ADDRESSING MULTIPLE FACETS OF BIAS AND UNCERTAINTY IN CONTINENTAL-SCALE BIODIVERSITY DATABASES

Elisa Marchetto[1,*], Martina Livornese[1,*], Francesco Maria Sabatini[1,2], Enrico Tordoni[3], Daniele Da Re[4],[5] Jonathan Lenoir[6], Riccardo Testolin[1], Giovanni Bacaro[7], Roberto Cazzolla Gatti[1], Alessandro Chiarucci[1], Giles M. Foody[8], Lukáš Gábor[9,10], Quentin Groom[11], Jacopo Iaria[1], Marco Malavasi[12], Vítězslav Moudrý[9], Diletta Santovito[1], Petra Šímová[9], Piero Zannini[1], and Duccio Rocchini[1,9]

[1]Alma Mater Studiorum - University of Bologna, Department of Biological, Geological and Environmental Sciences, via Irnerio 42, 40126 Bologna, Italy
[2]Czech University of Life Sciences Prague, Department of Forest Ecology, Faculty of Forestry and Wood Sciences, Kamýcka 129, 165 21 Prague, Czech Republic
[3]University of Tartu, Institute of Ecology and Earth Science, J. Liivi 2, 50409 Tartu, Estonia
[4]University of Trento, Center Agriculture Food Environment, Via Edmund Mach, 1, 38098 San Michele all'Adige, Italy
[5]Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy
[6]UMR CNRS 7058 Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), Université de Picardie Jules Verne, 1 rue des Louvels, 80037 Amiens, France
[7]University of Trieste, Department of Life Sciences, Via L. Giorgieri 10, 34127 Trieste, Italy
[8]University of Nottingham, School of Geography, University Park, Nottingham NG7 2RD, UK
[9]Czech University of Life Sciences Prague, Faculty of Environmental Sciences, Department of Spatial Sciences, Kamýcka 129, 16500 Praha - Suchdol, Czech Republic
[10]Yale University, New Haven, CT 06520, United States
[11]Meise Botanic Garden, Nieuwelaan 38, 1860 Meise, Belgium
[12]University of Sassari, Department of Chemistry, Physics, Mathematics and Natural Sciences, Via Vienna 2, 07100 Sassari, Italy

**Authors equally contributed to the manuscript*

Corresponding author:
Duccio Rocchini, duccio.rocchini@unibo.it
ORCID ID 0000-0003-0087-0594
BIOME Lab, Department of Biological, Geological and Environmental Sciences (BiGeA), Alma Mater Studiorum University of Bologna, Piazza di Porta S. Donato, 1, 40126 Bologna, Italy

Abstract. The availability of biodiversity databases is expanding at unprecedented rates. Nevertheless, species occurrence data can be intrinsically biased and contain uncertainties that impact the accuracy and reliability of biodiversity estimates. In this study, we developed a reproducible framework to assess three dimensions of bias—taxonomic, spatial, and temporal—as well as temporal uncertainty associated with data collections. We utilized the vegetation plot data located in Europe, from sPlotOpen, an open-access database, as a case study. The metrics proposed for estimating bias include completeness of the species richness for taxonomic bias, Nearest Neighbor Index for spatial bias, and Pielou's index for temporal bias. Additionally, we introduced a new method based on a negative exponential curve to model the temporal decay in biodiversity data, aiming to quantify temporal uncertainty. Finally, we assessed the sampling bias considering the influence of various spatial variables (i.e, road density, human population count, Natura 2000 network and topographic roughness). We discovered that the facets of bias and the temporal uncertainty varied throughout Europe, as did the different roles played by spatial variables in determining biases. sPlotOpen showed a clustered distribution of the vegetation plots, and an uneven distribution in sampling completeness, year of sampling and temporal uncertainty. The facets of bias were significantly explained mainly by the presence of Natura 2000 network and marginally by the human population count. These results suggest that employing an efficient procedure to examine biases and uncertainties in data collections can enhance data quality and provide more reliable biodiversity estimates.

