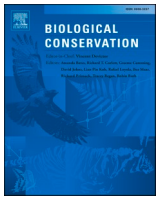
Biological Conservation 283 (2023) 110118

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon



The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently

Emilio García-Roselló [a], Jacinto González-Dacosta [a], Jorge M. Lobo [b,*]

[a] Departamento de Informática, University of Vigo, Campus As Lagoas, 32004 Ourense, Spain
[b] Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (C.S.I.C.), c/José Gutiérrez Abascal 2, 28006 Madrid, Spain

ARTICLE INFO

Keywords:
Completeness
Biodiversity data
Well-surveyed cells
Terrestrial classes
Climatic representativeness

ABSTRACT

The data collected by the Global Biodiversity Information Facility (GBIF), some 2.2 billion records, is arguably the largest international initiative to digitize and share primary biodiversity data. In this study, we examine the global distribution of completeness values discriminating those 30-minute cells that are likely to have reliable inventories for the most important terrestrial classes of Animalia and Plantae. The aim of this exploration is not only to show the biases and deficiencies in the biodiversity information collected so far, but also to estimate the climatic variability represented by these data in order to known their representativeness for conservation purposes. The results obtained show that information on biodiversity distribution is taxonomically and geographically biased towards regions and groups with more taxonomic resources and a longer naturalistic tradition. The amount of distributional data is very uneven across the different biological groups, and unrelated to the diversity they possess. The global patterns of completeness seem to be conditioned by the historical taxonomic, faunistic and floristic interest received by the different classes of organisms. In addition, well-surveyed global areas account for barely 1 % of global climate variability, leaving uncovered a large set of climatic conditions. All these results prevent us from relying exclusively on the available primary information on the distribution of organisms to identify biodiversity patterns and/or design conservation proposals. Given that the biodiversity crisis demands urgent action, biases and gaps in primary biodiversity information cannot be an excuse and conservation decisions must be made considering a broad set of criteria based on existing scientifically proven knowledge and techniques capable of providing the necessary answers.

1. Introduction

The recently concluded 15th Conference of the Parties (COP15) to the UN Convention on Biological Diversity adopted a historic global biodiversity agreement to protect 30 % of the Earth's land and water by 2030. This ambitious agenda has been seriously criticised as naïve and unrealistic (Cuff, 2022) and will be difficult to achieve without greater collaboration than the currently existing between all social stakeholders (Chan et al., 2022). Such international and societal involvement is crucial to obtain reliable and freely accessible information on the taxonomic and evolutionary identity of biodiversity elements, as well as data on their geographical distribution. This is and will be one of the key requirements for identifying priority areas for biodiversity conservation, as well as for designing effective conservation and restoration plans. The Global Biodiversity Information Facility (GBIF; http://www.gbif.org) is undoubtedly the largest international initiative to digitize and share the primary biodiversity data needed for these purposes. The information included in GBIF has been used in around 4000 peer-reviewed scientific publications (Heberling et al., 2021). A recent study examining the degree of data integration in global biodiversity databases showed that GBIF "serves as a central aggregator at a global scale that ingest species occurrence data from many databases" (Feng et al., 2022). Unfortunately, a common feature of all these biodiversity databases is the biased structure of the taxonomic and geographic information they contain (see, for example, Dennis and Thomas, 2000; Yesson et al., 2007; Hortal et al., 2015; Meyer et al., 2016; Troudet et al., 2017; Monsarrat et al., 2019; Hughes et al., 2021, Sánchez-Fernández et al., 2022 or Chesshire et al., 2023). These biases are mainly due to the uneven distribution of taxonomic resources and the opportunistic and contingent nature of the survey effort undertaken. The taxonomic and distributional information on biodiversity was collected during >300 years, while its mass digitization and storage were launched just over 20 years ago (Nelson and

* Corresponding author at: Museo Nacional de Ciencias Naturales, C.S.I.C., c/ José Gutiérrez Abascal 2, 28006 Madrid, Spain.
E-mail address: jorge.lobo@mncn.csic.es (J.M. Lobo).

https://doi.org/10.1016/j.biocon.2023.110118
Received 27 February 2023; Received in revised form 23 April 2023; Accepted 29 April 2023
Available online 13 May 2023

Ellis, 2018). Although much work remains to be done, the result of the huge effort carried out so far shows gaps and biases in biodiversity information, which constitute a serious drawback when using it to select sites for protection or make conservation decisions on biodiversity (Grand et al., 2007; Christie et al., 2021). It is certainly important to know and measure those biases, in order to be able to assess uncertainties and make better-informed decisions.

Several studies have undertook this task, but most of them are limited to specific groups (e.g., terrestrial vertebrates in Meyer et al., 2015; plants in Meyer et al., 2016; insects in Rocha-Ortega et al., 2021 or animals in Hughes et al., 2021), or were specifically focused on taxonomic biases (i.e. Troudet et al., 2017). Here we describe the status of these biases and information gaps at the global level for those plant and animal groups representing the largest part of known biodiversity. For this, we use the largest amount of taxonomic and distributional GBIF information studied to date. Using these data, we examine and compare the distribution of occurrence information among the main classes of organisms, also showing the geographic distribution of the inventory completeness, and the location of sites that are likely to be well inventoried for each of these classes. Considering that climate is a fundamental factor in explaining biodiversity patterns (Pilowsky et al., 2022), we finally estimate the ability of well-surveyed sites to represent the climatic variability of the planet. This analysis could tell us to what extent the information compiled so far provides a representative picture of the global distribution of the different groups.

2. Methods

2.1. Used data

Only the data available in GBIF (https://www.gbif.org/) are used in this study because this database aggregates a large part of the occurrence data available in other biodiversity databases (Feng et al., 2022). To limit the problems derived from the differences between taxonomical classifications when using information from databases (see Stropp et al., 2022), we also used the GBIF Backbone Taxonomy (GBIF Secretariat, 2022; see https://hosted-datasets.gbif.org/datasets/backbone/) to derive a complete taxonomy of the Animalia and Plantae kingdoms. Only those species labelled as "accepted" in this database are included, excluding synonyms and subspecies (accessed August 18, 2022). We have focused on the class as a high-rank taxonomic category because of its wide use since its proposal by Linnaeus, and because of the number of terrestrial taxa it contains in comparison with other higher-ranking taxonomic categories (Phylums would be too few taxa and Orders too many; see Ruggiero et al., 2015). We aim to select those taxa with enough data to estimate the global distribution of the compiled information. To do that, those classes in which 50 % or more of their species inhabit marine biomes according to the World Register of Marine Species (WoRMS Editorial Board, 2022; see https://www.marinespecies.org/) and the Interim Register of Marine and Nonmarine Genera (Rees, 2022) were discarded. Among the remaining classes, those with <100 occurrences per species were rejected in order to maximize the generation of completeness results. In total, 12 classes of Animalia and 8 of Plantae were selected (Table 1). The resulting number of species in each one of the selected classes was compared to the number of recognized species for those classes according to several heterogeneous sources: Bánki et al. (2022), Bellinger et al. (1996-2022), BirdLife International (2018), Martin et al. (2007), Chapman (2009), Blick and Harvey (2011), Hopkin (1998), Enghoff et al. (2015), Sendra et al. (2021), Galli et al. (2018) and the web pages of the Mammal Diversity Database (https://www.mammaldiversity.org/), AmphibiaWeb (https://amphibiaweb.org/amphibian/speciesnums.html), The Reptile Database (http://www.reptile-database.org/), World Spider Catalog (https://wsc.nmbe.ch/), and AlgaeBase (https://www.algaebase.org/). The result shows that the total number of species in each class according to these heterogeneous sources is positively correlated with the total number of species indicated by GBIF ($r = 0.998$; $p < 0.001$).

