Global drivers of species variation in mobilized point-occurrence information

Carsten Meyer[1]*, Walter Jetz[2]*, Robert P. Guralnick[3], Susanne A. Fritz[4], Holger Kreft[1]*

[1] *Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences and Forest Ecology, University of Goettingen, Germany.* cmeyer2@uni-goettingen.de

[2] *Department of Ecology and Evolutionary Biology, Yale University, USA.* walter.jetz@yale.edu

[3] *Florida Museum of Natural History, University of Florida, Gainesville USA.*

[4] *Senckenberg Biodiversity & Climate Research Centre (BiK-F), Senckenberg Gesellschaft für Naturforschung & Goethe University Frankfurt, Germany.*

[5] *Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences and Forest Ecology, University of Goettingen, Germany.* hkreft@uni-goettingen.de

Author contributions: All authors designed this study, C.M., W.J. and S.A.F. compiled data, C.M. performed the analyses and wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

Keywords: Wallacean shortfall, geographical bias, taxonomic bias, data deficiency, knowledge gaps, occurrence records, biocollections, GBIF, survey effort, detectability

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 among species 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 evaluate 13 putative drivers of species-level variation in data availability. We find that data limitations are mainly linked to range size and shape, and the geography of socio-economic conditions. Surprisingly, species attributes related to detection and collection probabilities, such as body size or diurnality, are much weaker predictors of the amount and range coverage of available records. Our results highlight the need to prioritize range-restricted species and to address the key socio-economic drivers of data bias in data mobilization efforts and distribution modeling.

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 source for many large-scale analyses, but they incur high errors of commission towards 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.* 2012). Large amounts of digital occurrence records from field observations, museum specimens, and other sources have been mobilized via national and 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 have 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 are randomly distributed across the entire range of relevant dimensions and that data quality 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 (Dorazio 2014; Iknayan *et al.* 2014). While the reliability of range maps in relation to range size and species attributes has been assessed (Jetz *et al.* 2008), patterns and drivers of species-level variation in point-occurrence information remain largely ignored.

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, species-level variation 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 are fully reliant on high *range coverage* of species.

Many possible drivers of species-level variation in occurrence records have been suggested (see Box 1). 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 better detectable due to higher abundances (Dorazio 2007), or because they possess specific traits that make them more conspicuous, such as large body size or diurnal activity (Iknayan *et al.* 2014). 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 may 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 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 million geographically and taxonomically validated records mobilized via GBIF for 3,625 terrestrial mammal species (c. 72% of all extant species) with their expert range maps. We developed scale-independent metrics for *range coverage* and *geographical bias* to explore 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 different variables; described in detail in Box 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.

Materials and Methods

Measuring occurrence information

We overlaid 4,524,585 point occurrence records mobilized via GBIF (retrieved Oct 2012; see *Supplementary Information* (*SI*) *1*) with expert-drawn 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 particular point in time (Soberón & Peterson 2004). In contrast, range maps delimit the geographical distribution of known and assumed species occurrences, based on expert interpretation of different distribution data types (Graham & Hijmans 2006). 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 (Hurlbert & Jetz 2007). We matched taxonomies between records and range maps and used range map overlays to validate records geographically (*SI 1*). The final, rigorously cleaned dataset contained 2,849,058 records for 3,625 species.

In addition to simple *record count*, we then used records and range maps 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. 1, Fig. S1).

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

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

where $MinDistRP_i$ is the minimum distance of the *i*-th random point to its nearest record (Fig. S1; Fig. 1 for examples).

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, repeated this 1000 times, and each time calculated MMD. *Geographical bias* was then the standardized effect size, calculated as the difference between observed MMD and null model mean divided by the null model standard deviation:

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

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. 1, Fig. 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 tested three major classes of hypotheses related to species attributes, range geometry, and socio-economic factors, which were represented by 13 variables (Box 1) as potential drivers of whether species have any mobilized records (details in *SI 2*) as well as of *record count* and *range coverage*. We tested range size and 8 variables of within-range variation in socio-economic factors as potential drivers of *geographical bias* (details in *SI 3*).

