverage* in the global model). Also, the variation in species-level *geographical bias* explained by range size and within-range variation in socio-economic factors varied substantially with geographical focus (Fig. II.3.5, Table V.3.S5, *SI V.3.1.4*) and most variation in *geographical bias* could be explained by the models in zoogeographical realms with large numbers of mobilized records (partial R^2_{adj}: Nearctic: 0.24, Palaearctic: 0.24, Australasian: 0.44).

Species attributes overall showed only weak effects on *record count* and *range coverage* (Fig. II.3.4, Tables V.3.S3). Body mass and time since description showed relatively consistent negative and positive effects, respectively, across global and realm-level models. Positive effects of public interest emerged as relatively important based on sums of QAIC/AIC

weights. Threat status, diurnality, dietary level and foraging stratum showed inconsistent effects. Strong effects for these factors only emerged in the Afrotropical, Australasian, Neotropical, and global and Neotropical models, respectively (Fig. II.3.4).

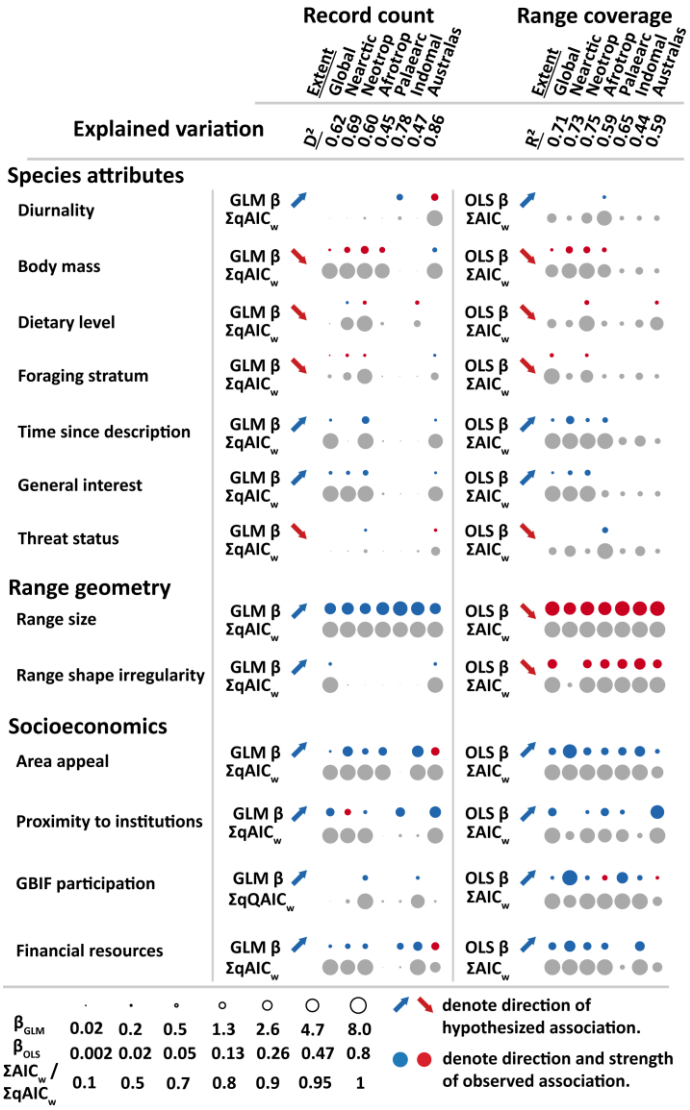


Figure II.3.4. Results from global and regional models of *record count* and *range coverage* for 3,625 mammal species. Effects on *record count* were tested in multiple generalized linear regression models with a quasi-Poisson distribution, those on *range coverage* in multiple ordinary least squares models. All possible model subsets were ranked based on QAIC/AIC scores; results are shown for the minimum adequate model. Arrow and bubble color denotes direction of expected and observed predictor-response relationships, respectively. Bubble size represents relative importance of variables, assessed by two different metrics: i) standardized coefficients of the minimum adequate models (GLM β and OLS β), and ii) sums of QAIC/AIC weights ($\sum QAIC_w$ and $\sum AIC_w$) across all model subsets. Partial adjusted deviance explained (D^2) and partial adjusted variance explained (R^2) have effects of the covariate 'mammal order' removed (Peres-Neto *et al.*, 2006). For details on hypotheses, methods, and results, see Box II.3.1, *Supporting Information*, Table V.3.S4.

Range geometry showed very strong effects on occurrence information. Range size consistently emerged as an important factor, with strong positive effects on *record count* and negative effects on *range coverage* (Fig. II.3.4) and *geographical bias* in the global and Neotropical models (Fig. II.3.5). Range size alone explained 7-38% of the variation in *record counts*, and 26-64% in *range coverage* (inferred from simple regressions). Range shape irregularity was an important constraint of *range coverage*, but only had minor positive effects on *record count* in the global and Australasian models (Fig. II.3.4, Table V.3.S4).

Socio-economic factors showed strong positive effects, particularly for *range coverage*, both from sums of QAIC/AIC weights and standardized coefficients (Fig. II.3.4). However, the strength of effects differed substantially between global and realm-scale models; and some noteworthy discrepancies emerged between effects on *record count* and *range coverage*. For instance, in the Nearctic and Palaearctic realms, GBIF participation greatly limited *range coverage* but not *record count*. Some significant negative effects emerged: for *record count* those of area appeal and financial resources in the Australasian and those of proximity to institutions in the Palaearctic, and for *range coverage* those of GBIF participation in the Afrotropical model. Relatively strong positive effects on *geographical bias* emerged for within-range variation in proximity to institutions in the Palaearctic and Australasian, GBIF participation in the Palaearctic, and available financial resources in the Neotropics (Fig. II.3.5).

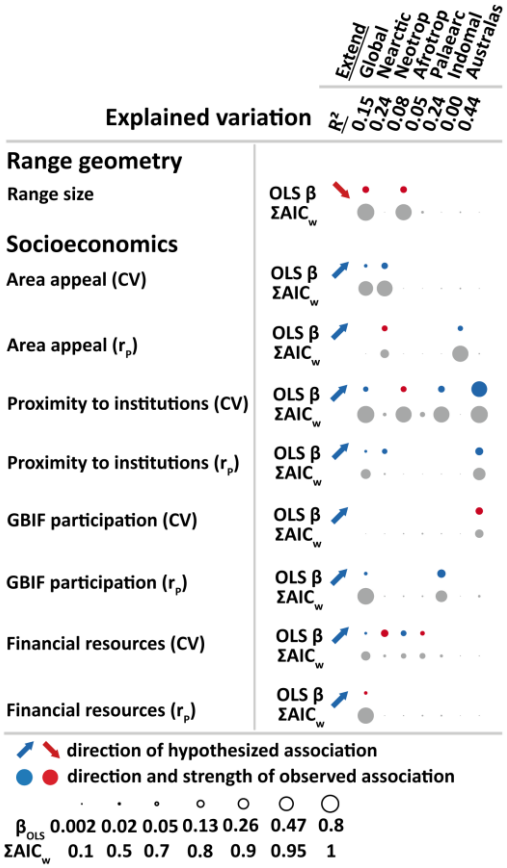


Figure II.3.5. Results from global and regional models of *geographic bias* for 3,625 mammal species. Effects were tested in multiple ordinary least squares models. All possible model subsets were ranked based on AIC scores; results are shown for the minimum adequate model (with AIC=0). Two metrics (*cv* and r_P) of within-range geographical variation were used (details in *SI 1.4*). Arrows, and bubble colors and sizes are as in Fig. 4; relative importance of predictors assessed by two metrics: i) standardized coefficients of minimum adequate models (OLS β), and ii) sums of AIC weights ($\sum AIC_w$) across all model subsets. Because *geographical bias* is highly correlated with *record count*, effects were tested with $\log_{10}$-transformed *record count* and mammal order included as fixed covariates. Partial adjusted variance explained (R^2) has effects of covariates 'mammal order' and '*record count*' removed (Peres-Neto *et al.*, 2006).

Variation partitioning confirmed that more variation in *record count* and *range coverage* was uniquely explained by geometry and socio-economic factors than by species attributes (except for *record count* in the Neotropical model, Fig. V.3.S4). The bulk of modeled variation in *record count* and *range coverage* potentially explained by species attributes was also explained by either range geometry or both range geometry and socio-economic factors (Fig. V.3.S4 B).

Discussion

Our analyses revealed strong species-level bias in globally mobilized mammal occurrence information, with *record counts*, *range coverage* and *geographical bias* scores differing among species by four to five orders of magnitude. A substantial proportion of mammal species (28%) had no GBIF records, and large parts of most mammal ranges were several hundred kilometers away from the closest record that provides direct evidence of occurrence, demonstrating considerable uncertainty regarding fine-scale distributions.

Global species-level bias appears to be largely a consequence of geographical data bias to North America, Australia and Western Europe (Meyer *et al.*, 2015). As expected, *range coverage* was primarily a function of *record count* relative to range size (Fig. II.3.3). However, even very high *record counts* only yield low range coverage scores if those records are *geographically biased* towards one range part, as is the case in many widespread Palaearctic species. Unsurprisingly, given the geographically aggregated and highly duplicated fashion in which occurrence information is collected and mobilized (Meyer *et al.*, 2015), *record count* itself was strongly positively correlated with *geographical bias*, and regions and mammal orders with well-sampled species often coincide with high *geographical bias* scores.

Species attributes

Multiple regression and variation partitioning analyses revealed a surprisingly minor role of species attributes in shaping occurrence information, although all variables that captured species attributes received limited support from multi-model inference. The most important species attribute was body mass, with relatively consistent negative effects on *record count* and *range coverage*. The poorer representation of large-bodied species, including primates and carnivores, might contradict such species' prominence in the scientific literature (Brooke *et al.*, 2014), trait datasets (Gonzalez-Suarez *et al.*, 2012) and monitoring data (Burton, 2012). A plausible explanation is that logistic difficulties in collecting and storing large specimens do not apply to less invasive field research. This points to a great potential for mobilizing occurrence information for underrepresented species from datasets and literature that were originally generated for other purposes (Jetz *et al.*, 2012a).

An alternative reason for negative effects of body mass might be that small-bodied mammals have higher abundances which lead to higher detectability. Surprisingly, however, this abundance hypothesis is otherwise not supported: dietary levels, an indicator of abundances

particularly when controlling for body size and habitat (Robinson & Redford, 1986), consistently showed weak effects (see also *SI V.3.1.5*). Using population density as a more direct measure of abundance for a subset of species also failed to support that hypothesis: there was only a weak positive relationship with *record count* and no effect on *range coverage* (see *SI V.3.1.5*). Traits associated with conspicuousness also failed to support the detectability hypothesis (Iknayan *et al.*, 2013) as both diurnality and foraging stratum showed only weak effects, contrasting results from regional studies (Burton, 2012; Chutipong *et al.*, 2014).

More consistent support accrued for the hypotheses that more records have accumulated for early-described species, and that species of public interest are more likely to have mobilized records (Tyler *et al.*, 2012). Against our expectation, we did not find a clear effect of current threat status. Scientific interest in threatened species (Tyler *et al.*, 2012) might be counter-balanced by legal or ethical impediments to specimen collection and data sharing (Whitlock *et al.*, 2010). While we cannot rule out a stronger role of species attributes at smaller scales, they are clearly not a major driver of species-level variation in range-wide mammal occurrence information.

Range geometry

In contrast to species attributes, range geometry had very strong effects on occurrence information. Range size was the single most important predictor, with a strong positive effect on *record count* and a strong negative effect on *range coverage*. At the global and Neotropical scale, range size was also an important predictor of *geographical bias*. Range shape irregularity was another important constraint to *range coverage*. These results support the notion that while large ranges are bound to overlap with more sampling locations (compare Garamszegi & Møller (2012)), large, irregular-shaped ranges constrain the detail with which a given number of records can cover a range. A few well-placed records can provide a high degree of *range coverage* for small-ranged species that is hardly attainable for large-ranged species. However, with a mean *range coverage* of -102km (median=-55km), even the lower range-size quartile of species did not provide the spatial detail needed for most conservation applications (typically sub-25km, Boitani *et al.* (2011)). Furthermore, occurrence records that could potentially be used in models to refine information were disproportionately scarcer for species in the lower range-size quartile (mean *record count*=23, median=0) compared to all species (mean *record count*=563, median=13). Most small ranges appeared better-covered not because of a detailed representation with records, but simply because any record within the range was automatically closer to any part within the range.

Socio-economic factors

Most key socio-economic drivers of assemblage-level occurrence information (Meyer *et al.*, 2015) also drive species-level information, reinforcing the need to address these factors to create an effective global data basis of species distributions. Mean endemism richness, used as a proxy for area appeal, had the most consistent effect (Soria-Auza & Kessler, 2008). In conjunction with clear positive effects of range size on *record count*, this demonstrates that despite increased collecting activity in endemism-rich areas, sampling to date has not resulted in better representation of range-restricted species in those regions. Consistent with previous suggestions, proximity to institutions (Dennis & Thomas, 2000), GBIF participation (Yesson *et al.*, 2007) and locally available financial resources (Soberón & Peterson, 2004) strongly limit species-level occurrence information, but we found that the importance of these factors differed substantially among realms.

Such realm-specific model differences demonstrate that different factors influence occurrence information in different regional contexts. For instance, the negative effect of area appeal on *record count* in Australasia was contrary to our expectation but has a plausible explanation: data collection and mobilization in endemism-rich northern Australasian countries (e.g. Solomon Islands, Papua New Guinea, East Timor) is in its infancy, whereas Australia has mobilized large numbers of records for most mammals, including those living in comparatively endemism-poor regions (Meyer *et al.*, 2015). As another example, most Palaearctic species have ranges that cover both non-GBIF-participating Asian countries and extremely data-rich Western or Northern European countries, causing strong effects of GBIF participation on *range coverage* and *geographical bias* but not on *record count*. Similarly, *geographical bias* in the Palaearctic realm is mainly driven by strong variation in the proximity of different parts of species' ranges to data-contributing institutions (Fig. II.3.5). These results show that considering the spatial extent and geographical focus of analyses is crucial for understanding bias in occurrence information.

Implications and conclusions

Our results have three major implications. First, species without records are not randomly distributed across orders and regions, nor is quality of available occurrence information constant across species. Without careful consideration of these biases, ecological models that

compare across species and include occurrence information violate statistical assumptions, potentially causing biased inference (Garamszegi & Møller, 2011).

Second, information gaps are particularly severe for range-restricted mammals, where detailed information would be urgently needed to confront future extinction risk (Fritz *et al.*, 2009; Boitani *et al.*, 2011). Data collectors, curators and observation programs should focus on further mobilizing data on range-restricted and threatened species to meet future conservation data needs (Cayuela *et al.*, 2009; Hermoso *et al.*, 2013).

Third, conventional species distribution modeling cannot provide a general remedy, due to high spatial pseudo-replication of records combined with poor spatial *coverage*. Even the 37% of represented mammals that have between 50 and 200 records, an often-cited range of minimum model requirements (Boitani *et al.*, 2011), typically have much fewer unique sampling locations (median=17), and a relatively low range coverage (median=-207 km). Modern hierarchical models can overcome many of these problems by explicitly incorporating models of site-specific survey effort or species-specific detectability (Iknayan *et al.*, 2013; Dorazio, 2014). As biases in mobilized occurrence information are mainly driven by geographical rather than species-specific factors, controlling for these biases by incorporating their site-specific socio-economic drivers may offer the most promising avenue for improving models.

Global point records on mammal distributions are rife with large-scale geographical and taxonomic gaps and biases, hampering species distribution modeling, conservation prioritization and other basic and applied research. All the while, expert range map information remains limited in spatial scale at which it can supplement (Jetz *et al.*, 2012a). To improve the data basis for such applications, the key socio-economic impediments to data availability need to be addressed, e.g. by prioritizing data mobilization in institutions near data gaps and fostering cooperation with data-sharing networks (see discussion in Meyer *et al.* (2015)). Researchers who collect or mobilize new occurrence information should consider possible synergies with global data priorities, e.g. through focusing on threatened, range-restricted or otherwise understudied species. Information metrics such as those developed for this study should be incorporated into online tools to allow researchers and funding agencies identify priority species for improving information. In the meantime, ecological models that account for present data limitations by explicitly incorporating the socio-economic drivers of data collection and mobilization could be a way of drawing improved inferences from accessible occurrence information.

Acknowledgements. We thank those active in collecting, sharing, curating, digitizing and mobilizing distribution data. We thank Tim Robertson and Jeremy Malczyk for help with data assembly, and the members of the Kreft lab for discussions of hypotheses and metrics. CM acknowledges funding from the Deutsche Bundesstiftung Umwelt (DBU). WJ and RPG acknowledge support from NSF (DBI 0960550, DEB 1026764, DEB1441737, DBI-1262600), NASA (NNX11AP72G) and the Yale Program in Spatial Biodiversity Science and Conservation. SAF acknowledges support from the LOEWE funding program of Hesse's Ministry of Higher Education, Research, and the Arts. HK acknowledges funding by the German Research Council (DFG) in the framework of the German Excellence Initiative within the Free Floater Program at the University of Göttingen.

Part III

Synopsis

Introduction

Detailed information on the species distributions is crucial for answering central questions in ecology (Brown *et al.*, 1996), evolutionary biology (Holt, 2003) and biogeography (Lomolino, 2004). Such information is also necessary for the effective allocation of conservation resources (Ferrier, 2002; Venter *et al.*, 2014). In particular, many questions 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 to inform policy-relevant indices of biodiversity change (Butchart *et al.*, 2010). Such detail must come directly from field data (Robertson *et al.*, 2010), or from species distribution modeling (Guisan & Thuiller, 2005) or downscaling approaches (Keil *et al.*, 2013).

Huge numbers of occurrence records from preserved specimens, field observations, literature, and other sources have been mobilized via international data-sharing networks, most notably that of the Global Biodiversity Information Facility (GBIF; Edwards (2000)). Such records provide the primary information on the taxonomic, geographical and temporal dimensions of species distributions, as they provide direct evidence that particular species occurred at particular locations at particular points in time (Soberón & 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 *digital accessible information* (DAI); originally referred to as DAK in Sousa-Baena *et al.* (2014a)).

Notwithstanding the increasing accessibility of occurrence information, global knowledge of species distributions remains extremely limited, a situation termed the 'Wallacean shortfall' (Lomolino, 2004). Most taxa and regions lack large-extent, fine-grain datasets, and existing information is often scattered across multiple sources (Jetz *et al.*, 2012a). Moreover, even available information is prone to many uncertainties arising from ambiguous scientific names (Jansen & Dengler, 2010), imprecisely geo-referenced sampling locations (Rocchini *et al.*, 2011) and old age of many record (Ladle & Hortal, 2013). Finally, because most occurrence records were collected opportunistically (Pyke & Ehrlich, 2010; ter Steege *et al.*, 2011), they inherit taxonomic, geographical and temporal biases (Nelson *et al.*, 1990; Dennis & Thomas, 2000). These biases hamper many important applications, including species distribution modeling (Guisan & Thuiller, 2005), macroecological analyses (Yang *et al.*, 2013) and conservation prioritization (Boitani *et al.*, 2011).

Geographical biases may be driven by biased field work, due to regional differences in accessibility (Freitag *et al.*, 1998; Dennis & Thomas, 2000), safety concerns (Brito *et al.*, 2013), lack of funding (Ahrends *et al.*, 2011) or preferential interest in endemism-rich, mountainous or protected areas (Soria-Auza & Kessler, 2008; Yang *et al.*, 2014). However, biases in DAI may also be caused by biased provision of existing information, due to regional differences in financial or institutional resources for digitization (Vollmar *et al.*, 2010), or poor scientific (Amano & 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 also species-specific factors like lower detectability of nocturnal (Burton, 2012) and 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 the study region of a given researcher intersects with a given range, which in turn affects the likelihood that this particular species is recorded.

The need for better baseline information on species distributions has been frequently emphasized (Boitani *et al.*, 2011; Beck *et al.*, 2012). Improving such information is closely linked to international targets in the framework of the United Nations Convention on Biological Diversity (Pereira *et al.*, 2013) and plays a central role in current discussions in the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES, 2015). However, limited funding and the scale of the Wallacean shortfall make it important to prioritize future data collection and mobilization efforts (Hobern *et al.*, 2013; Sousa-Baena *et al.*, 2014a). Effectively 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 these key factors in ecological models by explicitly incorporating them as variables (Dorazio, 2014; Fithian *et al.*, 2014). Previous studies of patterns and drivers of distribution information were limited in geographical (Ballesteros-Mejia *et al.*, 2013; Yang *et al.*, 2014) or taxonomic (Yesson *et al.*, 2007) scope, by the limited number of tested hypotheses, or by simplistic treatment of distribution information (Amano & Sutherland, 2013). No study has tested the generality of the various information-limiting factors globally across different taxonomic and spatial scales. The main goals of my PhD thesis therefore were to

a) provide the first global, detailed analyses of limitations in mobilized occurrence information for a large section of biodiversity,

b) better understand global taxonomic, geographical and temporal variation in different aspects of occurrence information,

c) better understand global drivers of this variation across different taxonomic groups and spatial scales, and to

d) create an empirical baseline for prioritizing data collection and mobilization efforts, for monitoring these activities, and for effectively accounting for data limitations in ecological models.

Methods

In chapter 1, I focused on land plants. I retrieved c. 120M records via GBIF in Jan 2014, standardized taxonomic information against comprehensive taxonomic databases and carried out plausibility checks of the indicated 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. I.2.1). 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 quantified their relationships using pairwise correlations and principal component analysis.

In chapter 2 (Meyer *et al*., 2015), I focused on terrestrial vertebrates to analyze two aspects of occurrence information at the level of geographical assemblages (Fig. I.2.1). I retrieved c. 183M records via GBIF in Oct 2012 (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 records 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 socio-economic drivers of global variation in these information aspects, separately for each vertebrate group at each of four spatial grain sizes. I used multi-model inference to quantify the relative importance of predictor variables.

In chapter 3, 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. I.2.1). 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 the range coverage and geographical bias metrics 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 and Discussion

To my knowledge, this thesis represents the first global analyses of different aspects of occurrence information (e.g. coverage, uncertainty) in the taxonomic, geographical and temporal dimensions (Fig. I.2.1), and is the first to systematically compare across different spatial scales and taxonomic groups. As expected, I found extensive gaps and biases. 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; compare Feeley & Silman (2010)), and many records fell outside of 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 to 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 including Central America and parts of the Andes (chapters I-II). In contrast, broad regions were without any mobilized occurrence records, particularly in Asia and non-Southern Africa. Surprisingly, there was no pronounced 'tropical data gap' (Collen *et al.*, 2008), neither in plants nor in vertebrates, as several temperate and arctic regions also emerged as extremely data scarce. I also found strong temporal variation in occurrence information (compare 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 leads to higher coverage of species assemblages (Soberón *et al.*, 2007), but also to lower opportunities for inference (chapter 2) and an underestimation of local data gaps (chapter 1). 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 aspects are naturally constrained by record quantity (chapters 1-3; compare Yang *et al.*, (2013)) and accordingly, show at least moderate positive pairwise associations (chapter 1). However, the generally positive relationship between data quantity and coverage aspects is disturbed by aggregation, duplication and biases in those records (chapters 1-3). In contrast, different aspects of data uncertainty generally showed poor correlations with one another as well as with coverage aspects (chapter 1).

I also provide the most comprehensive analyses of underlying causes of bias in occurrence information to date. Of twelve potential socio-economic drivers of assemblage-level record density and inventory completeness, only four received strong support across taxa and grain sizes (chapter 2). Endemism richness (Kier & Barthlott, 2001) generally had a strong effect, supporting the hypothesis that researchers preferentially survey regions where they can hope to find range-restricted species (Soria-Auza & Kessler, 2008). An effect of accessibility was mainly evident in strong positive effects of proximity of grid cells to record-contributing institutions (Moerman & Estabrook, 2006), while transportation infrastructure (Freitag *et al.*, 1998; Ballesteros-Mejia *et al.*, 2013) played a surprisingly minor role. 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 important for determining species-level variation in different aspects of DAI (chapter 3), but their relative importance differed substantially depending on the geographical focus of the analysis (global vs. realm-wide). This demonstrates that regional contexts determine which socio-economic factors are important causes of biases in occurrence information (compare Yang *et al.* (2014)). Interspecific variation in occurrence information was additionally strongly determined by range size and shape. This supports our hypothesis that while large ranges are bound to overlap with more sampling locations, large, irregular-shaped ranges constrain the detail with which a given number of records can cover a range. Against expectations, species attributes related to detection or collection probabilities received little support as predictors of species-level variation in occurrence information.

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 modeling. After more than a decade of intensive mobilization, DAI is still characterized by severe biases, gaps and uncertainties. Unless carefully accounted for, these limitations seriously impair research and conservation applications. The magnitude of data limitations shows that relying only on highest-quality records (Soberón & Peterson, 2004;

Feeley & Silman, 2010) or data-intensive modeling techniques (Feeley & Silman, 2011a) is unrealistic for many species and regions of particular conservation concern (chapters I-III). Further improving the ability of distribution modeling techniques to draw useful inference from low record numbers and to account for data bias and uncertainty (e.g. McInerny & Purves (2011)) should be a top priority. One promising way to account for biases is explicitly incorporating bias-causing factors into models (Dorazio, 2014; Fithian *et al.*, 2014), 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 identification of main 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 and fostering cooperation of large emerging economies with data-sharing networks (chapters 2-3), updating largely outdated information for non-Southern Africa and Southern Asia by carrying out novel surveys (chapter 1), as well as 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, cbd.int/sp/targets). They show that simple indicators like the number of GBIF-facilitated records (Tittensor *et al.*, 2014) cannot reliably show changes in coverage of species and areas, and even less so changes in data uncertainties. We 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.

In short, my thesis demonstrates tremendous taxonomic, geographical and temporal biases, gaps and uncertainties in digital accessible information on species occurrences. It constitutes the most comprehensive research on global patterns and drivers of these limitations to date, providing an empirical baseline for effectively using, improving, and monitoring the global information basis on species distributions.

Part IV

References

Literature cited

Abouheif E. (1999) A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research*, **1**, 895–909.

Ahrends A., Burgess N.D., Gereau R.E., Marchant R., Bulling M.T., Lovett J.C., Platts P.J., Wilkins Kindemba V., Owen N., Fanning E., & Rahbek C. (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.

Amori G. & Gippoliti S. (2000) What do mammalogists want to save? Ten years of mammalian conservation biology. *Biological Conservation*, **9**, 785–793.

Ariño A. (2010) Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, **7**, 81–92.

Ariño A.H., Chavan V., & King N. (2011) The Biodiversity Informatics Potential Index. *BMC Bioinformatics*, **12**, S4.

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.

Bebber D.P., Carine M.A., Wood J.R.I., Wortley A.H., Harris D.J., Prance G.T., Davidse G., Paige J., Pennington T.D., Robson N.K.B., & Scotland R.W. (2010) Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 22169–71.

Beck J., Ballesteros-Mejia L., Buchmann C.M., Dengler J., Fritz S.A., Gruber B., Hof C., Jansen F., Knapp S., Kreft H., Schneider A.-K., Winter M., & Dormann C.F. (2012) What's on the horizon for macroecology? *Ecography*, **35**, 673–683.

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.

Belmaker J. & Jetz W. (2011) Cross-scale variation in species richness-environment associations. *Global Ecology and Biogeography*, **20**, 464–474.

Berendsohn W.G. & Seltmann P.S. (2010) Using geographic and taxonomic metadata to set priorities in specimen digitization. *Biodiversity Informatics*, **7**, 120–129.

BirdLife International (2011) Taxonomic Checklist of the Birds of the World version 3 (2011). Available at http://www.birdlife.org.

Blumstein D.T. (2006) Developing an evolutionary ecology of fear: how life history and natural history traits affect disturbance tolerance in birds. *Animal Behaviour*, **71**, 389–399.

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**, 2623–32.

Bonfoh B., Raso G., Koné I., Dao D., Girardin O., Cissé G., Zinsstag J., Utzinger J., & Tanner M. (2011) Research in a war zone. *Nature*, **474**, 569–71.

Brito J.C., Godinho R., Martínez-Freiría F., Pleguezuelos J.M., Rebelo H., Santos X., Vale C.G., Velo-Antón G., Boratyński Z., Carvalho S.B., Ferreira S., Gonçalves D. V, Silva T.L., Tarroso P., Campos J.C., Leite J. V, Nogueira J., Alvares F., Sillero N., Sow A.S., Fahd S., Crochet P.-A., & Carranza S. (2013) Unravelling biodiversity, evolution and threats to conservation in the Sahara-Sahel. *Biological reviews of the Cambridge Philosophical Society*, **89**, 215–231.

Broennimann O. & Guisan A. (2008) Predicting current and future biological invasions: both native and invaded ranges matter. *Biology letters*, **4**, 585–9.

Brooke Z.M., Bielby J., Nambiar K., & Carbone C. (2014) Correlates of research effort in carnivores: body size, range size and diet matter. *PloS one*, **9**, e93195.

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.

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.

Brummitt N., Bachman S.P., Aletrari E., Chadburn H., Griffiths-Lee J., Lutz M., Moat J., Rivers M.C., Syfert M.M., & Nic Lughadha E.M. (2015) The sampled red list index for plants, phase II: ground-truthing specimen-based conservation assessments. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **370**, 20140015.

Burnham K.P. & Anderson. D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. Springer,

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., van Strien A., Scharlemann J.P.W., Almond R.E.A., Baillie J.E.M., Bomhard B., Brown C., Bruno J., Carpenter K.E., Carr G.M., Chanson J., Chenery A.M., Csirke J., Davidson N.C., Dentener F., Foster M., Galli A., Galloway J.N., Genovesi P., Gregory R.D., Hockings M., Kapos V., Lamarque J.-F., Leverington F., Loh J., McGeoch M.A., McRae L., Minasyan A., Hernández Morcillo M., Oldfield T.E.E., Pauly D., Quader S., Revenga C., Sauer J.R., Skolnik B., Spear D., Stanwell-Smith D., Stuart S.N., Symes A., Tierney M., Tyrrell T.D., Vié J.-C., & Watson R. (2010) Global biodiversity: indicators of recent declines. *Science*, **328**, 1164–1168.

Callmander M. (2011) The endemic and non-endemic vascular flora of Madagascar updated. *Plant Ecology and Evolution*, **144**, 121–125.

Cayuela L., Golicher D.J., Newton A.C., Kolb M., Arets E.J.M.M., Alkemade J.R.M., & Pérez A.M. (2009) Species distribution modeling in the tropics : problems , potentialities , and the role of biological data for effective species conservation. *Tropical Conservation Science*, **2**, 319–352.

Chauvel B., Dessaint F., Cardinal-Legrand C., & Bretagnolle F. (2006) The historical spread of Ambrosia artemisiifolia L. in France from herbarium records. *Journal of Biogeography*, **33**, 665–673.

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.

CIA (2013) GDP - per capita (PPP), The World Factbook. Available at https://www.cia.gov/library/publications/the-world-factbook/.

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.

CONABIO (2012) Dos Décadas de Historia, 1992-2012. Available at http://www.conabio.gob.mx/web/pdf/Conabio_Dos_Decadas_de_Historia_web.pdf.

Conference of the Parties to the Convention on Biological Diversity (2010) X/7. Examination of the outcome-oriented goals and targets and associated indicators and consideration of their possible adjustment for the period beyond 2010. Available at http://www.cbd.int/doc/meetings/sbstta/sbstta-14/draft/sbstta-14-10-draft-en.doc.

Costello M.J., May R.M., & Stork N.E. (2013) Response to comments on "Can we name Earth's species before they go extinct?". *Science (New York, N.Y.)*, **341**, 237.

Crase B., Liedloff A.C., & Wintle B.A. (2012) A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, **35**, 879–888.

Crisp M.D., Laffan S., Linder H.P., & Monro A. (2001) Endemism in the Australian Flora. *Journal of Biogeography*, **28**, 183–198.

Crosby M.R. & Magill R.E. (1988) *TROPICOS. A Botanical Database System at the Missouri Botanical Garden.* Missouri Botanical Garden, St. Louis.

Darwin C.R. & Wallace A.R. (1858) On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the proceedings of the Linnean Society of London. Zoology*, **3**, 45–62.

Dengler J., Jansen F., Glöckler F., Peet R.K., De Cáceres M., Chytrý M., Ewald J., Oldeland J., Lopez-Gonzalez G., Finckh M., Mucina L., Rodwell J.S., Schaminée J.H.J., & Spencer N. (2011) The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science. *Journal of Vegetation Science*, **22**, 582–597.

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.

Department of Commerce. National Oceanic and Atmospheric Administration (2014) 50 CFR Part 224. Endangered and Threatened Wildlife and Plants; Final Endangered Listing of Five Species of Sawfish Under the Endangered Species Act; Final Rule. **79 (239)**, .

Diniz-Filho J.A.F., Bastos R.P., Rangel T.F.L.V.B., Bini L.M., Carvalho P., & Silva R.J. (2005) Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. *Global Ecology and Biogeography*, **14**, 469–477.

Diniz-Filho J.A.F., Mauricio Bini L., Fernando Rangel T., Loyola R.D., Hof C., Nogués-Bravo D., & Araújo M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906.

Dorazio R.M. (2007) On the choice of statistical models for estimating occurrence and extinction from animal surveys. *Ecology*, **88**, 2773–2782.

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.

Dormann C.F., Elith J., Bacher S., Buchmann C., Carl G., Carré G., Marquéz J.R.G., Gruber B., Lafourcade B., Leitão P.J., Münkemüller T., McClean C., Osborne P.E., Reineking B., Schröder B., Skidmore A.K., Zurell D., & Lautenbach S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Dormann C.F., McPherson J.M., Araújo M.B., Bivand R., Bolliger J., Carl G., Davies R.G., Hirzel A., Jetz W., Daniel Kissling W., Kühn I., Ohlemüller R., Peres-Neto P.R., Reineking B., Schröder B., Schurr F.M., & Wilson R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Dormann C.F., Purschke O., García Márquez J.R., Lautenbach S., & Schröder B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–86.

Dutilleul P. (1993) Modifying the t test for assessing the correlation between two spatial processes. *Biometrics*, **49**, 305–314.

Edwards J.L. (2000) Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science*, **289**, 2312–2314.

Enke N., Thessen A., Bach K., Bendix J., Seeger B., & Gemeinholzer B. (2012) The user's view on biodiversity data sharing - Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, **11**, 25–33.

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. (2011a) Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, **17**, 1132–1140.

Feeley K.J. & Silman M.R. (2011b) The data void in modeling current and future distributions of tropical species. *Global Change Biology*, **17**, 626–630.

Ferrier S. (2002) Mapping Spatial Pattern in Biodiversity for Regional Conservation Planning: Where to from Here ? *Systematic biology*, **51**, 331–363.

Figueiredo E. & Smith G.F. (2010) The colonial legacy in African plant taxonomy. *South African Journal of Science*, **106**, 1–4.

Fithian W., Elith J., Hastie T., & Keith D.A. (2014) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, Early View.

Freitag S., Hobson C., Biggs H.C., & 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.

Fritz S. a, Bininda-Emonds O.R.P., & Purvis A. (2009) Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology Letters*, **12**, 538–549.

Frost D.R. (2012) Amphibian Species of the World: an Online Reference. Version 5.5 (2012). Electronic Database accessible at http://research.amnh.org/herpetology/amphibia/index.html. American Museum of Natural History, New York, USA.

Funk V.A. & Richardson K.S. (2002) Systematic data in biodiversity studies: use it or lose it. *Systematic biology*, **51**, 303–16.

Funk V.A., Zermoglio M.F., & Nasir N. (1999) Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana. *Biodiversity and Conservation*, **8**, 727–751.

Gamfeldt L., Snäll T., Bagchi R., Jonsson M., Gustafsson L., Kjellander P., Ruiz-Jaen M.C., Fröberg M., Stendahl J., Philipson C.D., Mikusiński G., Andersson E., Westerlund B., Andrén H., Moberg F., Moen J., & Bengtsson J. (2013) Higher levels of multiple ecosystem services are found in forests with more tree species. *Nature communications*, **4**, 1340.

