Science of the Total Environment 996 (2025) 180178

Contents lists available at ScienceDirect

Science of the Total Environment

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

Geographic biases undermine environmental representativeness of European biodiversity data

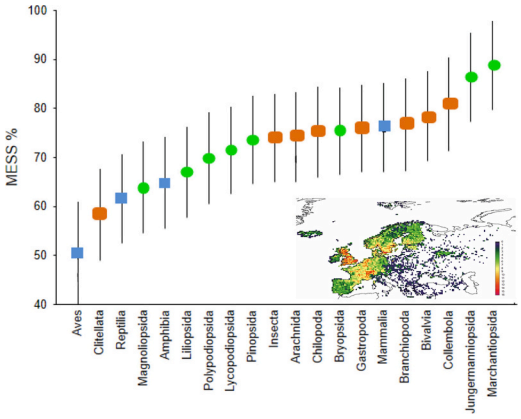
Emilio García-Roselló [a], Jacinto González-Dacosta [a], Jorge M. Lobo [b,*]

[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

HIGHLIGHTS

- Biodiversity data in Europe show strong geographical and taxonomic biases.
- Well-surveyed areas vary widely across taxonomic classes and regions.
- The most abundant and complete data are for vertebrates and vascular plants.
- Incomplete data reduce environmental representativeness and model reliability.
- Strategic, targeted sampling is needed to address biodiversity knowledge gaps.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Biodiversity inventories
Environmental coverage
Completeness of inventories
Geographical biases
MESS

ABSTRACT

Biases and gaps in biodiversity data lead to significant disparities in species descriptions and distribution patterns across taxonomic groups. Although various modelling approaches can help address these gaps, they require the available data to be environmentally representative. In this study, we use data from the Global Biodiversity Information Facility (GBIF) to examine geographical biases, data gaps, and spatial completeness patterns in species distribution records for the 20 main classes of terrestrial organism in Europe—the world's oldest region for taxonomic and natural history research. By identifying spatial units with complete inventories for each class, we assess their quantity, distribution, and capacity to represent the environmental variability of the European subcontinent. Our results reveal high spatial heterogeneity and substantial variation among taxa in the number of well-surveyed units. Vertebrates and vascular plants have several times more well-surveyed cells than invertebrates and mosses. In terms of environmental representativeness, the findings highlight the uncoordinated and opportunistic accumulation of biodiversity data and the urgent need for improved coverage. This situation raises concerns about the reliability of current biodiversity data for accurately characterizing species distributions and limits the effectiveness of species distribution models. Given the scale and urgency of the biodiversity crisis, waiting for complete and reliable data before taking conservation action is not a viable option.

* Corresponding author.
E-mail address: jorge.lobo@mncn.csic.es (J.M. Lobo).

https://doi.org/10.1016/j.scitotenv.2025.180178
Received 15 May 2025; Received in revised form 24 June 2025; Accepted 28 July 2025
Available online 30 July 2025

1. Introduction

Compiled biodiversity data are affected by significant biases and deficiencies, resulting in marked disparities in knowledge across taxonomic groups and regions, particularly concerning the proportion of described species and their known distributional ranges (Meyer et al., 2016; Titley et al., 2017; Hughes et al., 2021 or García-Roselló et al., 2023). When examining databases containing taxonomic and biogeographical information, some locations appear relatively well-surveyed, whereas others lack sufficient data or have none at all (Lomolino, 2004). These gaps and biases represent a major limitation when attempting to design effective conservation strategies based on primary data. The limitations of distributional data can be addressed through various modelling approaches that incorporate spatial, environmental, or biotic predictors to estimate the probable occurrence or abundance of species in under-surveyed regions (Guisan et al., 2017). According to statistical theory, any predictive model should be based on a sample that randomly or equitably captures the key characteristics of the complete dataset of interest (Soley-Guardia et al., 2024). In the context of species distributions, such models require sample datasets that, at a minimum, reflect the diverse environmental characteristics of the target geographic area. Consequently, although gaps in geographic distribution data pose a challenge for species distribution modelling (Lobo et al., 2018), ensuring that the dataset adequately represents the environmental variability of the study area is essential for generating robust predictions (Moudrý et al., 2024).

In this study, we use data sourced from GBIF (https://www.gbif.org/), the leading global platform for biodiversity information (Feng et al., 2022), to assess the geographical biases and gaps in species distribution records, and to examine the congruence of spatial completeness patterns across different taxonomic groups. By identifying spatial units most likely to contain comprehensive inventories for each class of terrestrial organisms, we provide insights into their number, distribution, and capacity to represent the environmental variability of the European subcontinent, currently the world region with the largest volume of biodiversity data and the longest standing tradition of biological research. Ultimately, we aim to assess the adequacy of the GBIF-compiled data for biogeographical and conservation purposes and to highlight regions where further data collection is needed for each taxonomic group.

2. Methods

2.1. Used data

All taxonomic and occurrence data in this study were exclusively sourced from GBIF, a freely accessible database that aggregates a substantial portion of occurrence records from other biodiversity repositories (Feng et al., 2022). The GBIF Backbone Taxonomy (see https://hosted-datasets.gbif.org/datasets/backbone/) was used to establish a comprehensive taxonomy of the kingdoms *Animalia* and *Plantae*, considering only species labelled as "accepted" within this database. The class level was chosen as the taxonomic grouping due to its widespread use and balanced representation of terrestrial taxa (Ruggiero et al., 2015). We selected the same 12 classes of *Animalia* and 8 classes of *Plantae* as used by García-Roselló et al. (2023), excluding classes in which 50 % or more of their species inhabit marine biomes, based on data from the World Register of Marine Species (WoRMS Editorial Board, 2022) and the Interim Register of Marine and Nonmarine Genera (Rees, 2022). Only georeferenced occurrence records labelled as "observations", "human observations", or "preserved specimens" were extracted from GBIF for terrestrial areas of the European subcontinent. For the purpose of this study, Europe is defined as extending east to the Urals and south to Turkey, including Middle East and Iceland, but excluding Greenland, islands located south of 34° latitude (e.g., Canary Islands), and islands west of longitude −24° (e.g., the Azores). The complete datasets were retrieved from GBIF.org (2023a, 2023b) and imported and managed by using ModestR software (https://www.modestr.es), which is specifically designed to handle and map large volumes of biodiversity records (García-Roselló et al., 2013). A basic quality control procedure was applied to the downloaded records, excluding records with: i) identical latitude and longitude values, ii) latitude or longitude equal to 0°, and iii) records from habitats other than terrestrial or freshwater ecosystems (García-Roselló et al., 2014). After filtering, a total of 682 million occurrences where retained, representing 115,170 different species (Table 1).