Key words: biodiversity; community composition; data quality; spatial bias; taxonomic bias; temporal bias; temporal uncertainty

Introduction

Biodiversity and ecosystem functioning are experiencing a widespread degradation globally. The main drivers of biodiversity decline are represented by an increase in the intensity of human activities such as land and sea-use, the exploitation of organisms and natural resources, atmospheric and water pollution as well as the introduction of alien species (IPBES 2019). Together with climate change, whose impact on biodiversity is expected to increase in the coming years (Di Marco et al. 2019), these factors pose a significant threat to the integrity of ecosystems and biodiversity. To monitor biodiversity change, we need records that capture the occurrence and/or co-occurrence (i.e. community composition) of species within specific time frames and geographical locations. These raw records, now increasingly available through global biodiversity collections such as the BIEN and sPlot database (Enquist et al. 2016; Bruelheide et al. 2019), play a crucial role in ecological research and represent essential sources of information for guiding and monitoring actions aimed at meeting global biodiversity targets (Boakes et al. 2010; Meyer et al. 2015). Their utility spans over a wide range of applications, including investigations into species redistribution (Jandt et al. 2022b), community reassembly (Bertrand et al. 2011), threat assessment and conservation planning (Ricci et al. 2024), as well as the study of invasive species propagation (Turbelin et al. 2017).

Since the 2000s, the number of publicly available biodiversity databases has risen, alongside their use (Ball-Damerow et al. 2019). Data availability alone, however, is not sufficient to ensure reliable ecological inferences. As a matter of fact, data quality should be considered and checked, both in terms of spatial and temporal representativeness (Wüest et al. 2020). One common issue with biodiversity databases relates to the way in which data are collected. Frequently, these databases contain opportunistic collections of data, which are characterized by uneven sampling effort and might hide subtle sources of bias and uncertainties (Daru and Rodriguez 2023; García-Roselló et al. 2023; Rocchini et al. 2023). When these limitations are not accounted for, our ability to describe and analyse biodiversity might be compromised (Hortal et al. 2015).

Bias and uncertainty are terms developed in the statistical literature, and refer to the theory of sampling (Walther and Moore 2005). Bias occurs when the sampling is unrepresentative of the target statistical population. It might depend on uneven sampling across geographic areas, taxonomic groups or time periods (Walther and Moore 2005). Uncertainty, on the other hand, refers to the lack of precision in measurements, which also affects the degree to which data can represent reality (Hortal et al. 2015). Biodiversity data are particularly prone to these problems, and considerations on the bias and uncertainty of the data acquire particular relevance across three specific dimensions: taxonomic, spatial and temporal (Meyer et al. 2016). While assessments of the limitations posed by the use of biodiversity databases do exist (Monsarrat et al. 2019; Colli-Silva et al. 2020; Ronquillo et al. 2020), most studies focus on one dimension at the time, commonly spatial or taxonomic (but see (Meyer et al. 2016) for a multidimensional approach), and often consider only bias but not their related uncertainty.

Taxonomic bias is a well-known issue in biodiversity research, where the study of specific taxa is favoured over others (Troudet et al. 2017) (e.g. vertebrates over invertebrates and vascular plants over bryophytes and lichens). As a result, biodiversity databases may over- and under-represent different taxonomic groups (García-Roselló et al. 2023). In the geographical space, taxonomic bias can be analysed using measures of inventory or sampling completeness, which estimate taxonomic coverage of the collected data within a given surface area (Chao and Jost 2012). Traditionally, sampling completeness is calculated using parametric or non-parametric estimators of the expected species richness within a given spatial unit and then computing the ratio of observed versus expected species richness (Chesshire et al. 2023). Alternatively, a metric of completeness is given by the final slope of Species Accumulation Curves for the investigated geographic unit (Yang et al. 2013; Girardello et al. 2019). Reliable methods for species richness estimation based on a combination of probabilistic and opportunistic data are now available (Chiarucci et al. 2018) but can hardly be applied only using opportunistically collected data.

Spatial bias arises when data distribution and density are uneven in space, as a result of an unbalanced sampling design (Tessarolo et al. 2014; Rocchini et al. 2023). The spatial distribution of collected data is often the result of socio-economic factors such as accessibility and the presence of road networks (Oliveira et al. 2016), uneven financial investments in research across regions (Meyer et al. 2015), but also the preference for sampling in nature protected areas hosting rare or charismatic species (Yang et al. 2014). The spatial distortion of the data resulting from these factors might yield inaccurate modelling outputs, especially when modelling species distribution (Bazzichetto et al. 2023; Rocchini et al. 2023).