Garamszegi L.Z. & Møller A.P. (2011) Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. *Systematic biology*, **60**, 876–80.

Garamszegi L.Z. & Møller A.P. (2012) Untested assumptions about within-species sample size and missing data in interspecific studies. *Behavioral Ecology and Sociobiology*, **66**, 1363–1373.

Gaston K.J. & Fuller R.A. (2009) The sizes of species' geographic ranges. *Journal of Applied Ecology*, **46**, 1–9.

GBIF (2011) GBIF Strategic Plan 2012-2016: Seizing the future. Available at http://www.gbif.org/orc/?doc_id=2792.

Gioia P. & Pigott J.P. (1998) Biodiversity assessment : a case study in predicting richness from the potential distributions of plant species in the forests of south-western Australia. *Journal of Biogeography*, **27**, 1065–1078.

Gonzalez-Suarez M., Lucas P.M., & Revilla E. (2012) Biases in comparative analyses of extinction risk: mind the gap. *The Journal of animal ecology*, **81**, 1211–22.

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.

Graham C.H. & Hijmans R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578–587.

Grantham H.S., Moilanen A., Wilson K.A., Pressey R.L., Rebelo T.G., & Possingham H.P. (2008) Diminishing return on investment for biodiversity data in conservation planning. *Conservation Letters*, **1**, 190–198.

Grinnell J. (1917) The Niche-Relationships of the California Thrasher. *The Auk*, **34**, 427–433.

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

Guisan A., Tingley R., Baumgartner J.B., Naujokaitis-Lewis I., Sutcliffe P.R., Tulloch A.I.T., Regan T.J., Brotons L., McDonald-Madden E., Mantyka-Pringle C., Martin T.G., Rhodes J.R., Maggini R., Setterfield S.A., Elith J., Schwartz M.W., Wintle B. a, Broennimann O., Austin M., Ferrier S., Kearney M.R., Possingham H.P., & Buckley Y.M. (2013) Predicting species distributions for conservation decisions. *Ecology letters*, **16**, 1424–1435.

Guisan A.A., Zimmermann N.E., Elith J., Graham C.H., Phillips S., & Peterson A.T. (2007) What Matters for Predicting the Occurrences of Trees : Techniques , Data , or Species ' Characteristics ? Published by : Ecological Society of America content in a trusted digital archive . We use information technology and tools to increase productivity. *Ecological Monographs*, **77**, 615–630.

Guralnick R. & Van Cleve J. (2005) Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. *Diversity and Distributions*, **11**, 349–359.

Hawkins B.A., Rueda M., & Rodríguez M.Á. (2008) What Do Range Maps and Surveys Tell Us About Diversity Patterns? *Folia Geobotanica*, **43**, 345–355.

Herberich E., Sikorski J., & Hothorn T. (2010) A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PloS one*, **5**, e9788.

Hermoso V., Kennard M.J., & Linke S. (2013) Data acquisition for conservation assessments: is the effort worth it? *PloS one*, **8**, e59662.

Hijmans R.J., Garrett K.A., Huaman Z., Zhang D.P., Schreuder M., & Bonierbale M. (2000) Assessing the Geographic Representativeness of Genebank Collections: the Case of Bolivian Wild Potatoes. *Conservation Biology*, **14**, 1755–1765.

Hobern D., Apostolico A., Arnaud E., Bello J.C., Canhos D., Dubois G., Field D., García E.A., Hardisty A., Harrison J., Heidorn B., Krishtalka L., Mata E., Page R., Parr C., Price J., & Willoughby S. (2013) Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age. Availble at: http://www.gbif.org/orc/?doc_id=5353.

Hochachka W.M., Fink D., Hutchinson R.A., Sheldon D., Wong W.-K., & Kelling S. (2012) Data-intensive science applied to broad-scale citizen science. *Trends in Ecology and Evolution*, **27**, 130–7.

Hof C., Araújo M.B., Jetz W., & Rahbek C. (2011) Additive threats from pathogens, climate and land-use change for global amphibian diversity. *Nature*, **480**, 516–9.

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

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

Hortal J., Jiménez-Valverde A., Gómez J.F., Lobo J.M., & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Hurlbert A.H. & Jetz W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 13384–9.

Iknayan K.J., Tingley M.W., Furnas B.J., & Beissinger S.R. (2013) Detecting diversity : emerging methods to estimate species diversity. *Trends in Ecology & Evolution*, **29**, 1–10.

Inouye D.W. (2014) IPBES: global collaboration on biodiversity and ecosystem services. *Frontiers in Ecology and the Environment*, **12**, 371–371.

Institute for Economics and Peace (2012) Global peace index 2008-2012. Sydney, Australia: Institute for Economics and Peace. Available at http://www.visionofhumanity.org/.

IPBES (2015) Guide on the production and integration of assessments from and across all scales (deliverable 2 (a)). Available at http://www.ipbes.net/images/documents/plenary/third/information/INF_4/IPBES_3_INF_4.pdf.

Isbell F., Calcagno V., Hector A., Connolly J., Harpole W.S., Reich P.B., Scherer-Lorenzen M., Schmid B., Tilman D., van Ruijven J., Weigelt A., Wilsey B.J., Zavaleta E.S., & Loreau M. (2011) High plant diversity is needed to maintain ecosystem services. *Nature*, **477**, 199–202.

ITIS Global Orrell T. (custodian) (2012) ITIS Global: The Integrated Taxonomic Information System (version Apr 11). In: Species 2000 & ITIS Catalogue of Life, 2012 Annual Checklist (Bisby F., Roskov Y., Culham A., Orrell T., Nicolson D., Paglinawan L., Bailly N., Appeltans W., Kirk P., Bourgoin.

IUCN (2010) IUCN Red List of Threatened Speciess. Version 2010.4. http://www.iucnredlist.org. Downloaded on 27 October 2010.

IUCN and UNEP (2009) The World Database on Protected Areas (WDPA). UNEP-WCMC. Cambridge, UK.

Jäger E.J. (1976) Areal- und Florenkunde (Floristische Geobotanik). *Progress in Botany*, **38**, 314–330. Reproduced in Frodin, D.G..(2001) *Guide to standard floras of the World: an annotated, geographically arranged systematic bibliography of the principal floras, enumerations, checklists and chorological atlases of different areas*. Cambridge University Press, UK.

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. (2012a) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution*, **27**, 151–159.

Jetz W., Sekercioglu C.H., & Watson J.E.M. (2008) Ecological Correlates and Conservation Implications of Overestimating Species Geographic Ranges. *Conservation Biology*, **22**, 110–119.

Jetz W., Wilcove D.S., & Dobson A.P. (2012b) Update of Jetz, W, D. S. Wilcove & Dobson, A.P. (2007): Projected impacts of climate and land-use change on the global diversity of birds. *PLoS Biology*, **5**, e157.

Jones K.E., Bielby J., Cardillo M., Fritz S.A., O'Dell J., Orme C.D.L., Safi K., Sechrest W., Boakes E.H., Carbone C., Connolly C., Cutts M.J., Foster J.K., Grenyer R., Habib M., Plaster C. a, Price S. a, Rigby E. a, Rist J., Teacher A., Bininda-Emonds O.R.P., Gittleman J.L., Mace G.M., & Purvis A. (2009) PanTHERIA: a

species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, **90**, 2648.

Kadmon R., Farber O., & Danon A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological applications : a publication of the Ecological Society of America*, **14**, 401–413.

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.

Kier G. & Barthlott W. (2001) Measuring and mapping endemism and species richness: a new methodological approach and its application on the flora of Africa. *Biodiversity and Conservation*, **10**, 1513–1529.

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.

King D.A. (2002) The scientific impact of nations - What different countries get for their research spending. *Nature*, **430**, 311–316.

Kissling W.D. & Carl G. (2008) Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, **17**, 59–71.

Kissling W.D., Field R., & Böhning-Gaese K. (2008) Spatial patterns of woody plant and bird diversity: functional relationships or environmental effects? *Global Ecology and Biogeography*, **17**, 327–339.

Knight A.J. (2008) "Bats, snakes and spiders, Oh my!" How aesthetic and negativistic attitudes, and other concepts predict support for species protection. *Journal of Environmental Psychology*, **28**, 94–103.

Kreft H. & Jetz W. (2007) Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 5925–30.

Kremen C., Cameron A., Moilanen A., Phillips S.J., Thomas C.D., Beentje H., Dransfield J., Fisher B.L., Glaw F., Good T.C., Harper G.J., Hijmans R.J., Lees D.C., Jr. E.L., Nussbaum R.A., Raxworthy C.J., Razafimpahanana A., Schatz G.E., Vences M., Vieites D.R., Wright P.C., & Zjhra M.L. (2008) Aligning Conservation Priorities Across Taxa in Madagascar with High-Resolution Planning Tools. *Science*, **320**, 222–226.

Küper W., Sommer J.H., Lovett J.C., & Barthlott W. (2006) Deficiency in African plant distribution data – missing pieces of the puzzle. *Botanical Journal of the Linnean Society*, **150**, 355–368.

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

Lavoie C. (2013) Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics*, **15**, 68–76.

Lenzen M., Moran D., Kanemoto K., Foran B., Lobefaro L., & Geschke A. (2012) International trade drives biodiversity threats in developing nations. *Nature*, **486**, 109–12.

Lepage D. (2012) Avibase - The World Bird Database v. 2012. Available from http://avibase.bsc-eoc.org. .

Lobo J.M. (2008) Database records as a surrogate for sampling effort provide higher species richness estimations. *Biodiversity and Conservation*, **17**, 873–881.

Lobo J.M., Romo H., & García-Barros E. (2006) Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, **6**, 873–885.

Lomolino M. (2004) Frontiers of Biogeography: new directions in the geography of. *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.

Longino J.T., Coddington J., & Colwell R.K. (2002) The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology*, **83**, 689–702.

MacArthur R.H. & Wilson E.O. (1967) *The theory of island biogeography.* Princeton University Press, Princeton, NJ.

Manceur A.M. & Kühn I. (2014) Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. *Methods in Ecology and Evolution*, **5**, 739–750.

Margules C.R. & Pressey R.L. (2000) Systematic conservation planning. *Nature*, **405**, 243–53.

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.

May R. (1997) The Scientific Wealth of Nations. *Science*, **275**, 793–796.

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.

Menke S.B., Holway D.A., Fisher R.N., & Jetz W. (2009) Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. *Global Ecology and Biogeography*, **18**, 50–63.

Meyer C., Kreft H., Guralnick R.P., & Jetz W. (2015) Global priorities for an effective information basis of biodiversity distributions. *PeerJ PrePrints*, **3**, e1057.

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.

Morueta-Holme N., Enquist B.J., McGill B.J., Boyle B., Jørgensen P.M., Ott J.E., Peet R.K., Símová I., Sloat L.L., Thiers B., Violle C., Wiser S.K., Dolins S., Donoghue J.C., Kraft N.J.B., Regetz J., Schildhauer M., Spencer N., & Svenning J.-C. (2013) Habitat area and climate stability determine geographical variation in plant species range sizes. *Ecology letters*, **16**, 1446–1454.

Murphey P.C., Guralnick R.P., Glaubitz R., Neufeld D., & Ryan J.A. (2004) Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain Informatics Initiative (Mapstedi). *PhyloInformatics*, **21**, 1–29.

Naidoo R. & Adamowicz W.L. (2001) Effects of Economic Prosperity on Numbers of Threatened Species. *Conservation Biology*, **15**, 1021–1029.

Nelson A. (2008) Travel time to major cities: A global map of Accessibility. Global Environment Monitoring Unit - Joint Research Centre of the European Commission, Ispra Italy. Available at http://gem.jrc.ec.europa.eu/.

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

NERC Centre for Population Biology - Imperial College (2010) Available at http://www.sw.ic.ac.uk/cpb/cpb/gpdd.html.

New York Botanical Garden (2014) Available at http://sciweb.nybg.org/science2/IndexHerbariorum.asp.

Newbold T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, **34**, 3–22.

O'Hara R.B. & Kotze D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.

Olson D.M., Dinerstein E., Wikramanayake E.D., Burgess N.D., Powell G.V.N., Underwood E.C., D'amico J.A., Itoua I., Strand H.E., Morrison J.C., Loucks C.J., Allnutt T.F., Ricketts T.H., Kura Y., Lamoreux J.F., Wettengel W.W., Hedao P., & Kassem K.R. (2001) Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, **51**, 933–938.

Osborne P.E. & Tigar B.J. (1992) Interpreting Bird Atlas Data Using Logistic Models: An Example From Lesotho, Southern Africa. *Journal of Applied Ecology*, **29**, 55–62.

Otegui J. (2012) Quality and Fitness-for-use Assessments on the Primary Data Indexed at the Global Biodiversity Information Facility (GBIF). Dissertation Thesis, 390p.

Otegui J., Ariño A.H., Encinas M.A., & Pando F. (2013) Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF). *PloS one*, **8**, e55144.

Palmer L. (2011) Show me the money. *Nature Climate Change*, **1**, 376–380.

Van Panhuis W.G., Paul P., Emerson C., Grefenstette J., Wilder R., Herbst A.J., Heymann D., & Burke D.S. (2014) A systematic review of barriers to data sharing in public health. *BMC public health*, **14**, 1144.

Parnell J.A.N., Simpson D.A., Moat J., Kirkup D.W., Chantaranothai P., Boyce P.C., Bygrave P., Dransfield S., Jebb M.H.P., Macklin J., Meade C., Middleton D.J., Muasya A.M., Prajaksood A., Pendry C.A., Pooma R., Suddee S., & Wilkin P. (2003) Plant collecting spread and densities: their potential impact on biogeographical studies in Thailand. *Journal of Biogeography*, **30**, 193–209.

Partow A. (2003) The Global Airport Database. Release Version 0.0.1. Available at http://www.partow.net/miscellaneous/airportdatabase/. .

Paton A.J. (2013) From Working List to Online Flora of All Known Plants—Looking Forward with Hindsight. *Annals of the Missouri Botanical Garden*, **99**, 206–213.

Pautasso M. & McKinney M.L. (2007) The botanist effect revisited: plant species richness, county area, and human population size in the United States. *Conservation biology : the journal of the Society for Conservation Biology*, **21**, 1333–40.

Pereira H.M., Belnap J., Brummitt N., Collen B., Ding H., Gonzalez-Espinosa M., Gregory R.D., Honrado J., Jongman R.H., Julliard R., McRae L., Proença V., Rodrigues P., Opige M., Rodriguez J.P., Schmeller D.S., van Swaay C., & Vieira C. (2010) Global biodiversity monitoring. *Frontiers in Ecology and the Environment*, **8**, 458–459.

Pereira H.M., Ferrier S., Walters M., Geller G.N., Jongman R.H.G., Scholes R.J., Bruford M.W., Brummitt N., Butchart S.H.M., Cardoso A.C., Coops N.C., Dulloo E., Faith D.P., Freyhof J., Gregory R.D., Heip C., Höft R., Hurtt G., Jetz W., Karp D.S., M. A. McGeoch D.O., Onoda Y., Pettorelli N., Reyers B., Sayre R., Scharlemann J.P.W., Stuart S.N., Turak E., Walpole M., & Wegmann M. (2013) Essential Biodiversity Variables. *Science*, **339**, 277–278.

Pereira H.M., Navarro L.M., & Martins I.S. (2012) Global Biodiversity Change: The Bad, the Good, and the Unknown. *Annual Review of Environment and Resources*, **37**, 25–50.

Peres-Neto P.R., Legendre P., Dray S., & Borcard D. (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, **87**, 2614–25.

Perry N. (2010) The ecological importance of species and the Noah's Ark problem. *Ecological Economics*, **69**, 478–485.

Phillips S.J., Dudík M., Elith J., Graham C.H., Lehmann A., Leathwick J., & Ferrier S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications : a publication of the Ecological Society of America*, **19**, 181–97.

Pigot A.L., Owens I.P.F., & Orme C.D.L. (2010) The environmental limits to geographic range expansion in birds. *Ecology letters*, **13**, 705–15.

Poliseli L. & Christoffersen M.L. (2012) Zoological Collections and the Effects of Scientific Territorialism, Sociological Landscape - Theories, Realities and Trends, Dr. Dennis Erasga (Ed.), ISBN: 978-953-51-0460-5, InTech, DOI: 10.5772/38363. Available from: http://www.intechopen.com/books/soc.

Prance G.T. (1977) Floristic Inventory of the Tropics: Where do we stand? *Annals of the Missouri Botanical Garden*, **64**, 659–684.

Pyke G.H. & Ehrlich P.R. (2010) Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological reviews of the Cambridge Philosophical Society*, **85**, 247–66.

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raven P.H. & Axelrod D.I. (1974) Angiosperm biogeography and past continental movements. *Annals of the Missouri Botanical Garden*, **61**, 539–673.

Reddy S. & Dávalos L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719–1727.

Rivers M.C., Taylor L., Brummitt N.A., Meagher T.R., Roberts D.L., & Lughadha E.N. (2011) How many herbarium specimens are needed to detect threatened species? *Biological Conservation*, **144**, 2541–2547.

Robertson M.P., Cumming G.S., & Erasmus B.F.N. (2010) Getting the most out of atlas data. *Diversity and Distributions*, **16**, 363–375.

Robinson J.G. & Redford K.H. (1986) Body size, diet, and population density of neotropical forest mammals. *The American Naturalist*, **128**, 665–680.

Rocchini D., Hortal J., Lengyel S., Lobo J.M., Jimenez-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.

Rodrigue J., Comtois C., & Slack B. (2006) *The Geography of Transport Systems*. Routledge (London & New York).

Rondinini C., Di Marco M., Visconti P., Butchart S.H.M., & Boitani L. (2014) Update or Outdate: Long-Term Viability of the IUCN Red List. *Conservation Letters*, **7**, 126–130.

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.

Schäfer A., Dallmeier-Tiessen S., Pfeiffenberger H., & et al. (The ODE Project) (2011) Ten Tales of Drivers and Barriers in Data Sharing. Available at http://epic.awi.de/.

Schmidt-Lebuhn A.N., Knerr N.J., & Kessler M. (2013) Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation*, **22**, 905–919.

Scholes R.J., Mace G.M., Turner W., Geller G.N., Jurgens N., Larigauderie A., Muchoney D., Walther B.A., & Mooney H.A. (2008) Toward a global biodiversity observing system. *Science (New York, N.Y.)*, **321**, 1044–5.

Schulman L., Toivonen T., & Ruokolainen K. (2007) Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *Journal of Biogeography*, **34**, 1388–1399.

Schurr F.M., Pagel J., Cabral J.S., Groeneveld J., Bykova O., O'Hara R.B., Hartig F., Kissling W.D., Linder H.P., Midgley G.F., Schröder B., Singer A., & Zimmermann N.E. (2012) How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography*, **39**, 2146–2162.

SCImago (2007) SJR — SCImago Journal & Country Rank. Retrieved March 12, 2013, from http://www.scimagojr.com.

Smith R.J., Veríssimo D., Leader-Williams N., Cowling R.M., & Knight A.T. (2009) Let the locals lead. *Nature*, **462**, 280–281.

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., Llorente J.B., & Oñate L. (2000) The use of specimen-label databases for conservation purposes : an example using Mexican Papilionid and Pierid butterflies. *Biodiversity and Conservation*, **9**, 1441–1466.

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**, 689–98.

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. (2014a) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions*, **20**, 369–381.

Sousa-Baena M.S., Garcia L.C., & Townsend Peterson a. (2014b) Knowledge behind conservation status decisions: Data basis for "Data Deficient" Brazilian plant species. *Biological Conservation*, **173**, 80–89.

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.

The World Bank (2012) Worldwide Governance Indicators. Available at http://info.worldbank.org/governance/wgi/.

Thomas C. (2009) Biodiversity Databases Spread, Prompting Unification Call. *Science*, **324**, 1632–1633.

Thuiller W., Albert C., Araújo M.B., Berry P.M., Cabeza M., Guisan A., Hickler T., Midgley G.F., Paterson J., Schurr F.M., Sykes M.T., & Zimmermann N.E. (2008) Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, **9**, 137–152.

Tittensor D.P., Walpole M., Hill S.L.L., Boyce D.G., Britten G.L., Burgess N.D., Butchart S.H.M., Leadley P.W., Regan E.C., Alkemade R., Baumung R., Bellard C., Bouwman L., Bowles-Newark N.J., Chenery a. M., Cheung W.W.L., Christensen V., Cooper H.D., Crowther a. R., Dixon M.J.R., Galli A., Gaveau V., Gregory R.D., Gutierrez N.L., Hirsch T.L., Hoft R., Januchowski-Hartley S.R., Karmann M., Krug C.B., Leverington F.J., Loh J., Lojenga R.K., Malsch K., Marques A., Morgan D.H.W., Mumby P.J., Newbold T., Noonan-Mooney K., Pagad S.N., Parks B.C., Pereira H.M., Robertson T., Rondinini C., Santini L., Scharlemann J.P.W., Schindler S., Sumaila U.R., Teh L.S.L., van Kolck J., Visconti P., & Ye Y. (2014) A mid-term analysis of progress toward international biodiversity targets. *Science*, **346**, 241–244.

TNRS (2014) The Taxonomic Name Resolution Service. iPlant Collaborative. Version 3.2 accessed Apr 2014. Available at http://tnrs.iplantcollaborative.org.

Tollefsen A.F., Strand H., & Buhaug H. (2012) PRIO-GRID: A unified spatial data structure. *Journal of Peace Research*, **49**, 363–374.

TPL (2014) The Plant List. Version 1.1; Published on the Internet (accessed January 2014). Available at: http://www.theplantlist.org/.

Tucker M. a., Ord T.J., & Rogers T.L. (2014) Evolutionary predictors of mammalian home range size: body mass, diet and the environment. *Global Ecology and Biogeography*, **23**, 1105–1114.

Tyler E.H.M., Somerfield P.J., Berghe E. Vanden, Bremner J., Jackson E., Langmead O., Palomares M.L.D., & Webb T.J. (2012) Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology. *Global Ecology and Biogeography*, **21**, 922–934.

U.S. National Committee for CODATA (1997) Bits of Power - Issues in Global Access to Scientific Data. *NATIONAL ACADEMY PRESS*, 249.

UNESCO Institute for Statistics (UIS) (2012) Science and technology report. Available at http://www.uis.unesco.org/ScienceTechnology/Pages/research-and-development-statistics.aspx.

US Geological Survey (1996) GTOPO30. Available at: http://www1.gsi.go.jp/geowww/globalmap-gsi/gtopo30/gtopo30.html.

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

Varela S., Anderson R.P., García-Valdés R., & Fernández-González F. (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, **37**, 1084–1091.

Venter O., Fuller R.A., Segan D.B., Carwardine J., Brooks T., Butchart S.H.M., Di Marco M., Iwamura T., Joseph L., O'Grady D., Possingham H.P., Rondinini C., Smith R.J., Venter M., & Watson J.E.M. (2014) Targeting Global Protected Area Expansion for Imperiled Biodiversity. *PLoS Biology*, **12**, e1001891.

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

Waldron A., Mooers A.O., Miller D.C., Nibbelink N., Redding D., & Kuhn T.S. (2013) Targeting global conservation funding to limit immediate biodiversity declines. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 12144–12148.

WCSP (2013) World Checklist of Selected Plant Families. Facilitated by the Royal Botanic Gardens, Kew. Available at: http://apps.kew.org/wcsp/.

Weigelt P., Jetz W., & Kreft H. (2013) Bioclimatic and physical characterization of the world's islands. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 15307–12.

Whitlock M.C. (2011) Data archiving in ecology and evolution: best practices. *Trends in Ecology and Evolution*, **26**, 61–5.

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

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**, 3–23.

Wikelski M. & Kays R. (2015) Movebank: archive, analysis and sharing of animal movement data. World Wide Web electronic publication.

Willis K.J., Araújo M.B., Bennett K.D., Figueroa-Rangel B., Froyd C.A., & Myers N. (2007) How can a knowledge of the past help to conserve the future? Biodiversity conservation and the relevance of long-term ecological studies. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **362**, 175–86.

Wilman H., Belmaker J., Simpson J., Rosa C. de la, Rivadeneira M.M., & Jetz W. (2014) EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology*, **95**, 2027.

Wilson D.E. & Reeder D.M. (2005) Mammal Species of the World. A Taxonomic and Geographic Reference (3rd ed). Available from http://www.press.jhu.edu.

Wisz M.S., Hijmans R.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.

World Health Organization (2012) Global Atlas of the Health Workforce. Available at: http://apps.who.int/globalatlas/.

Wright M.G. & Samways M.J. (1998) Insect species richness tracking plant species richness in a diverse flora: in the Cape Floristic South Africa Region, South Africa. *Oecologia*, **115**, 427–433.

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 & Biogeography*, **23**, 1284–1292.

Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White R.J., Jones A.C., Bisby F.A., & Culham A. (2007) How global is the global biodiversity information facility? *PloS one*, **2**, e1124.

Part V

Appendix

Supplementary information - Chapter 1

Multidimensional biases, gaps and uncertainties in global plant occurrence information

Carsten Meyer, Patrick Weigelt and Holger Kreft

SI V.1.1. Treatment of taxonomic information.

The basis for our taxonomic treatment was the comprehensive taxonomic information provided via The Plant List (TPL, 2014) and iPlant's Taxonomic Name Resolution Service (TNRS, 2014). In cases of conflicting information, we always gave TPL precedence.

The raw dataset as downloaded via GBIF contained 119,058,280 raw records (Fig. V.1.S1A). We first cleaned verbatim scientific names strings (Fig. V.1.S1B), by excluding name strings that would not be reliably linkable to accepted species names. For instance, we excluded cases where no species was identified (e.g. 'sp. nov.' or '*Sorbus* sp.', 'ined.', etc.) or where it was implied that the species identification was doubtful (e.g. 'cf.', 'aff.', 'à confirmer', '*Sorbus* ?*arnoldiana*', '*Oxyanthus* sp. possibly *unilocular*', etc.). We further excluded cases where a hybrid or a cultivated form was indicated (e.g. 'x', '<->', 'hybr.', 'hort.', 'cult.', 'var. "Ballerina"', etc.). We corrected wrong capitalizations of letters, and removed random punctuations and signs. These steps reduced 2,206,831 verbatim name strings to 1,552,901 interpretable names, including accepted species and subspecies names, synonyms, and spelling variants with or without author information.

We then performed the taxonomic standardization and validation. We first compared genus names against genus names listed in The Plant List (TPL, 2014) or iPlant's Taxonomic Name Resolution Service (TNRS, 2014). We corrected misspelled genus names where we were confident regarding the true genus (doubtful cases were excluded). We then compared each name string to all possible scientific names listed under that genus in TPL. For each resulting pair of verbatim name and TPL-listed scientific name, we calculated the orthographic distance between species epithets and between the entire name strings (e.g. including author

information), using an approximate string matching algorithm. This algorithm counts the total number of changes that have to be applied to one string in order to match another, and related that number to the entire length of the string. We then linked verbatim names via the best-matching TPL-listed name to the respective accepted species. For names that could not be matched to TPL-listed names or were not resolved to accepted species, we repeated these steps using taxonomic information from TNRS. We excluded all verbatim names that did not match names treated by TPL or TNRS as accepted names with no more than 25% orthographic distance, either directly or through a synonym. Overall, cleaning and validation led to an exclusion of 242,043 verbatim names strings (Fig. V.1.S1E); All remaining 1,964,788 verbatim name strings (89%) converged to 367,703 accepted species. These were further reduced to 229,218 accepted species (Fig. V.1.S1I) by applying our basic geographical filter (see Methods).

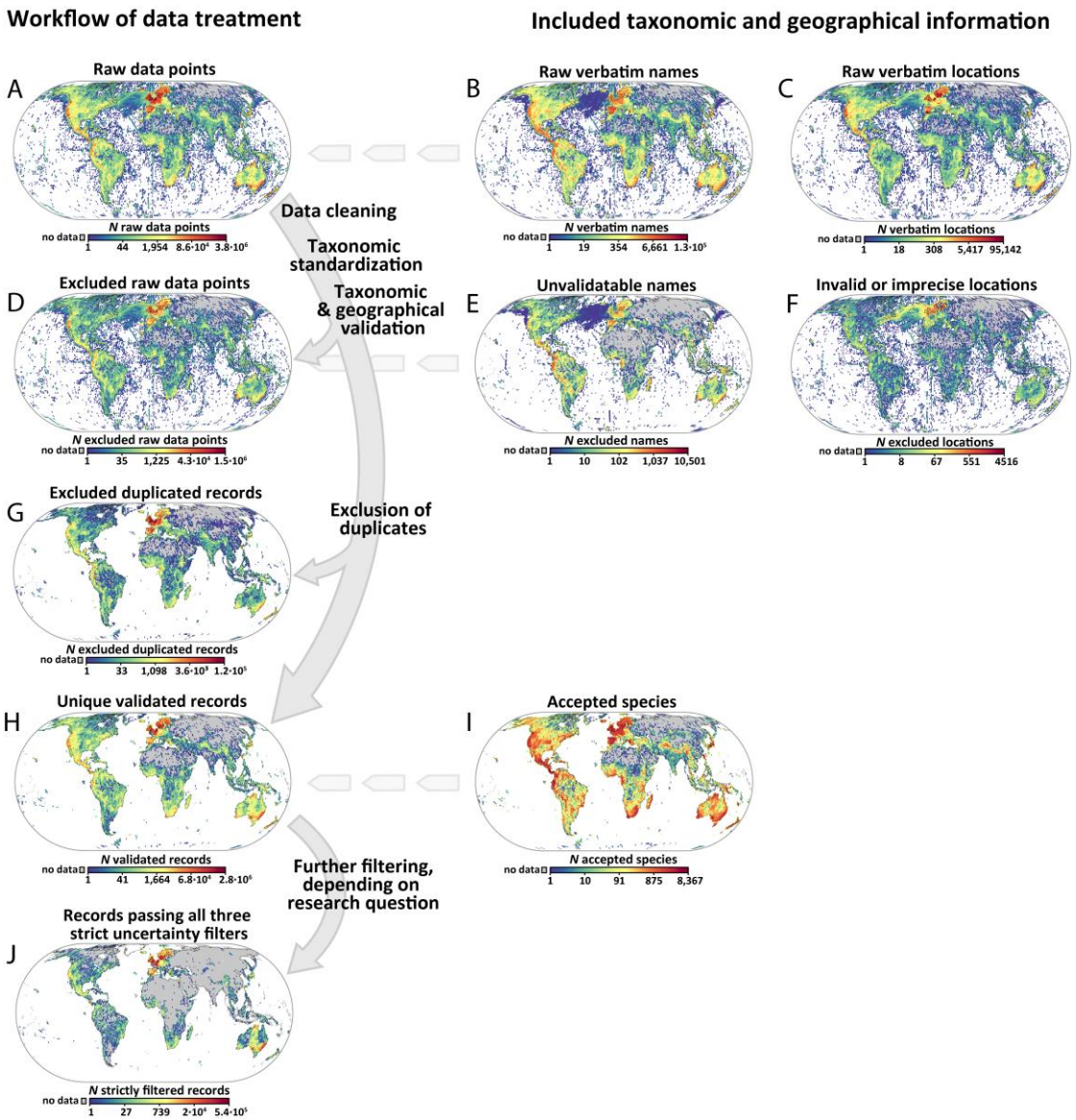


Figure V.1.S1. Workflow from raw mobilized data to usable occurrence records. Maps show spread of occurrence information for land plants, as mobilized via the Global Biodiversity Information Facility (GBIF), across 110 km x 110 km grid cells. We retrieved (A) 119,058,280 raw data via GBIF, including (B) 2,206,831 verbatim name strings and (C) 4,314,752 verbatim coordinate combinations. Data cleaning, taxonomic standardization and taxonomic and geographical validation led to the exclusion of (D) 12,538,809 raw data, including (E) 242,043 unvalidatable name strings and (F) 252,550 invalid or imprecise coordinate combinations. Remaining validated records were reduced to unique records, which led to the exclusion of (G) 24,899,514 duplicated species-location-year-month-combinations and left (H) 55,929,317 unique validated records including (I) 229,218 accepted species. Depending on research question, further filtering might be necessary, such applying our strict taxonomic, geographical and temporal filters (see *Methods*), which would leave (J) 9,295,847 strictly filtered records. For details, see *Methods* and *SI V.1.1*.

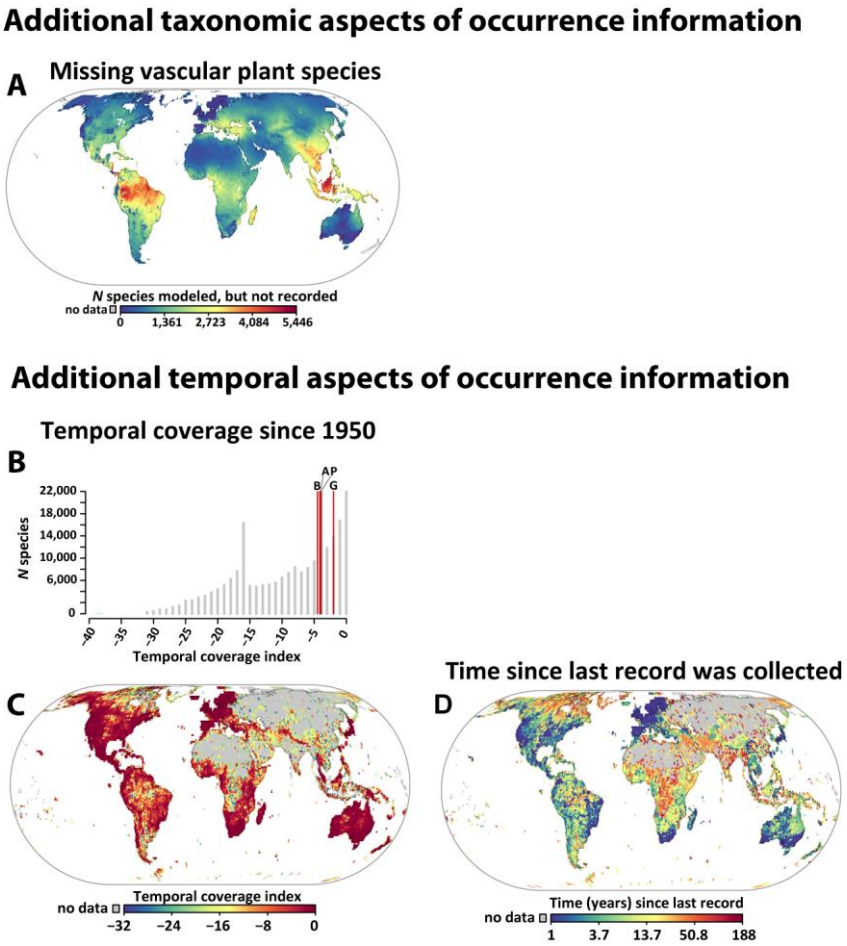


Figure V.1.S2. Additional taxonomic and temporal aspects of occurrence information mentioned in chapter 2. A) Geographical variation across 110 x 110 km grid cells in that portion of modeled vascular plant richness (Kreft & Jetz, 2007) that was missing from mobilized occurrence information. B) Frequency distribution across land plant species in scores of temporal coverage since 1950, calculated as the mean minimum Euclidean distance between all possible months between 1950 and 2010 to their respective closest months with records. C) Geographical variation in temporal coverage since 1950. D) Geographical variation in the time (in years) since the last mobilized record has been collected.