2.2. Completeness estimations

Accumulation curves, based on the complete set of GBIF records available for the species in each class, were used to estimate completeness values in the absence of reliable abundance or incidence data, which normally serve as the input for non-parametric estimators (Colwell et al., 2012). We considered the number of records as a proxy for survey effort, assuming that each record in the database represents an "occasion" on which a species was observed, following the approach and assumptions outlined by Lobo (2008) and Lobo et al. (2018). As more records (i.e., "occasions") accumulate, the rate at which new species are added is expected to decline, enabling the estimation of an asymptotic accumulation curve. Completeness is then calculated as the proportion of observed species relative to the predicted total number of species, as inferred from the asymptote. For each animal and plant class, accumulation curves were computed for every terrestrial grid cell in Europe, defined at a resolution of 30 arcminutes (approximately 55 × 55 km at the equator; $n = 7274$ cells). The exact estimator proposed by Ugland et al. (2003) was used, which avoids the employment of computationally intensive randomisation procedures to determine sample order. Arctic cells permanently covered by ice, as identified in the ISRIC database (https://data.isric.org), were excluded from the analysis. The resulting accumulation curves were fitted to a rational function as described by Flather (1996), and the extrapolated asymptotic values were used to estimate the probable total number of species per cell, assuming exhaustive sampling effort. A cell was considered well-surveyed if its completeness value reached at least 90 %. This entire process was conducted using the freely available ModestR software (www.modestr.es; García-Roselló et al., 2013 and 2023), which incorporates an optimized version of the R package *KnowBR* (Lobo et al., 2018; Guisande and Lobo, 2019). A detailed explanation of this process in ModestR is available at https://www.modestr.es/sweb/Manual_Tutorial.html.

2.3. Environmental representativeness

The multivariate environmental similarity surface (MESS) metric (see Elith et al., 2010) was used to assess the environmental representativeness of grid cells within the study area. MESS quantifies the environmental similarity of a given site relative to a set of reference spatial units. A maximum MESS value of 100 indicates that a site is the most typical within the environmental range of the reference sites. Lower values, approaching zero, denote increasingly atypical environments. Negative MESS values indicate that at least one environmental variable at a site falls outside the range observed in the reference sites, suggesting that the site represents a novel environment inadequately captured by the reference dataset. Therefore, a reasonable criterion is to consider only sites with non-negative MESS values as having environments adequately represented by the reference sites.

To characterize the environmental conditions of each cell, we used 19 bioclimatic variables and elevation data from WorldClim 2.1 (Fick and Hijmans, 2017), together with 16 bioclimatic and topographic variables from ENVIREM (Title and Bemmels, 2018). These variables were assumed to provide a comprehensive representation of the basic environmental conditions in each cell. All variables were resampled from

their original 5-arcminute resolution to 30 arcminutes to match the resolution of the completeness estimates. MESS values were then calculated for each cell across the study region for each *Animalia* and *Plantae* class. Reference sites were defined as cells with completeness percentages equal to or greater than a given threshold, with thresholds ranging from 50 % to 95 % in 5 % increments. The R package *modEvA* (Barbosa et al., 2013) was used for MESS calculations. The percentage of cells with non-negative MESS values was computed relative to the total number of considered cells for each class and completeness threshold, yielding 200 MESS percentage values (20 classes × 10 completeness thresholds; hereafter referred to as %MESS). To determine whether the number of environmental variables influenced %MESS values, we compared results obtained using all 36 variables with those derived from the scores of six principal components (PCs) with eigenvalues >1, based on a Principal Component Analysis of all environmental variables. The correlation between %MESS values from both approaches was positive and highly significant (Pearson's $r = 0.958$; $p < 0.001$; $n = 20$). Based on these results, we opted to retain all environmental variables in the MESS calculations.

2.4. Data analysis

To identify primary organism groups based on the extent and spatial distribution of their completeness values, a Cluster Analysis was performed on completeness data from all European terrestrial cells (30 arcminutes) across the 20 taxonomic classes. The recommended Ward's method was selected as the linkage rule, with squared Euclidean distances as the dissimilarity measure, as this approach progressively assigns greater weight to classes that are more dissimilar (Legendre and Legendre, 2012).

A General Linear Model was applied to examine the relationship between %MESS values (response variable) and the 20 taxonomic classes (categorical predictor), incorporating the logarithm of the number of occupied cells per class as a covariate. This analysis aimed to assess whether environmental representativeness varied among taxonomic groups after accounting for differences in data availability. If environmental representativeness differed across taxonomic groups independently of data quantity, this would indicate that survey efforts followed a specific environmental pattern in some or all groups. Model adequacy was evaluated by inspecting a normal probability plot of the residuals. All statistical analyses were conducted using StatSoft's STATISTICA v10.0 (StatSoft Inc., Tulsa, Oklahoma, USA).

3. Results

3.1. Completeness patterns

The quantity of data (number of GBIF records in Europe) and the geographical range they represent (number of occupied 30-arcminute cells) differ significantly among taxonomic classes (Table 1). These two metrics are positively correlated ($r = 0.58$; $p < 0.01$), and both show significant correlations with the number of well-surveyed cells ($r = 0.62$ and 0.70, respectively; $p < 0.01$). However, some classes—such as Insecta—despite their high species richness and record counts, have only around 30 well-surveyed cells (~0.4 % of the total). A similar disproportion between data quantity and well-surveyed cells is evident in Arachnida and Gastropoda, albeit to a lesser extent (Table 1). In contrast, Aves—despite moderate species richness—account for 46 % of all records and 40 % of well-surveyed cells (Table 1).

