Supplementary Information

Global drivers of species variation in mobilized occurrence information

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

Supplementary Material

SI 1 Mammal distribution data

We focused on records mobilized via the Global Biodiversity Information Facility (GBIF) as a representation of international efforts to mobilize biodiversity data. GBIF is by far the largest such effort in geographical and taxonomic scope (Edwards 2000; Graham *et al.* 2004). Data shared via GBIF represent the greatest body of mobilized species occurrence records, based on centuries' worth of museum specimens, citizen science observations, surveys, literature and other sources. We received 5,376,737 geo-referenced mammal records from GBIF in October 2012, that were provided to GBIF by 115 data publishers, including small institutions with a distinct taxonomic and geographical focus, large internationally active research museums, and citizen science programs (for an overview see Meyer *et al.* (2015)). 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, sirenians, and also excluding three terrestrial species with largely marine ranges - polar bear and two otter species; *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 used in IUCN range map and Red List data (IUCN 2010) partly differ from the taxonomy of Wilson & Reeder (2005), which underlies 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. 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 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 Global Orrell T. (custodian) 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 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 valid 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 2 Analyses of taxonomic bias, relative taxonomic and geographical species-level biases, and species without any mobilized records

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 S2 A) and to quantify the relative effects of zoogeographical realm and order memberships (Table S2 B). We also used nested type-III *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 S2 A) and whether they are more clearly distributed among zoogeographical realms than among mammal orders (Table S2 B). We found significant higher-taxonomic and realm-specific bias, i.e., missing species were not randomly distributed among orders or geographical regions (Table 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 *range coverage* models (Table S6). We expected the same relationships as with *record count* (Box 1). 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. Time since description and public interest are relatively weak positive predictors. Similar to our results for *record count* and *range coverage*, we found that whether or not species had any mobilized records was best predicted by range geometry and socio-economic factors: range size, area appeal, proximity to institutions and financial resources (Table S6).

SI 3 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 S5, Fig. 5). Our rationale was 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 values 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 4 Additional tests for effects of abundance-related traits

For a given body mass, 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 percentaged 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 has comparatively weak effects as the standardized coefficients are still smaller compared to those of range geometry and socio-economic factors (which remained similar to the original model; Table S4).

In another supplementary analysis, we tested for effects of a more direct measure of abundance by including population density in global models of *record count* 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 ($\beta_{GLM}=0.38$, $P=0.007$) 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: $\beta_{GLM}=0.29$, $P=0.03*$; foraging stratum: $\beta_{GLM}=-0.55$, $P=0.0005$, public interest: $\beta_{GLM}=0.67$, $P<<0.001$, range size: $\beta_{GLM}=3.78$, $P<<0.001$, range shape irregularity: $\beta_{GLM}=0.56$, $P<0.001$, area appeal: $\beta_{GLM}=0.63$, $P=0.004$, proximity to institutions: $\beta_{GLM}=1.96$, $P<<0.001$, GBIF participation: $\beta_{GLM}=-0.98$, $P<<0.001$, financial resources: $\beta_{GLM}=1.00$, $P<<0.001$). Population density was not retained in the minimum adequate model of *range coverage*.

SI 5 Testing for spatial and phylogenetic autocorrelation

We tested for spatial autocorrelation in model residuals, using Moran's I (Dormann *et al.* 2007). 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 (with Moran's I values ≤ 0.2). We also 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. S3).

SI 6 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.* 2014). 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 socio-economic hypotheses tested by 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 further investigated.

While data on further detectability-related traits like, e.g., coloration, fossoriality or vagility were not available, consistently weak effects of the tested species attributes lead us to conclude that detectability does not greatly impact the availability of 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 differences in occurrence information.