Occurrences were downloaded from GBIF (complete datasets can be retrieved from GBIF.org, 2022a, 2022b) and imported to ModestR (García-Roselló et al., 2013; see https://www.modestr.es/). Only georeferenced occurrence records labelled as "observations", "human observations" or "preserved specimens" were retrieved. Next, a simple data cleaning was performed. From the total of downloaded occurrences, we excluded those with i) the same latitude and longitude, ii) 0° latitude or longitude, and iii) occurrences in habitats other than those corresponding to terrestrial or freshwater ecosystems (see García-Roselló et al., 2014 for details). Finally, almost 1.9 billion occurrences belonging to 1.67 million of different species were used in subsequent analyses (Table 1), which corresponds to 99.6 % and 97.3 % of total occurrences of Animalia and Plantae available in GBIF, respectively.

2.2. Completeness estimations

The number of occurrences for each one of the selected classes and

Table 1
Animal and plant selected classes to examine the geographical distribution of completeness including their percentage of marine species (%Mspp), total number of species according to diverse sources (Tspp), total number of species with occurrence data according to GBIF (Tspp-Occ), total number of occurrences available in GBIF (TOcc), total number of terrestrial occurrences (Occ), and mean number of occurrences in those species with occurrence data in GBIF (Occ/sp).

Kingdom	Phylum	Class	%Mspp	Tspp	Tspp-Occ	TOcc	Occ	Occ/spp
Animalia	Annelida	Clitellata	13.6	8000	441	184,749	183,332	419
Animalia	Arthropoda	Arachnida	1.8	102,248	32,493	6,285,304	6,176,439	193
Animalia	Arthropoda	Branchiopoda	6.8	1418	786	217,313	177,472	276
Animalia	Arthropoda	Chilopoda	2.3	3141	1184	196,961	192,701	166
Animalia	Arthropoda	Collembola	0.7	9400	1655	318,828	303,968	193
Animalia	Arthropoda	Insecta	0.0	1,053,578	326,783	138,416,945	134,844,612	424
Animalia	Chordata	Amphibia	0.0	8489	6636	6,079,483	6,033,204	916
Animalia	Chordata	Aves	5.8	11,121	11,123	1,408,294,127	1,362,781,118	126,611
Animalia	Chordata	Mammalia	2.1	6495	5752	20,137,575	18,391,715	3501
Animalia	Chordata	Reptilia	0.9	11,690	9740	5,582,428	5,304,872	573
Animalia	Mollusca	Bivalvia	33.6	9773	7913	1,741,159	768,060	220
Animalia	Mollusca	Gastropoda	49.8	72,239	42,747	4,356,824	2,991,687	102
Plantae	Bryophyta	Bryopsida	0.3	12,953	11,528	8,306,792	8,032,508	641
Plantae	Marchantiophyta	Jungermanniopsida	0.0	6696	6238	1,892,858	1,816,934	303
Plantae	Marchantiophyta	Marchantiopsida	0.0	492	442	210,475	201,857	476
Plantae	Tracheophyta	Liliopsida	0.0	80,111	61,998	62,590,224	61,193,814	1010
Plantae	Tracheophyta	Lycopodiopsida	0.2	1434	1249	710,723	698,594	569
Plantae	Tracheophyta	Magnoliopsida	13.1	260,930	228,708	215,896,979	211,924,249	944
Plantae	Tracheophyta	Pinopsida	1.6	615	700	4,760,871	4,721,903	6801
Plantae	Tracheophyta	Polypodiopsida	0.7	12,092	10,972	8,954,777	8,760,974	816

for each terrestrial world cell of 30 arcminutes (approximately 55 × 55 km at the equator; n = 66,448 cells) was assumed to act as a surrogate of survey effort (see Lobo, 2008; Lobo et al., 2018). In each cell a record-by-species matrix was built which is subsequently used to estimate the relationship between the accumulated number of species and the number of occurrences. The so derived accumulation curves allow estimating the degree of completeness of each cell. To do this, accumulation curves were calculated according to the exact estimator of Ugland et al. (2003). All those Arctic or Antarctic cells permanently covered by ice according to ISRIC database (https://data.isric.org) were not considered. The obtained species accumulation curves were subsequently adjusted to the rational function, which has been selected because it offers good comparative results and requires fewer parameters than other functions (see Flather, 1996), and is also capable of adjusting to the data on a good number of occasions (Pelayo-Villamil et al., 2018). The extrapolated asymptotic values were subsequently used to estimate both the probable number of species in each cell and their completeness (i.e., the percentage of species that has been inventoried over the total predicted). Besides completeness, the final slope of the accumulation curve and the ratio between the number of occurrences and the observed species were also calculated. These three parameters are calculated for all the cells with at least one species recorded for each taxonomic class. Those cells with completeness values equal to or higher than 90 %, slope values equal or lower than 0.01(one species added each 100 occurrences), and ratios between occurrences and species equal to or higher than 25 were selected as probable well-surveyed cells (WSCs; see Lobo et al., 2018). This complete process was carried out using a modified version of the R package KnowBR (Lobo et al., 2018; Guisande and Lobo, 2019) which was optimized and integrated into the freely available ModestR software (www.modestr.es; García-Roselló et al., 2013, 2023). This application allows quick estimation of the completeness of species inventories at different resolutions, for large datasets, and in an unlimited number of territorial units simultaneously, be they cells or arbitrary shapes. A detailed explanation of how to perform this process in ModestR can be found in https://www.modestr.es/sweb/documents/tutorial_stepbystep/PermalinkTutorials.php?tutorial=26.

2.3. Data treatment

First, we estimate the geographical distribution of the mean and standard deviation (*sd*) of the calculated completeness values for all the considered taxonomic classes. Mean values allow us to examine the general pattern of the available biodiversity information, while *sd* values help to discriminate spatial units with reliable inventories for a high proportion of taxonomic classes from those with high completeness values for only some groups.