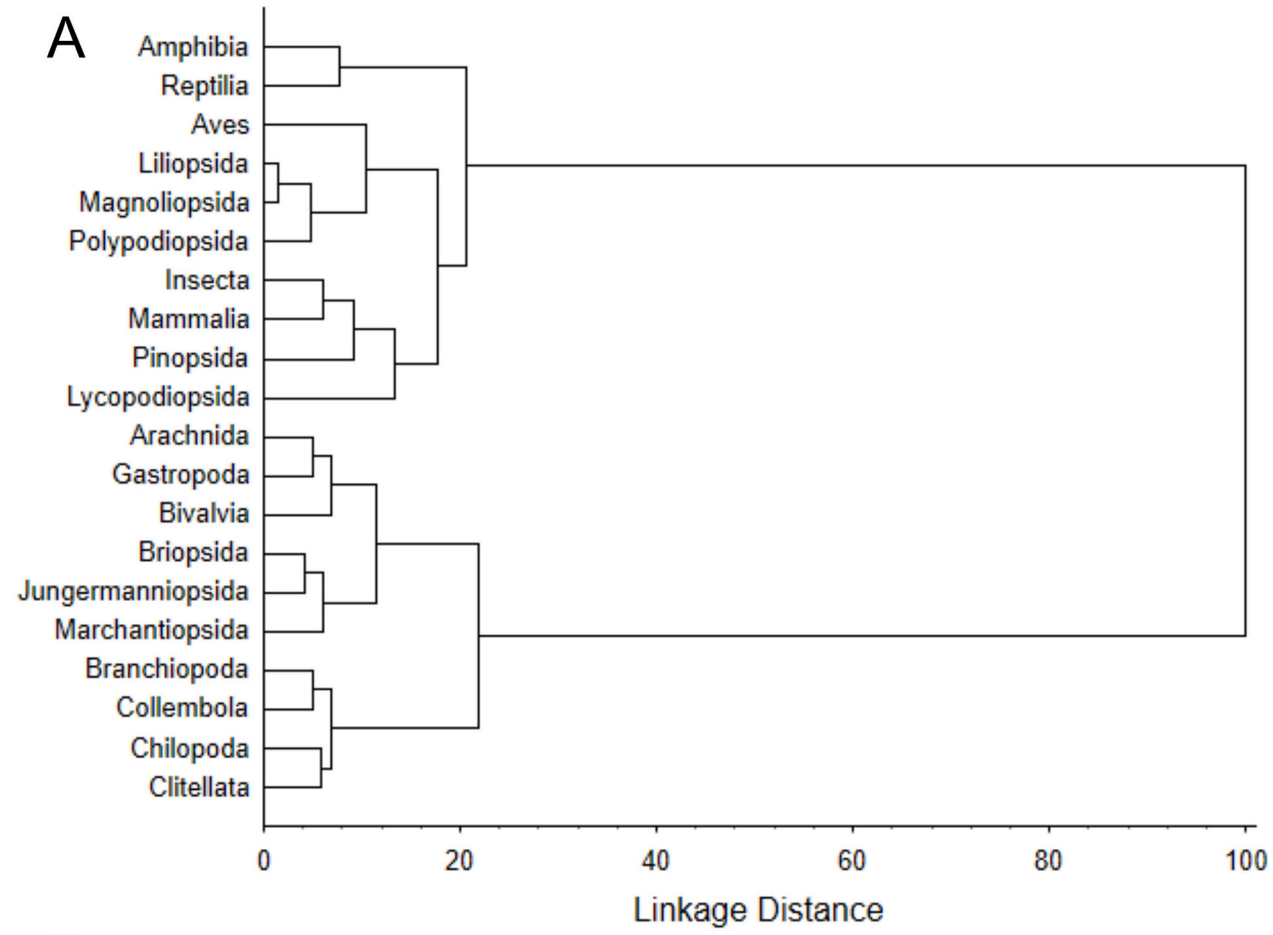
Geographical variation in completeness does not follow a uniform pattern across classes, but instead reveals two distinct groups (Fig. 1A; Fig. S1 in Supplementary Material). One group comprises vascular plants (Lycopodiopsida, Polypodiopsida, Pinopsida, Magnoliopsida and Liliopsida), Insecta, and Chordata, characterized by high data availability (average: 66.7 million records) and a large number of well-surveyed cells (average: 1403). The other group, consisting of mosses, liverworts (Marchantiopsida, Jurgermanniopsida and Bryopsida), and the remaining invertebrate classes, is characterized by lower data availability (average: 1.4 million records) and fewer well-surveyed cells (average: 168). Classes with higher data availability tend to show high completeness values in northern and central Western Europe, or more broadly along the western part of the continent (Fig. S1 in Supplementary Material). However, some classes exhibit highly localised completeness patterns. Liverworts and mosses show high completeness only in the United Kingdom, Benelux, and southern Sweden. Invertebrate classes such as Collembola, Chilopoda and Branchiopoda exhibit scattered completeness patterns, with high values restricted to isolated locations in northern and central Western Europe (Fig. S1 in Supplementary Material). Due to this disparity, no single European 30-arcminute cell can be considered well-surveyed for all 20 taxonomic classes simultaneously (Fig. 1B).

3.2. Environmental representativeness

The geographical distribution of negative MESS values was mapped for

Table 1
Number of GBIF records in Europe (Records) for each of the organism classes considered, total number of species in GBIF with occurrences in European subcontinent (Species), number of 30-arcminute cells occupied by the species of each class (Cells), number of cells with completeness values equal to or higher than 90 % (C90%), and their corresponding MESS value (%MESS).