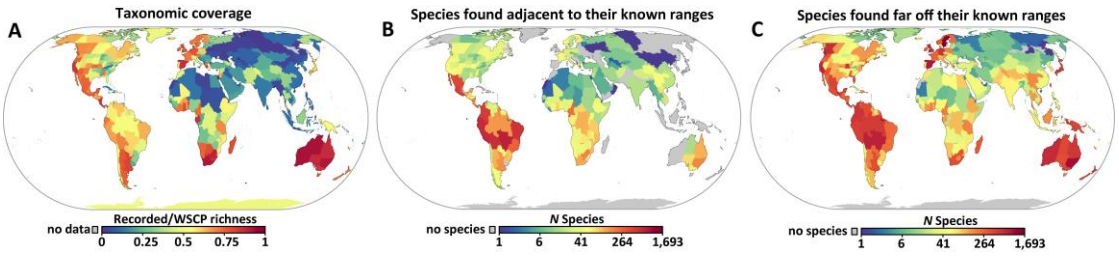


Figure V.1.S3. *Taxonomic coverage* and of native and non-native species for selected families of seed plants. Species records for a subset of the global seed plant flora (105,031 species, c. 34% of all) were geographically validated against 'botanical country' checklists from the World Checklist of Selected Plant Families (WCSP, 2013)). A) *Taxonomic coverage* of native seed plant species of selected families, based on geographically validated records. B-C) Number of species represented by occurrence records outside their known native ranges: B) Species recorded immediately adjacent to their native ranges; C) Species recorded far off their native ranges. Botanical countries are level-3 regions of the Biodiversity Information Standards (TDWG, formerly International Working Group on Taxonomic Databases). Color scales are the same in B-C.

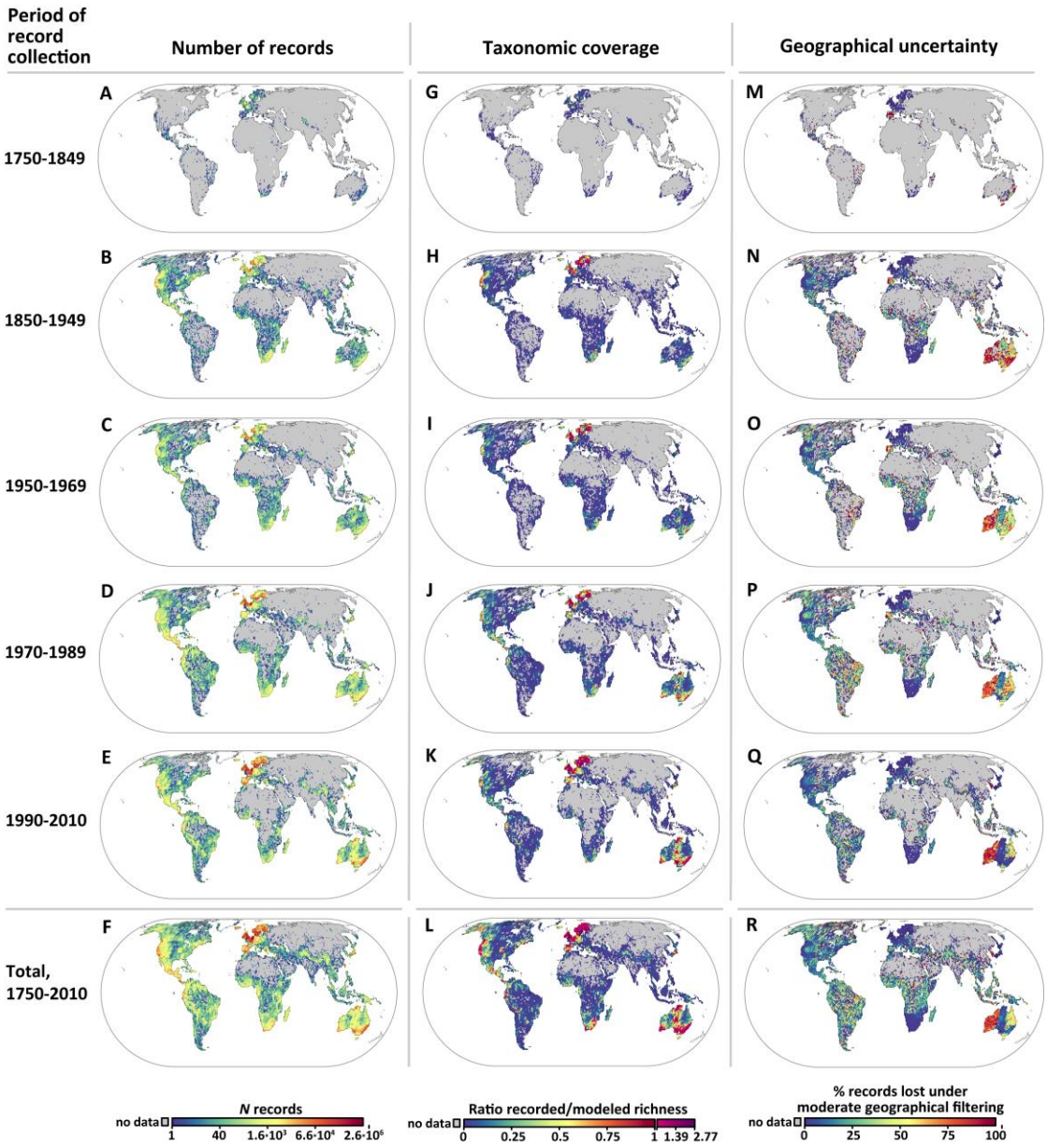


Figure V.1.S4. Spatio-temporal patterns in digital accessible occurrence information. Maps show geographical patterns of three exemplary aspects of vascular plant occurrence information across five time periods between 1750 and 2010 in, and for the entire time span. (A-F) Record number; (G-L) *Taxonomic coverage* for vascular plants (recorded richness (GBIF) / modeled richness (Kreft & Jetz, 2007)); values >1 mean larger recorded than modeled richness; note that mobilized records include non-native species whereas the model predicts native species richness; (M-R) Percentages of records excluded by moderate *geographical uncertainty* filtering (see *Methods*). Color scales are the same in (A-F), (G-L), (M-R).

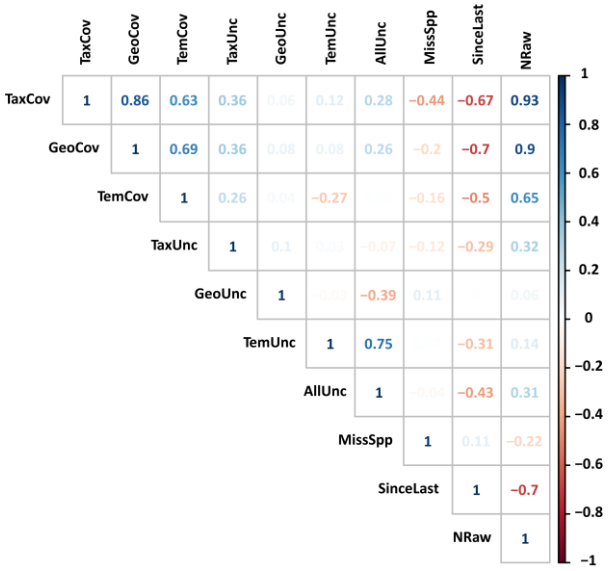


Figure V.1.S5. Relationships between 9 aspects of occurrence information and the number of raw data. Pairwise Spearman-rank correlations between geographical patterns of different occurrence information metrics at the level of 110 x 110 km grid cells. **TaxCov**: *taxonomic coverage*, calculated as the ratio between recorded richness and richness modeled by (Kreft & Jetz, 2007); **GeoCov**: *geographical coverage*, estimated as the number of sampling locations per 10^4 km² land area; **TempCov**: *temporal coverage* since 1750, estimated as the mean minimum Euclidean distance between all possible months between 1750 and 2010 to their respective closests month with records; **TaxUnc**: percentage of records lost under moderate taxonomic filtering; **GeoUnc**: percentage of records lost under moderate geographical filtering; **TempUnc**: percentage of species lost under moderate temporal filtering; **AllUnc**: percentage of species lost with all three moderate filters applied (see *Methods* for information on filters). **MissSpp**: number of species that are not recorded but expected based on the environment-richness; **SinceLast**: Time (in years) since the last mobilized record was recorded. **NRaw:** number of raw data mobilized via GBIF, included to test whether this simple surrogate is a good indicator of different occurrence information metrics. All correlations based on z-transformed variables.

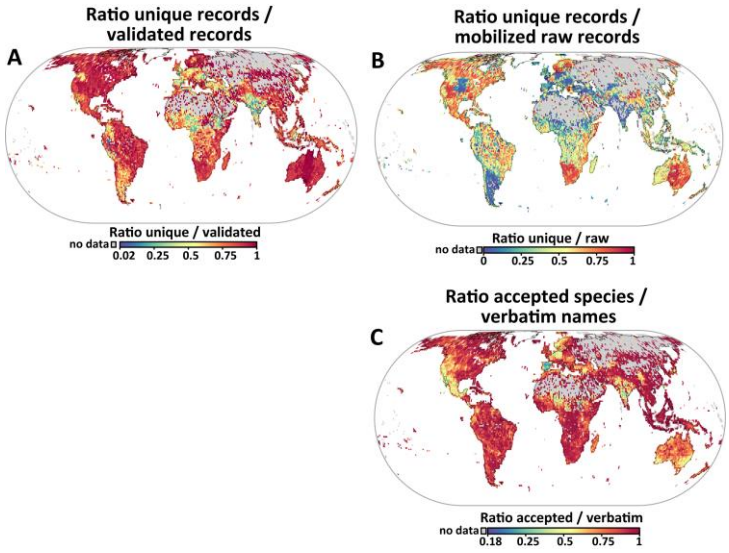


Figure V.1.S6. Different aspects of effectiveness of mobilized occurrence information. (A) Ratio between numbers of unique validated records and overall validated records (including duplicated species-location-month combinations); Cooler colors denote areas where duplicates make up a high percentages of all validated occurrence information. (B) Ratio between numbers of unique validated records and mobilized raw records (including invalid and duplicated records); Cooler colors denote areas where large percentages of all mobilized records were excluded as invalid records or duplicates. (C) Ratio between numbers of accepted species and verbatim scientific name strings (including invalid names and synonyms); cooler colors denote areas where many verbatim name string relative to number of species are in use, and more effort is thus necessary for interpreting those name strings.

Supplementary information - Chapter 2

Global priorities for an effective information basis of biodiversity distributions

Carsten Meyer, Holger Kreft, Robert P. Guralnick and Walter Jetz

Methods

1.A. Species distribution data

Range Data

We considered all species of terrestrial birds (excluding pelagic feeders, N = 9,712; BirdLife International (2011), terrestrial mammals (excluding cetaceans, pinnipeds and sirenians; N = 5,270; IUCN (2010), and amphibians (N = 6,188; (Frost, 2012). We projected expert based extent-of-occurrence range maps for these 21,170 species (IUCN, 2010; Jetz *et al.*, 2012b) into an equal area projection and overlaid them with four nested equal-area grids with grain sizes of c. 110 km, 220 km, 440 km, and 880 km, respectively, at the equator. These range maps were originally drawn by species experts based on a variety of data sources, including point records, local inventories, and atlas and literature data. We considered a grid cell as occupied by a species, if any portion of its range map overlapped with it, and chose 110 km as the finest resolution to minimize false presences (Hurlbert & Jetz, 2007; Hawkins *et al.*, 2008; Hortal, 2008). We excluded 110 km grid cells that did not have at least 30% land area unless they included oceanic islands, in order to minimize effects of area and imprecise range maps while keeping most range-restricted species in the analyses. We further excluded grid cells of which the majority of the land area overlapped with mangrove biomes. This led to the exclusion of 51 narrow endemics near coast lines (not included in the above species count). We overlaid the gridded range maps to define expert-opinion species richness.

Point occurrence records

We focused on records aggregated by the Global Biodiversity Information Facility (GBIF) as a representation of international efforts to mobilize biodiversity data into 'digital accessible information' (DAI; originally referred to as DAK in Sousa-Baena *et al.* (2014a)). GBIF is by far the largest such effort in geographic and taxonomic scope (Edwards, 2000; Graham *et al.*, 2004) and GBIF-facilitated data have been used to assess progress towards Aichi target 19 (Tittensor *et al.*, 2014) We received 192,637,611 geo-referenced records for birds, mammals and amphibians from GBIF in October 2012, of which we extracted 192,463,144 records with potentially sensible geographic coordinates (Longitude: -180° – +180°, Latitude: -90° – +90°) reported with a precision of at least one tenth of a degree. We excluded 8,861,041 records that did not have either a binomial or trinomial scientific name, 278,107 records for which the 'basis of record' field did not indicate 'preserved specimen', 'observation', or 'unknown' (most of which are observation records), and 9,865 records that were reportedly collected before the year 1850, leaving 183,488,598 records. We validated these taxonomically and geographically (see below), which left 157,086,248 records for further analyses.

Taxonomic and geographic validation of records

We then matched the taxonomies of records and range maps. To maximize the amount of records that would pass taxonomic standardization, we combined information on accepted names and synonyms from seven existing taxonomic databases (see below). We accepted species delimitations following BirdLife International (2011) for birds, IUCN (2010) for mammals, and Frost (2012) for amphibians. To each accepted species name, we linked further scientific names fully or partly included in the respective species concept from the above and four further databases (Wilson & Reeder, 2005; IUCN, 2010; ITIS, 2012; Lepage, 2012), including synonyms, subspecies, and common typographical variants. Via this "synonym table", we linked records to the accepted species. We excluded records likely referring to domesticated forms. We inferred the taxonomic identities of records with ambiguous scientific names (such as *pro parte* synonyms) from spatial overlays with the range maps of all accepted species to which the name could potentially refer. In further analyses, we only used records of which the species identity could be unambiguously determined because they fell inside the gridded range maps (at 110 km grain) of only one accepted species. This led to the exclusion of 13.9 to 29.0% "false" or unclear records (see table below). By validating localities of records against expert-opinion range

maps, we ensure that records are biologically plausible and do not refer to zoo or invasive animals outside of their native ranges. We note that this approach may lead to the exclusion of "good" records collected outside of range maps if the maps are inaccurate. While coordinate transposition of geographically false records and "fuzzy matching" of names would have decreased the number of excluded records marginally (Yesson *et al.*, 2007; Otegui, 2012) this would also have increased the uncertainty associated with the validity of records (Yesson *et al.*, 2007).

The table below shows results of the geographic and taxonomic validation of records.

Taxonomic group	*N* records	Linkable to DB	Not accepted name	Ambiguous name	Validated records
Birds	177,067,882	176,698,744	16,830,672	26,210,816	152,429,094
	(100%)	(99.8%)	(9.5%)	(14.8%)	(86.1%)
Mammals	4,725,561	4,708,363	625,540	308,662	3,355,082
	(100.0%)	(99.6%)	(13.2%)	(6.5%)	(71.0%)
Amphibians	1,695,155	1,689,766	416,666	642,943	1,302,072
	(100.0%)	(99.7%)	(24.6%)	(37.9%)	(76.8%)
Total	183,488,598	183,096,873	17,872,878	27,162,421	157,086,248
	(100%)	(99.8%)	(9.7%)	(14.8%)	(85.6%)

Results of the geographic and taxonomic validation of records: Of the geo-referenced specimen and observation data with a binomial or trinomial scientific names that passed initial filtering (see '*N* records'), between 99.6 and 99.8% could be linked to our taxonomic database (see 'Linkable to DB'). Between 9.5 and 24.6% of records are stored under a name that is not an accepted species name according to our three "master" taxonomies, e.g., a synonym or subspecies name, and thus required taxonomic name standardization (see 'Not accepted name'). 6.5 to 37.9% of records had ambiguous names, i.e., accepted names or synonyms that could refer to more than one accepted species, and thus required combined taxonomic and geographic inference to determine the most parsimonious species identity (see 'Ambiguous name'). 71.0 to 86.1% of records remained after taxonomic and geographic validation, i.e., the record could be confidently assigned to one accepted species, and was also collected within the presumed current distribution of that species (see 'Validated records').

Record density and inventory completeness

We overlaid the validated records with the same grids as the range maps. For each grid cell, we then calculated record density as the number of records per 10,000 km² land area and inventory completeness as the percentage of expert-opinion species richness documented by records.

1.B. Geographic and socio-economic variables explaining inventory completeness

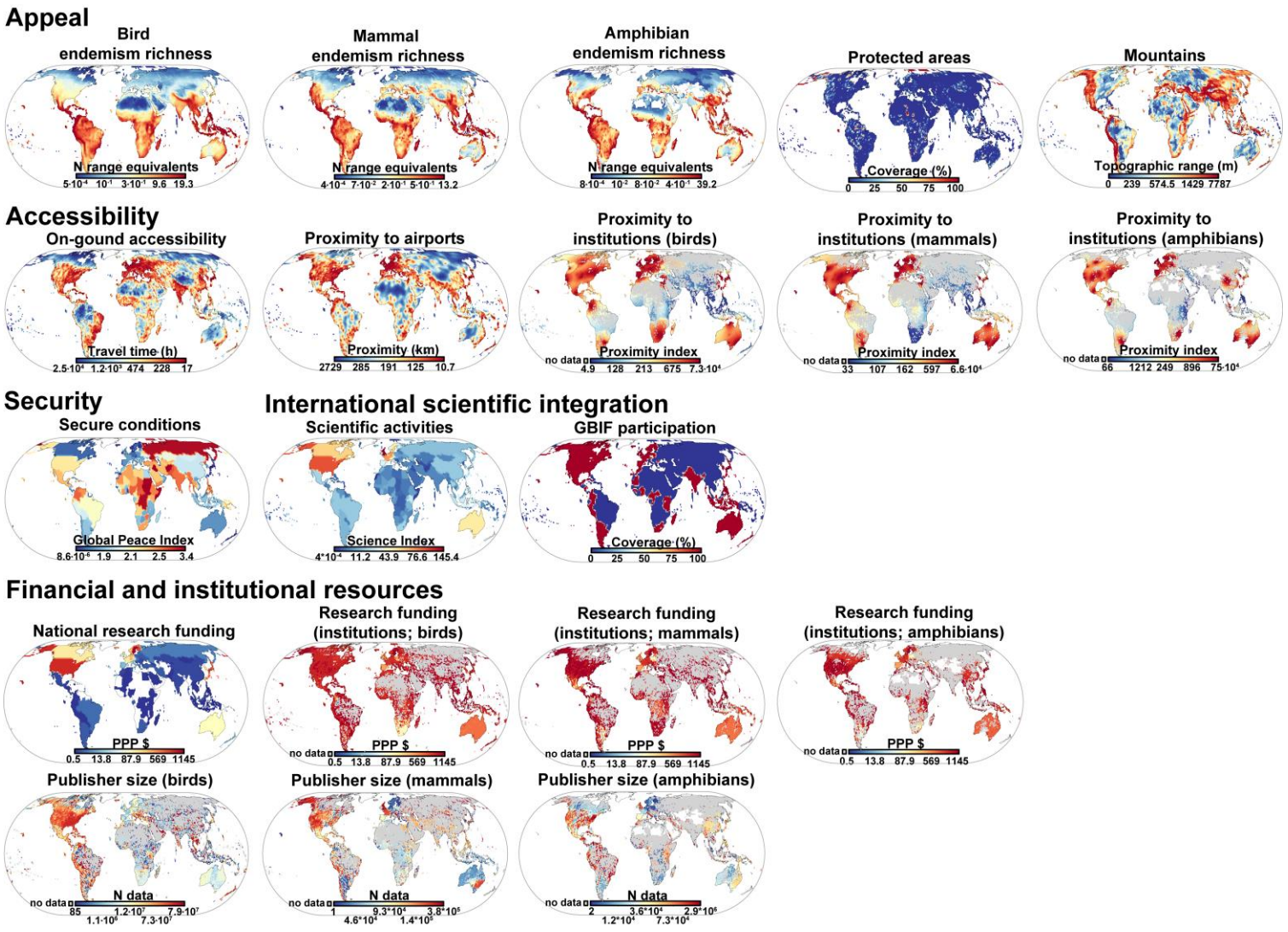
We analyzed the relationships of twelve different geographic and socio-economic factors with record density and inventory completeness. These represent a wide range of existing hypotheses that can be categorized into five broader categories: 1) appeal, 2) accessibility, 3) security, 4) international scientific integration, and 5) financial and institutional resources (for details see maps and discussion of variables below). We limited collinearity among predictor variables by only including variables with Pearson's correlation coefficients ≤0.7 (Dormann *et al.*, 2013).

Most data were available at spatial grains ≤0.25° and aggregated as arithmetic means for the grid cells. We created a few variables from country-level data sets, namely security, national research funding, integration into scientific activities, and GBIF participation (see below). We assumed that the effects of these factors on biodiversity sampling and data mobilization efforts would be similar throughout a given country, and thus used the same value for each grid cell within the country. For grid cells overlaying several countries, we calculated the arithmetic mean of the respective country values weighted by the proportion of land area that falls within each country. We based the definition of country boundaries and the calculation of land area on the polygons of the GADM database (www.gadm.org/version1). We assigned disputed areas to the country currently having *de facto* administrative control.

The figure on the opposite page shows predictor variables mapped at the 110 km grain.

Endemism richness

Areas with specific biodiversity features are naturally interesting to ecologists and several authors have suggested that collectors frequent areas where they can expect to find many or rare species (Dennis & Thomas, 2000; Kier *et al.*, 2005; Küper *et al.*, 2006; Lobo *et al.*, 2006; Soria-Auza & Kessler, 2008; Boakes *et al.*, 2010). To test whether there is global support for this "diversity tracking" hypothesis (Lobo *et al.*, 2006), we used endemism richness (Kier & Barthlott, 2001), as it combines aspects of both species richness and species' range-sizes within an assemblage. Endemism richness is calculated as the sum of the inverse global range sizes of all species present in a grid cell. We estimated the range of each species as the sum of 110 km grid cells overlaying the respective range map polygon (IUCN, 2010; Jetz *et al.*, 2012a). We assumed a taxonomic focus of most collectors to at least class-level and therefore used avian, mammalian, and amphibian endemism richness, respectively, to predict inventory completeness of the three vertebrate classes. Note that a focus on rare species during sampling (Guralnick & Van Cleve, 2005; ter Steege *et al.*, 2011) or a possible emphasis on type



Maps of predictor variables used to model point record density and inventory completeness (110 km grain). For details on variables and source datasets see *Methods*. The Global Peace Index (variable 'Secure conditions') has high values for insecure conditions and we multiplied values with -1 to test for effects of secure conditions. The variable 'Research funding (institutions)' describes the mean research funding of the countries where the providers of records for a given grid cell are situated. The variable 'Publisher size' describes the mean size (contributed data volume) of the providers of records for a given grid cell.

specimens during digitization could also lead to range-restricted species being disproportionately represented in mobilized data and thus to data being biased towards high endemism areas.

Mountains

Mountains could also draw a special attention of collectors because of their scenic beauty or their elevational habitat gradients and, accordingly, high species turnover and the presence of "mountain specialists" (Parnell *et al.*, 2003; Lobo *et al.*, 2006; Sánchez-Fernández *et al.*, 2008; Soria-Auza & Kessler, 2008; Yang *et al.*, 2014). Conversely, it has been reported that mountains are relatively neglected by collecting efforts in some areas due to their poor accessibility (Funk *et al.*, 1999; Funk & Richardson, 2002). To test for effects of mountains on inventory completeness and record density, we calculated the topographic range within each grid cell as the difference between the minimum and maximum altitude, based on data from the GTOPO-30 digital elevation model (US Geological Survey, 1996).

Protected areas

Protected areas could attract collectors because they may promise "pristine" habitats in otherwise altered landscapes or represent strongholds of rare or sought-after species (Freitag *et al.*, 1998; Parnell *et al.*, 2003; Reddy & Dávalos, 2003; Sánchez-Fernández *et al.*, 2008; Boakes *et al.*, 2010; Martin *et al.*, 2012; Ballesteros-Mejia *et al.*, 2013; Yang *et al.*, 2014). If developed for ecotourism or management, they may also provide the most straightforward access points to ecosystems (Ballesteros-Mejia *et al.*, 2013). To model the effect of protected areas, we calculated the proportion of the land area in each grid cell covered by protected areas of International Union for Conservation of Nature categories I to IV (IUCN and UNEP, 2009). Preliminary analyses demonstrated that using an alternative predictor variable based on all (IUCN and UNEP, 2009) protected areas (thus including more protected areas, e.g. from China) did not alter our conclusions.

On-ground accessibility

Some of the most frequently tested hypotheses regarding sampling bias revolve around the on-ground accessibility of areas to researchers, especially via roads (e.g., the "highway effect" (Soberón *et al.*, 2000) or "road-map effect" (Crisp *et al.*, 2001)). Because the time needed to

access an area on the ground has to be traded off against time spent sampling, collectors often choose to sample close to human population centers (Osborne & Tigar, 1992; Freitag *et al.*, 1998; Funk *et al.*, 1999; Parnell *et al.*, 2003; Reddy & Dávalos, 2003; Diniz-Filho *et al.*, 2005; Kier *et al.*, 2005; Küper *et al.*, 2006; Lobo *et al.*, 2006; Schulman *et al.*, 2007; Boakes *et al.*, 2010) or on-ground transportation routes like roads, railways, navigable rivers and coasts lines (Freitag *et al.*, 1998; Gioia & Pigott, 1998; Funk *et al.*, 1999; Soberón *et al.*, 2000; Hijmans *et al.*, 2000; Crisp *et al.*, 2001; Reddy & Dávalos, 2003; Kadmon *et al.*, 2004; Küper *et al.*, 2006; Lobo *et al.*, 2006; Schulman *et al.*, 2007; Newbold, 2010; Ballesteros-Mejia *et al.*, 2013). These effects have been documented mainly at local to regional spatial scales. While most studies found negative relationships between distance to urban areas and transportation routes, (Yang *et al.*, 2014) have found that in China, the opposite is true at the county scale, i.e. sampling intensity and inventory completeness are negatively correlated with both road and human population density. To test whether on-ground accessibility influences data availability at the global scale, we used the 'Travel time to major cities' dataset (Nelson, 2008), which provides estimates of the time needed to travel to cities with a population >50,000, and which combines data on urban areas, roads, railroads, navigable rivers, shipping lanes, habitat types, etc. We calculated mean values for every grid cell, and reversed arithmetic signs, so that higher numbers in our index corresponded to greater accessibility.

Proximity to airports

Since ecologists often have to travel long distances to their study areas, it is possible that regions more accessible by air travel have been better sampled and therefore have higher record density and inventory completeness (Funk *et al.*, 1999; Ballesteros-Mejia *et al.*, 2013). To estimate the accessibility of areas by air travel, we used data on the locations of >9,300 airports and airfields (Partow, 2003). Areas close to several airports should be more accessible to researchers, and we therefore calculated the mean distance of every grid cell centroid to the five closest airports. Again, we reversed arithmetic signs to create an index where large values correspond to close proximity to airports.

Proximity to research institutions

If sampling is mainly carried out by staff of specimen-housing institutions, then time and money constraints could lead collectors to focus on areas nearby their homes or home institutions, and correspondingly, to administrative areas with research institutions being more thoroughly sampled (Freitag *et al.*, 1998; Funk *et al.*, 1999; Dennis & Thomas, 2000;

Moerman & Estabrook, 2006; Pautasso & McKinney, 2007; Sánchez-Fernández *et al.*, 2008; Ahrends *et al.*, 2011; Yang *et al.*, 2014). This effect has been mostly documented for plants (hence, the "botanist effect"; Moerman & Estabrook, 2006), but it can be hypothesized for any group of organisms.

At the global scale, different aspects complicate testing this hypothesis: First, specimen-housing institutions often have a strong geographical and taxonomic focus. So not all institutions in close proximity to a given grid cell should be considered as potential samplers of its biodiversity. For instance, an institution specializing in bird migrations is unlikely to collect amphibians in a nearby wetland. We therefore created an index based on the distances to those institutions that currently focus or have focused on sampling the respective vertebrate class in the broader geographic region surrounding a grid cell. For a given focal grid cell and vertebrate class, we identified data publishers (i.e., institutions) that contributed records from within 750 km of the grid cell centroid. We geo-located these publishers (to at least 50 km accuracy) and calculated their distance (in km) to the grid cell centroid. When simply calculating the mean distance to those publishers weighted by their relative contribution, we found that the many large European and North American institutions had an overarching effect on the index, and all grid cells in the southern hemisphere emerged as remote, even if situated in close proximity to "southern" institutions. We therefore calculated the proximity of grid cells to the relevant publishers as the weighted mean of inverse distances or "proximities" (in km; multiplied by 10^8 for easier scaling):

$$10^8 * \sum_{i=0}^{n} \left(\frac{\mathrm{RelContrib}i}{\mathrm{D}i}\right)$$

where $\mathrm{RelContrib}_i$ is the relative contribution of the *i*-th publisher to the records from the area and D_i the distance (in km). This index has high values when the majority of data within an area are provided by publishers in close proximity. In preliminary analyses we also calculated the weighted mean of $\log_{10}$-transformed and square root-transformed distances, which yielded very similar results, so we used the best performing index based on AIC.

Our approach differs from that of Amano & Sutherland (Amano & Sutherland, 2013), who tested for the effect of the distance to data aggregators (e.g., the GBIF headquarters in Copenhagen, Denmark) rather than data publishers, and found only a negligible effect for GBIF-enabled data. However, while the big biodiversity data aggregators like GBIF, VertNet, SpeciesLink or eBird provide the infrastructure for linking biodiversity data, they are themselves not responsible for the amount or informational content of the data (this lies with distributed data providers). We therefore excluded data for which the indicated publisher itself is an international data aggregator from the calculation of our index.

Secure conditions

Human hazards associated with armed conflicts, territorial disputes, low levels of public safety or political instability can discourage scientific activities (Bonfoh *et al.*, 2011; Brito *et al.*, 2013) and have been reported or hypothesized to have adverse effects on biodiversity data collection and data administration activities, such that more data are available for areas characterized by secure conditions (Funk & Richardson, 2002; Küper *et al.*, 2006; Collen *et al.*, 2008; Hortal *et al.*, 2008; Boakes *et al.*, 2010; Amano & Sutherland, 2013; Otegui *et al.*, 2013). To test this hypothesis, we used the Global Peace Index (GPI; Institute for Economics and Peace, 2012), which is probably the most inclusive existing index describing the overall state of security within a country (Amano & Sutherland, 2013). We note that this index has several drawbacks. First, it is aggregated at the country level, while real levels of security can vary within countries. It is unclear at which spatial scales security levels would deter collecting efforts (i.e., depending on their risk tolerance and detail of available information, foreign collectors could avoid particular low-security parts of a country or entire geo-political regions). As a further drawback, even though we calculated the mean GPI score across several years, the index is only available for the time period between 2008 and 2012 and may not reflect real or perceived security levels in the 1950s through 1980s where many of the specimen records have been collected. In preliminary analyses, we found that an index of the frequency of armed conflicts from 1946 to 2008, created from more fine-scale data (Tollefsen *et al.*, 2012) was consistently a very poor predictor of record density and inventory completeness for all taxa and spatial grains (results not shown). A third potential drawback is that the GPI is not only based on factors affecting the level of personal safety within a country, but also on the level of militarization, which may be unimportant to collectors. However, potential alternative country-level measures of perceived personal safety that we tested in preliminary analyses ('political stability and absence of violence' (The World Bank, 2012), 'control of corruption' (The World Bank, 2012), physician density (World Health Organization, 2012)) were highly collinear with the GPI, so we restricted our main analyses to this measure. Because high GPI values stand for low levels of security, we reversed arithmetic signs of GPI values with after $\log_{10}$-transformation to create an index of secure conditions, and accordingly hypothesized a positive relationship with both record density and inventory completeness.

Scientific activities

Low levels of record density and inventory completeness in specific countries may also be due to a lack of scientific capacity or expertise (Collen *et al.*, 2008; Boakes *et al.*, 2010), or be the

result of a delayed start and poor international integration into the communication of ecological science due to linguistic reasons (Amano & Sutherland, 2013). Conversely, countries whose researchers actively engage in the communication of science through peer-reviewed publication and are internationally well-integrated through collaborations may also mobilize and share more data via international networks like GBIF. To estimate this integration of a country into international scientific communication and collaborations (or "globalization of science"; Amano & Sutherland (2013)), we used data on peer-reviewed primary literature from the *SCImago Journal & Country Rank*, which assembles publication ranks based on *Elsevier*'s *Scopus* database (SCImago, 2007). We extracted the H-index for every country based on peer-reviewed papers published between 1996 and 2011 in the field 'Ecology, Evolution, Behavior and Systematics', and multiplied it with the proportion of papers resulting from international collaborations, i.e., with authors' home institutions situated in at least two countries.

GBIF participation

Although GBIF represents by far the largest international effort facilitating access to point records, many data holders currently do not share their data or only make them accessible via smaller, mostly national networks. Not sharing available biodiversity data internationally due to, e.g., political, economic, or legal reasons has been identified as a key factor limiting scientific progress (U.S. National Committee for CODATA, 1997), and the availability of readily accessible biodiversity data from many parts of the world (Yesson *et al.*, 2007; Thomas, 2009). One of the main strategic goals of GBIF for the coming years therefore is winning the support and cooperation of as yet non-participating countries (GBIF, 2011). To test whether cooperation of countries with GBIF is important in limiting biodiversity information from their territories, we used the proportion of the land area within each grid cell that is covered by a GBIF-participating country (as of April 2013, information from GBIF website).

National research funding

Locally available financial resources have been shown to be an important factor limiting scientific activities in developing countries (May, 1997; King, 2002) and are thus a frequently hypothesized reason for low availability of biodiversity data (Soberón & Peterson, 2004; Collen *et al.*, 2008; Newbold, 2010; Ahrends *et al.*, 2011; Martin *et al.*, 2012; Amano & Sutherland, 2013). To estimate the financial resources that are potentially available for

biodiversity research, we gathered information on the per capita gross domestic expenditure (in purchase power parity dollars) on research and development (GERD; Palmer (2011); UNESCO Institute for Statistics (UIS), 2012). Most other studies have used measures of economic activity such as per capita GDP. Although biodiversity-related funding only makes up a tiny fraction of GERD, research and development spending is generally more closely tied to scientific activities and scientific output than GDP-based measures (May, 1997), and we believe it to be a better proxy for resources that are available for biodiversity studies. We assumed that research grants are mostly available from national funding institutions, and that every grid cell within a country has a similar likelihood of obtaining money for biodiversity data collection and mobilization. We therefore assigned the same GERD value to every grid cell within a country. We restricted our models to those grid cells with at least 70% of their land area covered by countries with available GERD data, which led to the exclusion of some grid cells, particularly in Africa and Asia (see maps of included grid cells and predictor variables above). Preliminary analyses in which we replaced GERD by per capita GDP (CIA, 2013) as an estimate of research funding and thus included more grid cells showed that it was indeed a poorer predictor of both record density and inventory completeness, but otherwise did not alter our conclusions.

Research funding of institutions

Data collection within a particular area as well as their mobilization is often carried out by staff of foreign research institutions. Therefore research funding available in the countries of those institutions that actually contribute data from that area may be a more plausible limiting factor for DAI than locally available funding. A survey on the challenges involved in specimen digitization among the natural history community (Vollmar *et al.*, 2010) found funding to institutions (or related institutional aspects such as technical infrastructure or number and expertise of staff) to be the main factor limiting specimen digitization and biodiversity data mobilization (see also Collen *et al.* (2008)). To test whether this factor limits record density and inventory completeness globally, we created an index based on GERD data in data publisher countries (see above, GERD data available for all 31 countries with data publishers that have contributed records used in this study). We linked to every data publisher the GERD value (in purchase power parity dollars) of the country where it is located. For each grid cell, we then calculated the mean GERD of data publishers, weighted by their relative contribution to the records from the respective grid cell:

$$\sum_{i=0}^{n} (\mathrm{RelContrib}i * \mathrm{GERD}i)$$

where $RelContrib_i$ is the relative contribution of the *i*-th publisher to the records from the grid cell and $GERD_i$ the GERD in the country where the *i*-th publisher is located. We acknowledge that research institutions within a given country may differ in their ability to attract funding, and chances of securing funding for data mobilization may depend more on the existence of specific funding programs (such as the National Science Foundation's 'Advancing Digitization of Biodiversity Collections' initiative) than on among-country differences in GERD.