Completeness values for all the taxonomic classes and spatial cells were submitted to a Principal Component Analysis (PCA) to find the uncorrelated components that identify closely related taxonomic classes with respect to geographic variation in their completeness. With this analysis, we aim to find out whether there is any taxonomic and/or geographical pattern in the distribution of existing biodiversity information. Completeness proportions were previously submitted to a logit transformation to fulfill linear modeling assumptions (Warton and Hui, 2011). Historical climatic information for each terrestrial 30 minute cell was also obtained from WorldClim2 (Fick and Hijmans, 2017; see https://www.worldclim.org) and ENVIREM (Title and Bemmels, 2018; see https://envirem.github.io). In total, 35 climatic variables representing temperature, precipitation and evapotranspiration values are used. After standardizing all these variables to mean zero and standard deviation one to avoid measurement scale differences, a new PCA was carried out to extract the two main components representing terrestrial climatic variability. Once extracted, the climate space covered by these components is used to estimate i) the proportion of the global terrestrial climatic conditions covered by the WSCs (*%Clim*), and ii) the proportions of the overall range of these two first climate principal components represented by these WSCs (*%PC1* and *%PC2*). To do that, the factor scores were rescaled between 0 and 1 and the corresponding climatic square space bounded by these two components was divided into a number of cells equal to the number of terrestrial cells considered in the geographical analyses (n = 66,448 cells). Determining the location of the WSCs within this climate space allows us to calculate *%Clim* as the number of cells with unique climate combinations that appear to be well-surveyed, and *%PC1* or *%PC2* as the proportion of the full ranges of these climate components represented by the values in the WSCs (maximum-minimum values). These two measures have been used to estimate the ability of the set of available WSCs to serve as a representative sample of the global climatic conditions.

3. Results

The number of species belonging to each class and the number of occurrences for each species are uncorrelated ($r = 0.05$, ns). Most part of the selected occurrences belong to the species of the class Aves (≈74 %), followed by Magnoliopsida (≈12 %), Insecta (≈7 %), Liliopsida (≈3 %) and Mammalia (≈1 %). The occurrences of all other considered classes represent <1 % of the total number of occurrences downloaded, so there are huge disparities in the number of occurrences per species among the different classes (Table 1).

3.1. Global distribution of completeness

The mean completeness values of the 20 selected classes are geographically biased (Fig. 1); high mean completeness values are located in areas across Western Europe, North America, eastern Australia, and north-eastern Asia. These regions are thus the only ones having reliable local inventories for a high proportion of biodiversity. Other areas in the same or nearby regions, as well as in Central and Southern Africa, Asia and South America, may show highly variable completeness values depending on taxonomic classes (see Fig. S1 in Supplementary material) as indicated by the distribution of the standard deviation of completeness values (Fig. 1). At the same time, large parts of Asia, Africa and South America have very few or no records at all, and very low completeness values. The relationship between mean and *sd* completeness values reveals a clear pattern in which only a few global cells would have high completeness values for all classes. Thus, <1 % of total cells have completeness values above 75 % and the completeness is below 10 % in a high proportion of cells (57 %). In fact, about 12 % of the total cells have no occurrence data for any class, and almost 44 % of the total cells have <100 occurrences in total.

3.2. Completeness patterns

The overall variance of the completeness values of all classes can be summarised in three uncorrelated factors, able to explain respectively 47.5 % (Pc1), 9.5 % (Pc2) and 5.4 % (Pc3) of total variability. The first component brings together all classes with the highest amount of information (98.6 % of occurrences and 96.1 % of total species): all terrestrial vertebrate animals, plant angiosperms, as well as insects and arachnids (Fig. 2). The positive score values of this first component are related to the existence of high completeness values for all these groups in Europe, North and Central America, Australia and some African, South American or Asian areas. All the remaining invertebrate groups are grouped in the second component, showing a clear distribution pattern confined to central and northern Europe (Fig. 2). Finally, the third component comprises the geographical distribution of the completeness values of the remaining plant classes, including conifers, ferns, liverworts, mosses and club-mosses. In this case, the pattern is much less aggregated, including areas of northern Europe and North America, Japan or New Zealand (Fig. 2).

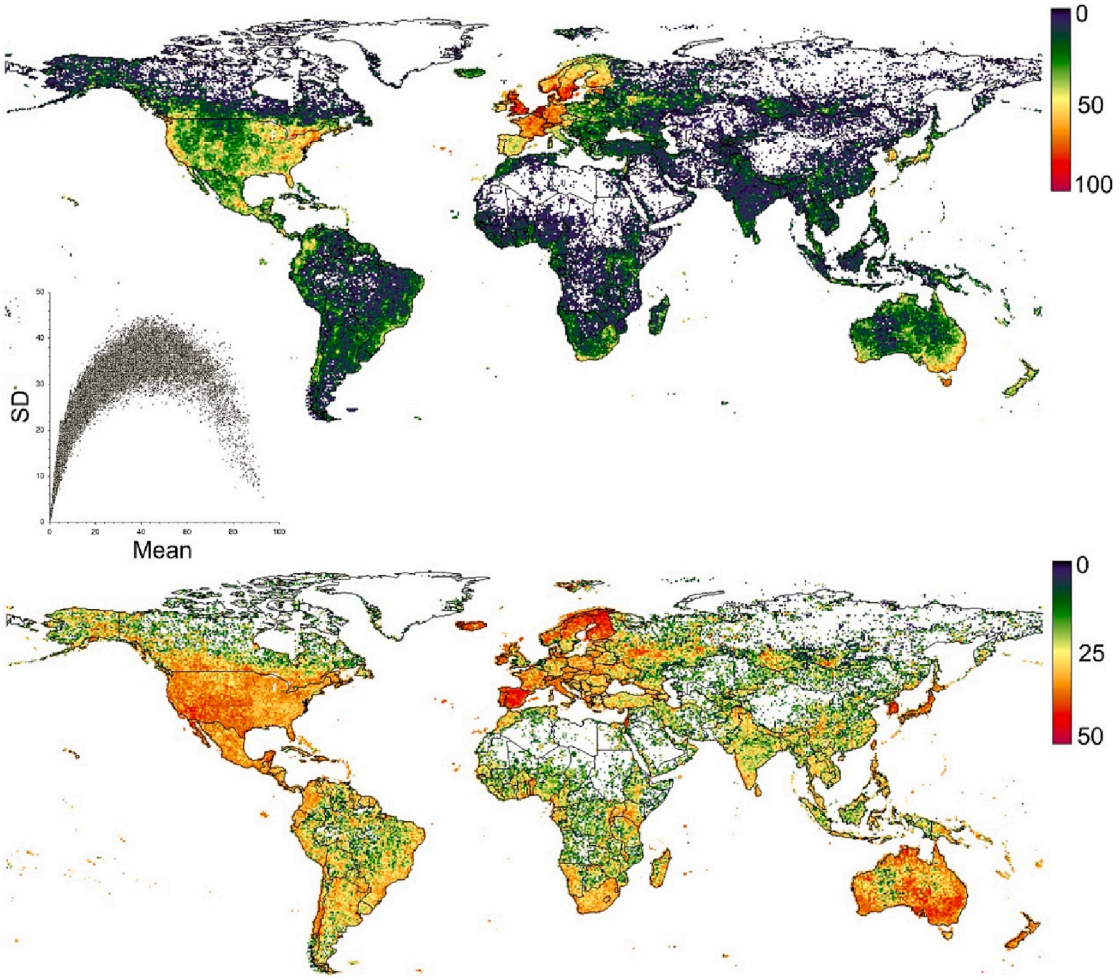


Fig. 1. Global distribution of the mean (upper panel) and standard deviation (lower panel) values of completeness for the 20 selected taxonomic classes (Table 1) at a 30 arcminute cell resolution, showing a bi-plot of the relationship between these two parameters.