Kingdom	Phylum	Class	Records	Species	Cells	C90%	%MESS
Animalia	Annelida	Clitellata	175,120	75	1758	145	24.04
Animalia	Arthropoda	Arachnida	3,341,408	6832	4692	67	25.49
Animalia	Arthropoda	Branchiopoda	131,999	189	1472	123	38.89
Animalia	Arthropoda	Chilopoda	98,033	48	2365	91	28.83
Animalia	Arthropoda	Collembola	188,920	909	1627	35	32.53
Animalia	Arthropoda	Insecta	92,958,044	57,646	6038	29	19.14
Animalia	Chordata	Amphibia	2,246,053	184	5258	1573	80.96
Animalia	Chordata	Aves	318,198,812	1663	6740	2908	83.17
Animalia	Chordata	Mammalia	13,320,866	595	5548	1270	80.31
Animalia	Chordata	Reptilia	1,492,802	371	4771	1375	70.17
Animalia	Mollusca	Bivalvia	344,941	875	3154	202	41.02
Animalia	Mollusca	Gastropoda	1,946,575	4301	4365	317	49.05
Plantae	Bryophyta	Bryopsida	6,771,091	1722	4915	274	22.83
Plantae	Marchantiophyta	Jungermanniopsida	1,354,944	489	3480	222	68.20
Plantae	Marchantiophyta	Marchantiopsida	135,452	94	3060	199	59.75
Plantae	Tracheophyta	Liliopsida	52,222,164	5919	6578	1643	68.23
Plantae	Tracheophyta	Lycopodiopsida	501,040	83	3929	1191	73.66
Plantae	Tracheophyta	Magnoliopsida	176,528,550	32,111	6718	1383	63.01
Plantae	Tracheophyta	Pinopsida	4,287,183	310	5058	899	77.15
Plantae	Tracheophyta	Polypodiopsida	5,969,427	754	5555	1755	84.93



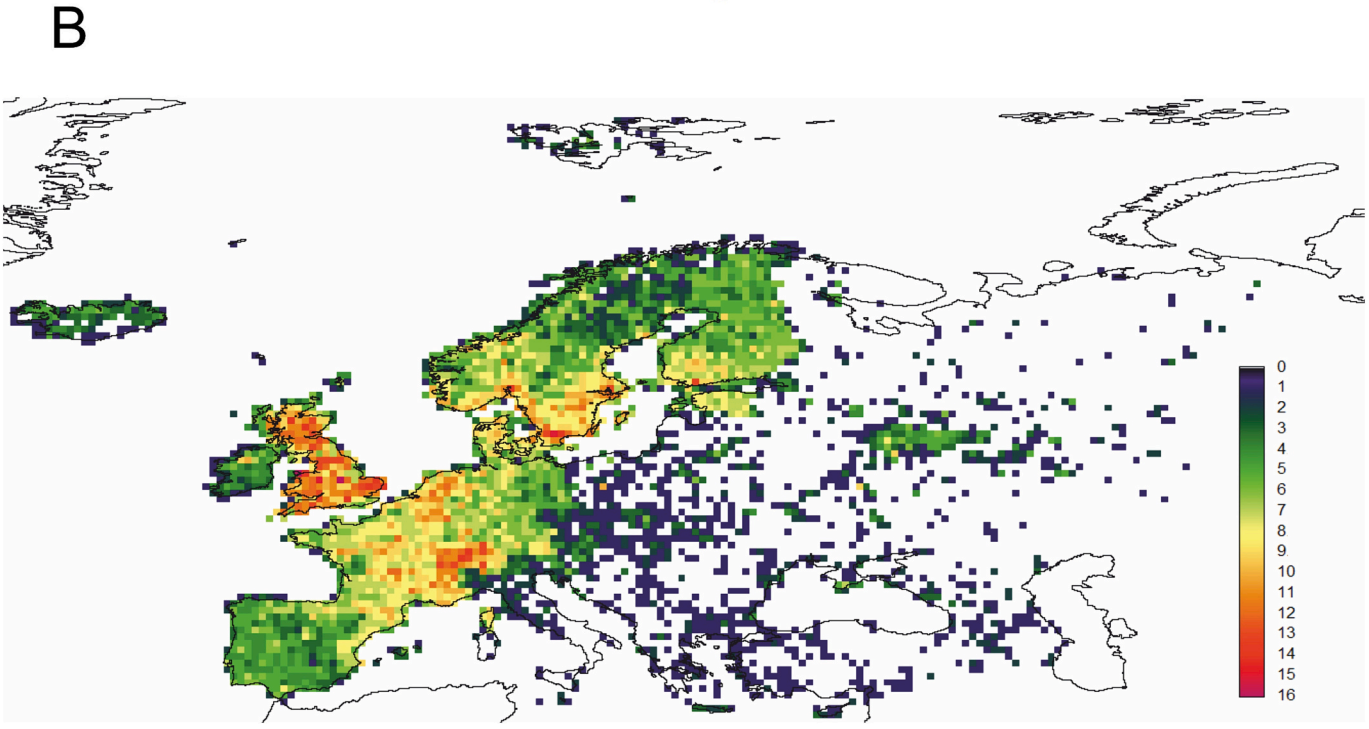


Fig. 1. (A) Results of a Cluster Analysis based on the completeness values of 30-arcminute cells across Europe for the twenty taxonomic classes considered. Squared Euclidean was used as the distance measure, and Ward's method as the linkage criterion. (B) Geographic distribution of the number of classes with completeness values equal to or higher than 90 % in European 30-arcminute cells.

each taxonomic class using reference cells with completeness values ≥90 % (see Fig. S2 in Supplementary Material). In classes with a high proportion of well-surveyed cells, unrepresented environmental conditions are mainly found in northernmost and easternmost Europe. As the number of well-surveyed cells decreases, these unrepresented areas expand into the southern Mediterranean. In classes with fewer database records and fewer well-surveyed cells, large portions of the continent lack environmental representation (Fig. S2 in Supplementary Material).

The percentage of occupied cells decreases linearly as completeness threshold values increase (Fig. 2A). This rate of decline varied across classes, from −76.9 cells per 1 % increase in completeness (Magnoliopsida) to −13.25 cells (Branchiopoda), with a mean rate of −46.44 ± 8.9 cells per 1 % increase (95 % CI; $n = 20$). Similarly, %MESS values declined with increasing completeness, though more heterogeneously manner (Fig. 2B). The rate of decline in %MESS per 1 % increase in completeness ranged from −2.10 in Insecta to −0.34 in Amphibia (mean: −1.02 ± 0.25). While both the number of occupied cells and their environmental coverage decrease as completeness increases, their relationship follows a curvilinear trend (Fig. 2C). This relationship shows a monotonic increase until an inflection point is reached, after which the curve stabilizes and approaches an asymptote. Consequently, environmental representativeness increases rapidly with the number of occupied cells, though always at a lower rate than would be expected from randomly selected cells (Fig. 2C). On average, a 1 % increase in the number of occupied cells corresponds to a 2.6 % increase in environmental representativeness.