With the many integrated data sources, GBIF covers by far the largest share of globally mobilized point occurrence information. However, we acknowledge that several other global and regional data mobilization initiatives provide access to digital data, but are not yet integrated into GBIF (e.g., speciesLink; http://splink.cria.org.br). Additionally, several regions have digital or non-digital data that are not openly shared, and are thus *de facto* inaccessible to most data users. Accordingly, the data limitations demonstrated here should not be considered 'knowledge gaps', but gaps in occurrence information that is both digital and easily accessible in a standard format (Sousa-Baena *et al.* 2014).

Supplementary References

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.

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. (2012). Untested assumptions about within-species sample size and missing data in interspecific studies. *Behav. Ecol. Sociobiol.*, 66, 1363–1373.

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 Ecol. Evol.*, 19, 497–503.

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.

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.*

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L., *et al.* (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90, 2648.

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

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.

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

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

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

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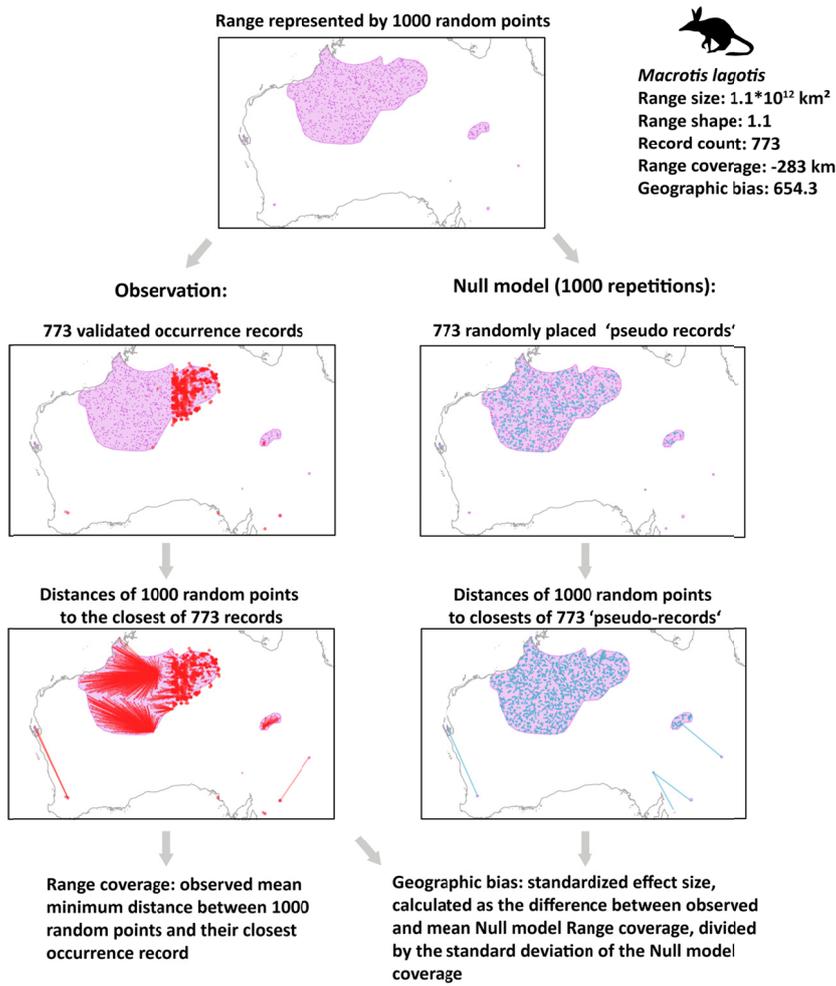
Figure S1. Visualization of calculation of *range coverage* and *geographical bias* in mobilized occurrence records.