For being aggregated over long time periods, considerations on biodiversity data should take into account the temporal dimension. This aspect is gaining attention as reliably estimating biodiversity loss and change in time stand as a paramount challenge in ecological research (Jandt et al. 2022b). However, surveys are often not conducted systematically over time, leading to collections characterized by uneven data coverage and large temporal gaps where no record is present.

Like bias, uncertainty is present in all the components of biodiversity data and can stem from various sources. For instance, in the taxonomic dimension uncertainty may arise from imprecise or equivocal species names (Stropp et al. 2022), whereas in the geographic space, positional inaccuracy of survey locations is recognized as a contributor to the overall uncertainty in the data (Gábor et al. 2020). While these aspects of taxonomic and spatial uncertainty are routinely considered in macroecological research, the uncertainty derived from the temporal dimension of the data is often neglected. Natural communities are not constant over time and exhibit spatial and/or compositional shifts in response to natural variability and/or human-induced alteration in land use, climate and introduction of alien species (Newbold et al. 2015). Because of the dynamism of ecological systems, the information associated with any data on the occurrence of a certain species or species assemblage in a specific area inevitably decays with time (Tessarolo et al. 2017). Understanding this process of information decay becomes particularly relevant when biodiversity records are used in conservation planning, where accurate and up-to-date knowledge is essential (Boitani et al. 2011).

Given all the above factors, it is important to recognize the different limitations of biodiversity databases and identify new approaches to tackle them. Here, we showcase how different aspects of bias and uncertainty can be quantified. As an example, we used vegetation plot data in Europe from the openaccess database sPlotOpen (Sabatini et al. 2021b). We assessed four specific aspects of error through the use of different metrics: taxonomic bias, spatial bias, temporal bias and temporal uncertainty, so to explore the geographical pattern of these sources of error. Finally, we explored how these sources of error relate to a set of geographic variables, namely human population count and road density, the occurrence of protected areas and topographic roughness. The ultimate goal is to provide a workflow (Fig. 1) that can be generalized and applied to other biodiversity databases, regardless of the spatial scale of the analysis.