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;

Box 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 collect, transport and store.

iii) Foraging stratum: Terrestrial species are more detectable 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) Time 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 of great public interest (e.g., due to commercial, medicinal, aesthetic, psychological, or cultural reasons; see Knight (2008); Perry (2010); Tyler *et al.* (2012)).

vii) Threat status: Despite higher interest in 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 (e.g., arboreal vs. terrestrial), 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 sampling locations.

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*, as random points in such ranges would be increasingly less likely to be close to a given record.

ix) Range shape irregularity: The same geographical 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 limiting 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 less frequently, making it likely that rare species evade detection (Dennis & Thomas 2000).

xii) GBIF participation: Political commitment to international data sharing limits 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.

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 (Wilman *et al.* 2014) on **iv)** main foraging stratum on an ordinal scale: 1=terrestrial (including, e.g., 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. 1 for examples). Ratios increase from 1 (perfect circle) as range 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:

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

where $\mathrm{RelProp}_i$ 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 overlapping 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). For details on models of whether or not species have any record, models of *geographical bias*, tests for spatial and phylogenetic autocorrelation, additional models of omitted variables, and limitations of this study, see *SI 2-6*). All analyses were performed in R 2.15.2–3.1.2 (R Core Team 2014).

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 2*, Table S2 A). We therefore included 'mammal order' as a fixed control variable 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. S3, for details see *SI 5*). We used multi-model inference (Burnham & Anderson. 2002) to assess model support and identify minimum adequate models 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* would yield positive effects. We limited collinearity by only including explanatory variables with generalized variance inflation factors $\leq$10 (Dormann *et al.* 2013; Table 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%.

Results

Patterns in occurrence information

3,625 or 72% of the 5,057 mammal species considered had at least one validated record (see Fig. S2, *SI 2* 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. 2 A-C, Table S1). Globally, the mean *record count* per species was 563 (SD=3,073, median=13, Table S1) and *record count* had a heavily right-skewed distribution. *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 (Fig. 2 D-F, see also *SI 2*, Table S1, *ANOVA* results in Table 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. 2 D-E), which also had above-average *geographical bias* scores (Fig. 2 F, Table 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 5*, Fig. S3).

Accordingly, occurrence information differed more strongly among geographical realms (Fig. 2 G-I) than among mammal orders (Fig. 2 D-F, *SI 2*, Tables S1, S2 B). Most mammal assemblages in the Nearctic, northern Neotropical, western and northern Palaearctic and Australasian realms were characterized by species with above-average *record counts*, whereas Madagascar and the south-eastern Palaearctic and Indomalayan realms had mostly species with below-average *record counts* (Fig. 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. 2 G-I). Species without GBIF-facilitated records had highest concentrations in Southern China and South-East Asia (Fig. S2).

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

Predictors of occurrence information

Record count and *range coverage* were well-predicted by a combination of species attributes, range geometry and socio-economic factors. Together, these variables explained 62% and 71%, respectively, of the variation in *record count* and *range coverage* in the global model, and between 44% and 86% in realm-scale models (Fig. 4). All 13 predictor variables showed at least weak effects in some of the models (Fig. 4, Table 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. 5, Table S5, *SI 3*); most variation in *geographical bias* was explained in zoogeographical realms with large numbers of mobilized records (partial R^2_{adj}: Nearctic: 0.24, Palaearctic: 0.24, Australasian: 0.44).

Of the three major groups of tested variables, range geometry and socio-economic factors emerged as the most important factors driving *record count* and *range coverage* (Fig. 4, Tables S3). Variation partitioning confirmed that more variation in either metric was uniquely explained by range geometry and socio-economic factors than by species attributes (except for *record count* in the Neotropical model, Fig. 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. S4 B).

Overall, most species attributes showed only weak yet often significant effects on *record count* and *range coverage* (Fig. 4, Tables S3). Body mass and time since description showed relatively consistent negative and positive effects, respectively, across global and realm-scale 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, respectively, in the Afrotropical, Australasian, Neotropical, and global and Neotropical models (Fig. 4).

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. 4) and on *geographical bias* in the global and Neotropical models (Fig. 5). Range size alone explained 7-38% of the variation in *record count*, and 26-64% of variation 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. 4, Table S4).

Socio-economic factors showed strong positive effects, particularly for *range coverage*, both from sums of QAIC/AIC weights and standardized coefficients (Fig. 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 unexpected negative effects emerged. For instance, higher *record counts* were associated with lower area appeal and financial resources in the Australasian, and with lower proximity to institutions in the Palaearctic model, and *range coverage* was negatively associated with 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, for GBIF participation in the Palaearctic, and for financial resources in the Neotropical realm (Fig. 5).