Publisher size

By definition, larger research institutions have larger quantities of data. Additionally, they often have more resources available for sampling and curatorial activities as well as more and highly specialized staff, combining a greater variety of research foci and taxonomic expertise than smaller institutions (Poliseli & Christoffersen, 2012). Some large North American and European institutions are also reported to have more important collections from Africa, Asia and South America than smaller local institutions because they were involved in extensive biodiversity inventory programs in those regions (Lavoie, 2013). Accordingly, data provided by these institutions should include specimens of more and rarer species (Longino *et al.*, 2002; Guralnick & Van Cleve, 2005; Boakes *et al.*, 2010; Lavoie, 2013), leading to higher levels of inventory completeness in regions where they are or have been active. On the other hand, Chauvel *et al*. (2006) also highlight the value of specific information added only by smaller institutions. Yesson *et al*. (2007) suggested that a focus on large institutions would most efficiently fill gaps in global, digital accessible information, and a focus on the largest North American and European collections is part of GBIF's strategic plan for 2012-2016 (GBIF, 2011). To test whether the size of contributing institutions is limiting record density and inventory completeness in their focal areas, we created an index based on the mean size of institutions that are active within a particular grid cell, weighted by their relative contributions:

$$\sum_{i=0}^{n} (\mathrm{RelContrib}i * \mathrm{V}i)$$

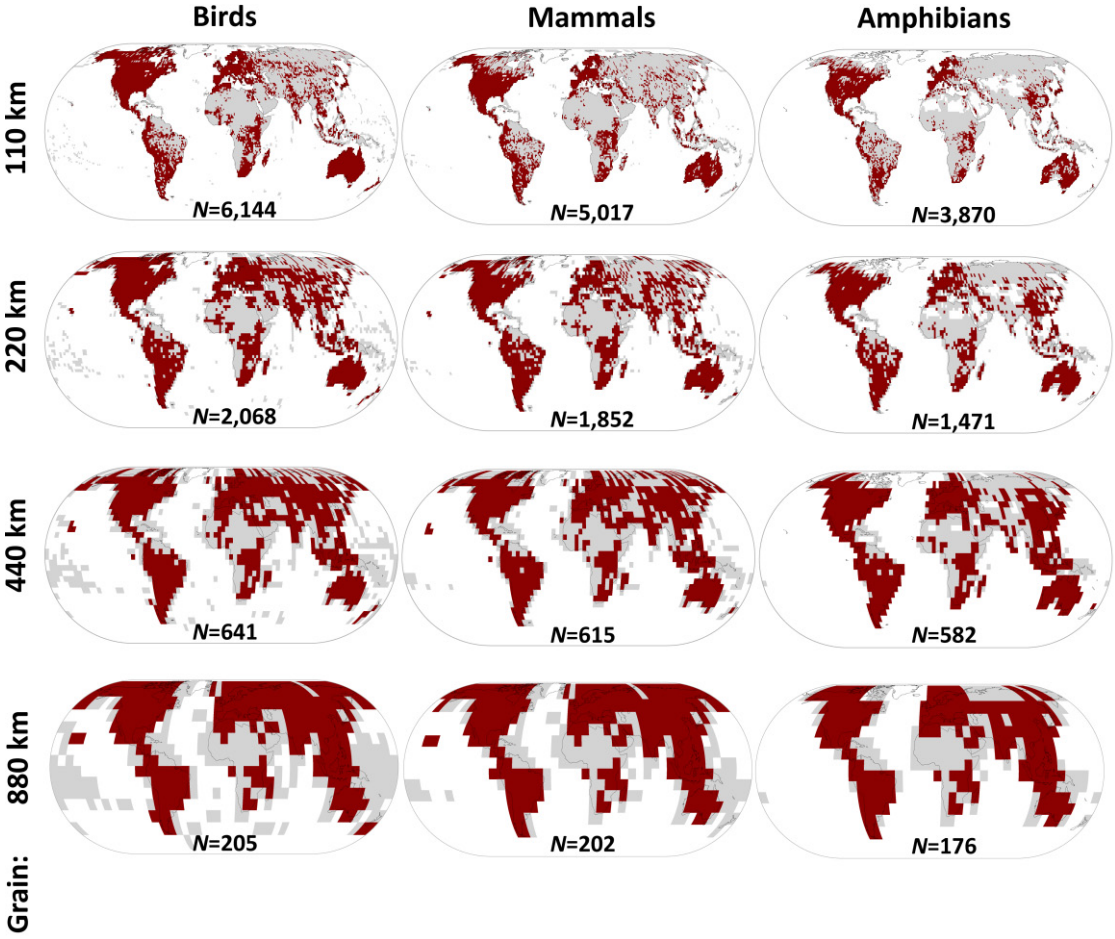
where $RelContrib_i$ is the relative contribution of the *i*-th publisher to the records from the grid cell and V_i the total data volume that the *i*-th publisher contributed to GBIF (as of Oct 2012). We acknowledge that different institutions have advanced to different degrees in terms of mobilizing their data into DAI (Ariño, 2010), which could potentially bias our estimation of publisher size. However, no reliable information of the size of all institutions that contribute

data to GBIF is currently available (compare Ariño (2010)). Record counts of data publishers are summarized in Table V.2.S7.

1.C. Statistical methods

We compared the mean completeness among regions using max-*t* tests (Herberich *et al.*, 2010), and *P*-values were adjusted to geographically effective degrees of freedom following Dutilleul (Dutilleul, 1993).

We investigated the effects of the predictor variables on record density and inventory completeness with simple and multiple regression analyses and built regression models separately for amphibians, birds and mammals at each of four spatial grains (110 km, 220 km, 440 km, 880 km). Because some explanatory variables were calculated using information from the records (e.g., 'Proximity to institutions'), we only included grid cells with at least one record (see figure below).



Grid cells selected for models of point record density and inventory completeness. Dark red cells were considered in models, grey cells were not considered although the taxonomic group is present because they either had no records or no data for all predictor variables was available. At the bottom part of each map the number of grid cells in the respective models (N) is shown.

Before entering the models, record density as well as all predictor variables were $\log_{10}(x + k)$-transformed, with a variable-specific constant k added to each value x, so that the smallest value before $\log_{10}$-transformation equaled 1 (Belmaker & Jetz, 2011). Predictor variables with values bound between 0 and 1 ('Protected areas', 'GBIF participation') were arcsine-square root-transformed before $\log_{10}$-transformation. To account for bias due to area-effects, we included the $\log_{10}$-transformed land area within each grid cell as a covariate in all multiple regression models (highly significant in all cases).

We modeled effects on record density with non-spatial linear models (ordinary least squares) as well as "spatial" simultaneous autoregressive models (SAR) of the error type, which account for spatial autocorrelation (SAC) in the residuals (Kissling & Carl, 2008), using functions from the R package *spdep*. We used non-spatial and spatial GLMs with a binomial distribution and a logit link to model effects on inventory completeness, which entered the model as a composite variable: *cbind*('species covered by GBIF', 'species not covered but presumed present') in *R* terminology. The spatial GLMs were formed by first running a given non-spatial model, and then calculating the 'residuals autocovariate' (RAC) using the *spdep*-function *autocov_dist*, based on a specific neighborhood structure (a list of neighborhood cells to each grid cell) and the residuals of the non-spatial model. The RAC was then entered in the model as a covariate and accounted for SAC in the model residuals (Crase *et al.*, 2012), similar to an error-type SAR. We used the global Moran's I test to determine the degree of SAC (Belmaker & Jetz, 2011). Significant SAC in model residuals often persisted in the spatial models but was reduced by about one order of magnitude compared to non-spatial models (see Moran's I values in Table V.2.S3).

To represent simple associations of predictor and response variables, we ran single-predictor models (non-spatial and not including log-transformed land area as a covariate) and report the coefficient of determination and deviance explained, respectively, for OLS and GLMs (Fig. V.2.S3, Tables V.2.S3-5). We assessed model fit of the minimum adequate models (MAMs) as the % deviance explained (D^2) in the case of RAC models (spatial binomial GLMs; Table V.2.S3 b) and as Pseudo-R^2 in the case of SAR models (Table V.2.S3 b). To test for potential country effects that would remain after controlling for the main 12 predictor variables, we added countries as an additional factor to the spatial MAMs and assessed the increase in model fit (Table V.2.S4).

Long computation times due to the large amount of predictor variables and high numbers of grid cells made it unfeasible to run all possible spatial models. For both inventory completeness and sampling effort, we instead first ran all possible non-spatial multiple-

regression models. We then identified all model subsets that would likely be among the minimum adequate spatial models (with a ΔAIC <10 to the MAM) and only re-ran those models as spatial models.

Both SAR and RAC models require defining a neighborhood structure that defines the distance over which SAC occurs in model residuals. For each grain, we identified the range of distances that would define a neighborhood structure with a median of 8 (~ one cell row) to 24 (~ two cell rows) neighbor cells around focal cells. We then re-ran all candidate model subsets as spatial models for each of five different neighborhood structures based on five distances within that range: for the 110 km grain 200, 250, 300, 350, and 400 km, for the 220 km grain 400, 500, 600, 700, and 800 km, for the 440 km grain 800, 1,000, 1,200, 1,400, and 1,600 km, and for the 880 km grain 1,600, 2,100, 2,600, 3,100, and 3,600 km.

We also investigated interactions and non-linear effects, and although many were significant, accounting for them did not greatly alter model fit or parameter estimates of the main effects in preliminary analyses. To maintain as much simplicity as possible with twelve predictor variables, we therefore decided to focus on the main effects.

Relative importance of predictor variables

For each taxon and grain, we identified the minimum adequate spatial models based on AIC scores. We report the standardized coefficient (β) of the most strongly supported spatial MAM (i.e., with lowest AIC score) in Fig. II.2.3 and Fig. V.2.S3, and where applicable, the range of the standardized coefficient among all potential spatial MAMs (with ΔAIC <2 to the lowest AIC score) in Tables V.2.S3-5. Where the model with the lowest AIC score did not include a factor, we report the standardized coefficient of the "second-best" model (if among the potential MAMs, Tables V.2.S16-S23). If none of the potential MAMs had a particular factor, it was left blank in Figures II.1.3 and V.1.S3.

As an alternative measure of relative importance, and considering all possible subsets of the full non-spatial model as experimental units, we carried out ANOVAs with a response variable consisting of the AIC scores of all possible models and predictor variables formed as dummy-variables coding for every factor whether or not it is in the respective model. The percentage of the total Sums of Squares (% SS) attributable to each factor corresponds to their relative importance (compare Dormann *et al.* (2008); Diniz-Filho *et al.* (2009)).

1.D. Limitations of this study

Biodiversity data sources

With GBIF and the many integrated data sources (see Table V.2.S7) we cover by far the largest share of global digital accessible information on biodiversity. However, several global and regional data mobilization initiatives provide access to digital data, but do not currently make their data accessible via GBIF. Further, several regions have digital or non-digital data that are not shared. We fully acknowledge many data collation programs play important roles in facilitating biodiversity analyses and progress towards Aichi target 19. Several initiatives address data types that inform about other aspects of critical relevance for conservation, such as species' abundances (NERC Centre for Population Biology - Imperial College, 2010), ranging behaviour (Wikelski & Kays, 2015), or conservation status (IUCN, 2010).

Explanatory variables

A general shortcoming of our study is that we had to rely on fairly recent socio-economic datasets. We investigated time series of collected data volumes per 5-year period which showed that the majority of records (i.e., including both observation and specimen records) have been collected in recent decades, but specimens in particular were often collected several decades ago (median recording year for amphibians: 1979; for mammals: 1989; for birds: 2007). We implicitly assumed that among-region differences in factors relating to field sampling, like on-ground accessibility, protected areas, and levels of research funding, have on average been similar at the times when data were collected. As digitization and sharing of these records happened mostly within the last decade, record age does not affect our conclusions regarding the main factors currently limiting DAI. However, spatiotemporal changes in sampling activities in relation to historical factors (e.g. roads, reserves) is a needed area of further study.

With the factors included in this study, we attempted to cover a wide range of existing hypotheses on the drivers of data bias and inventory completeness in global DAI. However, we note that original collection, digitization, mobilization, and sharing of data may be influenced by further contemporary and historical socio-economic factors, such as political systems and agendas, levels of bureaucracy and international cooperation, policies of funding agencies, and legal aspects (U.S. National Committee for CODATA, 1997; Küper *et al.*, 2006; Vollmar *et al.*, 2010; Schäfer *et al.*, 2011), information technological capacity (Ariño *et al.*, 2011), lingua franca (Schulman *et al.*, 2007; Amano & Sutherland, 2013), colonial history (Figueiredo & Smith, 2010; Ballesteros-Mejia *et al.*, 2013; Lavoie, 2013), traditions of natural

history institutions and personal preferences of collectors and curators (Pyke & Ehrlich, 2010), as well as attitudes of countries and data owners towards data-sharing (Enke *et al.*, 2012; van Panhuis *et al.*, 2014). Most of these effects are difficult to quantify, and existing country-level datasets are often highly collinear. Some of these effects, however, may become visible in the form of country effects, not least because data mobilization to GBIF is organized via national nodes. However, many countries have experienced extreme political transitions as well as changes in their sovereign territory over the course of time when data have been collected, and effects of modern country identities on record density and inventory completeness may be difficult to interpret for many parts of the world. We therefore decided not to perform hierarchical mixed effects models with countries as a random factor, but instead only assess the increase in model fit if a 'country' factor was added to the minimum adequate multi-predictor models.

Supplementary Figures

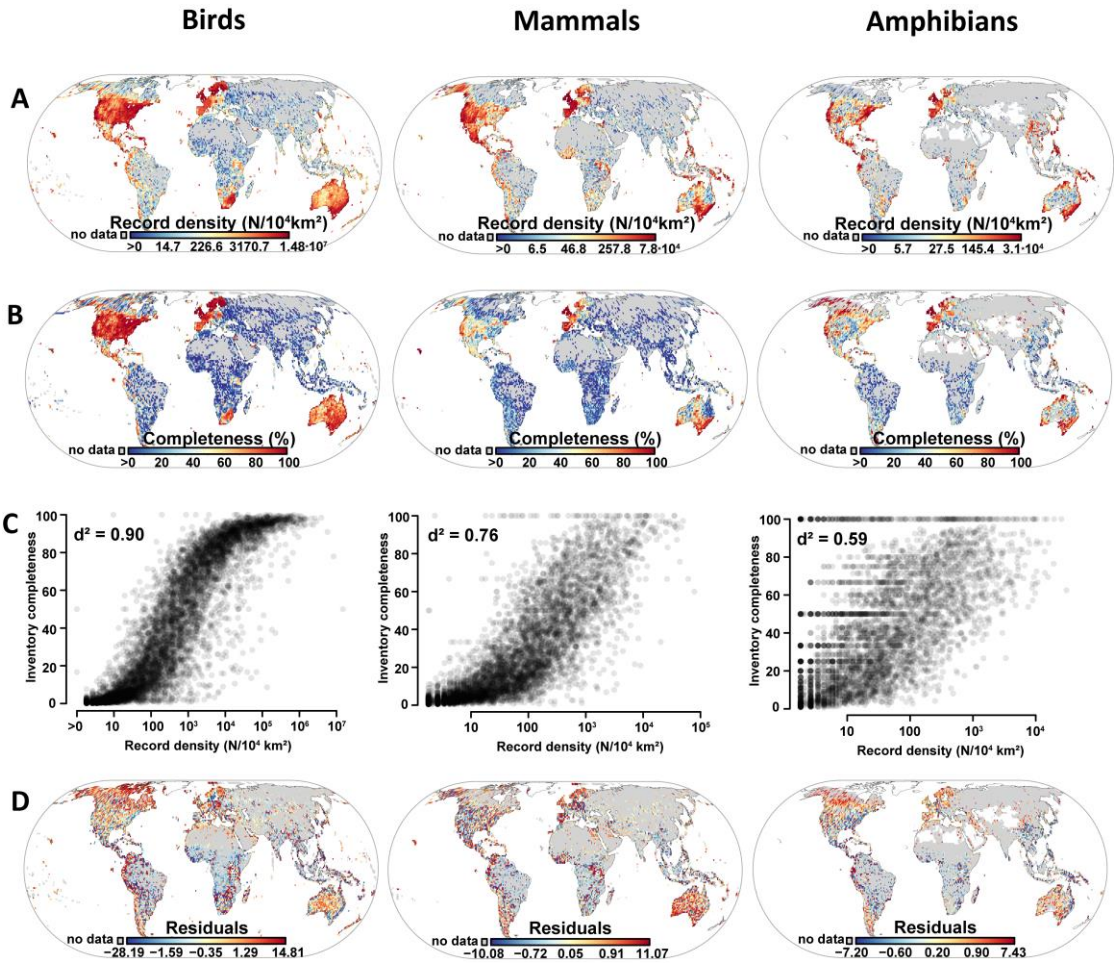


Figure V.2.S1. Relationships between record density and inventory completeness in global 'digital accessible information' for three vertebrate groups at the 110 km grain. A) Record density, B) Inventory Completeness, C) Scatter plots of relation between inventory completeness and record density with deviance explained (d^2) based non non-zero grid cells, D) Spatial arrangement of residuals of a binomial generalized linear model (logit link) explaining inventory completeness with record density. Red values indicate higher, blue values lower inventory completeness than expected from record density.

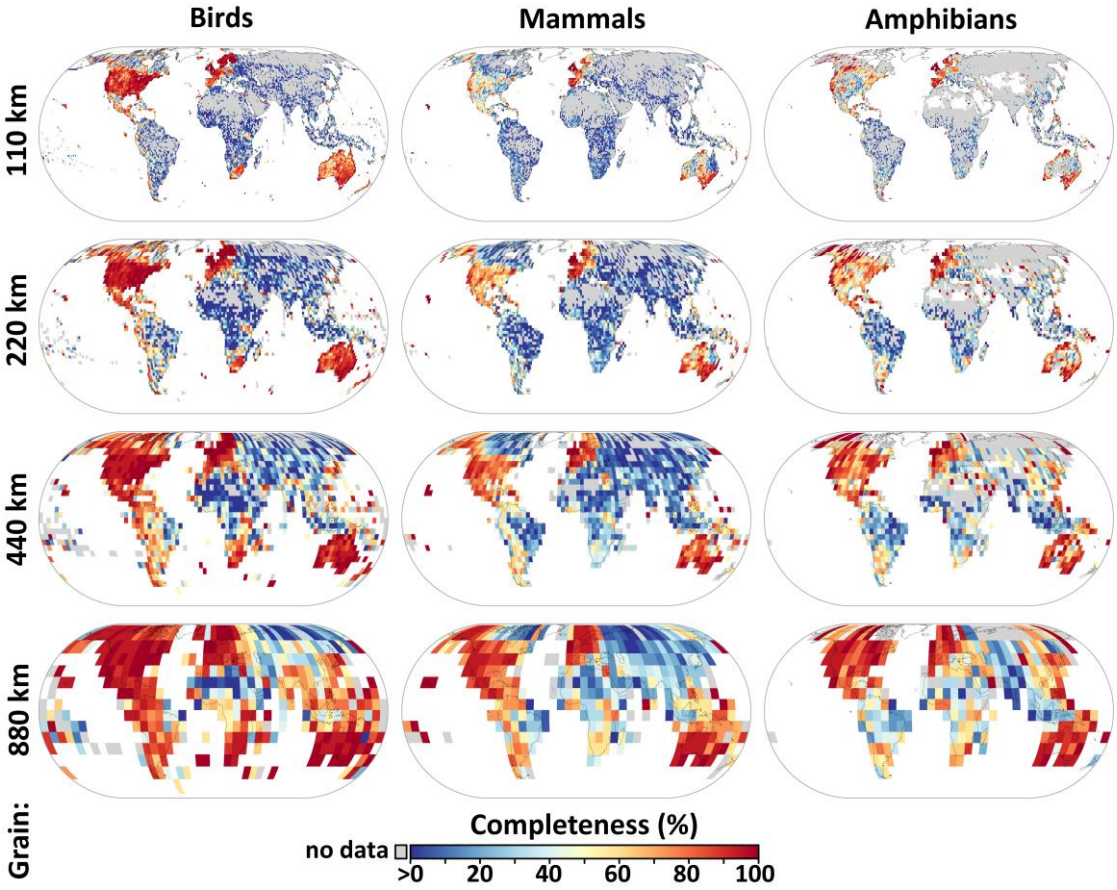


Figure V.2.S2. Spatial variation in record-based inventory completeness for three vertebrate taxa at four spatial grains. Grey grid cells show areas within the global range of the taxonomic group with no mobilized records.

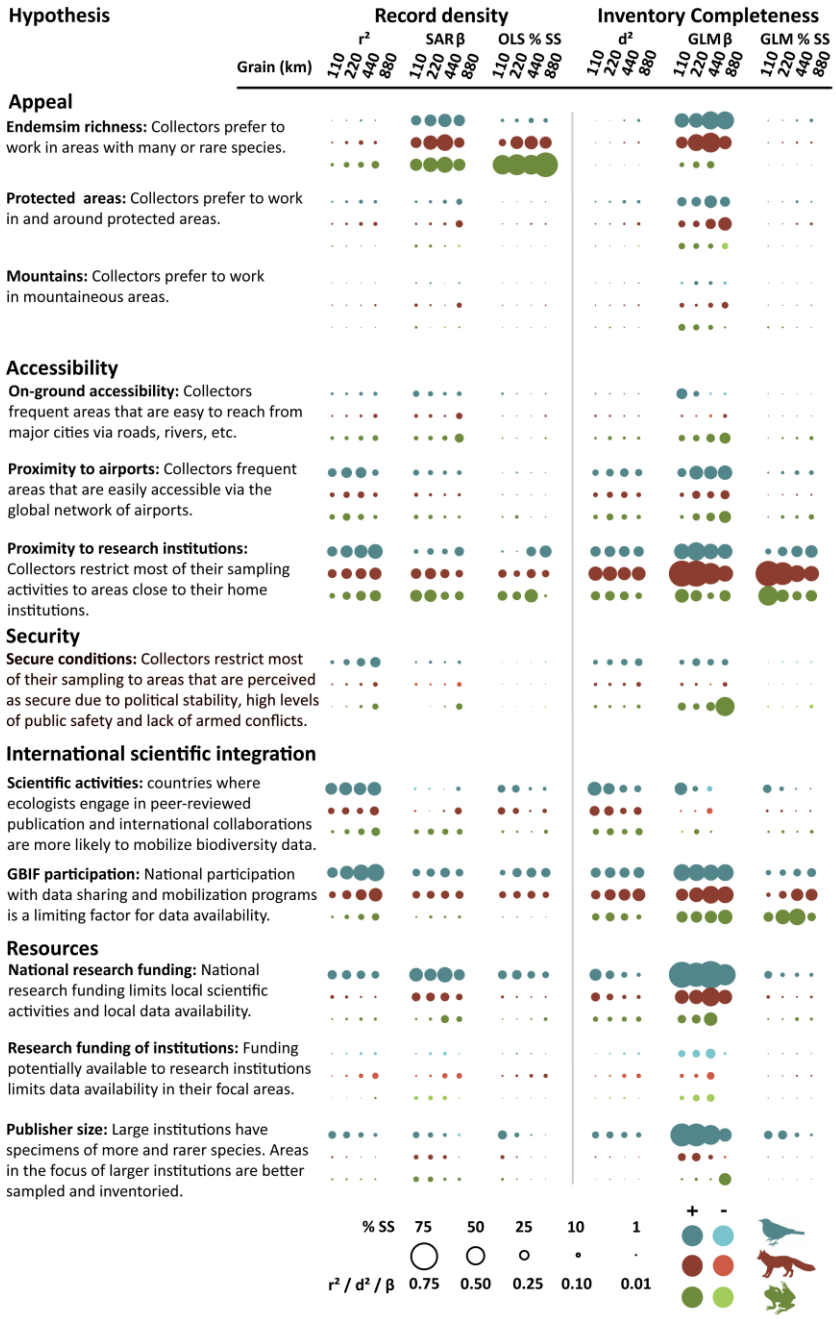


Figure V.2.S3. Determinants of point record density and inventory completeness. Effects were tested in simple and multiple regression models. All model subsets were ranked based on AIC scores and subsets with ΔAIC <10 re-run as spatial models, by accounting for spatial autocorrelation in model residuals. For record density, we used ordinary least squares models and simultaneous autoregressive models (SAR β and OLS % SS). For inventory completeness, we used spatial and non-spatial generalized linear models with a binomial distribution and a logit link (GLM β and GLM % SS). Bubble size represents the strength of predictor-response relationships. Vertebrate groups are represented by color, with shading denoting the direction of the relationship. We show predictor strength for record density using three different metrics: i) the coefficient of determination in simple regressions (r^2), ii) the standardized coefficients of the reduced subset of the spatial multi-predictor model with the lowest AIC score (blank cells indicate variables that were not included in these models) (SAR β), and iii) the percentage each predictor has in the total Sums of Squares (OLS % SS) of a type III ANOVA. For the latter we used AIC values of all possible model subsets as the response variable and dummy-variables coding whether or not a predictor is in the respective model as explanatory variables. We show predictor strength for inventory completeness using three different metrics analogous to those for record density: i) the deviance explained in simple generalized linear regression models (d^2), ii) the standardized coefficients of the reduced spatial multiple generalized linear regression models with the lowest AIC score (GLM β), and iii) the percentage each predictor has in the total Sums of Squares (GLM % SS) of a type III ANOVA.

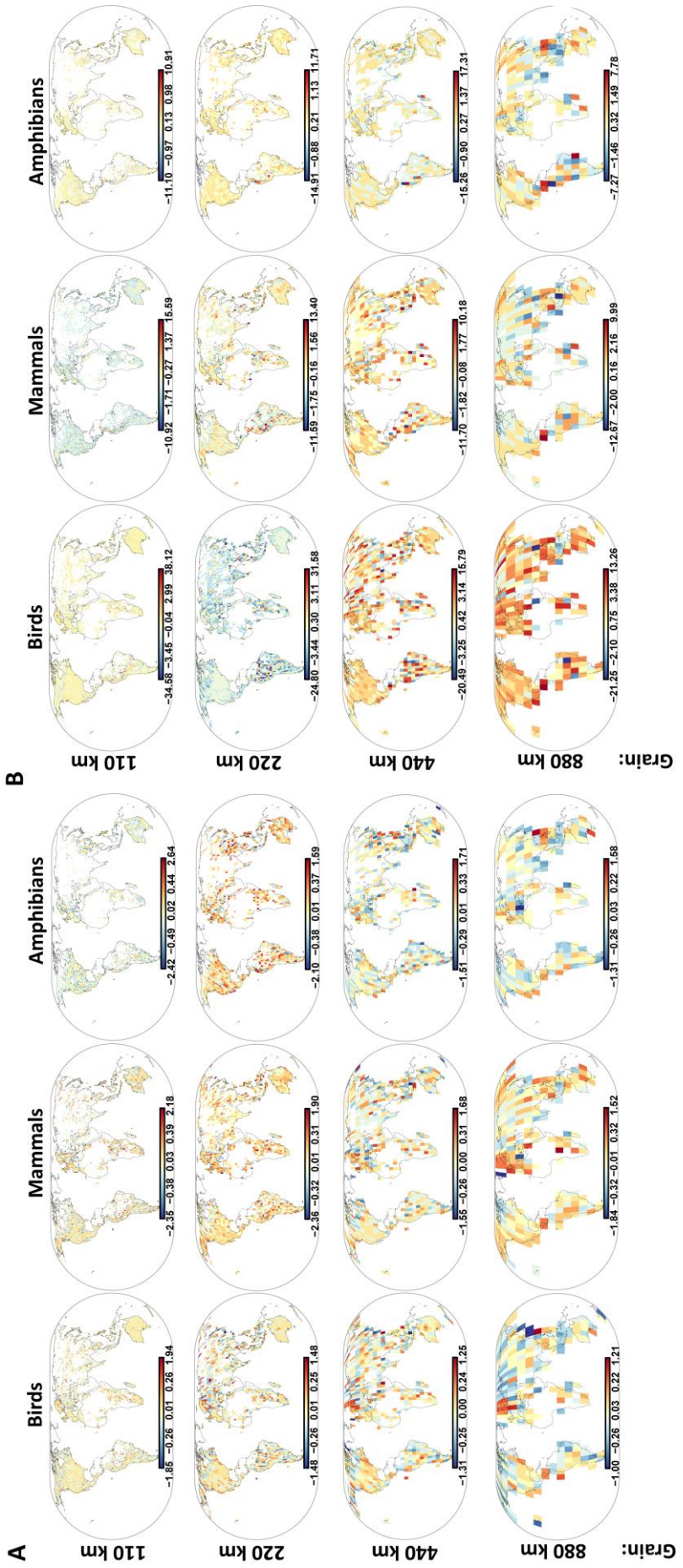


Figure V.2.S4. Spatial arrangement of residuals in **A)** simultaneous autoregressive models of point record density and **B)** spatial binomial generalized linear models (logit link) of inventory completeness. Red values indicate higher, blue values lower inventory completeness than expected from the predictor variables.

Supporting tables

Table V.2.S1. Global correlations between **a)** record density and inventory completeness (based on grid cells with at least one record) and **b)** species richness evident in mobilized occurrence point records ($SR_{records}$) and expected true species richness based on expert-opinion range maps (SR_{expert}). For each taxonomic group and spatial grain (km), the median record density (*N* records/10^4 km²), the median inventory completeness, the Spearman's rank coefficient (r_s), and the number of grid cells (*N* cells) are shown. Asterisks behind r_s represent *P*-values corrected for spatial autocorrelation(Dutilleul, 1993).

a) correlations between record density and inventory completeness

	Grain (km)	median record density	median inventory completeness	r_S	*N* cells
Birds	110	8.61	0.03	0.91***	7,378
	220	48.11	0.22	0.89***	2,863
	440	115.87	0.47	0.85***	1,007
	880	304.46	0.65	0.78***	350
Mammals	110	0.81	0.01	0.82***	5,885
	220	5.66	0.08	0.84***	2,447
	440	14.76	0.24	0.87***	888
	880	33.39	0.43	0.84***	300
Amphibians	110	0.00	0.00	0.57***	4,346
	220	1.83	0.16	0.57***	1,863
	440	4.13	0.36	0.56***	699
	880	13.81	0.50	0.60***	251

b) correlations between GBIF richness and expert richness

	Grain (km)	median $SR_{records}$	median SR_{expert}	r_S	*N* cells
Birds	110	4	193	0.35**	11,757
	220	34	205	0.58***	3,575
	440	83.5	228.5	0.79***	1,136
	880	157	274.5	0.91***	372
Mammals	110	1	52	0.28*	11,522
	220	5	57	0.49***	3,415
	440	16	69	0.69***	1,037
	880	39	92	0.83***	323
Amphibians	110	0	10	0.39***	10,002
	220	2	12	0.61***	2,973
	440	5	16	0.81***	919
	880	14	29	0.91***	280

Table V.2.S2. Variation in 110 km inventory completeness (%) for all three vertebrate groups combined (*N* = 21,170 species) among a) biomes, b) realms, c) biome-realm-combinations (following Olson *et al.* (2001)), and d) countries. Within biomes, realms are ordered from highest to lowest median completeness. Within broad geographical regions, countries are ordered from highest to lowest median completeness. Grouping of countries into geographical regions is for orientation only and does not reflect any view of the authors. Some countries are missing because they did not overlay the majority of the land area of any grid cell. Country codes (ISO 3166 standard) are the same as in Fig. II.2.5.

a) Variation among biomes

Biome	*N* cells	Min	Max	Mean	SD	Median
Tropical & Subtropical Moist Broadleaf Forests	2,214	0.0	100.0	14.1	20.2	3.2
Tropical & Subtropical Dry Broadleaf Forests	374	0.0	96.7	22.7	23.8	14.6
Tropical & Subtropical Coniferous Forests	62	0.4	80.9	46.7	21.8	51.2
Flooded Grasslands & Savannas	75	0.0	86.9	13.1	20.3	1.5
Tropical & Subtropical Grasslands, Savannas & Shrublands	1,637	0.0	100.0	14.4	23.5	1.7
Deserts & Xeric Shrublands	2,369	0.0	96.3	17.8	27.5	0.7
Mediterranean Forests, Woodlands & Scrub	325	0.0	96.1	47.6	31.0	52.2
Temperate Broadleaf & Mixed Forests	1,129	0.0	96.1	38.7	34.6	32.3
Temperate Conifer Forests	320	0.0	88.6	45.2	31.7	58.6
Montane Grasslands & Shrublands	410	0.0	72.1	11.8	19.8	1.5
Temperate Grasslands, Savannas & Shrublands	830	0.0	100.0	29.9	32.0	11.3
Boreal Forests/Taiga	1,317	0.0	94.1	15.9	25.5	0.5
Tundra	775	0.0	100.0	20.5	26.3	3.9

b) Variation among realms

Realm	*N* cells	Min	Max	Mean	SD	Median
Nearctic	1,727	0.0	94.1	49.9	25.6	58.8
Neotropics	1,715	0.0	86.9	19.8	23.2	8.9
Afrotropics	1,817	0.0	100.0	10.6	18.1	1.6
Palearctic	4,539	0.0	96.1	10.0	22.2	0.0
Indomalay	890	0.0	80.0	9.6	14.4	2.1
Australasia	985	0.0	96.3	53.1	29.3	62.3
Oceania	178	0.0	100.0	22.8	31.0	0.0

c) Variation among biome-realm combinations

Biome	Realm	*N* cells	Min	Max	Mean	SD	Median
Tropical &	Australasia	261	0.0	92.9	16.7	20.5	5.3
Subtropical Moist	Neotropics	799	0.0	86.5	16.8	22.3	4.6
Broadleaf Forests	Afrotropics	311	0.0	79.5	11.7	16.8	3.7
	Palearctic	44	0.0	19.6	4.0	4.7	2.8
	Indomalay	645	0.0	80.0	10.2	15.2	2.2
	Oceania	154	0.0	100.0	19.6	29.3	0.0
Tropical &	Nearctic	3	45.7	67.3	54.5	11.4	50.5
Subtropical Dry	Oceania	19	0.0	96.7	43.0	32.3	47.4
Broadleaf Forests	Neotropics	175	0.0	83.6	32.6	24.3	31.5
	Australasia	32	0.0	30.2	12.2	9.3	13.6
	Afrotropics	23	0.0	69.9	19.4	23.5	7.4
	Indomalay	122	0.0	51.0	8.1	11.8	2.0
Tropical &	Neotropics	32	6.6	80.9	55.3	20.2	60.0
Subtropical	Nearctic	22	16.8	74.4	44.2	16.7	43.2
Coniferous Forests	Indomalay	8	0.4	40.0	19.3	16.3	20.4
Flooded Grasslands	Neotropics	23	0.0	86.9	29.1	26.0	26.3
& Savannas	Indomalay	2	0.8	24.4	12.6	16.7	12.6
	Palearctic	19	0.0	36.7	5.2	10.4	0.7
	Afrotropics	31	0.0	45.1	6.1	12.6	0.5
Tropical &	Nearctic	8	67.9	86.4	74.2	5.6	73.0
Subtropical	Australasia	192	2.1	92.3	64.9	13.8	65.3
Grasslands,	Indomalay	1	36.6	36.6	36.6	-	36.6
Savannas &	Oceania	5	0.0	100.0	43.3	46.5	33.3
Shrublands	Neotropics	275	0.0	65.3	9.2	13.8	2.3
	Afrotropics	1,156	0.0	72.5	6.7	13.5	0.8
Deserts & Xeric	Australasia	297	20.3	96.3	65.9	14.6	67.6
Shrublands	Nearctic	198	3.7	87.3	59.8	16.9	64.4
	Afrotropics	214	0.0	72.2	19.8	22.5	9.2
	Neotropics	125	0.0	84.3	20.7	25.1	8.1

Variation among biome-realm combinations (continued)