The slope of the relationship between %MESS values and the number of occupied cells varies among taxonomic classes. Invertebrate classes such as Collembola, Chilopoda, Branchiopoda, Arachnida and Clitellata exhibit higher increases in %MESS, while vertebrates and vascular plants show lower increases (Fig. 3). These differences appear to be linked to the amount of available data. The logarithm of number of database records and the logarithm of the number of occupied cells per class are strongly and positively correlated ($r = 0.86$; $p < 0.0001$). Furthermore, the rate of increase in %MESS values is negatively correlated with the logarithm of database records ($r = -0.655$; $p = 0.002$) and with the logarithm of the number of occupied cells per class ($r = -0.884$; $p < 0.001$). Thus, classes with fewer records and fewer occupied cells show a greater gain in environmental representativeness per added cell. Even after controlling for the number of occupied cells, %MESS values differ significantly among taxonomic classes ($F_{19,\ 179} = 16.79$; $p < 0.001$; Fig. 3), being lowest in Aves and highest in liverworts (Marchantiopsida and Jugermanniopsida).

4. Discussion

4.1. Biases and deficiencies in completeness

Our results clearly demonstrate substantial heterogeneity in the spatial distribution of completeness values, as well as notable variation among high taxa in the number of spatial units that can be considered well-surveyed. These findings are not new; geographical and taxonomic biases in biodiversity databases have been widely documented across both plant (Meyer et al., 2016) and animal taxa (Titley et al., 2017), initially through pioneering regional studies such as those on British butterflies (Dennis and Thomas, 2000), and later through global databases such as GBIF (Troudet et al., 2017; Hughes et al., 2021; García-Roselló et al., 2023). Nonetheless, empirically confirming these biases remain crucial, especially in regions with long-standing taxonomic and naturalist traditions. The quality of existing biodiversity data—or what could feasibly be compiled in the near future—raises concerns about its capacity to provide a reliable understanding of biodiversity patterns, a prerequisite for effective conservation planning (Rocchini et al., 2023; Hughes et al., 2024). This limitation largely stems from the uncoordinated and opportunistic nature of biodiversity data collection, compounded by the scarcity of taxonomic expertise and effort in many of the

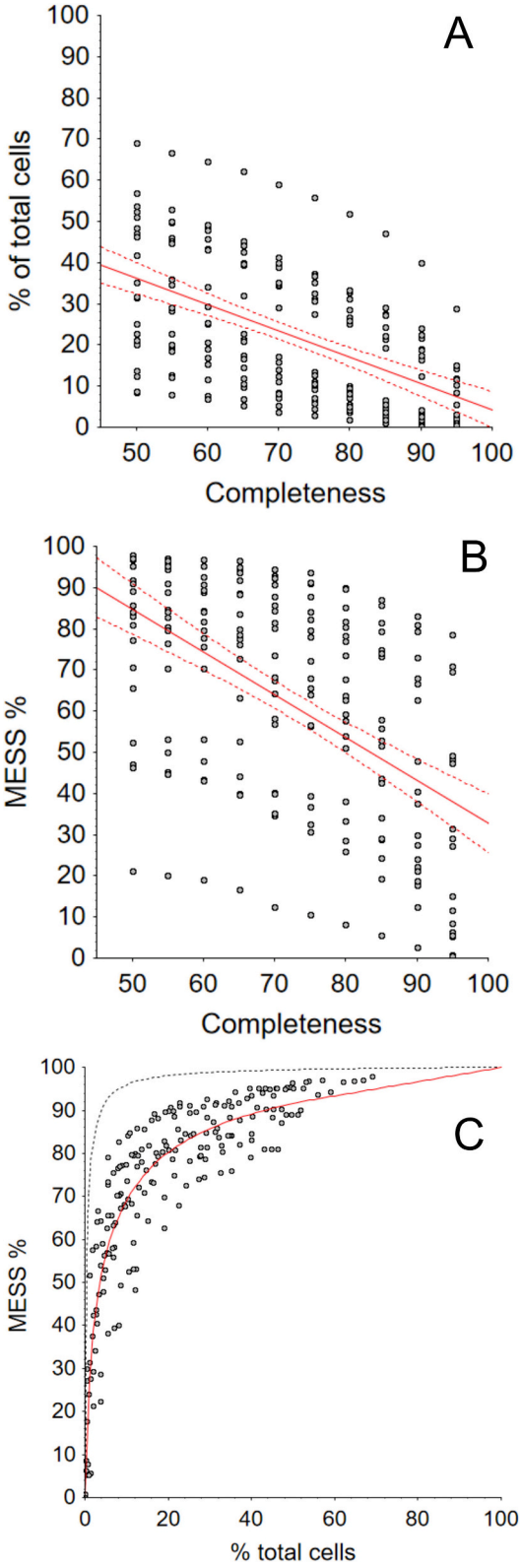


Fig. 2. Variation in the percentage of European 30-arcminute cells (A) and %MESS values (B) according to eight different completeness thresholds for the twenty animal and plant classes considered. The continuous red line represents the linear regression, and the dashed lines indicate the 95 % confidence intervals. The relationship between %MESS and cell percentages across all data (C) is also shown, where the continuous red line represents the fit to a Morgan-Mercer-Flodin growth model ($r = 0.923$), and the black dashed line shows the variation in %MESS with an increasing number of random cells (100 repetitions).