Discussion

Our analyses revealed strong species-level differences in the quantity and quality of 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 mobilized records, and large parts of most mammal ranges were several hundred kilometers away from the closest record that provides direct evidence of occurrence, revealing considerable uncertainty regarding fine-scale species distributions.

Global species-level differences appear to largely result from geographical data bias towards well-sampled 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. 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 clustered 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 coincided with high *geographical bias* scores.

Species attributes

Surprisingly, multiple regression and variation partitioning analyses revealed a minor role of species attributes in shaping occurrence information, although all variables that captured species attributes received some 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 charismatic groups like primates and carnivores, might contradict such species' prominence in the scientific literature (Brooke *et al.* 2014) and monitoring data (Burton 2012). As most mammal records in GBIF are from biocollections, a plausible explanation may be the greater logistic difficulties of collecting and storing large specimens. Improving the accessibility of occurrence datasets based on less-invasive field research, such as visual or auditory observations (Hoffmann *et al.* 2010), may effectively complement currently-mobilized information for these underrepresented species. Other species attributes associated with collection probabilities consistently emerged as weaker but significant determinants of *record count* and *range coverage*. For instance, more mobilized records exist for early-described species and for species of public interest (Tyler *et al.* 2012). Against our expectation, current threat status had little effects on *record count* or *range coverage*, possibly because greater scientific interest in threatened species (Tyler *et al.* 2012) is counter-balanced by legal or ethical impediments to specimen collection and data sharing (Whitlock *et al.* 2010).

It is often assumed that more records are available for species that are better detectable due to their higher abundances or higher conspicuousness (Iknayan *et al.* 2014). Accordingly, the negative effects of body mass might also be due to higher abundances of small-bodied mammals. However, a strong role of abundance is otherwise not supported: dietary level, which covaries with abundance particularly when controlling for body mass and habitat (Robinson & Redford 1986), consistently showed weak effects on *record count* and *range coverage* (see also *SI 4*). Similarly, population density only had a weak positive relationship with *record count* and no relationship with *range coverage* (see *SI 4*). Other traits associated with conspicuousness also failed to support the detectability hypothesis, as both diurnality and foraging stratum showed only weak effects, contrasting results from regional studies (Burton 2012; Chutipong *et al.* 2014). While we cannot rule out a stronger role of species attributes at smaller spatial 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 scale and in the Neotropics, 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 are severely constrained in the detail with which a given number of records could cover them. In contrast, 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 achieve 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 truly detailed representation with records, but simply because any record within the range was automatically closer to any other point 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 information basis of species distributions. Mean endemism richness, used as a proxy for area appeal (Soria-Auza & Kessler 2008), had the most consistent effects. In conjunction with clear positive effects of range size on *record count*, this demonstrates that despite increased collection 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 research institutions (Dennis & Thomas 2000), GBIF participation (Yesson *et al.* 2007), and locally available financial resources (Soberón & Peterson 2004) strongly limit species-level occurrence information.

However, 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. 5). Together, these results show that the spatial extent and geographical focus of analyses is crucial for understanding the causes of bias in occurrence information.

Implications and conclusions

Our results have three main implications: First, species without mobilized 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 among species and include occurrence information as a predictor variable (e.g., range size as a predictor of extinction risk) violate statistical assumptions and increase the potential for biased inference (Garamszegi & Møller 2011). Second, information gaps are particularly severe for range-restricted species, for which detailed information would be urgently needed to confront future extinction risk (Fritz *et al.* 2009; Boitani *et al.* 2011) and for which commonly-used range map information most strongly overestimates fine-scale occurrences (Jetz *et al.* 2008). Third, conventional species distribution modelling (Guisan & Thuiller 2005) cannot provide a general remedy for overcoming data limitations, due to high spatial pseudo-replication of records combined with poor spatial *coverage*. Even the 37% of represented mammals that had between 50 and 200 records, an often-cited range of minimum model requirements (Boitani *et al.* 2011; Feeley & Silman 2011), typically had much fewer unique sampling locations (median=17), and a relatively low *range coverage* (median=-207 km). Modern hierarchical models can address problems of biased records, by explicitly incorporating models of site-specific survey effort or species-specific detectability (Dorazio 2014; Iknayan *et al.* 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.