3.3. Well-surveyed cells

Around 17.8 % of total cells (n = 11,797) were assigned as well-surveyed for at least one class (Fig. 3). Most part of these cells (68 %) appear as well-surveyed in only one taxonomic class. Not one single 30′ cell can be considered as well-surveyed for all the classes at the same time, and only one cell, located in the United Kingdom, appears as well-surveyed in sixteen taxonomic classes at the same time. The geographical distribution of the number of well-surveyed cells (Fig. 3) shows a clearly biased pattern in which only some European countries as United Kingdom, Switzerland, Belgium, The Netherlands or Sweden show well-surveyed cells for a high number of taxonomic classes.

3.4. Climatic representativeness

Two PCA components account for 79.2 % of the total variability existing in the 35 considered climatic variables (56.3 % and 22.9 %, respectively). The first component represents a gradient of temperature and thermal stability because it is positively related with several temperature variables, such as mean temperature of the coldest quarter (factor loading = 0.92) and minimum temperature of the coldest month (0.93), but negatively with temperature seasonality (−0.92) and continentality (−0.93). The second component represents a dryness gradient, being positively related with precipitation variables (annual precipitation; factor loading = 0.92) and negatively with the aridity index (−0.74) (Fig. 4).

The cells considered as well-surveyed for any taxonomic class cover a 45.3 % of the climatic space represented by these two uncorrelated components, and a high proportion of the complete range of values in both components (88.7 % for PC1 and 78.2 % for PC2; see Fig. 4). When only those cells considered as well-surveyed for at least five different classes are taken into account, the percentage of climatic representativeness diminishes to 6.2 %, and the covered percentage of component ranges decrease to 72.6 % for PC1 and 46.0 % for PC2. The percentage of climatic representativeness diminishes to 0.9 % if only the cells recognized as well-surveyed for ten or more taxonomic classes are considered. In this case, the encompassed percentage of component ranges diminishes to 41.9 % for PC1 and 21.3 % for PC2 leaving uncovered the vast majority of locations with high temperature and precipitation values (high PC1 and PC2 values) (Fig. 4).

The number of cells that can be considered as well-surveyed differs among classes, highlighting the comparatively large number of cells with relatively reliable inventories in the case of birds (Table 2). The number of well-surveyed cells is positively and highly correlated with the climatic representativeness of these cells ($r = 0.996$; $p < 0.001$), but not so much with the average of the climatic ranges covered by the two principal components ($r = 0.59$; $p = 0.006$).

4. Discussion

The present study involved the analysis of an unprecedented amount of data stored in GBIF, focusing on terrestrial species and georeferenced occurrences. We have described i) the comparative distribution of occurrence information for the main classes of terrestrial organisms, ii)

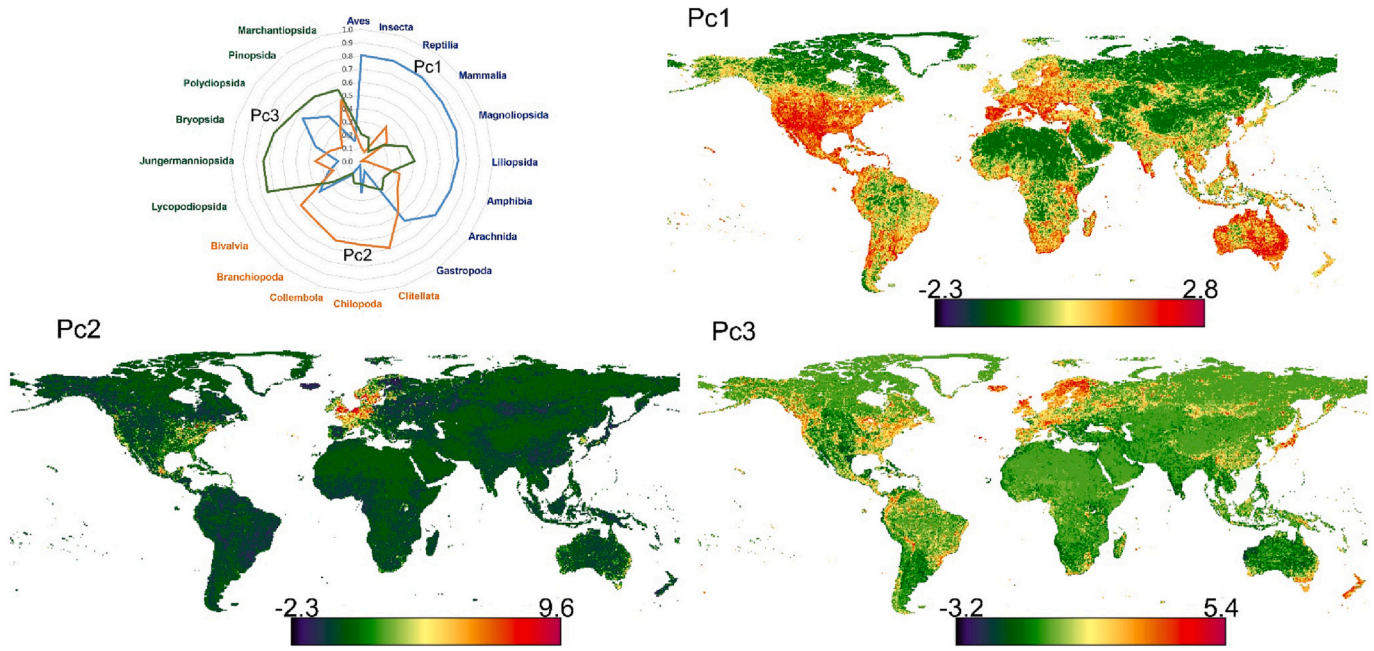


Fig. 2. Diagram showing the factor loadings (Varimax normalized) of the three principal components (Pc1, Pc2 and Pc3) for the different considered taxonomic classes. The colour of the classes corresponds to the colour of the component with a higher loading value. Pc1, Pc2 and Pc3 maps represent the geographical variation of the factor scores at a resolution of 30 arcminutes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