Biome	Realm	*N* cells	Min	Max	Mean	SD	Median
	Indomalay	90	0.0	43.9	5.6	10.1	0.7
	Palearctic	1,445	0.0	71.4	2.4	7.7	0.0
Mediterranean	Australasia	67	50.9	93.0	78.6	10.3	81.7
Forests,Woodlands	Nearctic	17	7.7	88.1	71.4	20.6	78.9
& Scrub	Afrotropics	8	57.5	78.7	69.8	6.2	71.2
	Neotropics	15	0.0	81.3	51.3	21.8	54.7
	Palearctic	218	0.0	96.1	35.1	28.7	27.2
Temperate Broadleaf	Australasia	73	0.0	94.4	79.1	16.5	82.4
& Mixed Forests	Nearctic	236	9.4	87.5	70.3	9.7	71.5
	Neotropics	43	0.0	79.5	42.2	25.9	47.5
	Indomalay	13	0.0	40.8	17.0	16.6	9.1
	Palearctic	764	0.0	96.1	25.2	32.0	6.9
Temperate Conifer	Nearctic	192	3.9	85.8	66.5	14.8	70.9
Forests	Palearctic	127	0.0	88.6	13.5	22.1	1.9
	Indomalay	1	0.2	0.2	0.2	-	0.2
Temperate	Australasia	49	48.4	90.4	76.8	10.3	79.1
Grasslands,	Nearctic	249	12.7	87.9	67.1	11.0	68.9
Savannas &	Afrotropics	5	0.0	100.0	45.9	49.7	15.5
Shrublands	Neotropics	144	0.0	75.2	18.2	18.1	10.6
	Palearctic	383	0.0	44.3	4.0	8.3	0.4
Montane Grasslands	Australasia	6	66.2	72.1	68.7	2.0	68.5
& Shrublands	Afrotropics	66	0.0	70.1	34.4	27.9	32.7
	Neotropics	62	0.0	63.8	22.7	16.6	21.2
	Palearctic	276	0.0	39.7	2.7	5.6	0.3
Boreal Forests/Taiga	Nearctic	438	0.0	94.1	30.4	23.6	26.3
	Palearctic	879	0.0	91.6	8.6	23.3	0.0
Tundra	Australasia	8	0.0	64.3	37.4	21.0	41.7
	Nearctic	364	0.0	89.1	32.8	23.4	31.8
	Palearctic	384	0.0	94.4	8.1	22.1	0.0

d) Variation among countries

GeoRegion	Country	Code	*N* cells	Min	Max	Mean	SD	Median
South America	Ecuador	ECU	30	0.0	84.3	52.6	22.3	58.6
	Falkland Islands (Islas Malvinas)	FLK	11	0.0	59.6	35.7	20.3	41.8
	Chile	CHL	76	0.0	81.3	36.9	25.5	39.0
	Peru	PER	108	0.0	78.3	29.6	20.3	31.2
	Bolivia	BOL	86	0.0	64.4	23.3	16.6	22.2
	Suriname	SUR	11	0.0	50.1	17.7	15.2	20.6
	Guyana	GUY	20	0.3	65.3	20.1	16.7	18.6
	French Guiana	GUF	6	0.0	29.0	15.1	11.8	18.3
	Paraguay	PRY	31	0.5	67.3	19.5	16.1	17.9
	Uruguay	URY	16	0.6	51.3	20.2	16.7	16.0
	Colombia	COL	98	0.0	68.4	19.0	19.2	13.4
	Venezuela	VEN	80	0.0	64.6	17.7	17.8	10.9
	Brazil	BRA	704	0.0	54.1	5.0	9.9	0.5
Central America/Caribbean	British Virgin Islands	VGB	1	79.9	79.9	79.9	-	79.9
	Puerto Rico	PRI	6	30.8	82.4	70.2	19.7	77.4
	Costa Rica	CRI	6	35.7	86.5	69.9	19.2	76.2
	Belize	BLZ	2	75.3	76.9	76.1	1.1	76.1
	El Salvador	SLV	2	67.7	75.5	71.6	5.5	71.6
	Virgin Islands	VIR	1	70.0	70.0	70.0	-	70.0
	Dominican Republic	DOM	6	64.0	73.2	67.7	3.6	66.4
	Dominica	DMA	1	64.9	64.9	64.9	-	64.9
	Guatemala	GTM	10	34.5	78.2	62.1	12.2	64.6
	Jamaica	JAM	5	53.9	73.3	62.3	7.3	61.8
	St. Vincent and the Grenadines	VCT	1	60.6	60.6	60.6	-	60.6
	Cayman Islands	CYM	3	47.2	84.3	64.0	18.8	60.5
	Netherlands Antilles	ANT	1	57.3	57.3	57.3	-	57.3
	St. Lucia	LCA	1	56.1	56.1	56.1	-	56.1
	Antigua and Barbuda	ATG	1	55.7	55.7	55.7	-	55.7
	Grenada	GRD	1	55.0	55.0	55.0	-	55.0
	Mexico	MEX	182	0.0	87.3	52.4	19.5	54.8
	Panama	PAN	11	10.2	76.5	45.8	23.3	53.8
	Barbados	BRB	1	53.7	53.7	53.7	-	53.7
	Haiti	HTI	3	36.5	58.3	47.5	10.9	47.8
	Martinique	MTQ	2	42.9	49.7	46.3	4.8	46.3
	Honduras	HND	12	0.0	61.2	42.9	17.9	46.0

Variation among countries (continued)

GeoRegion	Country	Code	*N* cells	Min	Max	Mean	SD	Median
	Trinidad and Tobago	TTO	2	34.1	57.6	45.8	16.7	45.8
	St. Kitts and Nevis	KNA	1	43.9	43.9	43.9	-	43.9
	Montserrat	MSR	1	41.0	41.0	41.0	-	41.0
	Nicaragua	NIC	13	0.0	63.6	37.5	21.5	42.7
	Cuba	CUB	16	1.1	61.2	36.8	15.9	40.4
	Guadeloupe	GLP	2	0.0	78.3	39.1	55.3	39.1
	Bahamas, The	BHS	22	0.0	80.9	33.9	26.5	33.4
	Anguilla	AIA	1	24.5	24.5	24.5	-	24.5
	Turks and Caicos Islands	TCA	4	0.0	39.7	17.6	20.2	15.3
Northern America	United States	USA	848	0.0	100.0	64.4	18.9	69.8
	Bermuda	BMU	1	45.5	45.5	45.5	-	45.5
	Canada	CAN	827	0.0	85.6	35.5	24.7	35.0
North/West Europe	Ireland	IRL	9	87.3	96.1	92.9	2.8	93.5
	Denmark	DNK	7	81.5	90.8	85.3	3.3	84.6
	Sweden	SWE	41	73.5	90.2	84.0	3.8	84.4
	Finland	FIN	30	74.2	91.6	84.1	4.4	84.2
	United Kingdom	GBR	33	16.7	94.2	83.6	14.5	88.1
	Norway	NOR	28	65.1	90.0	83.3	5.3	84.0
	Belgium	BEL	2	82.5	85.1	83.8	1.8	83.8
	France	FRA	49	66.4	89.1	79.9	5.2	81.3
	Spain	ESP	61	0.0	96.1	71.0	19.6	75.4
	Germany	DEU	29	41.1	81.5	68.8	10.4	71.1
	Switzerland	CHE	4	49.7	76.2	67.0	11.8	71.0
	Iceland	ISL	11	51.4	80.3	68.3	8.9	69.4
	Netherlands	NLD	3	62.3	81.8	70.4	10.1	67.3
	Austria	AUT	5	16.1	68.1	56.9	22.8	67.0
	Portugal	PRT	17	0.0	72.2	49.5	21.6	55.6
	Malta	MLT	1	26.2	26.2	26.2	-	26.2
	Italy	ITA	35	0.0	51.1	24.9	14.2	22.9
	Svalbard	SJM	29	0.0	94.4	27.4	29.1	17.6
	Greenland	GRL	13	0.0	65.5	17.8	21.3	9.5
	Faroe Islands	FRO	4	0.0	36.8	12.7	16.8	7.0
	Jan Mayen	SJM	3	0.0	11.1	3.7	6.4	0.0
East/South-East Europe	Estonia	EST	5	9.7	85.9	64.5	31.3	75.4
	Slovakia	SVK	4	62.4	68.5	65.4	2.5	65.4
	Poland	POL	27	16.0	85.7	55.1	19.7	60.9
	Hungary	HUN	7	11.4	63.3	42.8	21.3	49.8
	Cyprus	CYP	1	45.3	45.3	45.3	-	45.3
	Czech Republic	CZE	8	19.5	71.7	44.3	21.8	39.6
	Latvia	LVA	7	2.3	48.3	26.1	15.0	31.7
	Greece	GRC	14	12.0	47.3	30.0	12.0	31.2
	Bosnia and Herzegovina	BIH	13	0.5	36.2	18.9	13.9	24.6
	Croatia	HRV	3	12.4	21.2	17.8	4.7	19.9
	Macedonia	MKD	2	13.4	23.9	18.7	7.4	18.7
	Montenegro	MNE	1	18.0	18.0	18.0	-	18.0
	Slovenia	SVN	2	14.6	19.7	17.2	3.6	17.2
	Bulgaria	BGR	9	0.0	36.6	18.7	13.2	16.4
	Lithuania	LTU	5	6.5	32.7	17.1	10.6	13.6
	Moldova	MDA	2	9.5	14.9	12.2	3.8	12.2
	Albania	ALB	4	2.3	40.3	16.6	16.7	11.8
	Romania	ROU	20	0.3	45.8	13.7	13.5	10.8
	Ukraine	UKR	49	0.3	65.3	6.6	11.9	1.2
	Byelarus	BLR	20	0.0	22.8	2.2	5.2	0.4
Australia/Oceania	Wake Island	UMI	1	100.0	100.0	100.0	-	100.0
	Norfolk Island	NFK	1	92.9	92.9	92.9	-	92.9
	Nauru	NRU	1	80.0	80.0	80.0	-	80.0
	Australia	AUS	660	2.1	96.3	69.4	14.6	70.9
	Western Samoa	WSM	2	36.1	94.1	65.1	41.0	65.1
	New Zealand	NZL	40	0.0	88.1	54.9	24.9	65.0
	Guam	GUM	1	47.4	47.4	47.4	-	47.4
	Northern Mariana Islands	MNP	7	0.0	94.4	42.3	32.9	33.3
	Papua New Guinea	PNG	82	0.0	61.6	21.1	18.3	20.1
	Solomon Islands	SLB	29	0.0	73.7	22.8	21.9	18.0
	Niue	NIU	1	15.4	15.4	15.4	-	15.4
	New Caledonia	NCL	18	0.0	78.7	20.6	27.3	4.0
	Cook Islands	COK	14	0.0	88.9	21.0	32.0	0.0
	French Polynesia	PYF	30	0.0	44.0	9.0	13.8	0.0
	Kiribati	KIR	29	0.0	66.7	6.0	15.8	0.0
	Micronesia, Federated States of	FSM	38	0.0	100.0	18.0	34.3	0.0

Variation among countries (continued)

GeoRegion	Country	Code	*N* cells	Min	Max	Mean	SD	Median
	Pitcairn Islands	PCN	2	0.0	0.0	0.0	0.0	0.0
	Tokelau	TKL	3	0.0	0.0	0.0	0.0	0.0
	Tonga	TON	11	0.0	55.6	11.0	21.5	0.0
	Tuvalu	TUV	4	0.0	0.0	0.0	0.0	0.0
	US Minor Outlying Islands	UM	4	0.0	0.0	0.0	-	0.0
	Wallis and Futuna	WLF	3	0.0	0.0	0.0	0.0	0.0
Tropical Asia	Cocos (Keeling) Islands	CCK	1	50.0	50.0	50.0	-	50.0
	Bhutan	BTN	3	35.5	38.8	37.3	1.7	37.6
	Sri Lanka	LKA	7	3.0	50.4	32.5	17.8	37.6
	British Indian Ocean Territory	IO	6	0.0	40.0	18.3	15.7	22.5
	Philippines	PHL	72	0.0	62.5	20.4	19.0	17.7
	Malaysia	MYS	37	0.0	55.3	20.8	16.4	17.3
	Cambodia	KHM	17	0.0	33.0	12.1	9.1	13.6
	Nepal	NPL	10	0.2	39.0	16.2	16.1	13.3
	Thailand	THA	44	0.0	49.0	15.7	15.7	10.3
	Vietnam	VNM	28	0.0	40.6	10.0	10.7	6.3
	Lao People's Democratic Republic	LAO	17	0.0	31.6	8.5	9.0	4.5
	India	IND	276	0.0	60.8	8.1	12.3	1.9
	Myanmar	MMR	61	0.0	29.6	4.2	5.7	1.7
	Indonesia	IDN	316	0.0	50.3	6.2	10.0	1.3
	Bangladesh	BGD	12	0.0	39.7	7.0	13.3	0.8
	Pakistan	PAK	69	0.0	43.9	2.8	7.0	0.3
	Maldives	MDV	15	0.0	44.4	7.7	16.3	0.0
	Spratly Islands	PG	4	0.0	0.0	0.0	0.0	0.0
Temperate Asia	Korea, Republic of	KOR	13	16.7	71.3	46.1	17.3	46.4
	Taiwan	TWN	8	0.0	79.2	42.3	36.9	53.3
	Japan	JPN	78	0.0	70.3	22.7	19.5	16.2
	Korea, Democratic People's Republic of	PRK	11	0.0	64.1	8.1	18.7	1.9
	Kyrgyzstan	KGZ	12	0.0	3.4	1.3	1.0	1.2
	Tajikistan	TJK	12	0.0	3.1	1.1	1.0	1.0
	Mongolia	MNG	124	0.0	31.8	4.5	7.3	0.7
	China	CHN	774	0.0	44.2	2.8	6.1	0.2
	Kazakhstan	KAZ	224	0.0	44.3	3.0	8.6	0.0
	Russia	RUS	1,456	0.0	81.1	2.0	7.1	0.0
	Turkmenistan	TKM	38	0.0	4.8	0.6	1.2	0.0
	Uzbekistan	UZB	37	0.0	14.9	0.9	2.6	0.0
Greater Middle East	Israel	ISR	3	71.4	80.7	76.8	4.9	78.5
	United Arab Emirates	ARE	7	18.1	72.2	61.4	19.2	68.3
	Qatar	QAT	1	40.2	40.2	40.2	-	40.2
	Kuwait	KWT	1	36.3	36.3	36.3	-	36.3
	Morocco	MAR	34	4.5	48.4	23.7	13.4	25.5
	Tunisia	TUN	17	1.5	39.9	17.1	12.5	15.7
	Jordan	JOR	8	0.0	71.1	26.5	31.9	9.9
	Turkey	TUR	67	0.0	54.9	11.9	12.9	7.2
	Armenia	ARM	2	2.0	12.0	7.0	7.1	7.0
	Georgia	GEO	8	1.8	12.6	6.5	3.6	6.0
	Syrian Arab Republic	SYR	17	0.0	30.5	7.3	9.2	5.3
	Egypt	EGY	81	0.0	71.0	10.9	14.9	5.0
	Oman	OMN	27	0.0	52.4	9.5	14.0	3.8
	Iran, Islamic Republic of	IRN	137	0.0	27.3	3.5	5.1	1.5
	Afghanistan	AFG	51	0.0	18.3	3.1	4.3	0.9
	Azerbaijan	AZE	7	0.0	4.6	1.2	1.6	0.9
	Iraq	IRQ	35	0.0	30.4	4.3	8.2	0.9
	Algeria	DZA	194	0.0	25.0	1.8	3.8	0.0
	Libya	LBY	133	0.0	9.1	0.5	1.6	0.0
	Saudi Arabia	SAU	163	0.0	60.4	1.3	5.7	0.0
	Yemen	YEM	38	0.0	8.2	0.9	2.0	0.0
	Western Sahara	ESH	25	0.0	24.6	1.4	4.9	0.0
Sub-Saharan Africa	St. Helena	SHN	4	0.0	100.0	62.5	47.9	75.0
	Swaziland	SWZ	1	64.5	64.5	64.5	-	64.5
	South Africa	ZAF	104	2.1	100.0	56.7	16.7	61.5
	Sao Tome and Principe	STP	2	44.4	60.9	52.7	11.6	52.7
	Reunion	REU	1	52.2	52.2	52.2	-	52.2
	Lesotho	LSO	3	49.4	54.8	51.6	2.8	50.6
	Rwanda	RWA	2	38.1	52.9	45.5	10.5	45.5
	Mauritius	MUS	2	16.7	73.1	44.9	39.9	44.9
	Cape Verde	CPV	8	0.0	65.7	30.8	26.8	31.5
	Burundi	BDI	3	5.0	50.0	28.2	22.6	29.6

Variation among countries (continued)

GeoRegion	Country	Code	*N* cells	Min	Max	Mean	SD	Median
	Malawi	MWI	11	0.9	34.6	20.5	12.6	26.9
	Uganda	UGA	19	3.2	60.9	24.9	17.6	20.2
	Zimbabwe	ZWE	32	0.5	55.0	19.3	15.6	15.7
	Comoros	COM	2	11.5	19.6	15.6	5.8	15.6
	Namibia	NAM	66	0.0	63.1	20.2	16.4	15.6
	Botswana	BWA	46	0.0	61.6	20.8	18.1	13.8
	Liberia	LBR	8	0.7	47.5	20.1	17.0	13.7
	Equatorial Guinea	GNQ	4	2.2	37.5	16.7	15.5	13.6
	Ghana	GHA	21	0.8	40.7	15.0	13.4	12.0
	Madagascar	MDG	54	0.0	69.9	17.2	18.9	11.0
	Senegal	SEN	18	0.2	50.6	14.9	14.6	10.1
	Sierra Leone	SLE	6	4.9	30.4	13.5	10.2	9.5
	Kenya	KEN	48	0.0	69.6	19.4	21.3	9.1
	Benin	BEN	11	2.1	18.3	7.8	5.1	6.1
	Tanzania, United Republic of	TZA	76	0.0	54.5	12.6	15.5	5.9
	Guinea-Bissau	GNB	2	0.5	9.9	5.2	6.6	5.2
	Ivory Coast	CIV	27	0.0	36.7	5.8	7.1	4.9
	Gabon	GAB	21	0.0	29.0	7.2	8.6	4.7
	Togo	TGO	5	0.8	9.8	4.8	3.4	4.7
	Burkina Faso	BFA	22	0.0	13.7	4.9	4.5	3.6
	Cameroon	CMR	41	0.0	38.5	9.1	11.0	3.3
	Mayotte	MYT	1	2.3	2.3	2.3	-	2.3
	Zambia	ZMB	57	0.0	49.2	9.2	13.4	1.8
	Congo, Democratic Republic of	COD	194	0.0	67.7	6.7	11.6	1.7
	Mozambique	MOZ	67	0.0	70.1	5.6	12.9	1.7
	Guinea	GIN	22	0.0	44.4	8.1	13.6	0.9
	Congo, Republic of	COG	27	0.0	21.9	2.5	5.0	0.8
	Ethiopia	ETH	93	0.0	36.9	4.1	7.8	0.8
	Angola	AGO	101	0.0	60.7	3.3	7.7	0.7
	Eritrea	ERI	9	0.0	19.2	3.8	7.2	0.3
	Nigeria	NGA	72	0.0	26.0	1.7	4.5	0.3
	Central African Republic	CAF	51	0.0	22.7	0.6	3.2	0.0
	Chad	TCD	103	0.0	11.2	0.5	1.8	0.0
	Djibouti	DJI	3	0.0	0.3	0.1	0.1	0.0
	Mali	MLI	101	0.0	6.7	0.4	0.9	0.0
	Mauritania	MRT	81	0.0	7.6	0.5	1.1	0.0
	Niger	NER	98	0.0	10.7	0.7	1.9	0.0
	Seychelles	SYC	11	0.0	79.5	11.9	24.2	0.0
	Somalia	SOM	57	0.0	15.6	1.3	3.0	0.0
	Sudan	SDN	204	0.0	37.4	1.3	3.9	0.0

Table V.2.S3. Model fits and spatial autocorrelation for **a)** inventory completeness (RAC models) and **b)** record density (SAR models). Values are given for the model subset with the lowest AIC score. In a) model fit is expressed by the deviance explained (D^2). The degree of spatial autocorrelation (global Moran's I) in model residuals is compared between the minimum adequate spatial model subset (see 'Moran's I_{sp}') and the corresponding non-spatial model (see 'Moran's I_{nsp}'). Asterisks denote significant spatial autocorrelation (.: P<0.1; *: P<0.05; **: P<0.01; ***: P<0.001). In b) model fit is expressed by pseudo-R^2 values, calculated as the squared Pearson correlation coefficient between fitted and observed values (Kissling & Carl, 2008). Fitted values of SAR models can be partitioned additively into trend (non-spatial smooth) and signal (spatial smooth). We calculated both a pseudo-R^2 for the fitted values including the spatial component ('R^2_{sp}'), and a pseudo-R^2 for the trend excluding the spatial component, which represents the part of the variation explained by the predictors (in the context of SAR models hereafter 'R^2_{nsp}'). R^2 values of potential minimum adequate models (subsets with $\Delta AIC < 2$) never differed by more than 0.004. The degree of spatial autocorrelation (global Moran's I) in model residuals is compared between the minimum adequate spatial model (see 'Moran's I_{sp}') and the corresponding non-spatial (OLS) model (see 'Moran's I_{nsp}'). Asterisks denote significant spatial autocorrelation (.: P<0.1; *: P<0.05; **: P<0.01; ***: P<0.001).

a) Inventory completeness.

Taxon	Grain (km)	D^2	Moran's I_{nsp}	Moran's I_{sp}
Birds	110	0.78	0.067***	0.007***
	220	0.76	0.057***	0.003***
	440	0.77	0.040***	-0.003
	880	0.74	0.012	-0.012
Mammals	110	0.70	0.081***	0.006***
	220	0.75	0.079***	0.006***
	440	0.77	0.061***	-0.003
	880	0.73	0.030***	-0.006
Amphibians	110	0.57	0.062***	0.008***
	220	0.64	0.066***	0.008***
	440	0.60	0.064***	0.001
	880	0.60	0.059***	-0.005

b) Record density.

Taxon	Grain (km)	R^2_{sp}	R^2_{nsp}	Moran's I_{nsp}	Moran's I_{sp}
Birds	110	0.82	0.62	0.086***	0.006***
	220	0.83	0.70	0.069***	0.006***
	440	0.85	0.78	0.047***	0.007**
	880	0.86	0.82	0.025***	0.005
Mammals	110	0.66	0.41	0.068***	0.005***
	220	0.76	0.53	0.070***	0.007***
	440	0.80	0.59	0.060***	0.004.
	880	0.76	0.71	0.030***	0.006
Amphibians	110	0.58	0.38	0.063***	0.006***
	220	0.69	0.53	0.062***	0.005***
	440	0.77	0.59	0.060***	0.002
	880	0.83	0.70	0.046***	-0.000

Table V.2.S4. Influence of adding a) country identity of grid cells as a factor and b) record density to the minimum adequate model of inventory completeness. D^2_{MAM} is the deviance explained by the minimum adequate model. In a): $D^2_{MAM+Country}$ is the deviance explained when adding a country factor to the minimum adequate model. $D^2_{Country}$ is the deviance explained by a model containing only country membership as factor. The percentage of cross-country variation that is already captured by the minimum adequate model (% of cross-country variation already in D^2_{MAM}) was calculated as: $100 / D^2_{Country} * (D^2_{Country} - (D^2_{MAM+Country} - D^2_{MAM}))$. $\%D^2_{\text{ added by Country}}$ is the additional deviance explained by adding a country factor to the minimum adequate model (as percent of total D^2); in b): D^2_{MAM+RD} is the deviance explained when adding $\log_{10}$-transformed record density to the minimum adequate model. D^2_{RD} is the deviance explained by a model containing only $\log_{10}$-transformed record density as an explanatory variable. The percentage of the deviance explained by the MAM that is also attributable to differences in record density (% of D^2_{MAM} in ΔRD) was calculated as: $100 / D^2_{MAM} * (D^2_{MAM} - (D^2_{RD} - D^2_{MAM+RD}))$. $\%D^2_{\text{ added by RD}}$ is the additional deviance explained by adding record density to the minimum adequate model (as percent of total D^2_{MAM+RD}).

a) Adding country identity to MAM.

Taxon	Grain (km)	D^2_{MAM}	$D^2_{MAM+Country}$	$D^2_{Country}$	% of cross-country variation already in D^2_{MAM}	% of D^2 added by Country
irds	110	0.78	0.80	0.68	97.2	2.4
Mammals	110	0.70	0.73	0.64	94.7	4.6
Amphibians	110	0.57	0.62	0.55	92.1	7.1

b) Adding record density to MAM.

Taxon	Grain (km)	D^2_{MAM}	D^2_{MAM+RD}	D^2_{RD}	% of D^2_{MAM} in ΔRD	% of D^2 added by RD
Birds	110	0.78	0.94	0.90	94.2	5.8
	220	0.76	0.94	0.89	94.3	5.7
	440	0.77	0.92	0.88	95.2	4.8
	880	0.74	0.86	0.82	95.2	4.8
Mammals	110	0.70	0.88	0.76	83.7	16.3
	220	0.75	0.89	0.79	86.9	13.1
	440	0.77	0.89	0.81	89.0	11.0
	880	0.73	0.87	0.79	89.8	10.2
Amphibians	110	0.57	0.76	0.59	69.1	30.9
	220	0.64	0.79	0.64	76.8	23.2
	440	0.60	0.80	0.63	72.3	27.7
	880	0.60	0.76	0.57	68.0	32.0

Table V.2.S5. The effects of socioeconomic and geographic factors on a) – d) inventory completeness and e) – h) data density. The twelve predictor variables were endemism richness (EndRich), protected area coverage (ProtAreas), mountains (Mountains), on-ground accessibility (GroundAcc), proximity to airports (ProxAirp), proximity to data-contributing institutions (ProxInst), secure conditions (Security), participation with GBIF (GBIFpartic), scientific activities (ScientActiv), nationally available research funding (FundLocal), research funding in countries with contributing institutions (FundInst), and size of contributing institutions (PublSize). Three comparative measures were used: for inventory completeness (a – d): 1) the deviance explained from simple regressions (d^2), 2) standardized regression coefficients from the reduced spatial generalized linear model with the lowest AIC score (GLM β; a range of coefficients is given if several model subsets have ΔAIC < 2 to the "best" model), and 3) the percentage each predictor has in the total Sums of Squares of an ANOVA, where the AIC values of all possible non-spatial models enter as the response variable and dummy-variables coding whether or not a predictor is in the respective model as explanatory variables (% SS); for inventory completeness (e – h): 1) the coefficient of determination from simple ordinary least squares regressions (r^2), 2) standardized regression coefficients from the reduced simultaneous autoregressive model with the lowest AIC score (SAR β), and 3) the percentage each predictor has in the total Sums of Squares of an ANOVA, where the AIC values of all possible non-spatial models enter as the response variable and dummy-variables coding whether or not a predictor is in the respective model as explanatory variables (% SS) the. Asterisks denote significant spatial autocorrelation (.: P<0.1; *: P<0.05; **: P<0.01; ***: P<0.001).

a) Inventory completeness at 110 km.

Taxonomic group	Predictor	d^2	GLM β (range)	z-value	% SS
Birds	EndRich	0.01***	0.32***	127.21	0.01
	ProtAreas	0.03***	0.19***	80.96	0.01
	Mountains	0.00***	-0.03***	-11.83	0.00
	GroundAcc	0.03***	0.23***	72.45	0.00
	ProxAirp	0.16***	0.18***	57.32	0.03
	ProxInst	0.29***	0.35***	121.61	0.15
	Security	0.12***	0.08***	27.93	0.01
	GBIFpartic	0.27***	0.38***	134.25	0.13
	ScientActiv	0.39***	0.27***	56.93	0.22
	FundLocal	0.34***	0.61***	126.60	0.21
	FundInst	0.01***	-0.13***	-59.17	0.00
	PublSize	0.18***	0.53***	173.70	0.22
Mammals	EndRich	0.00	0.25***	47.92	0.01
	ProtAreas	0.02***	0.13***	26.50	0.00
	Mountains	0.01***	0.07***	14.28	0.00
	GroundAcc	0.05***	0.02*	2.25	0.00
	ProxAirp	0.12***	0.07***	10.51	0.00
	ProxInst	0.40***	0.61***	87.06	0.72
	Security	0.07***	-0.04***	-6.34	0.00
	GBIFpartic	0.25***	0.26***	38.41	0.10
	ScientActiv	0.27***	-0.01	-0.80	0.06
	FundLocal	0.24***	0.30***	29.94	0.08
	FundInst	0.02***	-0.06***	-12.24	0.01
	PublSize	0.02***	0.15***	25.34	0.01
Amphibians	EndRich	0.00***	0.08***	10.93	0.00
	ProtAreas	0.01***	0.12***	14.35	0.01
	Mountains	0.01***	0.13***	15.32	0.04
	GroundAcc	0.06***	0.12***	11.71	0.01
	ProxAirp	0.11***	0.05***	5.11	0.02
	ProxInst	0.25***	0.30***	29.13	0.56
	Security	0.07***	-0.16***	-16.03	0.03
	GBIFpartic	0.19***	0.24***	28.50	0.26
	ScientActiv	0.16***	-0.02		0.04
	FundLocal	0.13***	0.17*** (0.17 - 0.18)	14.26	0.03
	FundInst	0.01***	-0.07***	-8.40	0.01
	PublSize	0.00***	0.02	1.63	0.00

b) Inventory completeess at 220 km.

Taxonomic group	Predictor	d^2	GLM β (range)	z-value	% SS
Birds	EndRich	0.00***	0.32***	76.97	0.01
	ProtAreas	0.04***	0.20***	50.28	0.01
	Mountains	0.01***	0.06***	16.83	0.00
	GroundAcc	0.02***	0.10***	21.44	0.00
	ProxAirp	0.21***	0.30***	59.08	0.07
	ProxInst	0.30***	0.42***	95.26	0.25

Inventory completeess at 220 km (continued)

Taxonomic group	Predictor	d²	GLM β (range)	z-value	% SS
	Security	0.15***	-0.16***	-31.89	0.02
	GBIFpartic	0.29***	0.38***	89.00	0.16
	ScientActiv	0.32***	0.10***	13.75	0.11
	FundLocal	0.23***	0.56***	84.87	0.11
	FundInst	0.02***	-0.15***	-43.89	0.01
	PublSize	0.19***	0.53***	108.44	0.24
Mammals	EndRich	0.01***	0.38***	45.10	0.02
	ProtAreas	0.02***	0.12***	16.44	0.00
	Mountains	0.02***	0.03***	4.03	0.00
	GroundAcc	0.04***	-0.01	-1.05	0.00
	ProxAirp	0.16***	0.16***	16.52	0.02
	ProxInst	0.41***	0.61***	61.87	0.65
	Security	0.07***	-0.03***	-3.41	0.00
	GBIFpartic	0.31***	0.33***	34.11	0.20
	ScientActiv	0.26***	-0.01	-0.71	0.05
	FundLocal	0.18***	0.32***	24.10	0.04
	FundInst	0.03***	-0.06***	-8.16	0.01
	PublSize	0.02***	0.17***	17.41	0.01
Amphibians	EndRich	0.00.	0.12***	11.04	0.00
	ProtAreas	0.00***	0.10***	8.10	0.00
	Mountains	0.03***	0.11***	8.77	0.03
	GroundAcc	0.09***	0.12***	8.79	0.04
	ProxAirp	0.18***	0.14***	9.36	0.08
	ProxInst	0.26***	0.24***	15.90	0.34
	Security	0.06***	-0.12***	-8.01	0.01
	GBIFpartic	0.24***	0.24***	16.93	0.41
	ScientActiv	0.19***	0.09***	3.58	0.07
	FundLocal	0.12***	0.18***	8.22	0.03
	FundInst	0.02***	-0.13***	-9.76	0.00
	PublSize	0.01***	0.04*	2.44	0.00

c) Inventory completeness at 440 km.

Taxonomic group	Predictor	d²	GLM β (range)	z-value	% SS
Birds	EndRich	0.03***	0.41***	53.94	0.04
	ProtAreas	0.08***	0.27*** (0.27 - 0.28)	36.99	0.06
	Mountains	0.01***	0.05***	9.14	0.00
	GroundAcc	0.02***	-0.02*	-2.53	0.00
	ProxAirp	0.23***	0.29*** (0.28 - 0.29)	32.86	0.11
	ProxInst	0.30***	0.34***	46.12	0.31
	Security	0.17***	-0.12***	-14.12	0.02
	GBIFpartic	0.28***	0.36***	52.56	0.20
	ScientActiv	0.21***	-0.09***	-9.68	0.03
	FundLocal	0.15***	0.65***	63.13	0.10
	FundInst	0.05***	-0.20***	-31.61	0.02
	PublSize	0.16***	0.45***	53.53	0.11
Mammals	EndRich	0.02***	0.45***	30.81	0.04
	ProtAreas	0.05***	0.22***	16.70	0.04
	Mountains	0.02***	0.08***	7.28	0.01
	GroundAcc	0.03***	-0.03*	-2.10	0.00
	ProxAirp	0.17***	0.13***	8.87	0.03
	ProxInst	0.37***	0.49***	32.15	0.47
	Security	0.07***	0.00	0.24	0.00
	GBIFpartic	0.33***	0.40***	28.94	0.34
	ScientActiv	0.17***	-0.11***	-6.08	0.02
	FundLocal	0.10***	0.43***	22.43	0.03
	FundInst	0.10***	-0.15***	-11.96	0.03
	PublSize	0.00***	0.07***	5.01	0.00
Amphibians	EndRich	0.00	0.14***	8.48	0.00
	ProtAreas	0.00.	0.09***	4.75	0.00
	Mountains	0.02***	0.08***	4.38	0.01
	GroundAcc	0.07***	0.17***	9.15	0.05
	ProxAirp	0.14***	0.17***	8.27	0.08
	ProxInst	0.24***	0.12***	5.98	0.25
	Security	0.05***	-0.18***	-8.22	0.02
	GBIFpartic	0.20***	0.27***	14.35	0.46
	ScientActiv	0.14***	0.01	0.42	0.04
	FundLocal	0.12***	0.30***	11.11	0.09
	FundInst	0.04***	-0.15*** (-0.16 - .0.15)	-8.89	0.01
	PublSize	0.01***	0.01		0.00

d) Inventory completeness at 880 km.

Taxonomic group	Predictor	d²	GLM β (range)	z-value	% SS
Birds	EndRich	0.07***	0.41***	31.67	0.08
	ProtAreas	0.08***	0.21***	14.05	0.03
	Mountains	0.02***	-0.04***	-3.67	0.00
	GroundAcc	0.02***	-0.03*	-2.23	0.00
	ProxAirp	0.19***	0.32***	25.56	0.11
	ProxInst	0.28***	0.33*** (0.33 - 0.34)	24.05	0.34
	Security	0.20***	-0.12***	-8.07	0.02
	GBIFpartic	0.30***	0.38***	32.61	0.23
	ScientActiv	0.18***			0.02
	FundLocal	0.10***	0.49*** (0.48 - 0.49)	33.06	0.09
	FundInst	0.05***	-0.03*	-2.40	0.01
	PublSize	0.13***	0.29*** (0.28 - 0.29)	19.48	0.06
Mammals	EndRich	0.04***	0.29***	12.57	0.04
	ProtAreas	0.07***	0.29***	12.27	0.06
	Mountains	0.03***	0.12***	5.90	0.01
	GroundAcc	0.02***	0.06**	2.77	0.00
	ProxAirp	0.12***	0.18***	9.86	0.03
	ProxInst	0.39***	0.36***	17.67	0.42
	Security	0.11***	-0.07**	-2.96	0.00
	GBIFpartic	0.36***	0.38***	19.30	0.32
	ScientActiv	0.21***			0.03
	FundLocal	0.09***	0.31***	13.23	0.05
	FundInst	0.11***			0.03
	PublSize	0.01***	-0.02	-0.83	0.00
Amphibians	EndRich	0.00			0.00
	ProtAreas	0.02***	-0.11***	-3.80	0.00
	Mountains	0.00**	0.01		0.00
	GroundAcc	0.07***	0.23***	8.86	0.09
	ProxAirp	0.11***	0.26***	12.07	0.13
	ProxInst	0.17***	0.24***	9.84	0.31
	Security	0.09***	-0.44***	-14.46	0.09
	GBIFpartic	0.13***	0.32***	12.66	0.23
	ScientActiv	0.19***			0.08
	FundLocal	0.13***			0.06
	FundInst	0.00			0.00
	PublSize	0.00	0.27*** (0.26 - 0.27)	8.76	0.00

e) Record density at 110 km.

		r²	SAR β (range)	z-value	% SS
Birds	EndRich	0.01***	0.28***	14.66	0.07
	ProtAreas	0.04***	0.06***	7.47	0.00
	Mountains	0.00	0.03*	2.15	0.00
	GroundAcc	0.06***	0.16***	10.35	0.05
	ProxAirp	0.23***	0.15***	11.81	0.00
	ProxInst	0.28***	0.11*** (0.11 - 0.12)	5.11	0.04
	Security	0.09***	-0.04.	-1.93	0.00
	GBIFpartic	0.29***	0.21*** (0.21 - 0.22)	8.65	0.14
	ScientActiv	0.33***	-0.05 (-0.06 - -0.05)	-1.55	0.20
	FundLocal	0.25***	0.38*** (0.38 - 0.39)	10.73	0.24
	FundInst	0.01***	-0.03**	-3.05	0.01
	PublSize	0.21***	0.17***	22.82	0.24
Mammals	EndRich	0.03***	0.30***	17.79	0.19
	ProtAreas	0.03***	0.04***	4.13	0.01
	Mountains	0.01***	0.08***	4.78	0.01
	GroundAcc	0.03***	0.09***	4.38	0.01
	ProxAirp	0.12***	0.10***	5.29	0.01
	ProxInst	0.24***	0.29***	8.72	0.22
	Security	0.03***	0.06*	2.29	0.00
	GBIFpartic	0.18***	0.20***	6.66	0.18
	ScientActiv	0.17***	-0.05	-1.06	0.19
	FundLocal	0.09***	0.24***	4.99	0.06
	FundInst	0.02***	-0.04**	-3.00	0.03
	PublSize	0.04***	0.11***	11.10	0.09
Amphibians	EndRich	0.08***	0.34***	18.07	0.55
	ProtAreas	0.02***	0.07***	5.12	0.00
	Mountains	0.00***	0.05** (0.05 - 0.06)	2.62	0.00
	GroundAcc	0.08***	0.11***	4.78	0.02

Record density at 110 km (continued)

		r²	SAR β (range)	z-value	% SS
	ProxAirp	0.14***	0.11***	5.33	0.03
	ProxInst	0.14***	0.33*** (0.32 - 0.33)	9.99	0.26
	Security	0.01***			0.00
	GBIFpartic	0.05***	0.03 (0.03 - 0.04)	1.0	0.01
	ScientActiv	0.07***	0.12*** (0.09 - 0.12)	3.18	0.06
	FundLocal	0.03***	0.05	0.86	0.02
	FundInst	0.00.	-0.09*** (-0.10 - -0.09)	-5.53	0.00
	PublSize	0.04***	0.14***	9.26	0.06

f) Record density at 220 km.