In summary, 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. 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 (compare discussion in Meyer *et al.* (2015)). Researchers and institutions collecting or mobilizing new occurrence information should consider possible synergies with global data priorities, e.g., through focusing on threatened, range-restricted, or understudied species. Information metrics such as those developed in this study could be incorporated into online tools that allow researchers and funding agencies to identify priority species for improving information. Meanwhile, 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 less biased inference from accessible occurrence information.

Acknowledgements

We thank those individuals and institutions active in collecting, curating, digitizing, mobilizing, and sharing distribution data. We thank Tim Robertson and Jeremy Malczyk for help with data assembly, and the members of the Kreft lab for discussions about hypotheses and metrics. C.M. acknowledges funding from the Deutsche Bundesstiftung Umwelt (DBU), the German Academic Exchange Service (DAAD) and the Universitätsbund Göttingen. W.J. and R.P.G. acknowledge support from NSF (DBI 0960550, DEB 1026764, DEB1441737, DBI-1262600), NASA (NNX11AP72G) and the Yale Program in Spatial Biodiversity Science and Conservation. S.A.F. acknowledges support from the LOEWE funding program of Hesse's Ministry of Higher Education, Research, and the Arts. H.K. acknowledges funding by the German Research Foundation (DFG) in the framework of the German Excellence Initiative within the Free Floater Program at the University of Göttingen.

References

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

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

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

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? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 366, 2623–32.

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.

Burnham, K.P. & Anderson., D.R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach.* 2nd edn. Springer, New York.

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

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

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 Bull. Zool.*, 62, 521–535.

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

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. *Glob. Ecol. Biogeogr.*, 23, 1472–1484.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., *et al.* (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., *et al.* (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30, 609–628.

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

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

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. *Ecol. Lett.*, 12, 538–49.

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

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

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

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

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

Hoffmann, A., Decher, J., Rovero, F., Schaer, J., Voigt, C. & Wibbelt, G. (2010). Field Methods and Techniques for Monitoring Mammals. In: *Man. F. Rec. Tech. Protoc. All Taxa Biodivers. Invent. Monit.* Pensoft Publishers, Sofia, Bulgaria, pp. 482–529.

Hurlbert, A.H. & Jetz, W. (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci. U. S. A.*, 104, 13384–9.

Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014). Detecting diversity: emerging methods to estimate species diversity. *Trends Ecol. Evol.*, 29, 97–106.

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

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.*, 27, 151–159.

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

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

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

Meyer, C., Kreft, H., Guralnick, R.P. & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *PeerJ Prepr.*, 3, e1057. DOI: https://dx.doi.org/10.7287/peerj.preprints.856v1

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

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., *et al.* (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *Bioscience*, 51, 933–938.

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. *Ecol. Econ.*, 69, 478–485.

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

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/.

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

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

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

Soberón, J.M. & Peterson, A.T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 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. *Divers. Distrib.*, 14, 123–130.

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

Tyler, E.H.M., Somerfield, P.J., Berghe, E. Vanden, Bremner, J., Jackson, E., Langmead, O., *et al.* (2012). Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology. *Glob. Ecol. Biogeogr.*, 21, 922–934.

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.

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. *Proc. Natl. Acad. Sci. U. S. A.*, 110, 12144–12148.

Whitlock, M.C., McPeek, M.A., Rausher, M.D., Rieseberg, L. & Moore, A.J. (2010). Data archiving. *Am. Nat.*, 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. *Divers. Distrib.*, 11, 3–23.

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.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., *et al.* (2007). How global is the global biodiversity information facility? *PLoS One*, 2, e1124.