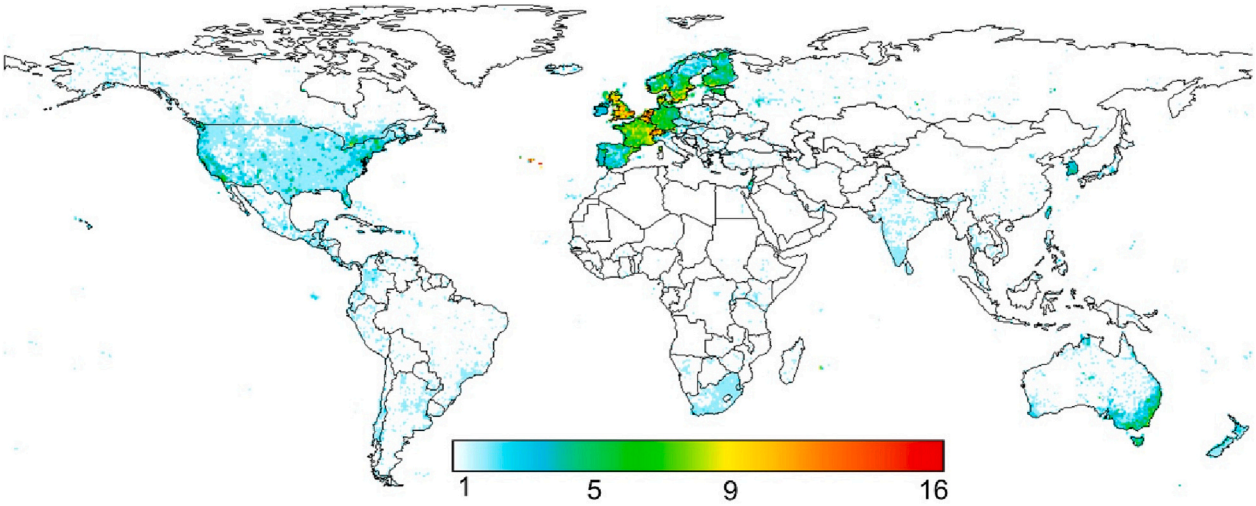


Fig. 3. Global distribution of the number of taxonomic classes whose inventories can be considered well-surveyed in each 30 minute cell.

the geographic distribution of inventory completeness, iii) the location of well-surveyed sites, and iv) the ability of well-surveyed sites to represent global climate variability. The results provide useful information for understanding how far we have come in our efforts to collect biodiversity information, and the potential for using these data for effective global conservation planning.

As expected, the information about the distribution of biodiversity shows obvious taxonomic biases, in this case exemplified by the GBIF data. The existence of data biases and shortcomings in favour of the most conspicuous biological groups and the best-studied territories is a clear and well-established fact in numerous studies (e.g. McRae et al., 2017; Titley et al., 2017; Troudet et al., 2017 or Freeman and Pennell, 2021). Our results clearly show again that the most diversified biological groups do not accumulate more information. Invertebrates, which account for the majority of terrestrial biodiversity (Stork, 2018), accumulate a comparatively smaller percentage of data than groups more appealing to humans such as vertebrates or vascular plants. Mean completeness values also reveal the existence of the often observed pattern of geographical bias towards those regions with more taxonomic resources and a long-standing naturalistic tradition (Dennis and Thomas, 2000; Yesson et al., 2007; Fattorini, 2013; Hortal et al., 2015; Ruete, 2015; Meyer et al., 2016; Monsarrat et al., 2019; Hughes et al., 2021; Titley et al., 2017; Rocha-Ortega et al., 2021; Sánchez-Fernández et al., 2022; Chesshire et al., 2023; García-Roselló et al., 2023). The global pattern derived from these mean completeness values shows that highly reliable inventories for a good number of classes of organisms can only be found in Europe, and to a lesser extent in North America, Australia or East Asia. Other European and non-European regions may offer reliable inventories for some specific taxonomic classes, but only a few regions in northern or central Europe can boast of hosting well-

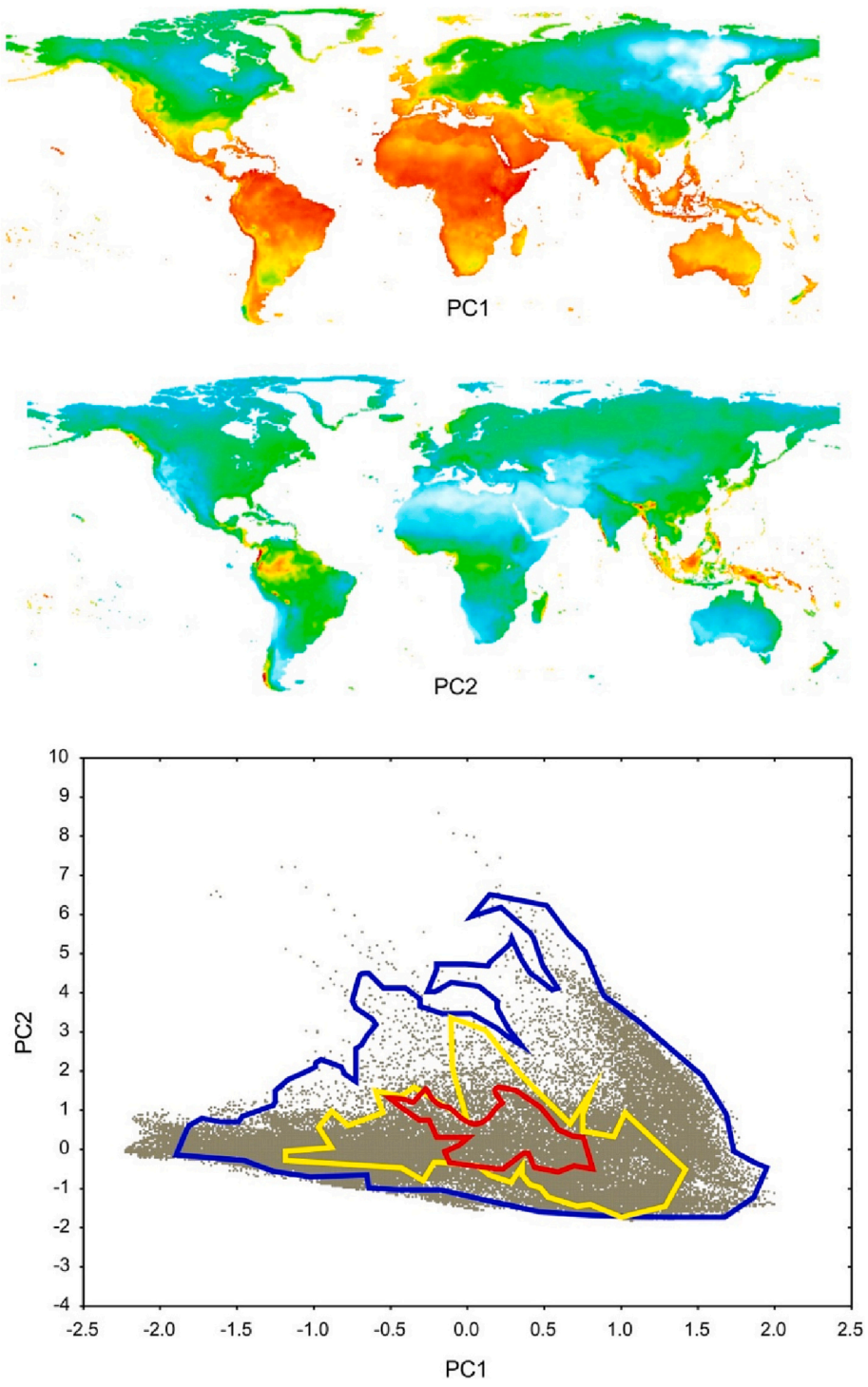


Fig. 4. World geographical distribution of the factor scores of the two first components (PC1 and PC2) of a principal component analysis carried out using thirty-five climatic variables at 30 min resolution, and bi-plot showing the complete factor scores of these components (grey points) and the outer contours of the corresponding scores when we consider all the well-surveyed cells for anyone taxonomic class (blue polygon), well-surveyed cells for five taxonomic classes at the same time (yellow polygon), and well-surveyed cells for ten taxonomic classes at the same time (red polygon). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