Taxonomic group	Predictor	r²	SAR β (range)	z-value	% SS
Birds	EndRich	0.02***	0.33*** (0.33 - 0.35)	10.83	0.09
	ProtAreas	0.06***	0.07***	5.62	0.00
	Mountains	0.00	0.03*	1.97	0.00
	GroundAcc	0.06***	0.13***	5.41	0.02
	ProxAirp	0.29***	0.13***	7.26	0.03
	ProxInst	0.34***	0.17***	6.34	0.02
	Security	0.16***	-0.07*	-2.51	0.00
	GBIFpartic	0.38***	0.22***	7.37	0.24
	ScientActiv	0.36***	-0.04	-1.07	0.20
	FundLocal	0.25***	0.34***	8.24	0.25
	FundInst	0.04***	-0.06***	-3.90	0.03
	PublSize	0.18***	0.13***	10.88	0.19
Mammals	EndRich	0.07***	0.40*** (0.40 - 0.42)	14.81	0.35
	ProtAreas	0.05***	0.06***	3.80	0.01
	Mountains	0.01***	0.04.	1.82	0.00
	GroundAcc	0.05***	0.08**	2.82	0.00
	ProxAirp	0.16***	0.07**	3.02	0.01
	ProxInst	0.29***	0.31*** (0.30 - 0.31)	8.24	0.17
	Security	0.05***	0.04	1.13	0.00
	GBIFpartic	0.24***	0.21***	5.45	0.20
	ScientActiv	0.19***	-0.00	-0.05	0.16
	FundLocal	0.07***	0.24***	4.65	0.04
	FundInst	0.04***	-0.05**	-3.09	0.05
	PublSize	0.02***	0.12***	9.04	0.03
Amphibians	EndRich	0.14***	0.41*** (0.41 - 0.42)	15.94	0.59
	ProtAreas	0.01***	0.07*** (0.06 - 0.07)	3.45	0.00
	Mountains	0.01**	-0.01	-0.24	0.00
	GroundAcc	0.09***	0.11*** (0.11 - 0.12)	3.54	0.01
	ProxAirp	0.20***	0.10***	4.03	0.08
	ProxInst	0.22***	0.34*** (0.33 - 0.35)	9.03	0.23
	Security	0.02***	0.00		0.00
	GBIFpartic	0.11***	0.07.	1.65	0.02
	ScientActiv	0.13***	0.16*** (0.12 - 0.20)	3.61	0.04
	FundLocal	0.05***	0.07	1.25	0.01
	FundInst	0.00	-0.11***	-4.62	0.00
	PublSize	0.04***	0.14*** (0.13 - 0.14)	6.18	0.02

g) Record density at 440 km.

Taxonomic group	Predictor	r²	SAR β (range)	z-value	% SS
Birds	EndRich	0.04***	0.37*** (0.37 - 0.38)	8.65	0.13
	ProtAreas	0.09***	0.10*** (0.09 - 0.10)	4.49	0.01
	Mountains	0.00	-0.00		0.00
	GroundAcc	0.08***	0.09* (0.07 - 0.09)	2.48	0.00
	ProxAirp	0.30***	0.10*** (0.10 - 0.11)	3.61	0.01
	ProxInst	0.37***	0.15*** (0.13 - 0.15)	4.39	0.26
	Security	0.22***	-0.04 (-0.06 - -0.04)	-1.24	0.00
	GBIFpartic	0.45***	0.26*** (0.24 - 0.27)	7.64	0.26
	ScientActiv	0.34***	-0.02	-0.34	0.08
	FundLocal	0.23***	0.43*** (0.40 - 0.44)	10.36	0.20
	FundInst	0.06***	-0.08** (-0.09 - -0.08)	-2.66	0.01
	PublSize	0.13***	0.09*** (0.08 - 0.09)	4.34	0.03
Mammals	EndRich	0.11***	0.47***(0.47 - 0.48)	10.43	0.37
	ProtAreas	0.09***	0.06* (0.06 - 0.07)	2.51	0.04
	Mountains	0.01**	0.01	-	0.00
	GroundAcc	0.06***	0.04 (0.04 - 0.06)	0.95	0.00
	ProxAirp	0.17***	0.05 (0.04 - 0.05)	1.58	0.00

Record density at 440 km (continued)

Taxonomic group	Predictor	r^2	SAR β (range)	z-value	% SS
	ProxInst	0.30***	0.25*** (0.24 - 0.25)	5.44	0.24
	Security	0.06***	0.04 (0.03 - 0.04)	0.92	0.00
	GBIFpartic	0.30***	0.29***	4.34	0.21
	ScientActiv	0.16***	0.04 (0.04 - 0.05)	0.64	0.04
	FundLocal	0.04***	0.23*** (0.21 - 0.24)	4.03	0.02
	FundInst	0.11***	-0.13***	-4.79	0.08
	PublSize	0.00	0.09***	3.99	0.00
Amphibians	EndRich	0.15***	0.45*** (0.45 - 0.46)	11.96	0.55
	ProtAreas	0.01*	0.03	1.08	0.00
	Mountains	0.00	-0.03	-0.84	0.00
	GroundAcc	0.12***	0.13** (0.13 - 0.14)	3.08	0.01
	ProxAirp	0.16***	0.06* (0.06 - 0.07)	2.06	0.01
	ProxInst	0.27***	0.22*** (0.21 - 0.22)	4.99	0.37
	Security	0.07***	-0.03	-0.74	0.00
	GBIFpartic	0.16***	0.05	0.98	0.01
	ScientActiv	0.17***	0.16** (0.13 - 0.16)	2.72	0.03
	FundLocal	0.08***	0.21***	3.35	0.03
	FundInst	0.00	-0.10*** (-0.11 - -0.10)	-3.26	0.00
	PublSize	0.03***	0.07* (0.07 - 0.08)	2.47	0.00

h) Record density at 880 km.

Taxonomic group	Predictor	r^2	SAR β (range)	z-value	% SS
Birds	EndRich	0.02.	0.31*** (0.30 - 0.33)	5.37	0.11
	ProtAreas	0.08***	0.15*** (0.15 - 0.16)	4.01	0.01
	Mountains	0.00	0.03 (0.02 - 0.03)	0.93	0.00
	GroundAcc	0.09***	0.11* (0.10 - 0.12)	2.38	0.02
	ProxAirp	0.17***	0.11** (0.11 - 0.12)	2.92	0.01
	ProxInst	0.42***	0.25*** (0.24 - 0.26)	5.83	0.33
	Security	0.28***	-0.08. (-0.09 - -0.07)	-1.67	0.02
	GBIFpartic	0.49***	0.20*** (0.20 - 0.28)	3.63	0.25
	ScientActiv	0.38***	0.10. (0.10 - 0.11)	1.82	0.09
	FundLocal	0.20***	0.31*** (0.31 - 0.39)	4.68	0.17
	FundInst	0.06***	-0.02		0.01
	PublSize	0.09***	-0.06 (-0.06 - -0.04)	-1.44	0.01
Mammals	EndRich	0.07***	0.28*** (0.26 - 0.28)	3.89	0.33
	ProtAreas	0.09***	0.19*** (0.17 - 0.19)	3.91	0.03
	Mountains	0.04**	0.13** (0.12 - 0.13)	2.82	0.02
	GroundAcc	0.10***	0.17** (0.13 - 0.18)	2.78	0.04
	ProxAirp	0.11***	0.07 (0.06 - 0.07)	1.34	0.00
	ProxInst	0.34***	0.20** (0.19 - 0.20)	2.94	0.19
	Security	0.11***	0.11. (0.11 - 0.11)	1.84	0.00
	GBIFpartic	0.38***	0.22** (0.16 - 0.22)	2.93	0.16
	ScientActiv	0.24***	0.18* (0.18 - 0.19)	2.41	0.11
	FundLocal	0.04**	0.17* (0.12 - 0.17)	2.15	0.03
	FundInst	0.16***	-0.14** (-0.15 - -0.14)	-2.74	0.10
	PublSize	0.008	0.005		0.00
Amphibians	EndRich	0.20***	0.34*** (0.32 - 0.37)	5.60	0.70
	ProtAreas	0.00	-0.05 (-0.05 - -0.03)	-1.05	0.00
	Mountains	0.00	0.03	0.65	0.00
	GroundAcc	0.14***	0.24*** (0.23 - 0.27)	4.36	0.03
	ProxAirp	0.10***	0.06 (0.06 - 0.07)	1.52	0.01
	ProxInst	0.30***	0.24*** (0.20 - 0.27)	4.62	0.06
	Security	0.16***	-0.16*** (-0.20 - -0.15)	-3.22	0.01
	GBIFpartic	0.19***	0.06 (0.06 - 0.12)	0.88	0.01
	ScientActiv	0.23***	0.15*	2.17	0.09
	FundLocal	0.07***	0.15. (0.15 - 0.22)	1.90	0.08
	FundInst	0.04**	-0.01		0.00
	PublSize	0.03*	0.07. (0.07 - 0.09)	1.77	0.00

Table V.2.S6. Top 50 countries based on number of species-grid cell combinations that are missing from country-wide completeness of 100% at the 110 km grain ('Non-inventoried species spp-cell'). Countries are ordered from highest to lowest percentage of non-inventoried species presences ('% of non-inventoried spp-cell).

Country	Non-inventoried spp-cell	% of non-inventoried spp-cell
Brazil	451,427	15.4
Russia	260,523	8.9
China	201,422	6.9
India	106,128	3.6
Indonesia	103,898	3.6
Congo, Democratic Republic of	98,291	3.4
Canada	74,129	2.5
Sudan	61,617	2.1
Colombia	61,122	2.1
USA	58,822	2.0
Peru	57,550	2.0
Argentina	51,619	1.8
Venezuela	50,096	1.7
Angola	47,694	1.6
Kazakhstan	45,568	1.6
Ethiopia	43,609	1.5
Tanzania	43,367	1.5
Bolivia	42,583	1.5
Australia	40,854	1.4
Myanmar	38,141	1.3
Nigeria	37,055	1.3
Zambia	34,246	1.2
Mozambique	34,066	1.2
Mexico	32,127	1.1
Iran, Islamic Republic of	27,411	0.9
Mali	24,308	0.8
Central African Republic	24,096	0.8
Kenya	23,930	0.8
Mongolia	23,835	0.8
Chad	23,608	0.8
Thailand	23,422	0.8
Cameroon	23,281	0.8
South Africa	19,359	0.7
Papua New Guinea	18,648	0.6
Malaysia	18,515	0.6
Niger	17,865	0.6
Namibia	17,842	0.6
Pakistan	17,135	0.6
Philippines	16,439	0.6
Zimbabwe	16,418	0.6
Turkey	16,375	0.6
Algeria	16,365	0.6
Vietnam	16,066	0.5
Somalia	15,873	0.5
Côte d'Ivoire	15,016	0.5
Botswana	14,617	0.5
Saudi Arabia	14,505	0.5
Congo, Republic of	13,677	0.5
Guyana	13,499	0.5
Paraguay	13,415	0.5

Table V.2.S7. Summary of a) bird, b) mammal, c) amphibian records contributed to GBIF by different data publishers and used in this study. Data publishers are ordered by decreasing number of contributed data. In the parentheses are percentages of overall data that passed geographic and taxonomic validation and were used in further analyses. Note that we applied a land area threshold of 30% at the 110 km grain, which resulted in the exclusion of some "good" data collected on or near the sea. We also excluded non-breeding ranges. Therefore percentages of excluded records do not necessarily allow conclusions on the quality of data provided by a particular publisher.

a) Publishers of bird records

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Avian Knowledge Network	USA	95,339,821 84,339,776 (88.5%)	-	95,339,821 84,339,776 (88.5%)	-
ArtDatabanken	Sweden	21,040,602 17,322,128 (82.3%)	-	21,040,602 17,322,128 (82.3%)	-
Birds Australia	Australia	10,969,497 9,803,262 (89.4%)	10,969,497 9,803,262 (89.4%)	-	-
BirdLife Finland	Finland	7,535,045 5,577,806 (74.0%)	55,638 35,060 (63.0%)	7,479,407 5,542,746 (74.1%)	-
South African National Biodiversity Institute	South Africa	6,792,022 6,120,569 (90.1%)	-	6,792,022 6,120,569 (90.1%)	-
UK National Biodiversity Network	UK	5,606,751 5,058,976 (90.2%)	5,606,751 5,058,976 (90.2%)	-	-
Danish Biodiversity Information Facility	Denmark	4,544,665 3,595,795 (79.1%)	25,333 17,626 (69.6%)	4,509,884 3,570,857 (79.2%)	9,448 7,312 (77.4%)
GBIF-Sweden	Sweden	4,237,991 3,968,443 (93.6%)	-	4,237,809 3,968,304 (93.6%)	182 139 (76.4%)
The Norwegian Biodiversity Information Centre (NBIC)	Norway	3,827,892 3,134,943 (81.9%)	-	3,827,892 3,134,943 (81.9%)	-
NSW Dpt. of Environment, Climate Change, and Water	Australia	2,601,841 2,109,362 (81.1%)	-	2,601,841 2,109,362 (81.1%)	-
Eremaea	Australia	1,207,943 1,068,708 (88.5%)	164,041 146,662 (89.4%)	1,043,902 922,046 (88.3%)	-
Canberra Ornithologists Group	Australia	1,159,524 965,904 (83.3%)	-	1,159,524 965,904 (83.3%)	-
Service du Patrimoine naturel,Musée national d'Histoire naturelle, Paris	France	960,908 909,673 (94.7%)	-	960,908 909,673 (94.7%)	-
University of Gdańsk, Bird Migration Research Station	Poland	667,168 601,202 (90.1%)	-	667,168 601,202 (90.1%)	-
National Biodiversity Data Centre	Ireland	647,220 358,159 (55.3%)	-	-	647,220 358,159 (55.3%)
Ocean Biogeographic Information System	OBIS	622,491 228,500 (36.7%)	976 186 (19.1%)	621,515 228,314 (36.7%)	-
Dpt. of Natural Resources, Environment (Northern Territory)	Australia	616,706 560,637 (90.9%)	-	616,706 560,637 (90.9%)	-
Dpt. of Environment and Natural Resources (South Australia)	Australia	586,633 527,342 (89.9%)	489 481 (98.4%)	585,597 526,360 (89.9%)	547 501 (91.6%)
Biologiezentrum Linz Oberösterreich	Austria	548,292 496,931 (90.6%)	548,292 496,931 (90.6%)	-	-
Finnish Museum of Natural History	Finland	513,504 340,535 (66.3%)	-	513,504 340,535 (66.3%)	-

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
GBIF-Spain	Spain	431,841 412,275 (95.5%)	-	429,746 410,541 (95.5%)	2,095 1,734 (82.8%)
Australian Antarctic Data Centre	Australia	400,449 108 (0.0%)	365,283 5 (0.0%)	35,166 103 (0.3%)	-
Bird Studies Canada	Canada	310,618 292,455 (94.2%)	-	310,618 292,455 (94.2%)	-
Arctos	USA	249,240 218,950 (87.8%)	-	-	249,240 218,950 (87.8%)
Yale University Peabody Museum	USA	196,614 169,340 (86.1%)	-	-	196,614 169,340 (86.1%)
University of Michigan Museum of Zoology	USA	173,337 147,644 (85.2%)	-	-	173,337 147,644 (85.2%)
KBIF Data Repository	Korea, Republic of	152,187 92,416 (60.7%)	149,984 91,626 (61.1%)	-	2,203 790 (35.9%)
Royal Ontario Museum	Canada	150,080 120,399 (80.2%)	-	-	150,080 120,399 (80.2%)
Israel Nature and Parks Authority	Israel / EU - BioCASE	134,076 101,540 (75.7%)	-	134,076 101,540 (75.7%)	-
Field Museum	USA	122,457 107,377 (87.7%)	-	-	122,457 107,377 (87.7%)
Canadian Biodiversity Information Facility	Canada	120,384 97,427 (80.9%)	120,384 97,427 (80.9%)	-	-
Museum of Comparative Zoology, Harvard University	USA	115,101 96,997 (84.3%)	-	-	115,101 96,997 (84.3%)
Australian Museum	Australia	107,389 86,946 (81.0%)	-	-	107,389 86,946 (81.0%)
Scientific Committee on Antarctic Research - Marine Biodiversity Information Network (SCAR-MarBIN)	International	104,527 8 (0.0%)	427 8 (1.9%)	104,100 0 (0.0%)	-
Canadian Museum of Nature	Canada	88,218 73,846 (83.7%)	-	-	88,218 73,846 (83.7%)
Comisión nacional para el conocimiento y uso de la biodiversidad (CONABIO)	Mexico	83,925 71,716 (85.5%)	65,111 55,757 (85.6%)	18,814 15,959 (84.8%)	-
University of Washington Burke Museum	USA	72,535 53,763 (74.1%)	-	-	72,535 53,763 (74.1%)
BeBIF Provider	Belgium	70,010 63,116 (90.2%)	41,033 35,940 (87.6%)	28,977 27,176 (93.8%)	-
TELDAP	Chinese Taipei	67,664 63,208 (93.4%)	-	67,664 63,208 (93.4%)	-
California Academy of Sciences	USA	63,523 54,871 (86.4%)	-	-	63,523 54,871 (86.4%)
Western Foundation of Vertebrate Zoology	USA	60,798 53,468 (87.9%)	-	-	60,798 53,468 (87.9%)
CSIRO	Australia	60,192 52,126 (86.6%)	12 0 (0.0%)	-	60,180 52,114 (86.6%)
Dpt. of Environment and Resource Management (Queensland)	Australia	58,653 31,921 (%)	-	58,287 31,611 (54.2%)	366 310 (84.7%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Taiwan Biodiversity Information Facility (TaiBIF)	Chinese Taipei	57,172 31,806 (55.6%)	-	57,172 31,806 (55.6%)	-
EMAN Provider	Canada	48,889 3,147 (6.4%)	48,889 3,147 (6.4%)	-	-
Natural History Museum, University of Oslo	Norway	48,659 16,262 (33.4%)	-	-	48,659 16,262 (33.4%)
Museum Victoria	Australia	45,922 36,033 (78.5%)	-	-	45,922 36,033 (78.5%)
Institute of Nature Conservation, Polish Academy of Sciences	Poland	45,373 44,633 (98.4%)	-	45,373 44,633 (98.4%)	-
British Antarctic Survey	UK	45,008 6 (0.0%)	30 1 (3.3%)	44,978 5 (0.0%)	-
GEO-Tag der Artenvielfalt	Germany	41,313 38,022 (92.0%)	-	41,291 38,007 (92.0%)	22 15 (68.2%)
Delaware Museum of Natural History	USA	39,111 35,247 (90.1%)	-	-	39,111 35,247 (90.1%)
South Australian Museum	Australia	36,888 30,382 (82.4%)	36,888 30,382 (82.4%)	-	-
San Diego Natural History Museum	USA	35,664 30,532 (67.5%)	-	-	35,664 30,532 (67.5%)
University of Kansas Biodiversity Institute	USA	35,334 23,868 (67.5%)	-	-	35,334 23,868 (67.5)
Natural History Museum of Los Angeles County	USA	33,933 28,805 (84.9%)	-	-	33,933 28,805 (84.9%)
UCLA-Dickey Collection	USA	32,931 29,428 (89.4%)	1 1 (100.0%)	-	32,930 29,427 (89.4%)
Royal Belgian Institute of Natural Sciences	Belgium	30,121 24,272 (80.6%)	-	-	30,121 24,272 (80.6%)
Borror Laboratory of Bioacoustics	USA	29,983 27,778 (92.6%)	-	29,983 27,778 (92.6%)	-
Western Australian Museum	Australia	29,417 22,222 (75.5%)	-	-	29,417 22,222 (75.5%)
Administración de Parques Nacionales, Argentina	Argentina	27,466 21,656 (78.8%)	-	27,466 21,656 (78.8%)	-
American Museum of Natural History	USA	27,008 22,643 (83.8%)	-	-	27,008 22,643 (83.8%)
Biodiversitäts-Monitoring Schweiz - BDMCH	Switzerland	26,721 26,480 (99.1%)	-	26,721 26,480 (99.1%)	-
Cornell University Museum of Vertebrates	USA	24,338 20,500 (84.2%)	-	-	24,338 20,500 (84.2%)
UNIBIO, IBUNAM	Mexico	22,090 19,614 (88.8%)	22,090 19,614 (88.8%)	-	-
James R. Slater Museum of Natural History	USA	20,978 18,094 (86.3%)	-	-	20,978 18,094 (86.3%)
Santa Barbara Museum of Natural History	USA	19,178 16,311 (85.1%)	-	-	19,178 16,311 (85.1%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Instituto de Investigación de Recursos Biológicos Alexander von Humboldt	Colombia	18,291 16,086 (87.9%)	17,047 14,845 (87.1%)	1,244 1,241 (99.8%)	-
Facultad de Ciencias, UNAM	Mexico	16,642 15,234 (91.5%)	16,642 15,234 (91.5%)	-	-
Conservation International	USA	15,678 14,433 (92.1%)	-	15,678 14,433 (92.1%)	-
Musée national d'histoire naturelle Luxembourg	Luxembourg	14,630 13,362 (91.3%)	-	14,630 13,362 (91.3%)	-
Instituto de Ciencias Naturales	Colombia	12,993 12,150 (93.5%)	-	-	12,993 12,150 (93.5%)
National Museum of Natural History	USA	12,824 7,005 (54.6%)	2 0 (0.0%)	-	12,822 7,005 (54.6%)
Museum fürNaturkunde Berlin	Germany	10,804 9,971 (92.3%)	-	10,779 9,946 (92.3%)	25 25 (100.0%)
Centre d'estudis de la neu i de la muntanya d'Andorra (CENMA), Institut d'Estudis Andorrans	Andorra	10,120 9,876 (97.6%)	-	10,120 9,876 (97.6%)	-
University of Nebraska State Museum	USA	9,581 8,310 (86.7%)	-	-	9,581 8,310 (86.7%)
Jagiellonian University, Institute of Environmental Sciences	Poland	8,460 7,898 (93.4%)	-	8,460 7,898 (93.4%)	-
Upper Silesian Museum, Bytom	Poland	8,403 5,241 (62.4%)	-	8,403 5,241 (62.4%)	-
Museo Argentino de Ciencias Naturales	Argentina	8,145 6,997 (85.9%)	-	-	8,145 6,997 (85.9%)
New Brunswick Museum	Canada	7,911 6,324 (79.9%)	7,911 6,324 (79.9%)	-	-
Bernice Pauahi Bishop Museum	USA	7,741 5,330 (68.9%)	-	-	7,741 5,330 (68.9%)
Corantioquia	Colombia	7,057 6,238 (88.4%)	-	7,057 6,238 (88.4%)	-
inatura – Erlebnis Naturschau Dornbirn	Austria	6,319 6,098 (96.5%)	6,319 6,098 (96.5%)	-	-
National Museum of Nature and Science, Japan	Japan	5,956 4,543 (76.3%)	-	-	5,956 4,543 (76.3%)
Ireland?	Ireland?	5,913 5,078 (85.9%)	-	-	5,913 5,078 (85.9%)
Queen Victoria Museum and Art Gallery	Australia	5,585 4,143 (74.2%)	5,585 4,143 (74.2%)	-	-
iNaturalist,org	USA	5,325 4,684 (88.0%)	-	5,325 4,684 (88.0%)	-
Netherlands Biodiversity Information Facility (NLBIF)	Netherlands	4,779 806 (16.9%)	-	-	4,779 806 (16.9%)
Isagen	Colombia	4,135 3,895 (94.2%)	11 11 (100.0%)	4,124 3,884 (94.2%)	-
Haus der Natur Salzburg	Austria	3,752 3,749 (99.9%)	3,752 3,749 (99.9%)	-	-

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
National Science Museum of Korea	Korea, Republic of	3,715 2,589 (69.7%)	2,660 1,909 (71.8%)	-	1,055 680 (64.5%)
Tasmanian Museum and Art Gallery	Australia	3,355 2,044 (60.9%)	1 1 (100.0%)	28 16 (57.1%)	3,326 2,027 (60.9%)
Senckenberg	Germany	3,116 2,618 (84.0%)	-	-	3,116 2,618 (84.0%)
University of Colorado Museum of Natural History	USA	3,068 2,515 (82.0%)	-	-	3,068 2,515 (82.0%)
Mokpo Museum of Natural History	Korea, Republic of	2,630 1,525 (58.0%)	2,605 1,514 (58.1%)	-	25 11 (44.0%)
Illinois State University	USA	2,457 2,006 (81.6%)	-	-	2,457 2,006 (81.6%)
Tall Timbers Research Station and Land Conservancy	USA	2,407 2,071 (86.0%)	2,407 2,071 (86.0%)	-	-
Natural History Museum, University of Tartu	Estonia	1,794 1,784 (99.4%)	-	-	1,794 1,784 (99.4%)
Citizen Science - ALA Website	Australia	1,543 1,458 (94.5%)	1,543 1,458 (94.5%)	-	-
Cincinnati Museum Center	USA	1,009 920 (91.2%)	1,009 920 (91.2%)	-	-
Wildlife Institute of India	India	752 606 (80.6%)	752 606 (80.6%)	-	-
PANGAEA - Publishing Network for Geoscientific and Environmental Data	Germany	673 240 (35.7%)	-	673 240 (35.7%)	-
National Chemical Laboratory (via OBIS)	International	647 285 (44.0%)	647 285 (44.0%)	-	-
European Molecular Biology Laboratory Australia	Australia	631 549 (87.0%)	631 549 (87.0%)	-	-
University of Alberta Museums	Canada	476 331 (69.5%)	-	-	476 331 (69.5%)
Ohio State University Insect Collection	USA	469 456 (97.2%)	-	-	469 456 (97.2%)
Wildlife Conservation Society - Madagascar Program (WCS - Mad)	Madagascar	469 460 (98.1%)	-	469 460 (98.1%)	-
Field Study Group of the Dutch Mammal Society	Netherlands	445 321 (72.1%)	-	445 321 (72.1%)	-
Museé national d'Histoire naturelle	France	209 164 (78.5%)	-	-	209 164 (78.5%)
New Mexico Biodiversity Collections Consortium	USA	199 177 (88.9%)	-	-	199 177 (88.9%)
SysTax	Germany	199 140 (70.4%)	199 140 (70.4%)	-	-
Wildlife Sightings	Canada	189 177 (93.7%)	-	189 177 (93.7%)	-
Queensland Museum	Australia	183 177 (96.7%)	-	-	183 177 (96.7%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Botanic Garden and Botanical Museum Berlin-Dahlem	Germany	163 108 (66.3%)	-	115 82 (71.3%)	48 26 (54.2%)
Jagiellonian University, Institute of Zoology	Poland	137 70 (51.1%)	-	137 70 (51.1%)	-
University of Navarra, Museum of Zoology	Spain	105 85 (81.0%)	-	-	105 85 (81.0%)
Gyeryonsan Natural History Museum	Korea, Republic of	53 23 (43.4%)	-	-	53 23 (43.4%)
University of Helsinki, Dpt. of Applied Biology	Finland	45 41 (91.1%)	-	45 41 (91.1%)	-
Michigan State University Museum	USA	9 9 (100.0%)	-	-	9 9 (100.0%)
Nicolaus Copernicus University of Toruń	Poland	6 6 (100.0%)	-	6 6 (100.0%)	-
Humboldt State University	USA	5 5 (100.0%)	-	-	5 5 (100.0%)
Mammal Research Institute, Polish Academy of Sciences	Poland	4 4 (100.0%)	44 4 (100.0%)	-	-
Sam Noble Oklahoma Museum of Natural History	USA	2 0 (0%)	2 0 (0%)	-	-
Jyvaskyla University Museum	Finland	1 1 (100.0%)	-	-	1 1 (100.0%)
University of Silesia, Herbarium KTU	Poland	1 1 (100.0%)	-	-	1 1 (100.0%)

b) Publishers of mammal records

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
UK National Biodiversity Network	UK	521,021 396,214 (76.0%)	521,021 396,214 (76.0%)	-	-
Arctos	USA	455,737 401,284 (88.1%)	-	-	455,737 401,284 (88.1%)
NSW Dpt. of Environment, Climate Change, and Water	Australia	375,532 306,596 (81.6%)	-	375,532 306,596 (81.6%)	-
Service du Patrimoine naturel, Musée national d'Histoire naturelle, Paris	France	334,434 258,876 (77.4%)	-	334,434 258,876 (77.4%)	-
Australian Antarctic Data Centre	Australia	289,554 0 (0%)	119,930 0 (0%)	169,624 0 (0%)	-
Ocean Biogeographic Information System (via OBIS)	International	262,463 2082(0.8%)	3,874 1(0.0%)	258,219 2081(0.8%)	370 0 (0%)
University of Kansas Biodiversity Institute	USA	159,667 144,186 (90.3%)	-	-	159,667 144,186 (90.3%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Field Museum	USA	156,235 132,015 (84.5%)	-	-	156,235 132,015 (84.5%)
Comisión nacional para el conocimiento y uso de la biodiversidad	Mexico	153,422 130,345 (85.0%)	147,755 125,501 (84.9%)	5,667 4,844 (85.5%)	-
South Australia, Department of Environment and Natural Resources	Australia	125,906 92,962 (73.8%)	31 20 (64.5%)	120,168 88,613 (73.7%)	5,707 4,329 (75.9%)
GBIF-Spain	Spain	103,041 87,740 (85.2%)	-	99,615 85,978 (86.3%)	3,426 1,762 (51.4%)
National Museum of Natural History	USA	98,159 82,376 (83.9%)	-	8 0 (0%)	98,151 82,376 (83.9%)
Mammal Research Institute, Polish Academy of Sciences	Poland	86,239 82,915 (96.1%)	753 747 (99.2%)	-	85,486 82,168 (96.1%)
Natural History Museum of Los Angeles County	USA	79,770 68,834 (86.3%)	-	-	79,770 68,834 (86.3%)
National Biodiversity Data Centre	Ireland	73,067 62,727 (85.8%)	-	-	73,067 62,727 (85.8%)
Australian Museum	Australia	71,124 54,736 (77.0%)	-	-	71,124 54,736 (77.0%)
BeBIF Provider	Belgium	69,848 62,665 (89.7%)	-	5,763 4,764 (82.7%)	64,085 57,901 (90.4%)
University of Navarra, Museum of Zoology	Spain	60,888 55,009 (90.3%)	-	1,878 1,858 (98.9%)	59,010 53,151 (90.1%)
Dpt. of Natural Resources, Environment, The Arts and Sport, Northern Territory of Australia	Australia	56,085 33,864 (60.4%)	-	56,085 33,864 (60.4%)	-
University of Washington Burke Museum	USA	53,415 37,178 (69.6%)	-	-	53,415 37,178 (69.6%)
James R. Slater Museum of Natural History	USA	49,585 45,673 (92.1%)	-	-	49,585 45,673 (92.1%)
Western Australian Museum	Australia	44,644 35,351 (79.2%)	-	-	44,644 35,351 (79.2%)
Scientific Committee on Antarctic Research - Marine Biodiversity Information Network (SCAR-MarBIN)	International	41,863 0 (0%)	41,739 0 (0%)	124 0 (0%)	-
Sam Noble Oklahoma Museum of Natural History	USA	36,269 25,681 (70.8%)	36,269 25,681 (70.8%)	-	-
Royal Belgian Institute of Natural Sciences	Belgium	32,736 29,287 (89.5%)	-	-	32,736 29,287 (89.5%)
CSIRO	Australia	31,727 25,205 (79.4%)	4,503 3,521 (78.2%)	-	27,224 21,684 (79.7%)
Israel Nature and Parks Authority	Israel / EU - BioCASE	30,754 25,909 (84.2%)	-	30,754 25,909 (84.2%)	-
UNIBIO, IBUNAM	Mexico	30,197 25,149 (83.3%)	30,197 25,149 (83.3%)	-	-
Museum Victoria	Australia	28,568 22,947 (80.3%)	-	-	28,568 22,947 (80.3%)
Louisiana State University Museum of Natural Science	USA	27,866 23,784 (85.4%)	-	-	27,866 23,784 (85.4%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Michigan State University Museum	USA	27,768 24,803 (89.3%)	-	-	27,768 24,803 (89.3%)
ArtDatabanken	Sweden	27,674 22,403 (81.0%)	-	27,674 22,403 (81.0%)	-
South Australian Museum	Australia	23,997 15,298 (63.7%)	23,456 15,062 (64.2%)	134 23 (17.2%)	407 213 (52.3%)
California Academy of Sciences	USA	23,411 18,965 (81.0%)	-	-	23,411 18,965 (81.0%)
Danish Biodiversity Information Facility	Denmark	21,549 10,863 (50.4%)	-	21,469 10,797 (50.3%)	80 66 (82.5%)
Administración de Parques Nacionales, Argentina	Argentina	19,136 13,891 (72.6%)	117 96 (82.1%)	12,739 8,035 (63.1%)	6,280 5,760 (91.7%)
Natural History Museum, University of Oslo	Norway	18,499 9,345 (50.5%)	362 16(4.4%)	1 0 (0%)	18,136 9,329 (51.4%)
The Norwegian Biodiversity Information Centre (NBIC)	Norway	18,314 15,914 (86.9%)	-	18,314 15,914 (86.9%)	-
British Antarctic Survey	UK	17,341 0 (0%)	15 0 (0%)	17,326 0 (0%)	-
UCLA-Dickey Collection (UCLA-Dickey)	USA	16,553 14,106 (85.2%)	-	-	16,553 14,106 (85.2%)
Museo Argentino de Ciencias Naturales	Argentina	14,514 10,265 (70.7%)	-	-	14,514 10,265 (70.7%)
Finnish Museum of Natural History	Finland	14,469 8,874 (61.3%)	-	14,469 8,874 (61.3%)	-
New York State Museum (NYSM)	USA	13,388 12,667 (94.6%)	-	-	13,388 12,667 (94.6%)
Yale University Peabody Museum	USA	11,881 9,565 (80.5%)	-	-	11,881 9,565 (80.5%)
Musée national d'histoire naturelle Luxembourg	Luxembourg	11,754 11,033 (93.9%)	-	11,754 11,033 (93.9%)	-
New Mexico Biodiversity Collections Consortium	USA	11,679 10,752 (92.1%)	-	-	11,679 10,752 (92.1%)
Santa Barbara Museum of Natural History	USA	9,633 7,773 (80.7%)	-	-	9,633 7,773 (80.7%)
PANGAEA - Publishing Network for Geoscientific and Environmental Data	Germany	7,884 3,526 (44.7%)	-	7,884 3,526 (44.7%)	-
American Museum of Natural History	USA	7,704 6,603 (85.7%)	-	-	7,704 6,603 (85.7%)
University of Colorado Museum of Natural History	USA	7,598 7,087 (93.3%)	-	-	7,598 7,087 (93.3%)
University of Warsaw, Dpt. of Ecology	Poland	6,834 6,673 (97.6%)	489 352 (72.0%)	6,345 6,321 (99.6%)	-
inatura – Erlebnis Naturschau Dornbirn	Austria	6,068 6,061 (99.9%)	6068 6061 (99.9%)	-	-
Museum of Comparative Zoology, Harvard University	USA	5,200 3,855 (74.1%)	-	-	5,200 3,855 (74.1%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Instituto de Ciencias Naturales	Colombia	4,985 4,500 (90.3%)	-	-	4,985 4,500 (90.3%)
Queen Victoria Museum and Art Gallery	Australia	4,693 4,087 (87.1%)	4693 4087 (87.1%)	-	-
Texas Cooperative Wildlife Collection	USA	4,586 4,326 (94.3%)	-	-	4,586 4,326 (94.3%)
Centre d'estudis de la neu i de la muntanya d'Andorra (CENMA), Institut d'Estudis Andorrans	Andorra	4,410 4,323 (98.0%)	-	4,410 4,323 (98.0%)	-
TELDAP	Chinese Taipei	3,643 3,405 (93.5%)	-	3,641 3,403 (93.5%)	2 2 (100.0%)
New Mexico Museum of Natural History and Science	USA	3,270 170(5.2%)	-	-	3,270 170(5.2%)
Cornell University Museum of Vertebrates	USA	2,983 2,733 (91.6%)	-	-	2,983 2,733 (91.6%)
Conservation International	USA	2,734 2,345 (85.8%)	-	2,734 2,345 (85.8%)	-
Tasmanian Museum and Art Gallery	Australia	2,710 1,273 (47.0%)	-	67 51 (76.1%)	2,643 1,222 (46.2%)
Bernice Pauahi Bishop Museum	USA	2,512 1,457 (58.0%)	-	-	2,512 1,457 (58.0%)
Avian Knowledge Network	USA	2,438 470 (19.3%)	-	2,438 470 (19.3%)	-
Field Study Group of the Dutch Mammal Society	Netherlands	2,167 2,010 (92.8%)	-	2,167 2,010 (92.8%)	-
GEO-Tag der Artenvielfalt	Germany	1,987 1,776 (89.4%)	-	1,987 1,776 (89.4%)	-
GBIF-Sweden	Sweden	1,961 898 (45.8%)	-	451 78 (17.3%)	1,510 820 (54.3%)
Instituto de Investigación de Recursos Biológicos Alexander von Humboldt	Colombia	1,910 1,820 (95.3%)	897 807 (90.0%)	1,013 1013 (100.0%)	-
Corantioquia	Colombia	1,735 1,168 (67.3%)	-	1,735 1,168 (67.3%)	-
Dutch Mammal Society	Netherlands	1,626 0 (0%)	-	1,626 0 (0%)	-
EMAN Provider	Canada	1,414 14(1.0%)	1,414 14(1.0%)	-	-
Museum für Naturkunde Berlin	Germany	1,404 652 (46.4%)	-	1,378 628 (45.6%)	26 24 (92.3%)
Institute of Research for Development	France	1,321 0 (0%)	-	1,321 0 (0%)	-
United States Geological Survey	USA	1,136 3(0.3%)	-	1,124 3(0.3%)	12 0 (0%)
Institute of Nature Conservation, Polish Academy of Sciences	Poland	1,113 825 (74.1%)	-	1,113 825 (74.1%)	-
Borror Laboratory of Bioacoustics	USA	1,041 426 (40.9%)	-	1,041 426 (40.9%)	-