Figures

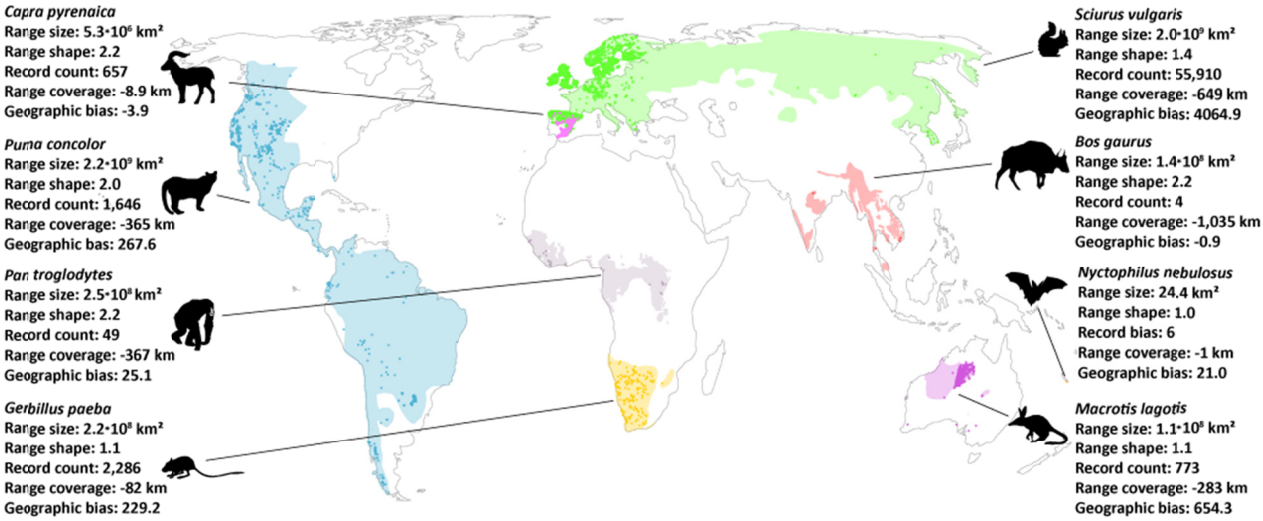

Figure 1: Occurrence information and range geometry for eight selected mammal species. Examples demonstrate variation in *record count*, *range coverage* by those records, and *geographical bias* in how records represent different range parts. Pale colours denote extent-of-occurrence range maps (IUCN 2010), brightly colored dots indicate locations of GBIF-facilitated occurrence records. 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.