surveyed cells for a good part of the biodiversity. Thus, 0.02 % of all 30′ world cells (n = 14) have mean completeness values equal to or higher than 90 %, and all of these cells are located in northern or central Europe (6 in United Kingdom; 3 in Sweden, 3 in Norway, 1 in Germany and 1 in the Netherlands). Similarly, only some central and north European countries harbour well-surveyed cells for a high number of taxonomic classes. Our results also show that those cells that can be considered well-surveyed for at least half of the taxonomic classes considered do not account for even 1 % of global climate variability and do not cover a large set of climatic conditions.

Interestingly, our PCA results indicate that different geographical patterns can be observed in the completeness values according to taxonomic classes. Vertebrates, vascular plant and insects are the groups that historically concentrate most of the naturalistic interest and data. As a consequence, they show a relatively similar geographical pattern in which moderate to high completeness values can be found in a wide variety of continents and regions, but mainly in Europe, North America, Australia, some regions of South America and tropical and East Asia. Clearly, within a single taxonomic class, the distribution patterns of available information may differ among its various groups. This is especially true for the most hyperdiverse terrestrial class, the insects, which can show a great heterogeneity of completeness patterns and

Table 2
Number of well-surveyed 30′ cells (WSCs) and, in brackets, percentage over total of terrestrial cells. %Clim represents the proportion of the Earth climate conditions covered by the well-surveyed cells in each one of the taxonomic classes, while %PC1 and %PC2 are the proportions of the global range of two first principal components of climatic variables represented by the corresponding well-surveyed cells.

	WSCs	%Clim	%PC1	%PC2
Aves	10,845 (16.3 %)	42.4	87.2	77.7
Amphibia	2445 (3.7 %)	12.7	79.5	61.6
Mammalia	1756 (2.6 %)	9.1	79.5	70.5
Polypodiopsida	1484 (2.2 %)	7.1	63.3	52.3
Reptilia	1480 (2.2 %)	7.9	69.4	68.2
Liliopsida	1298 (2.0 %)	6.0	55.1	43.8
Magnoliopsida	1293 (1.9 %)	6.1	84.6	47.0
Pinopsida	1050 (1.6 %)	5.6	83.5	51.2
Lycopodiopsida	676 (1.0 %)	3.6	62.0	50.0
Insecta	286 (0.4 %)	1.8	83.4	52.2
Bryopsida	202 (0.3 %)	1.2	43.8	25.0
Gastropoda	182 (0.3 %)	1.1	66.2	27.2
Arachnida	162 (0.2 %)	1.0	78.0	47.5
Jungermanniopsida	136 (0.2 %)	0.8	42.5	21.3
Marchantiopsida	98 (0.1 %)	0.6	44.7	18.0
Bivalvia	81 (0.1 %)	0.5	60.2	45.8
Clitellata	75 (0.1 %)	0.5	45.4	11.0
Collembola	72 (0.1 %)	0.5	64.0	11.8
Chilopoda	71 (0.1 %)	0.4	50.2	21.2
Branchiopoda	49 (0.1 %)	0.3	59.5	27.2

large information gaps (García-Roselló et al., 2023). The less surveyed plant classes show a much more concentrated pattern in Europe, North America and East Asia, while the completeness values of the remaining invertebrates follow a clearly aggregated pattern in central and northern Europe. Of course, the historical, biogeographical and environmental differences between the considered groups must influence the observed patterns of available information (Freeman and Pennell, 2021). However, our results suggest that the differences in global completeness patterns are importantly conditioned by the historical growth in the taxonomic, faunistic and floristic interest. Thus, when little information has yet been collected on the distribution of a taxon, it is aggregated at those localities where taxonomic resources are historically concentrated. However, as taxonomic knowledge accumulates, more data are collected at distant locations and reliable inventories begin to appear in other regions (see Sastre and Lobo, 2009). The origin of biodiversity information on the European continent is evident in the sequence of patterns of distribution of completeness provided by the three groups of organism classes obtained by PCA.

In conclusion, our results confirm i) that the available information on the distribution of biodiversity is unambiguously biased and scarce, ii) that the amount of information on the distribution of different biological groups is very uneven and unrelated to the diversity they possess, iii) that the geographical coverage of the data is also biased towards regions and countries, which, although not highly diverse, do have a long tradition of taxonomic studies, and iv) that the overall climatic representativeness of the sites with reliable inventories is very low. Although we do not doubt the increasing usefulness of GBIF data for the study of biodiversity patterns and the design of conservation proposals (see, for example, García-Roselló et al., 2015; Di Marco et al., 2019; Mokany et al., 2020), it is clear that these biases and shortcomings make it difficult to generate global recommendations on priority areas for biodiversity conservation based solely on the available primary information on the distribution of organisms. The spatial resolution used in our case may be considered too broad to select protected areas in some groups, and adequate in others, but it is clear that there is still a lot of information available to be incorporated. However, we doubt that the inclusion of more data would ostensibly change the observed pattern in the short to medium term. For example, the observed difference in the amount of information between classes such as Aves, which alone account for almost 75 % of all occurrences, and other more diverse classes, appears to have increased from previous results, suggesting that citizen science and societal preferences may have further amplified the differences in the amount of information between groups (Troudet et al., 2017).

A better understanding of these still poorly known aspects and well-designed public policies can help take measures to reduce these biases. However, it will be virtually impossible to obtain the taxonomic and biogeographic information needed to generate reliable conservation responses in a short time, without an innovative and integrative approach to the problem that involves pooling the results of a wide range of procedures. Conservation decisions can be partially congruent between groups (Critchlow et al., 2022) and various procedures can help estimate the diversity and distribution of organisms in the absence of comprehensive data. Regression and machine-learning methods may help to predict the distribution of species in some circumstances (Guisan et al., 2017), expert-based maps may provide complementary forecasts in absence of data (Merow et al., 2017), and approaches based on geodiversity (Zarnetske et al., 2019) or environmental diversity (Engelbrecht et al., 2016) can provide useful information when primary biodiversity data are lacking. The biodiversity crisis requires urgent decisions (Leadley et al., 2022) that cannot wait for improved taxonomic and geographic coverage of information. Biases and gaps detected must be used to design new surveys and recognise the location of areas of ignorance (Tessarolo et al., 2021), but also to take better informed and grounded conservation decisions. These conservation decisions can no longer be postponed and should be based on a broad set of criteria and information sources. Lack of information on biodiversity, or biased information, can never be an excuse for not making the courageous decisions that life on Earth needs. Biodiversity science possesses the scientifically proven knowledge and techniques capable of providing the necessary answers. Are our political and social leaders ready to take up this challenge or will we biologist continue to be the notaries of a dying world?