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
University of Michigan Museum of Zoology	USA	1,034 1,009 (97.6%)	-	-	1,034 1,009 (97.6%)
Isagen	Colombia	1,031 722 (70.0%)	10 6 (60.0%)	1,021 716 (70.1%)	-
iNaturalist.org	USA	1,018 833 (81.8%)	-	1,018 833 (81.8%)	-
Natural History Museum, University of Tartu	Estonia	996 914 (91.8%)	-	-	996 914 (91.8%)
Association for Nature WOLF	Poland	987 878 (89.0%)	-	987 878 (89.0%)	-
Illinois State University	USA	827 735 (88.9%)	-	-	827 735 (88.9%)
University of Alberta Museums	Canada	822 551 (67.0%)	-	-	822 551 (67.0%)
KBIF Data Repository	Korea, Republic of	806 622 (77.2%)	746 602 (80.7%)	-	60 20 (33.3%)
National Museum of Nature and Science, Japan	Japan	309 278 (90.0%)	-	-	309 278 (90.0%)
Ohio State University Insect Collection	USA	253 195 (77.1%)	-	-	253 195 (77.1%)
European Forest Institute	Finland	226 220 (97.3%)	226 220 (97.3%)	-	-
Wildlife Conservation Society - Madagascar Program (WCS - Mad)	Madagascar	189 173 (91.5%)	-	189 173 (91.5%)	-
University of Minnesota Bell Museum of Natural History	USA	172 172 (100.0%)	-	-	172 172 (100.0%)
Citizen Science - ALA Website	Australia	168 143 (85.1%)	168 143 (85.1%)	-	-
Queensland Museum	Australia	136 121 (89.0%)	-	-	136 121 (89.0%)
National Chemical Laboratory (via OBIS)	International	127 33 (26.0%)	127 33 (26.0%)	-	-
Haus der Natur Salzburg	Austria	108 108 (100.0%)	108 108 (100.0%)	-	-
Geocollections of Estonia	Estonia	67 3(4.5%)	67 3(4.5%)	-	-
Botanic Garden and Botanical Museum Berlin-Dahlem	Germany	46 37 (80.4%)	-	17 14 (82.4%)	29 23 (79.3%)
University of Helsinki, Dpt. of Applied Biology	Finland	39 35 (89.7%)	-	39 35 (89.7%)	-
Netherlands Biodiversity Information Facility (NLBIF)	Netherlands	34 3(8.8%)	-	-	34 3(8.8%)
National Science Museum of Korea	Korea, Republic of	31 20 (64.5%)	-	-	31 20 (64.5%)
Jagiellonian University, Institute of Zoology	Poland	30 17 (56.7%)	-	30 17 (56.7%)	-

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Staatliche Naturwissenschaftliche Sammlungen Bayerns	Germany	27 26 (96.3%)	-	-	27 26 (96.3%)
Wildlife Sightings	Canada	25 12 (48.0%)	-	25 12 (48.0%)	-
European Molecular Biology Laboratory Australia	Australia	22 11 (50.0%)	22 11 (50.0%)	-	-
Senckenberg	Germany	20 2 (10.0%)	-	-	20 2 (10.0%)
University of Nebraska State Museum	USA	11 7 (63.6%)	-	-	11 7 (63.6%)
Biologiezentrum Linz Oberösterreich	Austria	7 1 (14.3%)	7 1 (14.3%)	-	-
University of Silesia, Laboratory of Botanical Documentation - Herbarium KTU	Poland	6 6 (100.0%)	-	-	6 6 (100.0%)
Museum of Texas Tech University (TTU)	USA	5 5 (100.0%)	-	-	5 5 (100.0%)
Upper Silesian Museum, Bytom	Poland	4 3 (75.0%)	-	-	4 3 (75.0%)
South African National Biodiversity Institute	South Africa	4 1 (25.0%)	-	-	4 1 (25.0%)
University of Texas at El Paso	USA	2 1 (50.0%)	-	-	2 1 (50.0%)
IHAR	Poland	1 0 (0%)	-	1 0 (0%)	-
Nicolaus Copernicus University of Toruń	Poland	1 0 (0%)	-	1 0 (0%)	-
University of Gdańsk, Bird Migration Research Station	Poland	1 0 (0%)	-	1 0 (0%)	-
University of Gdańsk, Dpt. of Plant Taxonomy and Nature Conservation	Poland	1 0 (0%)	-	-	1 0 (0%)
Royal Ontario Museum	Canada	1 1 (100.0%)	-	-	1 1 (100.0%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)

c) Publishers of amphibian records

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
National Museum of Natural History	USA	233,924 198,468 (84.8%)	2 2 (100.0%)	-	233,922 198,466 (84.8%)
Arctos	USA	136,381 120,466 (88.3%)	-	-	136,381 120,466 (88.3%)
Museum of Comparative Zoology, Harvard University	USA	98,370 77,722 (79.0%)	-	-	98,370 77,722 (79.0%)
UK National Biodiversity Network	UK	96,559 94,502 (97.9%)	96,559 94,502 (97.9%)	-	-
California Academy of Sciences	USA	89,345 73,794 (82.6%)	-	-	89,345 73,794 (82.6%)
Australian Museum	Australia	85,814 71,155 (82.9%)	-	-	85,814 71,155 (82.9%)
NSW Dpt. of Environment, Climate Change, and Water	Australia	72,921 61,468 (84.3%)	-	72,921 61,468 (84.3%)	-
Chengdu Institute of Biology, Chinese Academy of Science	Chinese Taipei	58,164 48,396 (83.2%)	-	-	58,164 48,396 (83.2%)
Natural History Museum of Los Angeles County	USA	43,768 37,214 (85.0%)	-	-	43,768 37,214 (85.0%)
GBIF-Spain	Spain	38,174 35,623 (93.3%)	-	28,393 27,058 (95.3%)	9,781 8,565 (87.6%)
Museum Victoria	Australia	34,845 31,303 (89.8%)	-	-	34,845 31,303 (89.8%)
Comisión nacional para el conocimiento y uso de la biodiversidad	Mexico	28,282 20,774 (73.5%)	24,616 17,712 (72.0%)	3,666 3,062 (83.5%)	-
South Australia, Department of Environment and Natural Resources	Australia	25,147 23,611 (93.9%)	-	24,340 22,945 (94.3%)	807 666 (82.5%)
Musée d'histoire naturelle de la Ville de Genève - MHNG	Switzerland	24,894 22,218 (89.3%)	-	-	24,894 22,218 (89.3%)
Bird Studies Canada	Canada	24,856 18,852 (75.8%)	-	24,856 18,852 (75.8%)	-
Western Australian Museum	Australia	23,294 20,508 (88.0%)	-	-	23,294 20,508 (88.0%)
Royal Ontario Museum	Canada	23,182 19,307 (83.3%)	-	-	23,182 19,307 (83.3%)
ArtDatabanken	Sweden	18,660 16,196 (86.8%)	-	18,660 16,196 (86.8%)	-
University of Kansas Biodiversity Institute	USA	18,2533 24,438 (13.4%)	18,2533 24,438 (13.4%)	-	-
Canadian Museum of Nature	Canada	17,371 12,232 (70.4%)	-	-	17,371 12,232 (70.4%)
Service du Patrimoine naturel, Musée national d'Histoire naturelle, Paris	France	16,352 14,665 (89.7%)	-	16,352 14,665 (89.7%)	-

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Instituto de Ciencias Naturales	Colombia	14,626 12,749 (87.2%)	-	-	14,626 12,749 (87.2%)
Yale University Peabody Museum	USA	13,682 12,082 (88.3%)	-	-	13,682 12,082 (88.3%)
Museo Argentino de Ciencias Naturales	Argentina	13,055 10,249 (78.5%)	-	-	13,055 10,249 (78.5%)
South Australian Museum	Australia	13,031 11,000 (84.4%)	13,031 11,000 (84.4%)	-	-
New Mexico Biodiversity Collections Consortium	USA	12,049 10,257 (85.1%)	-	-	12,049 10,257 (85.1%)
Museum of Southwestern Biology, Division of Amphibians and Reptiles	USA	11,255 9,579 (85.1%)	11,255 9,579 (85.1%)	-	-
Dpt. of Natural Resources, Environment, The Arts and Sport, Northern Territory of Australia	Australia	10,808 9,334 (86.4%)	-	10,808 9,334 (86.4%)	-
San Diego Natural History Museum	USA	10,617 8,354 (78.7%)	-	-	10,617 8,354 (78.7%)
CSIRO	Australia	9,190 7,579 (82.5%)	6,290 4,950 (78.7%)	-	2,900 2,629 (90.7%)
Cornell University Museum of Vertebrates	USA	9,078 7,915 (87.2%)	-	-	9,078 7,915 (87.2%)
Alabama Museum of Natural History	USA	8,931 7,325 (82.0%)	8,931 7,325 (82.0%)	-	-
South African National Biodiversity Institute	South Africa	7,107 6,491 (91.3%)	-	-	7,107 6,491 (91.3%)
Musée national d'histoire naturelle Luxembourg	Luxembourg	6,997 5,320 (76.0%)	-	6,997 5,320 (76.0%)	-
Bernice Pauahi Bishop Museum	USA	6,853 5,251 (76.6%)	-	-	6,853 5,251 (76.6%)
EMAN Provider	Canada	6,639 5,090 (76.7%)	6,639 5,090 (76.7%)	-	-
TELDAP	Chinese Taipei	6,596 6,379 (96.7%)	-	6,596 6,379 (96.7%)	-
Royal Belgian Institute of Natural Sciences	Belgium	6,560 4,961 (75.6%)	-	-	6,560 4,961 (75.6%)
Danish Biodiversity Information Facility	Denmark	6,274 4,968 (79.2%)	498 498 (100.0%)	3,422 2,643 (77.2%)	2,354 1,827 (77.6%)
Natural History Museum, University of Oslo	Norway	6,221 5,065 (81.4%)	-	5,253 4,615 (87.9%)	968 450 (46.5%)
Sternberg Museum of Natural History	USA	5,110 3,447 (67.5%)	-	-	5,110 3,447 (67.5%)
Zoological Institute, Russian Academy of Sciences, St. Petersburg (via the Society for the Management of Electronic Biodiversity Data)	Russia	4,534 3,285 (72.5%)	4,534 3,285 (72.5%)	-	-
National Biodiversity Data Centre	Ireland	4,033 4,032 (100.0%)	-	-	4,033 4,032 (100.0%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
James R. Slater Museum of Natural History	USA	3,843 3,303 (85.9%)	-	-	3,843 3,303 (85.9%)
UNIBIO, IBUNAM	Mexico	3,490 257(7.4%)	418 257 (61.5%)	-	3,072 0 (0%)
Institute of Nature Conservation, Polish Academy of Sciences	Poland	3,185 2,377 (74.6%)	-	3,185 2,377 (74.6%)	-
Cincinnati Museum Center	USA	3,005 2,607 (86.8%)	3,005 2,607 (86.8%)	-	-
KBIF Data Repository	Korea, Republic of	3,396 3,074 (90.5%)	3,396 3,074 (90.5%)	-	-
University of Alberta Museums	Canada	2,679 2,413 (90.1%)	-	-	2,679 2413 (90.1%)
Finnish Museum of Natural History	Finland	2,514 791 (31.5%)	-	2,514 791 (31.5%)	-
Raffles Museum of Biodiversity Research	BioNET-ASEANET	2,439 2,089 (85.6%)	-	-	2,439 2089 (85.6%)
University of Colorado Museum of Natural History	USA	2,118 1,661 (78.4%)	-	-	2,118 1,661 (78.4%)
Administración de Parques Nacionales, Argentina	Argentina	2,010 1,699 (84.5%)	-	99 41 (41.4%)	1,911 1,658 (86.8%)
University of Warsaw, Dpt. of Ecology	Poland	1,945 3(0.2%)	-	1,945 3 (0.2%)	-
The Norwegian Biodiversity Information Centre (NBIC)	Norway	1,872 1,622 (86.6%)	-	1,872 1,622 (86.6%)	-
Sam Noble Oklahoma Museum of Natural History	USA	1,770 706 (39.9%)	1,770 706 (39.9%)	-	-
United States Geological Survey	USA	1,752 172(9.8%)	-	1,067 93 (8.7%)	685 79 (11.5%)
Haus der Natur Salzburg	Austria	1,741 818 (47.0%)	1,741 818 (47.0%)	-	-
Białowieża National Park	Poland	1,723 679 (39.4%)	-	1,723 679 (39.4%)	-
Conservation International	USA	1,460 1,159 (79.4%)	-	1,460 1,159 (79.4%)	-
Royal Museum for Central Africa, Belgium	Belgium	1,413 1,036 (73.3%)	-	-	1,413 1,036 (73.3%)
University of Nevada, Reno	USA	1,257 742 (59.0%)	-	-	1,257 742 (59.0%)
Redpath Museum, McGill University	Canada	1,113 919 (82.6%)	-	-	1,113 919 (82.6%)
GEO-Tag der Artenvielfalt	Germany	1,113 742 (66.7%)	2 0 (0%)	1,111 742 (66.8%)	-
Staatliches Museum für Naturkunde Stuttgart	Germany	1,107 758 (68.5%)	-	-	1,107 758 (68.5%)
Queensland Museum	Australia	871 852 (97.8%)	-	-	871 852 (97.8%)

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
Santa Barbara Museum of Natural History	USA	744 619 (83.2%)	-	-	744 619 (83.2%)
Borror Laboratory of Bioacoustics	USA	721 326 (45.2%)	-	721 326 (45.2%)	-
Queen Victoria Museum and Art Gallery	Australia	716 708 (98.9%)	716 708 (98.9%)	-	-
Senckenberg	Germany	615 552 (89.8%)	-	-	615 552 (89.8%)
Isagen	Colombia	611 557 (91.2%)	-	611 557 (91.2%)	-
iNaturalist.org	USA	565 479 (84.8%)	-	565 479 (84.8%)	-
University of Navarra, Museum of Zoology	Spain	525 466 (88.8%)	-	25 23 (92.0%)	500 443 (88.6%)
Israel Nature and Parks Authority	Israel / EU - BioCASE	485 338 (69.7%)	-	485 338 (69.7%)	-
Instituto de Investigación de Recursos Biológicos Alexander von Humboldt	Colombia	400 338 (84.5%)	-	400 338 (84.5%)	-
Netherlands Biodiversity Information Facility (NLBIF)	Netherlands	373 220 (59.0%)	6 6 (100.0%)	367 214 (58.3%)	-
GBIF-Sweden	Sweden	326 230 (70.6%)	-	104 84 (80.8%)	222 146 (65.8%)
Museum für Naturkunde Berlin	Germany	283 188 (66.4%)	-	283 188 (66.4%)	-
Avian Knowledge Network	USA	281 257 (91.5%)	-	281 257 (91.5%)	-
National Museum of Nature and Science, Japan	Japan	238 189 (79.4%)	-	-	238 189 (79.4%)
Milwaukee Public Museum	USA	215 102 (47.4%)	-	-	215 102 (47.4%)
Tasmanian Museum and Art Gallery	Australia	200 192 (96.0%)	-	4 3 (75.0%)	196 189 (96.4%)
Corantioquia	Colombia	141 91 (64.5%)	-	141 91 (64.5%)	-
Wildlife Conservation Society - Madagascar Program	Madagascar	139 120 (86.3%)	-	139 120 (86.3%)	-
Field Study Group of the Dutch Mammal Society	Netherlands	135 114 (84.4%)	-	135 114 (84.4%)	-
American Museum of Natural History	USA	110 0 (0%)	-	-	110 0 (0%)
Centre d'estudis de la neu i de la muntanya d'Andorra (CENMA), Institut d'Estudis Andorrans	Andorra	106 72 (67.9%)	-	106 72 (67.9%)	-
Citizen Science - ALA Website	Australia	63 42 (66.7%)	63 42 (66.7%)	-	-
inatura – Erlebnis Naturschau Dornbirn	Austria	55 42 (76.4%)	55 42 (76.4%)	-	-

Table V.2.S7 (continued)

Data publisher	Country	Records total / valid (% of total)	Unknown total / valid (% of total)	Observations total / valid (% of total)	Specimens total / valid (% of total)
University of Minnesota Bell Museum of Natural History	USA	43 43 (100.0%)	-	-	43 43 (100.0%)
Wildlife Sightings	Canada	30 29 (96.7%)	-	30 29 (96.7%)	-
Zoologisches Forschungsinstitut und Museum Alexander Koenig	Germany	2 0 (0%)	-	-	2 0 (0%)
Botanic Garden and Botanical Museum Berlin-Dahlem	Germany	2 1 (50.0%)	-	2 1 (50.0%)	-
European Molecular Biology Laboratory Australia	Australia	17 17 (100.0%)	17 17 (100.0%)	-	-
SysTax	Germany	167 150 (89.8%)	167 150 (89.8%)	-	-
Michigan State University Museum	USA	16 16 (100.0%)	-	-	16 16 (100.0%)
National Chemical Laboratory (via OBIS)	International	10 2 (20.0%)	10 2 (20.0%)	-	-
Geocollections of Estonia	Estonia	1 0 (0%)	1 0 (0%)	-	-
Carnegie Museums	USA	1 1 (100.0%)	-	-	1 1 (100.0%)

Supplementary information - Chapter 3

Global drivers of species-level variation in mobilized occurrence information

Carsten Meyer, Walter Jetz, Robert P. Guralnick and Holger Kreft

Supplementary Text

SI V.3.1.1 Mammal distribution data

We focused on records aggregated via the Global Biodiversity Information Facility (GBIF) as a representation of international efforts to mobilize biodiversity data and as GBIF is by far the largest such effort in geographical and taxonomic scope (Edwards, 2000; Graham *et al.*, 2004). We received 5,376,737 geo-referenced mammal records from GBIF in October 2012, of which we extracted 5,140,771 records with potentially sensible geographical coordinates (Longitude: -180° – +180°, Latitude: -90° – +90°) reported with a precision of at least 0.1 degree. We excluded 564,978 records that did not have either a binomial or trinomial scientific name, a further 50,369 records for which the 'basis of record' field did not indicate 'preserved specimen', 'observation', or 'unknown' (most of which are observation records), and 839 records that were reportedly collected before the year 1850, leaving 4,524,585 records. We validated these taxonomically and geographically (see below), which left 2,849,075 records for further analyses.

We used extent-of-occurrence range map polygons (IUCN, 2010) to delimit the current native ranges of the World's terrestrial mammals (excluding cetaceans, pinnipeds and sirenians; *N*=5,270). These range maps were originally drawn by species experts based on a variety of data sources, including occurrence records as well as inventory, survey, atlas and literature data, and represent the most complete and consistent data set available for mammal distributions globally. Species delimitations adopted by the IUCN for their range map and Red List data (IUCN, 2010) partly differ from the taxonomy (Wilson & Reeder, 2005) underlying

most trait and phylogenetic datasets. To link the two distribution data sets, we always adopted the more inclusive species concept, i.e., we merged range maps of species that are lumped by the taxonomy of Wilson & Reeder (2005), and averaged trait values and reduced nodes of the phylogenetic tree for species lumped by the IUCN. We excluded 3 terrestrial species with largely marine ranges (polar bear and two otter species). This resulted in a total of 5,057 accepted terrestrial mammal species. We focused our analyses on the 3,625 species with at least one validated record.

Species concepts followed by collectors and curators of records mobilized via aggregative data networks like GBIF are usually unknown. To account for this uncertainty, we combined all scientific names (including synonyms, subspecies and spelling variants) fully or partly included in our accepted species concepts from three existing taxonomic databases (Wilson & Reeder, 2005; IUCN, 2010; ITIS, 2012; compare Meyer *et al.*, (2015)). We used the resulting 'synonym table' to link GBIF records to our accepted species. We excluded records likely referring to domesticated forms. We inferred the taxonomic identities of GBIF records with ambiguous scientific names (such as *pro parte* synonyms) from spatial overlays with the range maps of 'candidate species', i.e., those accepted species to which the name could potentially refer. To validate records geographically and exclude ambiguous records, we reduced our dataset to those records that fell within a 50-km buffer around the range map of only one of its candidate species. We note that this approach may lead to the exclusion of 'good' occurrence records collected outside of range maps if the maps do not encompass the full extent of occurrence of the species or if ranges have contracted since the collection of records.

SI V.3.1.2 Testing for taxonomic bias and relative taxonomic and geographical species-level biases

We performed nested type III-*ANOVA*s to test whether occurrence information is biased towards species in certain mammal orders or families (Garamszegi & Møller, 2012; Table V.3.S2 A). We performed type III-*ANOVA*s to test for relative effects of zoogeographical realm and order memberships (Table V.3.S2 B).

SI V.3.1.3 Modeling whether or not species have any records mobilized via GBIF

We used similar nested *ANOVA*s to test whether missing species (i.e., species without any mobilized records remaining after validation) are randomly distributed across the mammal

taxonomy (Table V.3.S2 A), and whether they are more clearly distributed among zoogeographical realms than among mammal orders (Table V.3.S2 B). We found significant higher-taxonomic and realm-specific bias, i.e., missing species are not randomly distributed among orders or geographical assemblages (Table V.3.S2 B). We used generalized linear models (GLM) with a quasi-binomial distribution to model whether species have any records mobilized via GBIF, using the same 13 predictor variables as in the *record count* and *coverage* models (Table V.3.S6). We expected the same relationships as with *record count*. The directions of significant relationships are all in line with our hypotheses, but most hypotheses on species attributes found no or only limited support. There is a comparatively weak negative effect of foraging stratum, suggesting that flying or arboreal mammals are more likely to have no mobilized records. Years since description and public interest are relatively weak positive predictors. Similar to our results for *record count* and *range coverage*, we found species' having any mobilized records to be best predicted by range geometry and socio-economic factors: range size, area appeal, proximity to institutions and financial resources (Table V.3.S6).

SI V.3.1.4 Modeling socio-economic drivers of geographical bias

While it is difficult to hypothesize links between *geographical bias* and species attributes, *geographical bias* should be high if the geography of socio-economic conditions causes high *record counts* in some and comparably low *record counts* in other parts of the range. Rather than the range-wide means, we thus used two measures of within-range variation in socio-economic conditions to model *geographical bias* (Table V.3.S5, Fig. II.3.5). The rationale is that strong geographical variation in socio-economic factors within ranges should lead to high levels of data aggregation and *geographical bias* of sampling locations to those range parts where conditions are more favorable of record collection and mobilization. We sampled the four socio-economic factors at 100 random points within each range. We used the coefficient of variation (*cv*) among these local measurements as a measure of within-range variation in socio-economic conditions. Additionally, we calculated the Pearson's correlation coefficient between two distance matrices, one containing the Euclidean distances in socio-economic factors between all pairs of measurements at random points and the other containing the geographical great-circle distances (in km) between random points. This measure has high scores if high values of socio-economic factors are concentrated in one extreme of the range and low values in the other extreme. We did not $\log_{10}$-transform these measures, as resulting effects would be difficult to interpret.

SI V.3.1.5 Additional tests for effects of abundance-related traits

For a given body size, abundance in mammals is negatively correlated with dietary level (Robinson & Redford, 1986). However, this relationship may only show if additionally accounting for habitat (Robinson & Redford, 1986), therefore we tested whether coefficients of dietary level in the global minimum adequate models of *record count* and *range coverage* would decrease (i.e., show stronger negative effects) when additionally including habitat, calculated as percentage overlap of ranges with different biomes (Olson *et al.*, 2001). Dietary level is not retained in the original MAM of *record count*, but when including habitat as a fixed covariate in all candidate model subsets, it is retained in the MAM with a standardized coefficient of -0.31 (P=0.016). The standardized coefficient of dietary level in the model of *range coverage* decreased from -0.028 (P=0.08) to to -0.035 (P= 0.03). Thus, the hypothesis that dietary level affects occurrence information through its indirect effect on species abundances is not rejected, but nevertheless finds only limited support since the standardized coefficients are still smaller compared to those of range geometry and socio-economic factors (which remained similar to the original model (Table V.3.S4).

In another side analysis, we tested for effects of a more direct measure of abundance by including population density in global models of record number and *range coverage* for 844 species with available data (Jones *et al.*, 2009), along with the 13 original predictor variables. Population density showed a significant but weak positive effect (β_{GLM}=0.38, P=0.007) on record number. Thus the hypothesis that population density affects *record counts* is not rejected, but it too finds only limited support from the low relative importance compared to most other variables (dietary level: β_{GLM}=0.29, P=0.03*; foraging stratum: β_{GLM}=-0.55, P=0.0005, public interest: β_{GLM}=0.67, P<<0.001, range size: β_{GLM}=3.78, P<<0.001, range shape irregularity: β_{GLM}=0.56, P<0.001, area appeal: β_{GLM}=0.63, P=0.004, proximity to institutions: β_{GLM}=1.96, P<<0.001, GBIF participation: β_{GLM}=-0.98, P<<0.001, financial resources: β_{GLM}=1.00, P<<0.001). Population density was not retained in the minimum adequate model of *range coverage*.

SI V.3.1.6 Testing for spatial and phylogenetic autocorrelation

We tested for spatial autocorrelation in model residuals, using Moran's I. Because distances between ranges based on range centroids do not account for differences in range size, shape and overlap, we used a distance matrix that for each pair of species contained the mean distance between 100 random points of each range. Residual spatial autocorrelation was in

part significant, but generally low (up to 0.2). We tested for residual phylogenetic autocorrelation with Abouheif's adaptation of Moran's I, based on the phylogenetic tree of (Fritz *et al.*, 2009)). Residual phylogenetic autocorrelation was consistently non-significant or very low (Fig. V.3.S3).

SI V.3.1.7 Limitations of this study

To our knowledge, this is the most comprehensive assessment of drivers of species-level occurrence information to date, and the first to investigate the relative contribution of species attributes, range geometry, and socio-economic factors. We tested these three major groups of hypotheses using a large set of species- and site-specific factors, but acknowledge that, survey-specific factors like sampling method, observer experience, or seasonal changes in species abundances might also play a role (Iknayan *et al.*, 2013). To limit the number of hypotheses, we only included the four socioeconomic variables that were consistently important (across different spatial grain sizes) for predicting global record density and inventory completeness in mammals at the assemblage level (out of twelve tested socio-economic hypotheses (Meyer *et al.*, 2015)). However, given the strong effects of the geographical focus of the analysis in this study, we cannot rule out that globally unimportant socio-economic factors might be important for influencing regional occurrence information, which would have to be investigated further.

While data on further detectability-related traits like e.g. coloration, fossoriality or vagility were not available, consistently weak effects of the tested attributes lead us to conclude that detectability does not greatly impact global mammal occurrence information. The large proportion of variation in range coverage explained jointly by range geometry and socio-economic factors demonstrates that disentangling their separate influences remains difficult. However, our results clearly demonstrate a dominance of geographical over species-specific factors as drivers of species-level bias in occurrence information.

Supplementary Figures

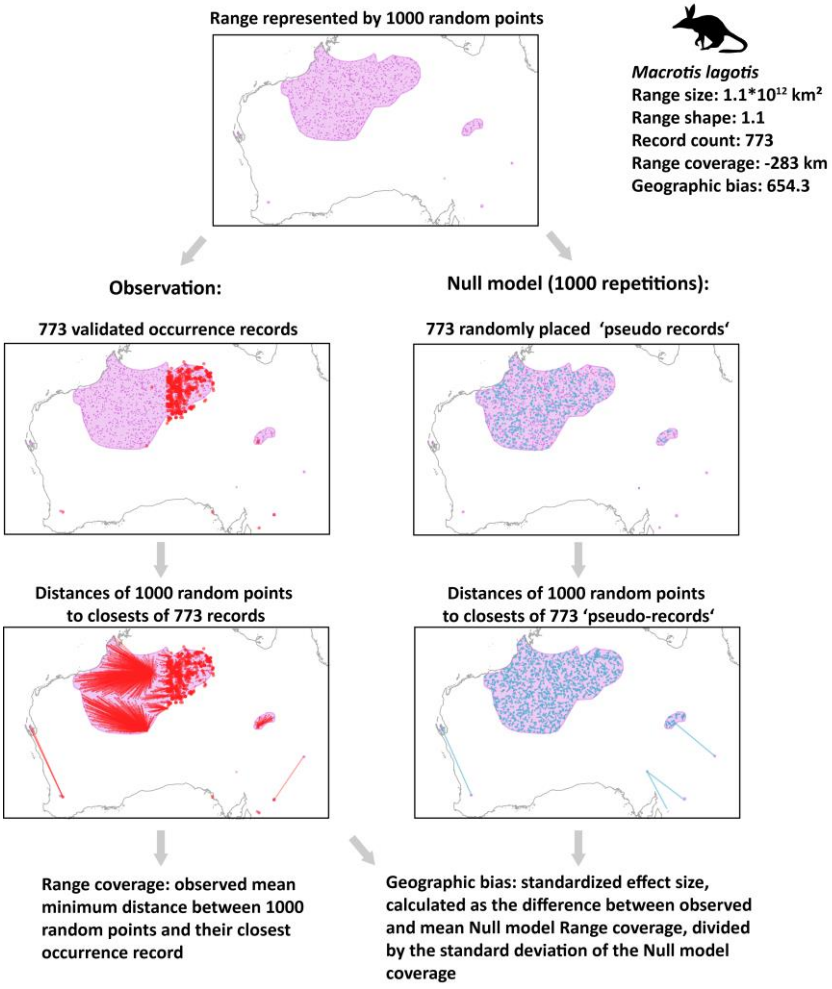


Figure V.3.S1. Visualization of calculation of *range coverage* and *geographical bias* in mobilized occurrence records.

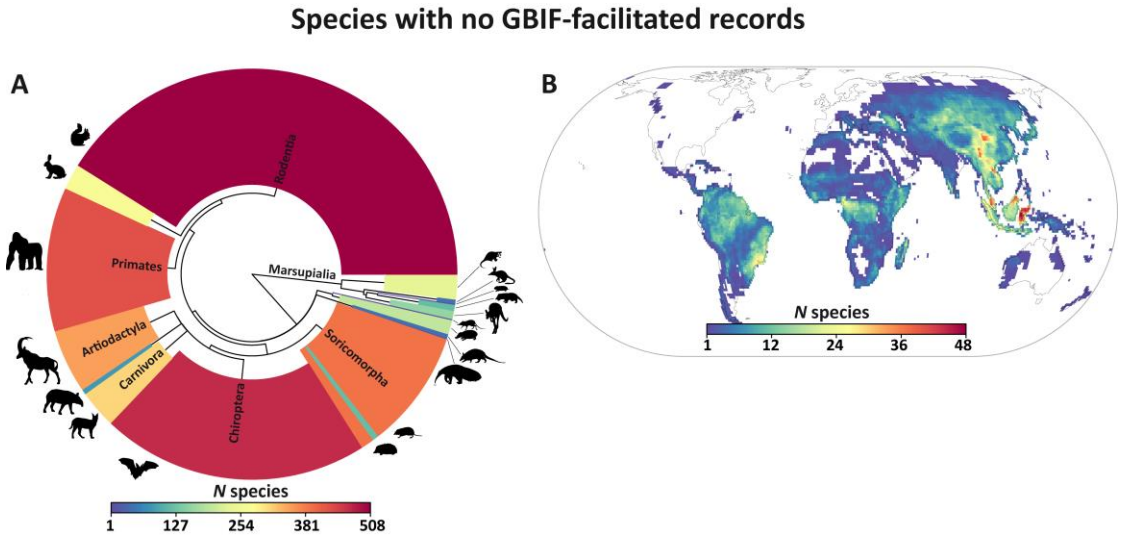


Figure V.3.S2. Global variation in number of species without any GBIF-facilitated records. Shaded areas at branch tips denote mammal orders, with widths proportional to the number of species. Labels within shaded orders in A) highlight the six most speciose orders. Silhouettes are for visual orientation. B) – the same represented as median per 110x110 km grid cell.