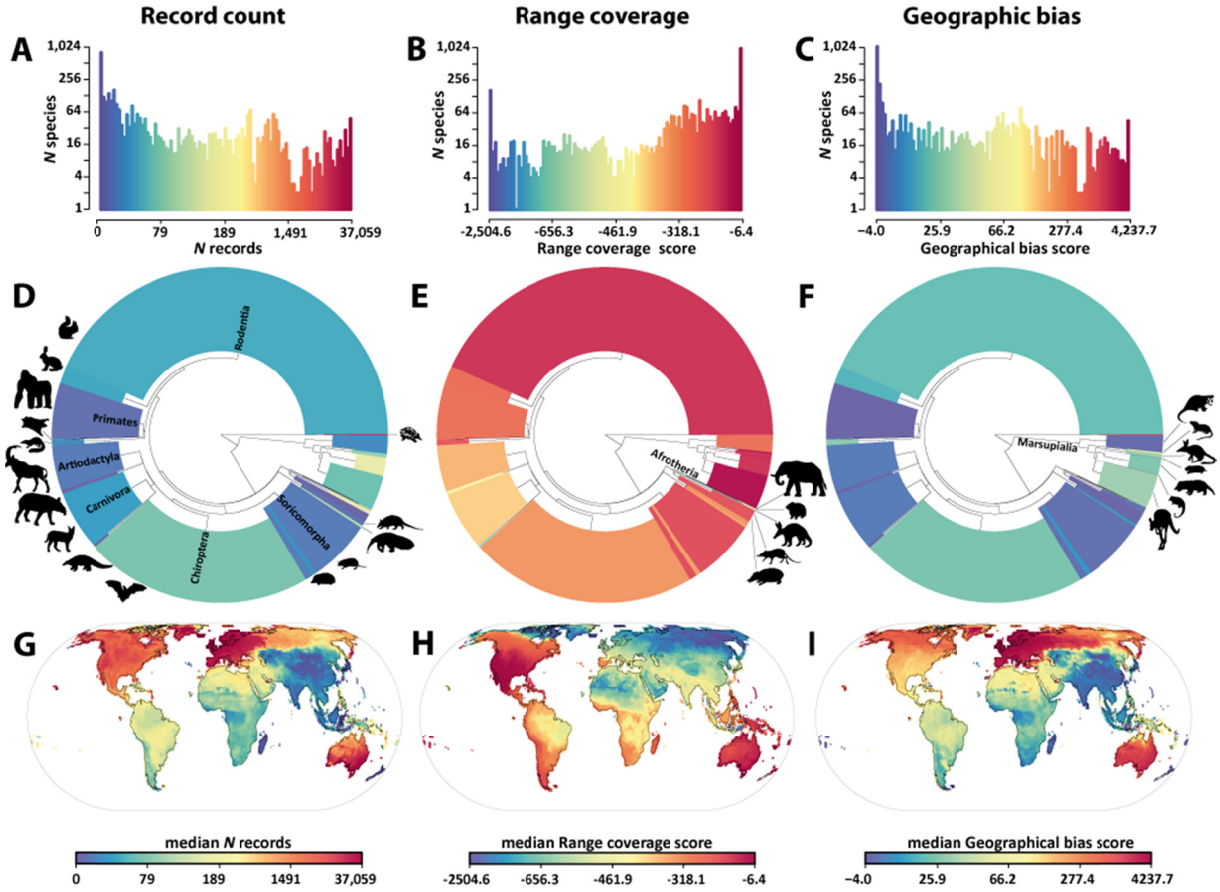

Figure 2. Species-level variation in *record count*, *range coverage* and *geographical bias* for 3,625 mammal species. Shown are frequency distributions of A) *record count*, B) *range coverage* and C) *geographical bias* across all species, median scores for each mammal order (D-F) and for each 110 km grid cell (G-I). Phylograms in D-F are based on Fritz *et al.* (2009) with order bars proportional to species number and colored following the scale in A-C. Labels and silhouettes in D-F are arranged for visual orientation. Note that in G-I grid cells show median metric values of the species occurring there and not information about records within grid cells. Color scales are calibrated on grid-cell percentiles and identical in each column of figures. Most species have few records (hence, the mostly cooler colors in D), yet most species also have small ranges which often have higher *range coverage* scores (warmer colors in E).

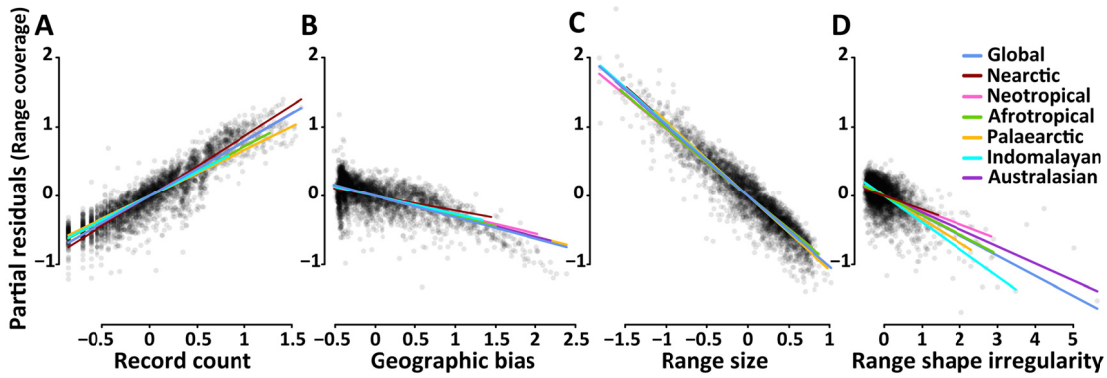

Figure 3. Effects of *record count*, *geographical bias* and range geometry on *range coverage*. Partial residuals show effects of A) *record count*, B) *geographical bias*, C) range size and D) range shape irregularity on *range coverage*, in each case controlling for the other three variables. Partial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