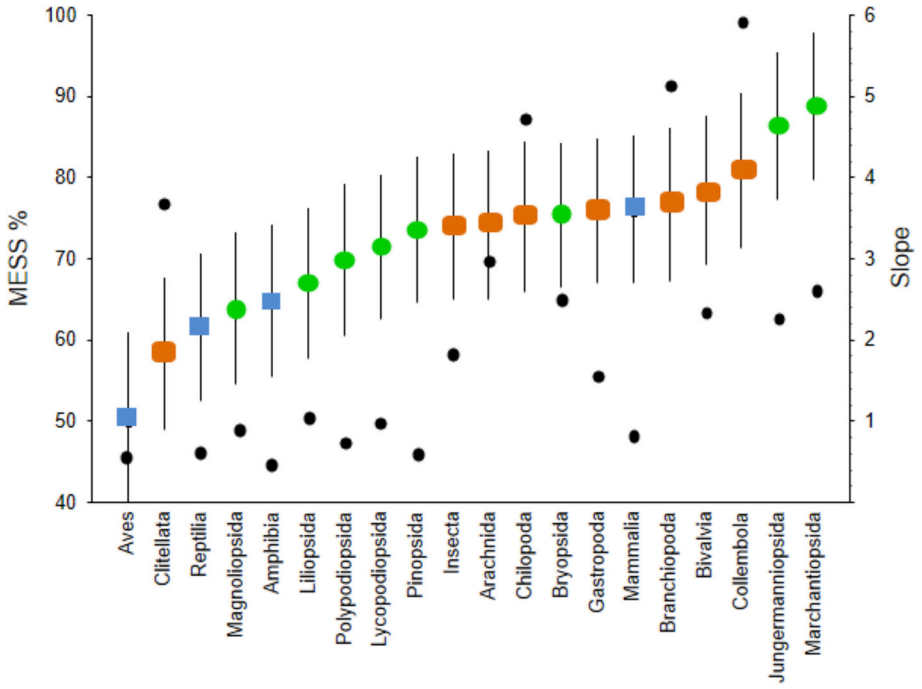


Fig. 3. Slope of the relationship between variation in %MESS and the increase in cell numbers for each of the twenty classes considered (black points). Least-squares adjusted means (in colour) are also shown, with the effect of cell numbers held constant at their mean value. Classes are arranged in ascending order according to their adjusted mean values. Vertical lines indicate 95 % confidence intervals, and colours correspond to vertebrate (blue squares), invertebrate (orange rectangles), and plant classes (green circles).

world's most biodiverse regions (Löbl et al., 2023).

Regarding geographical disparities, our European dataset reveals a clear latitudinal gradient in both the occurrence of well-surveyed cells and their environmental representativeness. This pattern reflects the well-documented increase in biodiversity towards the south (Myers et al., 2000), as well as the concentration of taxonomic resources and research infrastructure in the north (dos Santos et al., 2020). A pronounced longitudinal gradient is also evident, likely influenced in part by GBIF's limited coverage of Eastern Europe (Gaiji et al., 2013).

Taxonomic disparities are equally striking. The proportion of 30-arcminute cells with ≥90 % completeness ranges from 40 % in Aves to just 0.4 % in Insecta (29.3 % and 0.1 %, respectively, at ≥95 %). The 20 studied classes fall into two broad groups, mainly differentiated by data availability. Vertebrates and vascular plants have nearly 50 times more records and 8 times more well-surveyed cells than invertebrates and mosses. This taxonomic bias, in which more diverse groups such as invertebrates are underrepresented, is also spatially structured. Poorly surveyed groups achieve high completeness only in isolated parts of northern and central Western Europe, while better-surveyed groups show broader and more consistent patterns across the western part of the subcontinent.

4.2. Environmental representativeness

This taxonomic and spatial inequality in completeness is closely linked to the ability to generate reliable samples that adequately represent environmental variability. Our findings show that a relatively small number of 30-arcminute cells can capture a substantial proportion of total environmental variability, as indicated by %MESS values. For instance, selecting just 5 % of European cells at random is sufficient to represent around 92 % of total environmental variability. However, restricting this to cells with ≥90 % completeness (approximately 10.8 % of total cells, on average), the mean proportion of environmental variability represented drops to 54 % (range: 17–85 %, depending on taxonomic class). This figure declines further for completeness ≥95 % (5.9 % of total cells), where the mean %MESS value is only 34 %.

Our results also underscore the need to improve the representativeness of well-surveyed spatial units. Random selection of cells emerges as an efficient strategy for achieving environmental representation. The difference in %MESS values between cells with ≥90 % completeness and an equivalent number of randomly selected cells across Europe serves as a benchmark for how far a group's data fall short of optimal environmental coverage. This difference is negatively correlated with the number of occupied cells per class ($r = -0.525$; $p < 0.02$), indicating that broader spatial coverage enhances representativeness. Under-surveyed taxa, such as non-vascular plants and invertebrates, showed substantially lower %MESS values (mean: 35.8 %; range: 17.7 %–84.6 %) than vertebrates and vascular plants (mean: 75.4 %; range: 62.7–86.7 %). Nonetheless, due to the nature of the representativeness growth curve, these under-surveyed groups also exhibit greater potential gains in coverage with the addition of new data. While data volume is the principal factor influencing environmental representativeness, our results suggest that class-specific characteristics—such as spatial distribution and ecological traits—may also play a role.

These findings raise important concerns regarding the application of Species Distribution Models (Guisan et al., 2017) under conditions of data scarcity and low completeness, which can lead to unquantified false absences. Absence data are as critical as presence data in these models, and completeness information can assist in selecting and weighting absences based on their reliability (Lobo, 2016). A further key requirement for such models is that the response variable be adequately distributed across the environmental gradient of the study region. Failure to meet this condition leads to extrapolations beyond the range of environmental conditions used during model calibration (Jiménez-Valverde et al., 2013; Yackulic et al., 2013).