Species with no GBIF-facilitated records
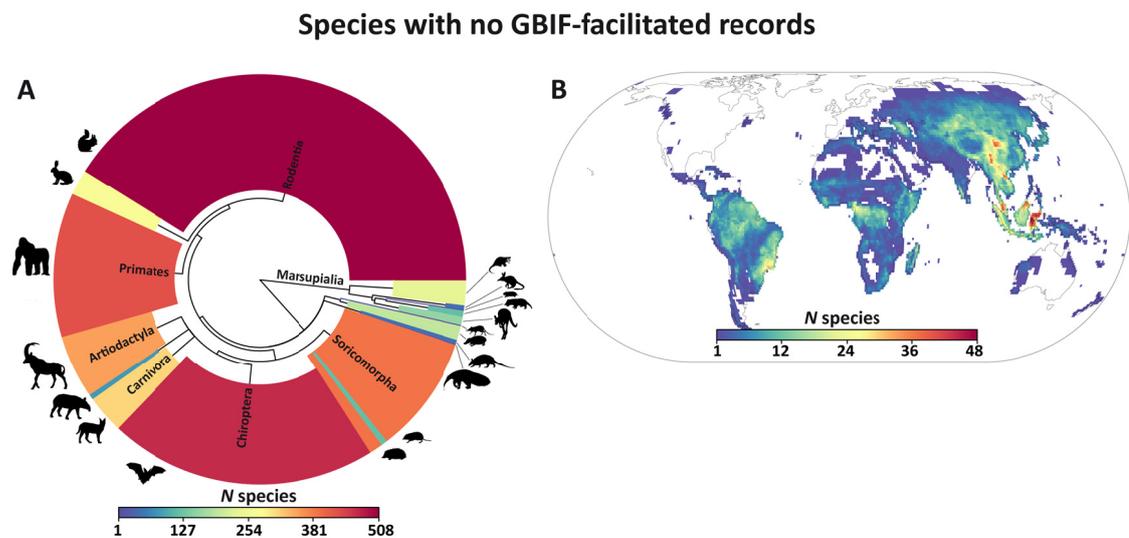
Figure S2. Global variation in number of species without any GBIF-facilitated records. Shaded areas at branch tips denote mammal orders, with widths proportional to the number of species. Labels within shaded orders in A) highlight the six most speciose orders. Silhouettes are for visual orientation. B) – the same represented as median per 110x110 km grid cell.
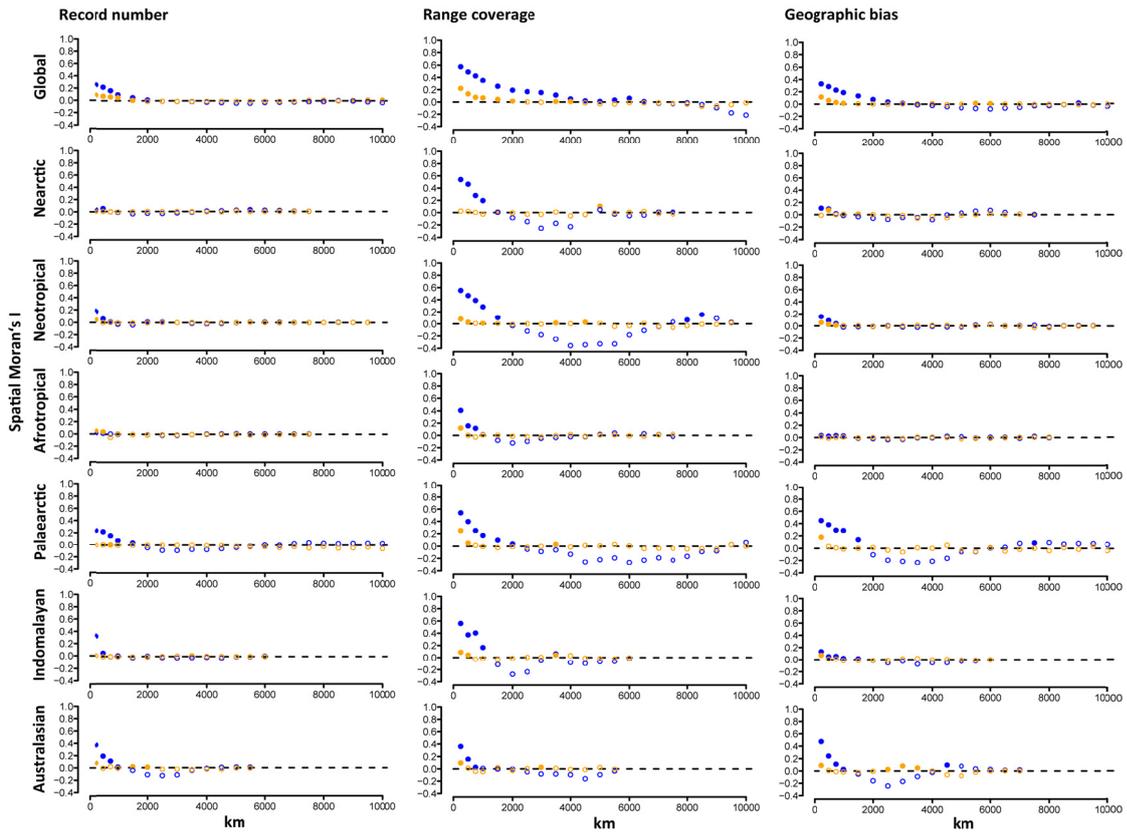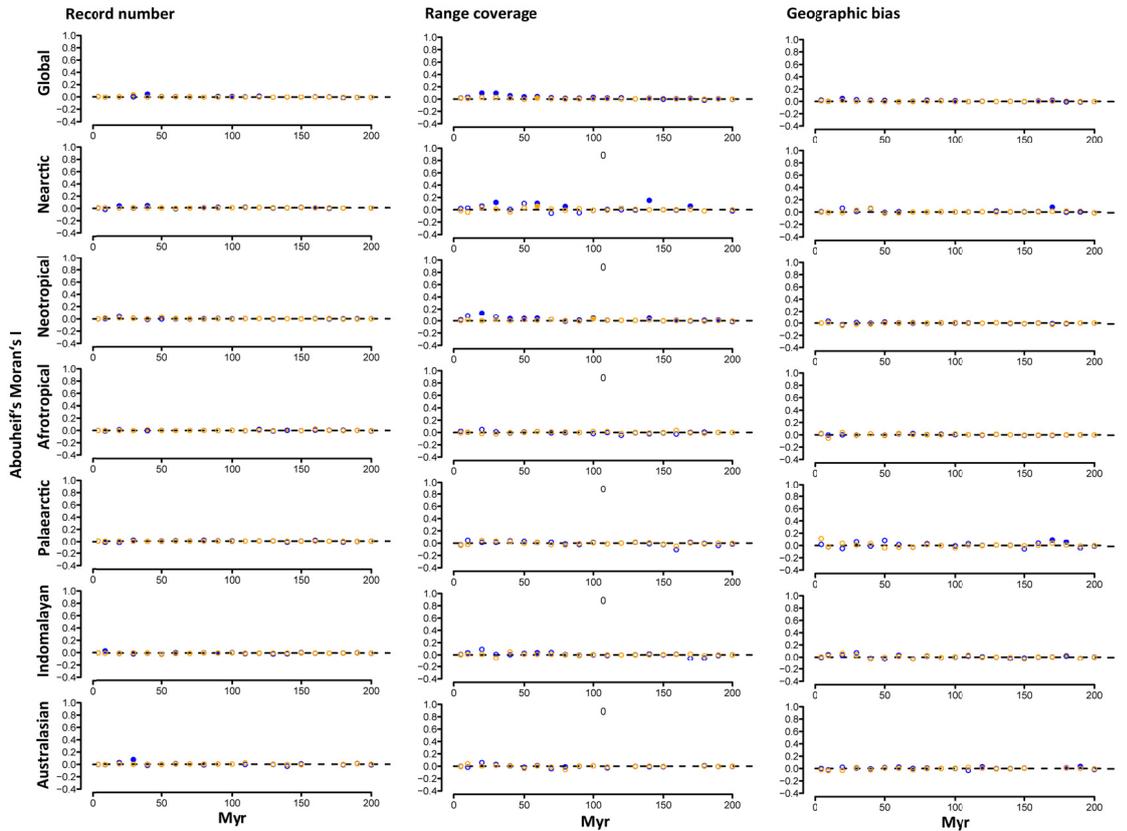
Figure S3. Correlograms of regression models of record count, range coverage and geographical bias. We tested for spatial autocorrelation using spatial Moran's I across different spatial distance classes (in km), and for phylogenetic autocorrelation using Abouheif's Moran's I (Abouheif 1999) across phylogenetic 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.
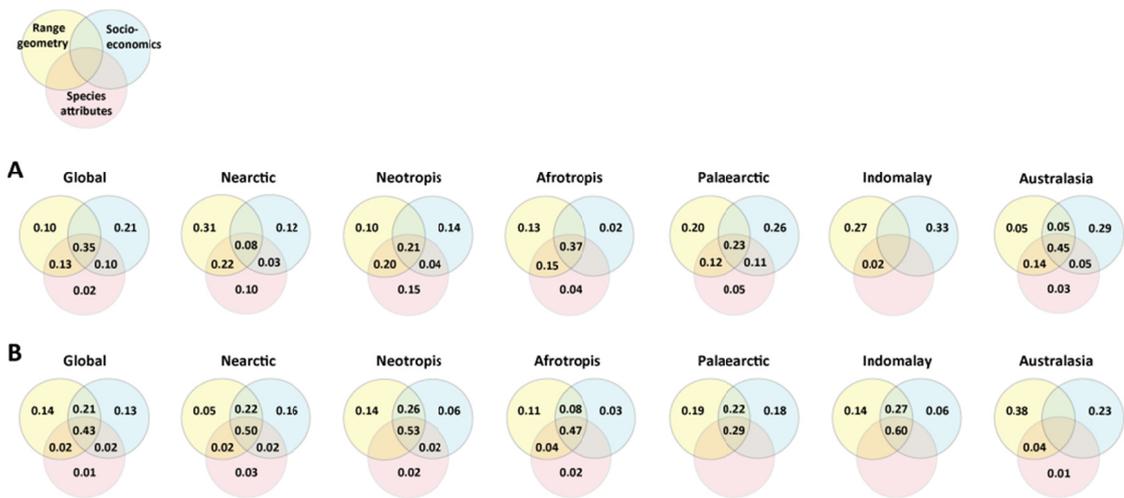
Figure S4. Results of variation partitioning. A) deviance partitioning of *record count*; B) variance partitioning of *range coverage*. Circles represent the three groups of hypotheses: upper left circle: geometry; upper right: socio-economics; lower: species attributes. Shown are the fractions of the 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 S1. Species 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

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 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 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
R²= 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
R²=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
R²=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
R²=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
R²=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
R²=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
R²=0.73	Range Size	-1.03***	0.03	-32.38
	Range shape irregularity	-0.25***	0.02	-15.80

Table 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 mass, 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 (D²) and partial adjusted variance explained (R²) 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	∆QAICw	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
D²=0.62	Time 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
D²=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
D²=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
	Time 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
D²=0.45	Dietary level	0.53.	0.29	1.82	0.67	2.1
	Time 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

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
D²=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
D²=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
D²=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
	Time 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
R²=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
	Time 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	Time since description	0.20***	0.04	4.71	1	1.8
R²=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
R²=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
	Time 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

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
R²=0.59	Time 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
R²=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	Time since description	0.06.	0.04	1.73	0.69	1.8
R²=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
R²=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 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 3* 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
R²=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
R²=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
R²=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
R²=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
R²=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
R²=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
R²=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 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 (D²) 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

D²=0.30

Predictor	GLM β	se	t	$\sum$QAICw	GVIF
Foraging stratum	-0.59***	0.15	-3.94	1	3.5
Time 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