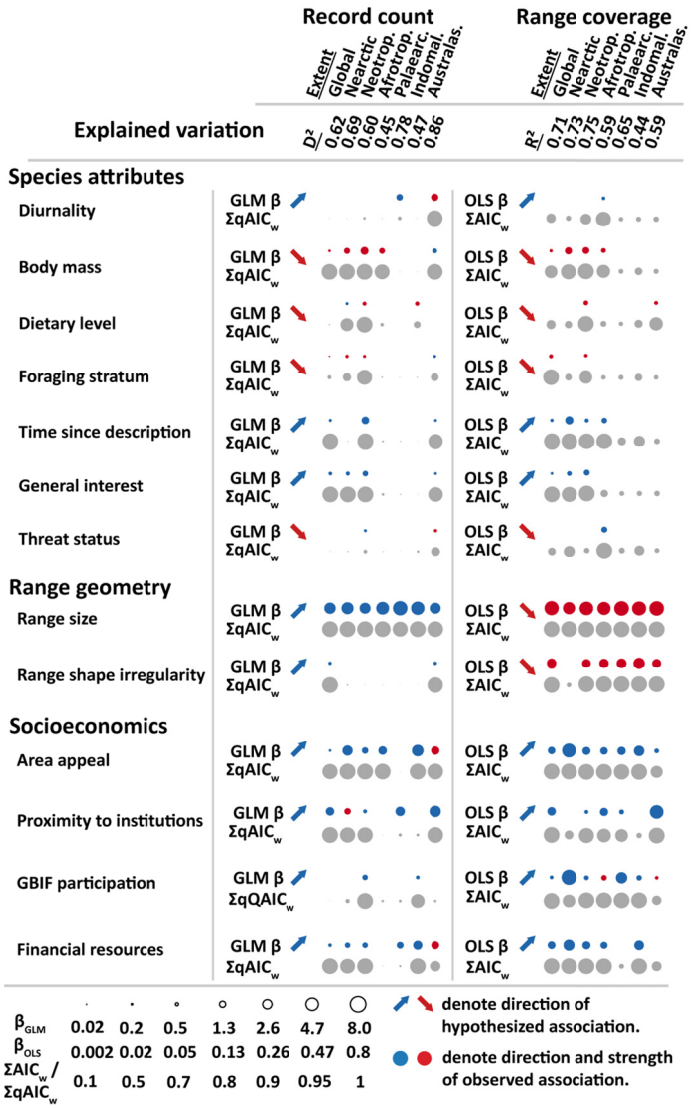

Fig. 4: Results from global and realm-scale models of *record count* and *range coverage* for 3,625 mammal species. Effects on *record count* were tested in multiple generalized linear regression models with a quasi-Poisson distribution, those on *range coverage* in multiple ordinary least squares models. 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 control variable 'mammal order' removed (Peres-Neto *et al.* 2006). For details on hypotheses, methods, and results, see Box 1, *Supporting Information*, Table S4.

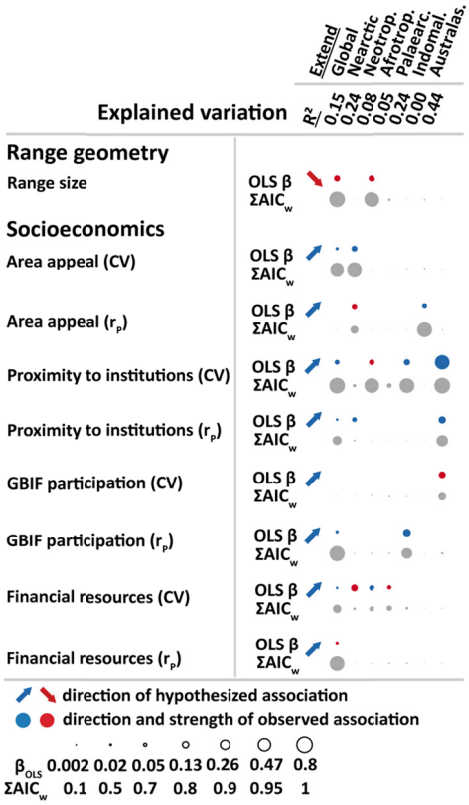

Fig. 5: Results from global and realm-scale models of *geographical bias* for 3,625 mammal species. Effects were tested in multiple ordinary least squares models. All possible model subsets were ranked based on AIC scores; results are shown for the minimum adequate model (with AIC=0). Two metrics (*cv* and r_P) of within-range geographical variation were used (details in *SI 3*). For other details see Fig. 4. Because *geographical bias* is highly correlated with *record count*, effects were tested with $\log_{10}$-transformed *record count* and mammal order included as fixed control variables. Partial adjusted variance explained (R^2) has effects of control variables 'mammal order' and '*record count*' removed (Peres-Neto *et al.* 2006).