4.3. Limitations and uncertainties

Our results are conditioned by the presence of an unknown—but potentially substantial—proportion of erroneous records in GBIF data,

which are often difficult to detect and correct (Zizka et al., 2020). It could be argued that the presence of such errors might undermine the validity of our findings. However, we believe the core conclusions would remain unchanged even if exhaustive data-cleaning procedures were applied to eliminate basic distributional errors. This is because erroneous records are more likely to inflate artificially the apparent environmental variability represented; thus, the actual level of environmental representativeness would probably be even lower if such errors were fully identified and removed. A related concern is that the data available through GBIF represent only an unknown fraction of all existing primary biodiversity data, and that incorporating additional sources could substantially alter the observed patterns. While this is likely true, the mere existence of data does not equate to its accessibility. For biodiversity information to be useful, it must be findable and openly accessible (Costello et al., 2013; Anderson et al., 2020). Although integrating data from multiple sources could reduce gaps and biases, such integration is still far from complete. Much of the existing information remains inaccessible or "hidden" (Hochkirch et al., 2021). We still lack a comprehensive catalogue of biodiversity databases (Blair et al., 2020), and we do not know the extent of redundancy among them (Feng et al., 2022). At present, we must rely on the data that are available. In this regard, GBIF remains the largest global provider of biodiversity data, aggregating information from numerous sources (Feng et al., 2022).

4.4. Implications for management

The compilation of existing primary biodiversity data and its open accessibility are essential for developing effective conservation strategies. However, the meaningful use of such data for conservation assessments and planning still requires both targeted sampling strategies and improved access to underutilised datasets (Jetz et al., 2012; Tulloch et al., 2013). Our results suggest that future data compilation should prioritise spatial units that remain poorly surveyed but have the potential to enhance the current environmental representativeness of well-surveyed areas. It is evident that northernmost and easternmost regions of Europe—and to a lesser extent, the Euro-Mediterranean region—require additional data collection. However, effective site selection must follow an iterative allocation process (Faith and Walker, 1996; Hortal and Lobo, 2005; Engelbrecht et al., 2016), incorporating additional criteria such as taxon-specific data availability and survey costs. This process should be stepwise and iterative, as the addition of each new spatial unit alters the environmental similarity landscape of the remaining cells. Although site selection must be carefully designed, immediate efforts should focus on systematically collecting essential data for underrepresented taxa and regions. Doing so will help to address existing biases and enable more robust biodiversity monitoring over time. Only through such efforts can we build a representative dataset capable of supporting reliable species distribution modelling and evidence-based biodiversity management. Given the magnitude and urgency of the biodiversity crisis (IPBES, 2019; Glaubrecht, 2023), delaying conservation actions until data are complete and error-free is not a viable option.

5. Conclusions

Our study demonstrates that biases and low completeness in primary biodiversity data are widespread across Europe and affect numerous taxonomic groups, particularly those with high diversity, such as invertebrates. The concentration of taxonomic expertise and research infrastructure in northern European countries, combined with the higher biodiversity levels towards the south, has produced a geographically contrasting pattern: poorly surveyed groups tend to achieve high completeness only in isolated areas of northern and central Western Europe, whereas better-surveyed groups display broader and more consistent patterns across the western part of the subcontinent. In addition to highlighting substantial biases and deficiencies in European biodiversity data, our findings cast doubt on the capacity of currently well-inventoried areas to adequately represent the environmental variability of the entire subcontinent. These results underscore the urgent need to improve the representativeness of biodiversity sampling by prioritising spatial units that are both well-surveyed and inclusive of a wide range of taxonomic groups, particularly the most diverse ones

CRediT authorship contribution statement

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

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We are indebted to the naturalists who, over the span of several decades, have compiled the data studied herein.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2025.180178.

Data availability

The used data are available in the GBIF web by mentioned DOis

References

Anderson, R.P., Araújo, M.B., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T., Soberón, J.M., 2020. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. Front. Biogeogr. 12 (3), e47839. https://doi.org/10.21425/F5FBG47839.

Barbosa, A.M., Real, R., Muñoz, A.R., Brown, J.A., 2013. New measures for assessing model equilibrium and prediction mismatch in species distribution models. Div. Distrib. 19, 1333–1338. https://doi.org/10.1111/ddi.12100.

Blair, J., Gwiazdowski, R., Borreli, A., Hotchkiss, M., Park, C., Perrett, G., Hanner, R., 2020. Towards a catalogue of biodiversity databases: an ontological case study. Biodivers. Data J. 8, e32765. https://doi.org/10.3897/BDJ.8.e32765.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R.L., Longino, J.T., 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J. Plant Ecol. 5, 3–21. https://doi.org/10.1093/jpe/rtr044.

Costello, M., Michener, W., Gahegan, M., Zhang, Z., Bourne, P., 2013. Biodiversity data should be published, cited, and peer reviewed. Trends Ecol. Evol. 28, 454–461. https://doi.org/10.1016/j.tree.2013.05.002.

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. https://doi.org/10.1023/A:1009690919835.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x.

Engelbrecht, I., Robertson, M., Stoltz, M., Joubert, J.W., 2016. Reconsidering environmental diversity (ED) as a biodiversity surrogacy strategy. Biol. Conserv. 197, 171–179. https://doi.org/10.1016/j.biocon.2016.02.028.

Faith, D., Walker, P.A., 1996. Environmental diversity: on the best-possible use of surrogate data for assessing the relative sets of areas. Biodivers. Conserv. 5, 399–415. https://doi.org/10.1007/BF00056387.

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. https://doi.org/10.1111/geb.13497.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.

Flather, C.H., 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. J. Biogeogr. 23, 155–168. https://doi.org/10.1046/j.1365-2699.1996.00980.x.

Gaiji, S., Chavan, V., Ariño, A.H., Otegui, J., Hobern, D., Sood, R., Robles, E., 2013. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. Biodiv. Informatics 8 (94–172). https://doi.org/10.17161/bi.v8i2.4124.

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. https://doi.org/10.1111/j.1600-0587.2013.00374.x.

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. https://doi.org/10.1111/2041-210X.12209.

García-Roselló, E., González-Dacosta, J., Lobo, J.M., 2023. The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently. Biol. Conserv. 283, 110118. https://doi.org/10.1016/j.biocon.2023.110118.

GBIF org, 2023a. GBIF Occurrence Download. https://doi.org/10.15468/dl.7dq5d9. Accessed 09 May 2023.

GBIF org, 2023b. GBIF Occurrence Download. https://doi.org/10.15468/dl.jnwejp. Accessed 09 May 2023.

Glaubrecht, M., 2023. On the end of evolution. Humankind and the annihilation of species. Zool. Scr. 52, 215–225. https://doi.org/10.1111/zsc.12592.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat suitability and distribution models, with applications in R. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/9781139028271.