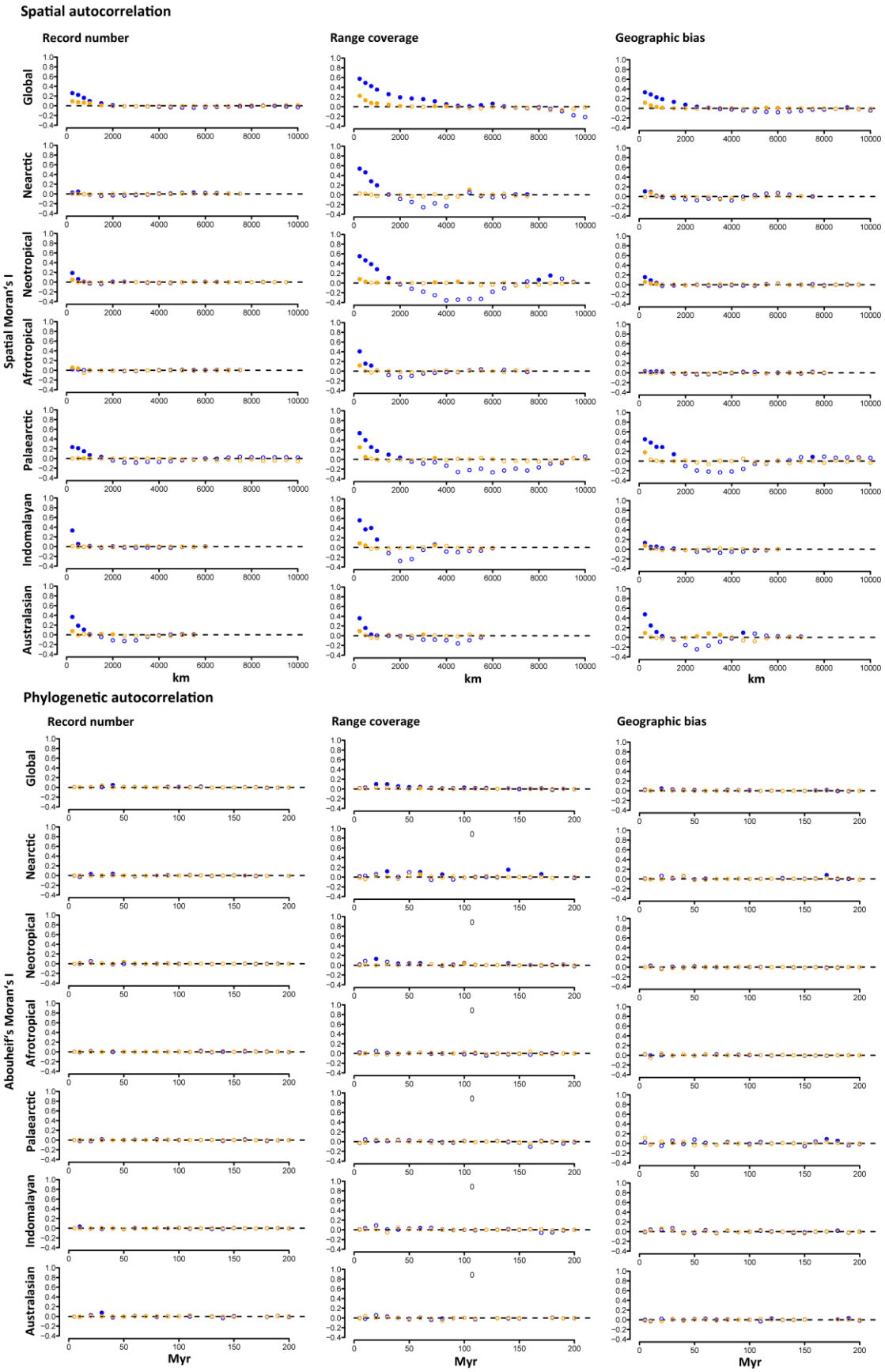


Figure V.3.S3. Correlograms of regression models of record count, range coverage and geographical bias. We tested for spatial autocorrelation using spatial Moran's I across different spatial distance classes (in km), and for phylogenetic autocorrelation using Abouheif's Moran's I (Abouheif, 1999) across phylogenetic distance classes (in Myr). Blue dots mark Moran's I values of the response variables, orange dots mark Moran's I values of model residuals. Solid dots denote significant, circles denote non-significant values. Note that strong autocorrelation only poses a problem in model residuals, not in response variables.

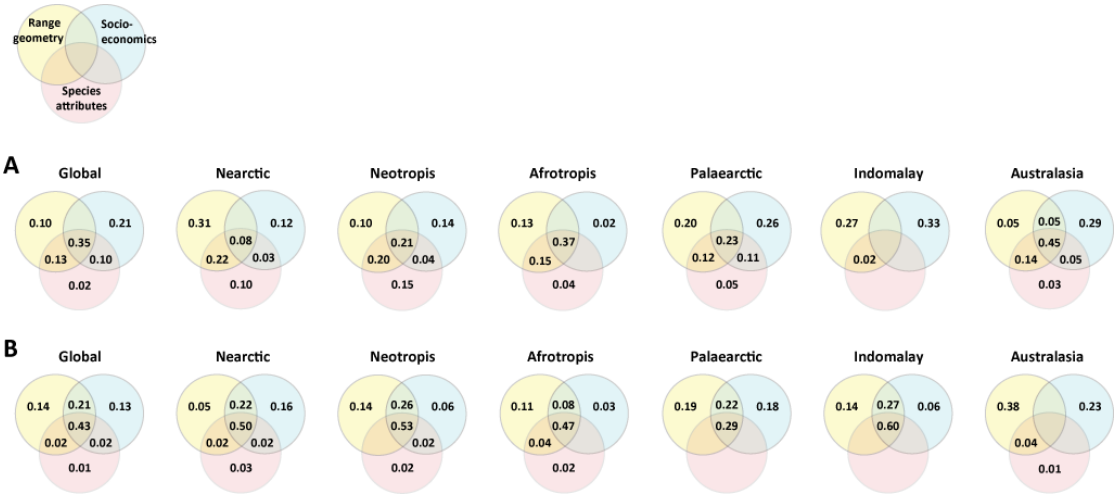


Figure V.3.S4. Results of variation partitioning. A) deviance partitioning of *record count*; B) variance partitioning of *range coverage*. Circles represent the three groups of hypotheses: upper left circle: geometry; upper right: socio-economics; lower: species attributes. Shown are the fractions of the total variation explained uniquely by one or jointly by two or all hypothesis groups. The factor 'Order' was included as a covariate in all models, and accounts for some of the explained variation. Accordingly, values can be compared among hypothesis groups but do not add up to the total explained variation.

Supplementary Tables

Table V.3.S1. Variation in a) record count, b) range coverage and c) geographical bias across zoogeographical realms and mammal orders.

a) Record count

Geographical focus	*N* species	Min	Max	Mean	SD	Median
Global	5,057	0	72,900	563.4	3,072.8	1
Nearctic	568	0	72,900	1,765.8	4,592.5	660
Neotropical	1,563	0	11,943	197.6	794.7	16
Afrotropical	1321	0	16,038	126.0	661.2	8
Palaearctic	840	0	55,910	1246.4	5,524.8	4
Indomalayan	943	0	2,149	43.2	137.0	4
Australasian	852	0	52,441	1292	4,924.1	24

Order	*N* species	Min	Max	Mean	SD	Median
Afrosoricida	51	0	95	11.1	19.3	1
Artiodactyla	225	0	30,756	281.8	2,203.7	9
Carnivora	239	0	52,225	904.9	5,028.6	22
Chiroptera	1,083	0	48,586	653.0	2,923.8	22
Cingulata	21	0	856	54.6	184.5	6
Dasyuromorphia	69	0	24,734	1,213.3	3,534.0	62
Dermoptera	2	30	81	55.5	36.1	56
Didelphimorphia	84	0	2,965	112.2	370.8	5.5
Diprotodontia	135	0	52,441	2,710.4	7,952.7	42
Erinaceomorpha	24	0	25,531	1,105	5,203.3	8
Hyracoidea	4	47	224	126.8	79.0	118
Lagomorpha	91	0	6,978	389.3	1,100.5	7
Macroscelidea	15	0	278	119.9	102.8	132
Microbiotheria	1	149	149	149	-	149
Monotremata	5	0	28,965	7,604.6	12,562.9	21
Notoryctemorphia	2	0	277	138.5	195.9	139
Paucituberculata	6	2	174	66.3	72.6	30.5
Peramelemorphia	18	0	6917	918.8	1,949.8	63
Perissodactyla	16	0	2,828	191.1	703.7	2.5
Pholidota	8	3	84	19.1	27.1	8
Pilosa	10	0	212	93.4	84.4	86
Primates	354	0	329	16.8	41.9	1
Proboscidea	2	5	81	43	53.7	43
Rodentia	2161	0	72,900	506.8	2,796.1	17
Scandentia	19	1	465	87.7	118.0	31
Soricomorpha	411	0	32,117	425.3	2,509.3	4
Tubulidentata	1	27	27	27	-	27

b) Range coverage

Geographical focus	*N* species	Min	Max	Mean	SD	Median
Global	3625	-1.0	-5,278.7	-313.7	378.1	-199.4
Nearctic	505	-2.0	-1,465.9	-102.7	141.7	-56.8
Neotropical	1166	-2.3	-2,727.6	-260.2	289.1	-177.4
Afrotropical	931	-6.8	-4,550.3	-374.5	383.4	-282.9
Palaearctic	545	-4.3	-5,278.7	-540.2	522.3	-419.3
Indomalayan	600	-2.1	-4,156.4	-412.7	453.1	-307.2
Australasian	628	-1.0	-1,612.4	-157.1	191.1	-107.2

Order	*N* species	Min	Max	Mean	SD	Median
Afrosoricida	51	-66.2	-759.0	-229.1	146.5	-172.1
Artiodactyla	225	-8.9	-2,313.7	-453.1	357.3	-397.3
Carnivora	239	-2.0	-2734.3	-603.4	534.3	-418.0
Chiroptera	1083	-1.0	-4,550.3	-452.6	485.1	-335.3
Cingulata	21	-161.0	-1,046.9	-432.4	249.7	-339.1
Dasyuromorphia	69	-8.7	-999.2	-187.0	205.0	-122.2
Dermoptera	2	-59.7	-443.2	-251.4	271.1	-251.4
Didelphimorphia	84	-6.4	-1,529.1	-315.2	285.8	-240.2
Diprotodontia	135	-4.4	-666.9	-106.7	108.9	-78
Erinaceomorpha	24	-38.7	-1,536.8	-471.4	383.5	-367.4
Hyracoidea	4	-316.1	-626.2	-466.3	127.6	-461.4
Lagomorpha	91	-2.5	-1,422.2	-303.7	307.4	-234.1

Range coverage (continued)

Geographical focus	*N* species	Min	Max	Mean	SD	Median
Macroscelidea	15	-62.2	-363.7	-166.3	83.9	-166.5
Microbiotheria	1	-38.5	-38.5	-38.5	-	-38.5
Monotremata	5	-117.4	-169.7	-150.6	22.8	-157.6
Notoryctemorphia	2	-99.6	-99.6	-99.6	-	-99.6
Paucituberculata	6	-7.7	-299.0	-95.5	103.4	-69.0
Peramelemorphia	18	-6.2	-1,137.1	-244.4	342.5	-124.5
Perissodactyla	16	-121.8	-5,278.7	-923.9	1,466.9	-449.5
Pholidota	8	-85.3	-938.7	-590.3	324.8	-598.1
Pilosa	10	-108.6	-596.9	-379.0	165.3	-370.9
Primates	354	-26.8	-989.9	-279.1	189.0	-240.0
Proboscidea	2	-370.4	-837.4	-603.9	330.2	-603.9
Rodentia	2161	-1.3	-2,154.9	-216.4	251.6	-128.2
Scandentia	19	-12.2	-813.0	-209.2	196.5	-176.0
Soricomorpha	411	-2.3	-3,256.0	-308.9	411.2	-152.7
Tubulidentata	1	-676.3	-676.3	-676.3	-	-676.3

b) Geographical bias

Geographical focus	*N* species	Min	Max	Mean	SD	Median
Global	3625	-6.9	7,254.6	116.1	380.0	16
Nearctic	505	-6.8	1,123.3	135	154.9	102.9
Neotropical	1166	-5.7	3,536.3	70.2	193.5	15.0
Afrotropical	931	-6.9	1,249.5	50.2	120.6	10.2
Palaearctic	545	-5.9	7,254.6	316.2	850.5	12.7
Indomalayan	600	-6.0	905.9	31.8	76.8	8.1
Australasian	628	-5.7	4,894.1	212.9	525.4	26.9

Order	*N* species	Min	Max	Mean	SD	Median
Afrosoricida	51	-2.8	142.7	14.8	30.0	1.7
Artiodactyla	225	-4.5	2,230.6	81.0	317.5	9.7
Carnivora	239	-6.8	5,678.1	121.2	488.8	9.7
Chiroptera	1083	-6.0	7,254.6	157.5	472.0	24.6
Cingulata	21	-1.1	153.8	17.2	37.6	6.1
Dasyuromorphia	69	-4.6	3,512.1	265.5	682.5	25.8
Dermoptera	2	9.8	25.8	17.8	11.3	17.8
Didelphimorphia	84	-4.3	364.6	31.0	62.7	6.2
Diprotodontia	135	-4.4	4,894.1	268.2	675.0	27.7
Erinaceomorpha	24	-2.7	1,220.8	103.9	302.3	14.4
Hyracoidea	4	-3.4	26.0	10.5	15.3	9.6
Lagomorpha	91	-4.4	2,436.4	101.2	346.6	18.6
Macroscelidea	15	-3.8	123.9	32.5	45.0	7.8
Microbiotheria	1	-3.7	-3.7	-3.7	-	-3.7
Monotremata	5	11.5	1,345.2	632.8	720.1	587.3
Notoryctemorphia	2	-4.0	-4.0	-4.0	-	-4.0
Paucituberculata	6	7.6	141.7	57.8	52.5	46.2
Peramelemorphia	18	-2.9	766.5	224.2	318.5	28.6
Perissodactyla	16	-2.1	680.6	70.3	202.7	4.7
Pholidota	8	-3.3	29.8	5.6	11.2	0.8
Pilosa	10	-5.1	87.6	31.2	33.5	17.3
Primates	354	-4.8	472.7	15.3	41.0	4.8
Proboscidea	2	0.8	14.4	7.6	9.6	7.6
Rodentia	2161	-6.9	5,253.4	103.8	302.1	22.5
Scandentia	19	-2.4	152.6	45.5	46.6	27.2
Soricomorpha	411	-4.6	3,276.3	100.3	362.8	7.5
Tubulidentata	1	5.8	5.8	5.8	-	5.8

Table V.3.S2. Taxonomic bias as well as relative geographical and taxonomic biases for different aspects of occurrence information. A) Results of nested type III-*ANOVA*s for higher-taxonomic bias of *record count*, *range coverage*, within-range *geographical bias* and species' presence of any mobilized records towards mammal orders and families. B) results of type III-*ANOVA*s for relative bias of *record count*, *range coverage*, within-range *geographical bias* and species' presence of any mobilized records towards zoogeographical realms and mammal orders.

	Factor	*F*	%SS
A) higher-taxonomic bias			
Record count	Order	5.03***	2.4
	Order:Family	3.84***	7.4
	Residuals		90.2
Range coverage	Order	21.30***	12.4
	Order:Family	3.19***	7.4
	Residuals		80.2
Geographical bias	Order	3.07***	2.1
	Order:Family	1.92***	5.2
	Residuals		92.8
Mobilization of any records	Order	6.01***	3.0
	Order:Family	2.73***	5.6
	Residuals		91.4
B) realm bias vs. order bias			
Record count	Order	3.31***	1.7
	Realm	30.99***	3.1
	Realm*Order	1.38.	1
	Residuals		94.3
Range coverage	Order	19.98***	10.7
	Realm	101.44***	10.5
	Realm*Order	8.84***	6.6
	Residuals		72.2
Geographical bias	Order	1.83**	1.3
	Realm	35.84***	4.7
	Realm*Order	2.32***	2.2
	Residuals		91.8
Mobilization of any records	Order	6.62***	3.3
	Realm	46.45***	4.4
	Realm*Order	2.37***	1.7
	Residuals		90.7

Table V.3.S3. The effects of *record count*, *geographical bias*, range size and range shape irregularity on *range coverage* at different spatial extents (global and realm-scale). Shown are the standardized regression coefficients (OLS β). Asterisks denote significant spatial effects (.: $P<0.1$; *: $P<0.05$; **: $P<0.01$; ***: $P<0.001$). All variables were $\log_{10}$-transformed and standardized.

Range coverage

Geographical focus	Predictor	β	se	*t*
Global	Record count	0.80***	0.01	77.79
N=3,353	Geographical bias	-0.31***	0.01	-32.60
R^2= 0.86	Range Size	-0.80***	0.01	-134.23
	Range shape irregularity	-0.29***	0.01	-40.12
Nearctic	Record count	0.87***	0.03	27.78
N=347	Geographical bias	-0.22***	0.03	-8.21
R^2=0.89	Range Size	-1.06***	0.02	-45.98
	Range shape irregularity	-0.20***	0.04	-5.64
Neotropical	Record count	0.72***	0.02	36.90
N=925	Geographical bias	-0.28***	0.02	-15.54
R^2=0.87	Range Size	-0.97***	0.01	-75.81
	Range shape irregularity	-0.21***	0.02	-13.26
Afrotropical	Record count	0.72***	0.02	32.43
N=737	Geographical bias	-0.30***	0.02	-13.70
R^2=0.81	Range Size	-0.98***	0.02	-56.14
	Range shape irregularity	-0.28***	0.01	-19.12
Palaearctic	Record count	0.67***	0.03	22.78
N=361	Geographical bias	-0.30***	0.02	-12.30
R^2=0.81	Range Size	-1.08***	0.03	-36.02
	Range shape irregularity	-0.34***	0.03	-12.55
Indomalayan	Record count	0.71***	0.03	23.40
N=408	Geographical bias	-0.27***	0.03	-7.81
R^2=0.89	Range Size	-1.05***	0.02	-56.20
	Range shape irregularity	-0.39***	0.02	-22.95
Australasian	Record count	0.80***	0.03	23.25
N=444	Geographical bias	-0.30***	0.03	-11.37
R^2=0.73	Range Size	-1.03***	0.03	-32.38
	Range shape irregularity	-0.25***	0.02	-15.80

Table V.3.S4. Effects of species traits, range geometry, and socio-economic factors on A) – G) record count and H) – N) range coverage at different spatial extents (global and realm-scale). The 14 predictor variables were Diurnality, Body size, Foraging stratum, Dietary level, Time since description, Threat status, Public interest, Threat status, Range size, Range shape irregularity, Endemism richness, Proximity to institutions, GBIF participation, Financial resources. Two comparative measures were used: for record count (A – G): 1) standardized regression coefficients from the reduced spatial generalized linear model with the lowest QAIC score (GLM β), and 2) the sum of QAIC weights across all possible model subsets (∑QAICw; Burnham & Anderson., 2002); for range coverage (H – N): 1) standardized regression coefficients from the reduced ordinary least squares model with the lowest AIC score (OLS β), and 2) the sum of AIC weights across all possible model subsets (∑AICw). GVIF/VIF are generalized variance inflation factors. Asterisks denote significant spatial effects (.: P<0.1; *: P<0.05; **: P<0.01; ***: P<0.001). Partial adjusted deviance explained (D^2) and partial adjusted variance explained (R^2) refer to the variation that is explained by the predictor variables, with effects of the covariate 'Order' partialled out (Peres-Neto *et al.*, 2006).

A) Record count

Geographical focus	Predictor	GLM β	se	t	ΔQAICw	GVIF
Global	Body mass	-0.40***	0.09	-4.30	1	4.6
N=3,353	Foraging stratum	-0.19*	0.09	-2.05	0.70	3.4
D^2=0.62	Years since description	0.47***	0.13	3.68	1	1.9
	Public interest	0.61***	0.06	10.54	1	2.2
	Range size	3.77***	0.18	21.30	1	4.1
	Range shape irregularity	0.55***	0.10	5.82	1	1.5
	Area appeal	0.40***	0.12	3.52	0.99	2.5
	Proximity to research institutions	2.12***	0.10	21.83	1	2.5
	Financial resources	0.59***	0.09	6.96	1	3.3
Nearctic	Body mass	-1.24***	0.23	-5.46	1	3.1
N=347	Foraging stratum	-0.55*	0.26	-2.12	0.86	2.7
D^2=0.69	Dietary level	0.47**	0.18	2.59	0.96	2.5
	Public interest	0.70***	0.12	5.92	1	1.8
	Range size	4.24***	0.36	11.86	1	4.8
	Range shape irregularity	0.58	0.43	1.33	0.47	1.7
	Area appeal	3.03***	0.34	9.02	1	3.1
	Proximity to research institutions	-1.36**	0.48	-2.82	0.99	1.7
	GBIF participation	3.14.	1.83	1.72	0.71	1.4
	Financial resources	1.05**	0.37	2.82	0.99	1.2
Neotropical	Diurnality	-0.27.	0.16	-1.71	0.61	1.6
N=925	Body mass	-1.88***	0.22	-8.37	1	3.9
D^2=0.60	Foraging stratum	-0.48***	0.15	-3.31	0.99	2.7
	Dietary level	-0.67***	0.16	-4.17	1	2.4
	Years since description	1.75***	0.24	7.28	1	1.9
	Public interest	1.09***	0.13	8.14	1	1.8
	Threat status	0.47*	0.23	2.00	0.69	1.9
	Range size	3.51***	0.32	11.06	1	5.9
	Area appeal	1.36***	0.25	5.44	1	3.4
	Proximity to research institutions	0.67***	0.19	3.42	0.99	2.8
	GBIF participation	1.07***	0.24	4.54	1	2.4
	Financial resources	1.08***	0.22	4.96	1	2.0
Afrotropical	Body mass	-1.16**	0.38	-3.07	0.99	2.8
N=737	Foraging stratum	-0.49	0.36	-1.36	0.49	3.6
D^2=0.45	Dietary level	0.53.	0.29	1.82	0.67	2.1
	Years since description	0.73	0.46	1.59	0.53	2.2
	Public interest	0.35	0.22	1.61	0.61	1.5
	Range size	4.92***	0.71	6.88	1	4.8
	Area appeal	2.15***	0.55	3.92	1	2.9
	GBIF participation	-0.80	0.49	-1.64	0.6	1.9

Record count (continued)

Geographical focus	Predictor	GLM β	se	t	ΔQAICw	GVIF
Palaearctic	Diurnality	1.36**	0.49	2.76	0.69	2.8
N=361	Public interest	0.71.	0.37	1.94	0.43	3.8
D^2=0.78	Range size	6.38***	1.78	3.59	0.99	7.7
	Proximity to research institutions	2.57*	1.13	2.29	0.66	7.1
	Financial resources	1.49*	0.59	2.54	0.56	2.6
Indomalayan	Dietary level	-0.73*	0.29	-2.50	0.82	4.0
N=408	Threat status	-0.48	0.3	-1.57	0.59	1.4
D^2=0.47	Range size	5.38***	0.61	8.86	1	10.0
	Area appeal	3.91***	0.56	7.04	1	5.0
	Proximity to research institutions	1.41.	0.81	1.76	0.63	1.5
	GBIF participation	0.63**	0.22	2.90	0.97	1.2
	Financial resources	2.44***	0.32	7.69	1	2.8
Australasian	Diurnality	-1.63***	0.39	-4.18	1	2.2
N=444	Body mass	0.89***	0.17	5.15	1	9.3
D^2=0.86	Foraging stratum	0.47*	0.19	2.42	0.85	7.3
	Threat status	-0.54*	0.23	-2.39	0.88	1.5
	Years since description	0.45**	0.15	3.06	0.98	1.6
	Public interest	0.43**	0.14	3.15	0.98	3.0
	Range size	3.55***	0.35	10.18	1	4.6
	Range shape irregularity	0.57***	0.15	3.73	1	2.3
	Area appeal	-2.07***	0.36	-5.74	1	4.1
	Proximity to research institutions	3.77***	0.23	16.09	1	3.5
	GBIF participation	-0.89	0.54	-1.65	0.61	4.3
	Financial resources	-1.89*	0.79	-2.40	0.91	6.3

B) Range coverage

Geographical focus	Predictor	OLS β	se	t	ΔAICw	GVIF
Global	Diurnality	0.02.	0.01	1.72	0.59	1.5
N=3,353	Body mass	-0.05**	0.02	-2.65	0.81	4.9
R^2=0.71	Foraging stratum	-0.07***	0.02	-4.03	1	3.6
	Dietary level	-0.03.	0.02	-1.74	0.57	3.4
	Years since description	0.06***	0.01	4.64	1	1.7
	Public interest	0.04***	0.01	3.73	1	1.7
	Range size	-0.62***	0.02	-38.60	1	2.0
	Range shape irregularity	-0.26***	0.01	-24.04	1	1.2
	Area appeal	0.17***	0.01	13.48	1	2.0
	Proximity to research institutions	0.22***	0.01	19.62	1	1.5
	GBIF participation	0.06***	0.01	5.56	1	1.6
	Financial resources	0.18***	0.01	15.61	1	1.8
Nearctic	Body mass	-0.16***	0.05	-3.39	0.95	4.4
N=347	Years since description	0.20***	0.04	4.71	1	1.8
R^2=0.73	Public interest	0.09**	0.03	2.80	0.96	2.0
	Threat status	0.07.	0.04	1.86	0.70	2.0
	Range size	-0.44***	0.05	-8.38	1	4.5
	Area appeal	0.59***	0.05	11.08	1	2.5
	Proximity to research institutions	0.09	0.07	1.41	0.55	1.4
	GBIF participation	0.72*	0.34	2.11	0.81	1.2
	Financial resources	0.36***	0.04	8.67	1	1.3
Neotropical	Diurnality	0.04.	0.02	1.83	0.68	1.7
N=925	Body mass	-0.15***	0.04	-4.08	1	4.4
R^2=0.75	Foraging stratum	-0.06*	0.03	-2.09	0.80	3.3
	Dietary level	-0.09***	0.03	-3.44	0.99	2.3
	Years since description	0.07***	0.02	3.94	1	1.5
	Public interest	0.12***	0.02	5.12	1	1.8
	Range size	-0.56***	0.03	-21.21	1	4.5

Range coverage (continued)

Geographical focus	Predictor	OLS β	se	t	ΔAICw	GVIF
	Range shape irregularity	-0.22***	0.02	-9.65	1	1.3
	Area appeal	0.19***	0.02	7.98	1	2.1
	Proximity to research institutions	0.07**	0.02	2.79	0.95	1.7
	GBIF participation	0.09***	0.02	4.39	1	1.5
	Financial resources	0.20***	0.03	7.14	1	1.6
Afrotropical	Diurnality	0.06**	0.02	2.83	0.94	1.6
N=737	Body mass	-0.09*	0.04	-2.21	0.84	8.5
R^2=0.59	Years since description	0.11***	0.03	4.03	1	2.1
	Threat status	0.12***	0.03	4.09	1	2.3
	Range size	-0.51***	0.04	-11.61	1	5.3
	Range shape irregularity	-0.26***	0.02	-11.04	1	1.4
	Area appeal	0.17***	0.03	5.45	1	3.1
	Proximity to research institutions	0.20*	0.09	2.36	0.87	2.1
	GBIF participation	-0.11***	0.03	-3.44	0.99	1.9
	Financial resources	0.16***	0.04	4.16	1	2.5
Palaearctic	Range size	-0.66***	0.05	-14.08	1	3.6
N=361	Range shape irregularity	-0.28***	0.04	-6.90	1	1.5
R^2=0.65	Area appeal	0.21***	0.06	3.34	0.98	1.9
	Proximity to research institutions	0.09*	0.04	2.48	0.81	3.7
	GBIF participation	0.37***	0.05	7.28	1	3.1
Indomalayan	Body mass	0.08.	0.05	1.68	0.47	6.4
N=408	Years since description	0.06.	0.04	1.73	0.69	1.8
R^2=0.44	Threat status	-0.05.	0.03	-1.76	0.62	2.1
	Range size	-0.60***	0.05	-11.53	1	4.9
	Range shape irregularity	-0.37***	0.03	-13.40	1	1.5
	Area appeal	0.32***	0.04	7.15	1	2.7
	Financial resources	0.28***	0.04	7.29	1	1.6
Australasian	Dietary level	-0.07**	0.03	-2.61	0.85	3.9
N=444	Range size	-0.65***	0.03	-19.13	1	2.9
R^2=0.59	Range shape irregularity	-0.23***	0.02	-12.06	1	1.3
	Area appeal	0.08**	0.03	2.63	0.76	3.0
	Proximity to research institutions	0.56***	0.04	15.06	1	3.3
	GBIF participation	-0.06*	0.03	-2.01	0.59	2.0

Table V.3.S5. The effects of range size and within-range gradients in socio-economic factors on within-range *geographical bias* in mobilized records. We modeled effects of within-range variation in socio-economic factors using two metrics per socio-economic factor: 1) the coefficient of variation (*cv*) and 2) the correlation coefficient between a euclidean socio-economic distance matrix and a geographical distance matrix (r_P; see *SI V.3.1.4* for explanation). The 9 predictor variables were range size, CV endemism richness, r_P endemism richness, CV proximity to institutions, r_P proximity to institutions, CV GBIF participation, r_P GBIF participation, CV locally available research funding, and r_P locally available research funding. Two comparative measures were used: 1) standardized regression coefficients from the reduced ordinary least squares model with the lowest AIC score (OLS β), and 2) the sum of AIC weights across all possible model subsets ($\sum AIC_w$). GVIF are generalized variance inflation factors. Asterisks denote significant spatial effects (.: P<0.1; *: P<0.05; **: P<0.01; ***: P<0.001). Partial adjusted variance explained (R^2) refers to the variation that is explained by the predictor variables, with effects of the covariates 'Order' and '*Record count*' removed (Peres-Neto *et al.*, 2006).

Geographical bias

Geographical focus	Predictor	OLS β	se	t	ΔAICw	GVIF
Global	Range size	-0.13***	0.02	-5.72	1	3.3
N=3,353	Endemism richness (cv)	0.05**	0.02	2.91	0.97	2.2
R^2=0.15	Proximity to research institutions (rP)	0.04*	0.02	2.45	0.88	1.9
	Proximity to research institutions (cv)	0.09***	0.02	5.67	1	1.7
	GBIF participation (rP)	0.05***	0.01	3.58	0.99	1.6
	Financial resources (rP)	-0.05***	0.01	-3.36	0.99	1.5
	Financial resources (cv)	0.04*	0.02	2.40	0.87	1.8
Nearctic	Endemism richness (rP)	-0.1**	0.04	-2.62	0.85	1.1
N=347	Endemism richness (cv)	0.12***	0.03	3.53	0.98	2.3
R^2=0.24	Proximity to research institutions (rP)	0.09*	0.04	2.06	0.53	2.0
	Financial resources (cv)	-0.15*	0.07	-2.1	0.61	2.0
Neotropical	Range size	-0.13***	0.03	-4.06	0.99	4.4
N=925	Proximity to research institutions (cv)	-0.1***	0.03	-3.53	0.99	1.8
R^2=0.08	Financial resources (cv)	0.1*	0.05	2.04	0.75	2.0
Afrotropical	Range size	-0.06	0.04	-1.49	0.60	3.9
N=737	Proximity to research institutions (cv)	-0.07	0.05	-1.63	0.74	1.8
R^2=0.05	GBIF participation (rP)	0.04	0.02	1.6	0.49	1.6
	Financial resources (cv)	-0.07*	0.03	-2.18	0.77	1.6
Palaearctic	Proximity to research institutions (cv)	0.13***	0.03	3.71	0.99	2.5
N=361	GBIF participation (rP)	0.18**	0.06	2.96	0.92	1.5
R^2=0.24	Financial resources (cv)	-0.07.	0.04	-1.79	0.62	2.4
Indomalayan	Endemism richness (rP)	0.09***	0.02	3.77	0.99	1.5
N=408	Endemism richness (cv)	-0.04.	0.03	-1.68	0.48	2.6
R^2=0.00						
Australasian	Proximity to research institutions (rP)	0.16**	0.06	2.73	0.93	2.1
N=444	Proximity to research institutions (cv)	0.65***	0.12	5.18	1	2.7
R^2=0.44	GBIF participation (rP)	-0.06	0.04	-1.58	0.58	4.6
	GBIF participation (cv)	-0.15*	0.07	-2.24	0.85	1.6

Table V.3.S6. Effects of species attributes, range geometry, and socio-economic factors on whether or not species have any mobilized records. Effects were tested in multiple generalize linear models with a quasi-binomial distribution and a logit link. All possible model subsets were ranked based on QAIC scores, results are shown for the minimum adequate model (with the lowest QAIC score). Two comparative measures were used: 1) standardized regression coefficients from the reduced spatial generalized linear model with the lowest QAIC score (GLM β), and 2) the sum of QAIC weights across all possible model subsets ($\sum QAIC_w$; Burnham & Anderson, 2002). GVIF are generalized variance inflation factors. Asterisks denote significant spatial effects (.: $P<0.1$; *: $P<0.05$; **: $P<0.01$; ***: $P<0.001$). Partial adjusted deviance explained (D^2) refer to the variation that is explained by the predictor variables, with effects of the covariate 'Order' removed (Peres-Neto *et al.*, 2006).

Has records

N= 4,934

D^2=0.30

Predictor	GLM β	se	*t*	∑QAICw	GVIF
Foraging stratum	-0.59***	0.15	-3.94	1	3.5
Years since description	0.55***	0.09	6.31	1	1.7
Public interest	0.68***	0.11	6.20	1	1.8
Range size	3.09***	0.13	22.97	1	3.2
Range shape irregularity	0.41***	0.08	5.28	1	1.2
Area appeal	0.93***	0.10	9.30	1	2.0
Proximity to institutions	1.24***	0.10	12.09	1	1.6
GBIF participation	0.17*	0.08	2.06	0.71	1.6
Financial resources	1.06***	0.10	10.50	1	1.7

Acknowledgements

I am truly grateful to several people who supported me. First, I would like to thank my supervisor, Prof. Holger Kreft for his commitment. Thank you for invaluable guidance, patience and encouragement. Thank you also for providing me with the same support and privileges as your employees and for trusting in my work even though I worked from another city during much of my program.

I am very grateful to Prof. Kerstin Wiegand for co-supervision. Thanks to Prof. Ulrich Brose who kindly agreed to be a member of my thesis committee.

Many thanks to the Deutsche Bundesstiftung Umwelt for funding me with a PhD scholarship, and especially to Hans-Christian Schaefer for his support. Thanks to the DAAD (German academic exchange service) for an additional short-term PhD scholarship for a research stay with Prof. Walter Jetz.

I am also grateful to Walter Jetz for his hospitality and for inviting me to work in his research lab. Thanks to Walter and Rob Guralnick for inviting me to participate in the Map of Life project, and to the larger Map of Life team for technical support and fruitful discussions during various stages.

Thanks to my co-authors, Susanne Fritz, Rob Guralnick, Walter Jetz, Holger Kreft and Patrick Weigelt for their valuable suggestions, tireless feedback and other contributions. I am also grateful to several colleagues for feedback and advice, in particular to Kathrin Böhning-Gaese, Carsten Dormann and Marten Winter. I acknowledge Rod Page and several anonymous referees for valuable comments which greatly improved chapter 2.

Many thanks to my lab mates in the Macroecology, Biodiversity and Conservation Biogeography Group, in particular to Patrick Weigelt, for devoting considerable time to sharing their expertise and giving useful suggestions, and for being such a friendly and supportive bunch. Special thanks to Yael Kisel and Patrick Weigelt for proof-reading.

I am grateful to my supervisor, the Universitätsbund Göttingen, the International Biogeography Society, Fairchild Tropical Botanic Garden, The British Ecological Society, and the Map of Life project for sponsoring conference and workshop travels.

I am most grateful to my family and friends for their many ways of support.