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biocon.2023.110118.

CRediT authorship contribution statement

Emilio García-Roselló: Conceptualization, Methodology, Writing – review & editing. **Jacinto González-Dacosta:** Methodology, Writing – review & editing. **Jorge M. Lobo:** Conceptualization, Formal analysis, Writing – original draft.

Declaration of competing interest

The authors have no personal or financial competing interests to declare.

Data availability

All the used data are included in the provided GBIF links.

Acknowledgements

We are indebted to the entomologists and naturalists who over the span of dozens of years have compiled the data studied herein. We appreciated the insightful comments and suggestions of an anonymous referee.

References

Bánki, O., Roskov, Y., Döring, M., Ower, G., Vandepitte, L., Hobern, D., Remsen, D., Schalk, P., DeWalt, R.E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Adlard, R., Adriaenssens, E.M., Aedo, C., Aescht, E., Akkari, N., Alexander, S., et al., 2022. Catalogue of Life Checklist (Annual Checklist 2022). Catalogue of Life. https://doi.org/10.48580/dfq8.

Bellinger, P.F., Christiansen, K.A., Janssens, F., 1996-2022. Checklist of the Collembola of the World. http://www.collembola.org.
BirdLife International, 2018. State of the World's Birds: Taking the Pulse of the Planet. BirdLife International, Cambridge, UK.
Blick, T., Harvey, M.S., 2011. Worldwide catalogues and species numbers of the arachnid orders (Arachnida). Arachnol. Mitt. 41–43.
Chan, S., Bauer, S., Betsill, M.M., Biermann, F., Boran, I., Bridgewater, P., Bulkeley, H., Bustamante, M.M.C., Deprez, A., Dodds, F., Hoffmann, M., Hornidge, A.K., et al., 2022. The global biodiversity framework needs a robust action agenda. Nat. Ecol. Evol. 7, 172–173.
Chapman, A.D., 2009. Numbers of Living Species in Australia and the World. Australian Biodiversity Information Services, Australia.
Chesshire, P.R., Fischer, E.E., Dowdy, N.J., Griswold, T.L., Hughes, A.C., Orr, M.C., Ascher, J.S., Guzman, L.M., Hung, K.-L.J., Cobb, N.S., McCabe, L.M., 2023. Completeness analysis for over 3000 United States bee species identifies persistent data gap. Ecography 2023, e06584.
Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2021. The challenge of biased evidence in conservation. Conserv. Biol. 35, 249–262.
Critchlow, R., Cunningham, C.A., Crick, H.Q.P., Macgregor, N.A., Morecroft, M.D., Pearce-Higgins, J.W., Oliver, T.H., Carroll, M.J., Beale, C.M., 2022. Multi-taxa spatial conservation planning reveals similar priorities between taxa and improved protected area representation with climate change. Biodivers. Conserv. 31, 683–702.
Cuff, M., 2022. COP15 aims "unrealistic". NewScientist 256, 7.
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.
Di Marco, M., Harwood, T.D., Hoskins, A.J., Ware, C., Hill, S.L., Ferrier, S., 2019. Projecting impacts of global climate and land-use scenarios on plant biodiversity using compositional-turnover modelling. Glob. Chang. Biol. 25, 2763–2778.
Engelbrecht, I., Robertson, M., Stoltz, M., Joubert, J.W., 2016. Reconsidering environmental diversity (ED) as a biodiversity surrogacy strategy. Biol. Conserv. 197, 171–179.
Enghoff, H., Golovatch, S., Short, M., Stoev, P., Wesener, T., 2015. Diplopoda - taxonomic overview. In: Minelli, A. (Ed.), Treatise on Zoology-Anatomy, Taxonomy, Biology. The Myriapoda, vol. 2. Leiden-Boston, Brill, pp. 363–453.
Fattorini, S., 2013. Regional insect inventories require long time, extensive spatial sampling and good will. PLoS One 8, e62118.
Feng, X., Enquist, B.J., Park, D.S., Boyle, B., Breshears, D.D., Gallagher, R.V., Lien, A., Newman, E.A., Burger, J.R., et al., 2022. A review of the heterogeneous landscape of biodiversity databases: opportunities and challenges for a synthesized biodiversity knowledge base. Glob. Ecol. Biogeogr. 31, 1242–1260.
Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. Int. J. Clim. 37, 4302–4315.
Flather, C.H., 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. J. Biogeogr. 23, 155–168.
Freeman, B.G., Pennell, M.W., 2021. The latitudinal taxonomy gradient. Trends Ecol. Evol. 36, 778–786.
Galli, L., Shrubovych, J., Bu, Y., Zinni, M., 2018. Genera of the Protura of the World: diagnosis, distribution, and key. ZooKeys 772, 1–45.
García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., Vaamonde, A., Granado-Lorencio, C., 2013. ModestR: a software tool for managing and analysing species distribution map databases. Ecography 36, 1202–1207.
García-Roselló, E., Guisande, C., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., González Vilas, L., González-Dacosta, J., Vaamonde, A., Granado-Lorencio, C., 2014. Using ModestR to download, import and clean species distribution records. Methods Ecol. Evol. 5, 708–713.
García-Roselló, E., Guisande, C., Manjarrés-Hernández, A., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Gozález-Vilas, L., Vari, R.P., Vaamonde, A., Granado-Lorencio, C., Lobo, J.M., 2015. Can we derive macroecological patterns from primary GBIF data? Glob. Ecol. Biogeogr. 24, 335–347.
García-Roselló, E., González-Dacosta, J., Guisande, C., Lobo, J.M., 2023. GBIF falls shorts of providing a representative picture of the global distribution of insects. Syst. Entomol. https://doi.org/10.1111/syen.12589.
GBIF.org, 2022a. GBIF Occurrence Download. https://doi.org/10.15468/dl.udrufp (accessed 09 September 2022).
GBIF.org, 2022b. GBIF Occurrence Download. https://doi.org/10.15468/dl.cqpa99 (accessed 09 September 2022).
GBIF Secretariat, 2022. GBIF Backbone Taxonomy. Checklist dataset. https://doi.org/10.15468/39omei. accessed via GBIF.org on 2022-09-9.
Grand, J., Cummings, M.P., Rebelo, T.G., Ricketts, T.H., Neel, M.C., 2007. Biased data reduce efficiency and effectiveness of conservation reserve networks. Ecol. Lett. 10, 364–374.
Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models, with Applications in R. Cambridge University Press, Cambridge, UK.
Guisande, C., Lobo, J.M., 2019. Discriminating well surveyed spatial units from exhaustive biodiversity databases. R package version 2.0. http://cran.r-project.org/web/packages/KnowBR.
Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. Proc. Natl. Acad. Sci. U. S. A. 118, e2018093118.
Hopkin, S.P., 1998. Collembola : the most abundant insects on earth. Antenna 22, 117–121.
Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M., Ladle, R.J., 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu. Rev. Ecol. Evol. Syst. 46, 523–549.
Hughes, A.C., Orr, M.C., Ma, K., Costello, M.J., Waller, J., Provoost, P., Yang, Q., Zhu, C., Qiao, H., 2021. Sampling biases shape our view of the natural world. Ecography 44, 1259–1269.
Leadley, P., Gonzalez, A., Obura, D., Krug, C.B., Londoño-Murcia, M.C., Millette, K.L., Radulovici, A., Rankovic, A., Shannon, L.J., Archer, E., et al., 2022. Achieving global biodiversity goals by 2050 requires urgent and integrated actions. One Earth 5, 597–603.
Lobo, J.M., 2008. Database records as a surrogate for sampling effort provide higher species richness estimations. Biodivers. Conserv. 17, 873–881.
Lobo, J.M., Hortal, J., Yela, J.L., Millán, A., Sánchez-Fernández, D., García-Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L., Guisande, C., 2018. KnowBR: an application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. Ecol. Indic. 91, 41–248.
Martin, P., Martinez-Ansemil, E., Pinder, A., Timm, T., Wetzel, M.J., 2007. Global diversity of oligochaetous clitellates (Oligochaeta; Clitellata) in freshwater. In: Balian, E.V., Lévêque, C., Segers, H., Martens, K. (Eds.), Freshwater Animal Diversity Assessment, Developments in Hydrobiology, vol. 198. Springer, Dordrecht.
McRae, L., Deinet, S., Freeman, R., 2017. The diversity-weighted living planet index: controlling for taxonomic bias in a global biodiversity indicator. PLoS One 12, e0169156.
Merow, C., Wilson, A.M., Jetz, W., 2017. Integrating occurrence data and expert maps for improved species range predictions. Glob. Ecol. Biogeogr. 26, 243–258.
Meyer, C., Kreft, H., Guralnick, R., Jetz, W., 2015. Global priorities for an effective information basis of biodiversity distributions. Nature Commun. 6, 8221.
Meyer, C., Weigelt, P., Kreft, H., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19, 992–1006.
Mokany, K., Ferrier, S., Harwood, T.D., Ware, C., Di Marco, M., Grantham, H.S., Venter, O., Hoskins, A.J., Watson, J.E.M., 2020. Reconciling global priorities for conserving biodiversity habitat. Proc. Natl. Acad. Sci. U. S. A. 117, 9906–9911.
Monsarrat, S., Boshoff, A.F., Kerley, G.I.H., 2019. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. Ecography 42, 125–136.
Nelson, G., Ellis, S., 2018. The history and impact of digitization and digital data mobilization on biodiversity research. Philos. Trans. R. Soc. B 374, 20170391.
Pelayo-Villamil, P., Guisande, C., Manjarrés, A.M., Fernanda Jiménez, L., Granado-Lorencio, C., García Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L., Lobo, J.M., 2018. Completeness of national freshwater fish species inventories around the world. Biodivers. Conserv. 27, 3807–3817.
Pilowsky, J.A., Colwell, R.K., Rahbek, C., Fordham, D.A., 2022. Process-explicit models reveal the structure and dynamics of biodiversity patterns. Sci. Adv. 8, eabj2271.
Rees, T., 2022. The interim register of marine and nonmarine genera. Available from, VLIZ. https://www.irmng.org. (Accessed 16 August 2022) (compiler).
Rocha-Ortega, M., Rodriguez, P., Córdoba-Aguilar, A., 2021. Geographical, temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. Ecol. Entomol. 46, 718–728.
Ruete, A., 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. Biodivers. Data J. 3, e5361.
Ruggiero, M.A., Gordon, D.P., Orrell, T.M., Bailly, N., Bourgoin, T., Brusca, R.C., Cvalier-Smith, T., Guiry, M.D., Kirk, P.M., 2015. A higher level classification of all living organisms. PLoS One 10, e0119248.
Sánchez-Fernández, D., Yela, J.L., Acosta, R., Bonada, N., García-Barros, E., Guisande, C., Heine, J., Millán, A., Munguira, M.L., Romo, H., Zamora-Muñoz, C., Lobo, J.M., 2022. Are patterns of sampling effort and completeness of inventories congruent? A test using databases for five insect taxa in the Iberian Peninsula. Insect Conserv. Divers. 15, 406–415.
Sastre, P., Lobo, J.M., 2009. Taxonomist survey biases and the unveiling of biodiversity patterns. Biol. Conserv. 142, 462–467.
Sendra, A., Jiménez-Valverde, A., Selfa, J., Reboleira, A.S.P.S., 2021. Diversity, ecology, distribution and biogeography of Diplura. Insect Conserv. Divers. 14, 415–425.
Stork, N.E., 2018. How many species of insects and other terrestrial arthropods are there on Earth? Annu. Rev. Entomol. 63, 31–45.
Stropp, J., Ladle, R.J., Emilio, T., Lessa, T., Hortal, J., 2022. Taxonomic uncertainty and the challenge of estimating global species richness. J. Biogeogr. 49, 1654–1656.
Tessarolo, G., Ladle, R.J., Lobo, J.M., Rangel, T.F., Hortal, J., 2021. Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models. Ecography 44, 1743–1755.
Title, P.O., Bemmels, J.B., 2018. ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. Ecography 41, 291–307.
Titley, M.A., Snaddon, J.L., Turner, E.C., 2017. Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. PLoS One 12, e0189577.
Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7, 9132.
Ugland, K.I., Gray, J.S., Ellingsen, K.E., 2003. The species-accumulation curve and estimation of species richness. J. Anim. Ecol. 72, 888–897.
Warton, D.I., Hui, F.K., 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92, 3–10.
WoRMS Editorial Board, 2022. World Register of Marine Species. Available from, VLIZ. https://www.marinespecies.org. (Accessed 16 August 2022).
Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A., Culham, A., 2007. How global is the Global Biodiversity Information Facility? PLoS One 2, e1124.
Zarnetske, P.L., Read, Q.D., Record, S., Gaddis, K.D., Pau, S., Hobi, M.L., Malone, S.L., Costanza, J., Dahlin, K.M., Latimer, A.M., Wilson, A.M., Grady, J.M., Ollinger, S.V., Finley, A.O., 2019. Towards connecting biodiversity and geodiversity across scales with satellite remote sensing. Glob. Ecol. Biogeogr. 28, 548–556.