Guisande, C., Lobo, J.M., 2019. KnowBR. Discriminating well surveyed spatial units from exhaustive biodiversity databases. R Package Version 2.0.. http://cran.r-project.org/web/packages/KnowBR.

Hochkirch, A., Samways, M.J., Gerlach, J., Böhm, M., Williams, P., Cardoso, P., Cumberlidge, N., Stephenson, P.J., Seddon, M.B., Clausnitzer, V., Borges, P.A.V., Mueller, G.M., Pearce-Kelly, P., Raimondo, D.C., Danielczak, A., Dijkstra, K.-D.B., 2021. A strategy for the next decade to address data deficiency in neglected biodiversity. Conserv. Biol. 35, 502–509. https://doi.org/10.1111/cobi.13589.

Hortal, J., Lobo, J.M., 2005. An ED-based protocol for optimal sampling of biodiversity. Biodivers. Conserv. 14, 2913–2947. https://doi.org/10.1007/s10531-004-0224-z.

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. https://doi.org/10.1111/ecog.05926.

Hughes, A.C., Dorey, J.B., Bossert, S., Qiao, H., Orr, M.C., 2024. Big data, big problems? How to circumvent problems in biodiversity mapping and ensure meaningful results. Ecography, e07115. https://doi.org/10.1111/ecog.07115.

IPBES, 2019. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES Secretariat, Bonn, Germany.

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. https://doi.org/10.1016/j.tree.2011.09.007.

Jiménez-Valverde, A., Acevedo, P., Barbosa, A.M., Lobo, J.M., Real, R., 2013. Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. Glob. Ecol. Biogeogr. 22, 508–516. https://doi.org/10.1111/geb.12007.

Legendre, P., Legendre, L., 2012. Numerical Ecology. Elsevier, The Netherlands.

Löbl, I., Klausnitzer, B., Hartmann, M., Krell, F.-T., 2023. The silent extinction of species and taxonomists—an appeal to science policymakers and legislators. Diversity 15, 1053. https://doi.org/10.3390/d15101053.

Lobo, J.M., 2008. Database records as a surrogate for sampling effort provide higher species richness estimations. Biodivers. Conserv. 17, 873–881. https://doi.org/10.1007/s10531-008-9333-4.

Lobo, J.M., 2016. The use of occurrence data to predict the effects of climate change on insects. Curr. Opin. Insect Sci. 17, 62–68. https://doi.org/10.1016/j.cois.2016.07.003.

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. https://doi.org/10.1016/j.ecolind.2018.03.077.

Lomolino, M.V., 2004. Conservation biogeography. In: Lomolino, M.V., Heaney, L.R. (Eds.), Frontiers of Biogeography: new directions in the geography of nature. Sinauer Associates, Sunderland, Massachusetts.

Meyer, C., Weigelt, P., Kreft, H., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19, 992–1006. https://doi.org/10.1111/ele.12624.

Moudrý, V., Bazzichetto, M., Remelgado, R., Devillers, R., Lenoir, J., Mateo, R.G., Lembrechts, J.J., et al., 2024. Optimising occurrence data in species distribution models: sample size, positional uncertainty, and sampling bias matter. Ecography (12), e07294. https://doi.org/10.1111/ecog.07294.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858. https://doi.org/10.1038/35002501.

Rees, T., 2022. The Interim Register of Marine and Non marine Genera. Available from. https://www.irmng.org. at VLIZ. Accessed 2022-8-16.

Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Márcia Barbosa, A., et al., 2023. A quixotic view of spatial bias in modelling the distribution of species and their diversity. npj Biodiv. 2, 10. https://doi.org/10.1038/s44185-023-00014-6.

Ruggiero, M.A., Gordon, D.P., Orrell, T.M., Bailly, N., Bourgoin, T., Brusca, R.C., Cavalier-Smith, T., Guiry, M.D., Kirk, P.M., 2015. A higher level classification of all living organisms. PLoS One 10, e0119248. https://doi.org/10.1371/journal.pone.0119248.

dos Santos, J.W., Correia, R.A., Malhado, A.C.M., Campos-Silva, J.V., Teles, D., Jepson, P., Ladle, R.J., 2020. Drivers of taxonomic bias in conservation research: a global analysis of terrestrial mammals. Anim. Conserv. 23, 679–688. https://doi.org/10.1111/acv.12586.

Soley-Guardia, M., Alvarado-Serrano, D.F., Anderson, R.P., 2024. Top ten hazards to avoid when modeling species distributions: a didactic guide of assumptions, problems, and recommendations. Ecography 2024, e06852. https://doi.org/10.1111/ecog.06852.

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. https://doi.org/10.1111/ecog.02880.

Titley, M.A., Snaddon, J.L., Turner, E.C., 2017. Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. PlosOne 12, e0189577. https://doi.org/10.1371/journal.pone.0189577.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7, 9132. https://doi.org/10.1038/s41598-017-09084-6.

Tulloch, A.I.T., Mustin, K., Possingham, H.P., Szabo, J.K., Wilson, K.A., 2013. To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. Div. Distrib. 19, 465–480. https://doi.org/10.1111/j.1472-4642.2012.00947.x.

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. https://doi.org/10.1046/j.1365-2656.2003.00748.x.

WoRMS Editorial Board, 2022. World Register of Marine Species. Available from. https://www.marinespecies.org. at VLIZ. Accessed 2022-8-16.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H., Veran, S., 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods Ecol. Evol. 4, 236–243. https://doi.org/10.1111/2041-210x.12004.

Zizka, A., Antunes Carvalho, F., Calvente, A., Baez-Lizarazo, R.M., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., Lambert Moreira, F.G., Santos, N.M.C., Santos, T.A.B., dos Santos-Costa, R.C., Serrano, F.C., Alves da Silva, A.P., de Souza Soares, A., Cavalcante de Souza, P.G., Calisto Tomaz, E., Vale, V.F., Vieira, T.L., Antonelli, A., 2020. No one-size-fits-all solution to clean GBIF. PeerJ 8, e9916. https://doi.org/10.7717/peerj.9916.