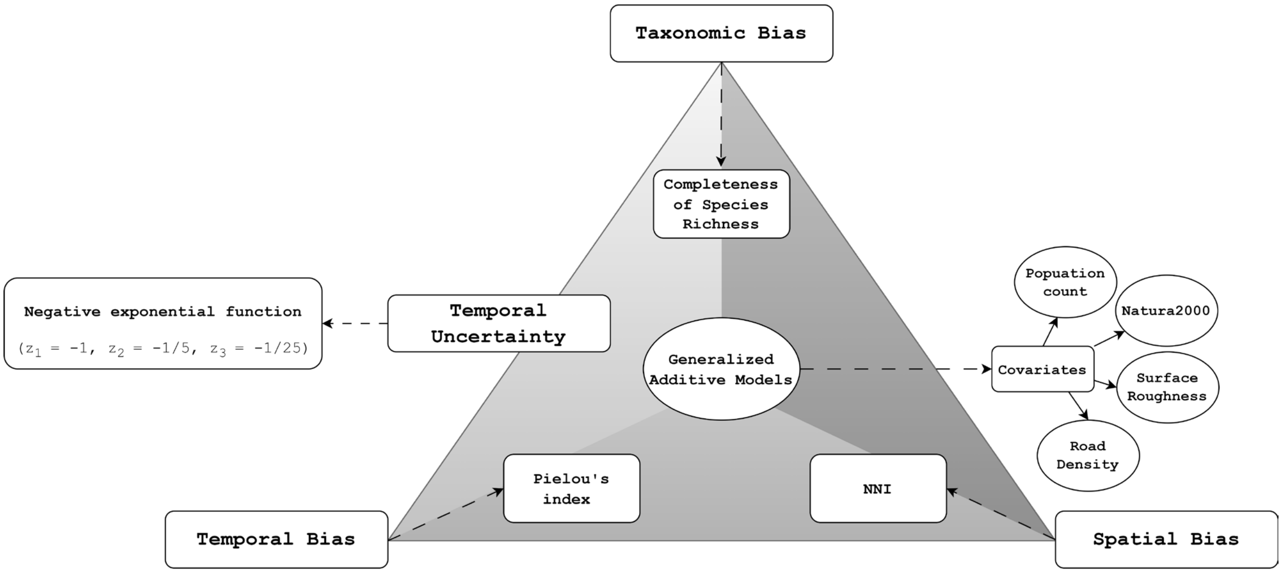

Figure 1: Methodological workflow to assess the presence of taxonomic, spatial and temporal shortfalls in biodiversity databases. The assessment of bias in raw data involves the following measurements: sampling completeness for taxonomic bias, Nearest Neighbour Index for spatial bias, Pielou's index for temporal bias. The temporal uncertainty is calculated using a negative exponential curve. The different facets of bias and the temporal uncertainty are computed for grid cells of 39.5 km. The response of biases to spatial variables is estimated by fitting generalized additive models.

Material and Methods

Data preparation

sPlotOpen is an open-access, stratified subset of the sPlot database. It includes only vascular plant species and was built based on climatic and soil variables as resampling strata (Sabatini et al. 2021b). The stratified resampling used to build sPlotOpen specifically focuses on maximizing the representativeness of the vegetation plot data in the environmental space, at the expense of the geographical space. After accessing sPlotOpen (March 2023 version 2.0, (Sabatini et al. 2021a)), we exclusively extracted data 1) located in Europe and within the boundaries of LAEA Europe coordinates system (WGS84 bounds: -16.1, 32.88, 40.18, 84.73), 2) having coordinates uncertainty lower than 250 m, and 3) with a year of recording equal to or greater than 1992. We did this to minimize errors coming from the inaccurate location of the plots, mainly deriving from possible errors of data georeferencing, and to be consistent with the year of establishment of the Natura 2000 network. This filtering phase reduced the data from 94,951 to 9,481 vegetation plots. We superimposed a grid of 0.5 degree resolution (EPSG:4326) over the European extent and projected it to LAEA Europe coordinates system (ETRS89-extended, EPSG:3035). Accordingly, the resolution of the grid cells was transformed from 0.5 degrees to 39.5 km. Finally, we assigned each vegetation plot to its corresponding grid cell.

Bias

We measured and represented three facets of bias (taxonomic, spatial and temporal) and we plotted them in a trivariate map (Appendix).

Taxonomic bias.—We represented the spatial distribution of the taxonomic bias, according to the taxonomic coverage of the vascular plants in sPlotOpen, in terms of completeness in species richness. Using Chao's formula to estimate the total number of species in a grid cell, we calculated the sample completeness as the ratio of the observed species in a sample to the true species richness (observed plus undetected) in the entire assemblage (Chao et al. 2020). We used the R package *iNEXT* (version 3.0.1) (Hsieh et al. 2016), to determine the species richness for each grid cell of 39.5 km. For each grid cell, the input data comprise the number of sampling units (T) (i.e., vegetation plots), the observed incidence frequencies and the Hill number q = 0. We set *k*, the equally spaced knots (samples sizes), to 5 and we removed to the input data all those grid cells containing two vegetation plots or less. The values of the completeness of species richness were calculated without considering that plots varied in size, within and across grid cells.

Spatial bias.—We estimated the degree of spatial bias (or geographical sampling bias) by estimating the spatial pattern of the plots locations within each grid cell through the Nearest Neighbor Index (NNI) (Clark and Evans 1954). We used the R package spatstat (version 3.0.8) and the package *spatstat.explore* (Revision: 1.21, Date: 2023/10/17). The NNI was computed using the function *clarkevansCalc* (Baddeley et al. 2016) and it evaluates whether the plots exhibit a clustered or random distribution. The NNI is expressed as the ratio of the observed average distance between each plot and its nearest neighbor and the expected average distance in a random distribution with the same number of plots. Values of the index less than one indicate clustering i.e., higher spatial bias, values around one a random distribution i.e., lower spatial bias, whereas values greater than one imply overdispersion (e.g., systematic distribution). We also modified the original *clarkevans.test* function in *clarkevans.test2* to calculate the grid-based NNI with Standardized Effect Size (NNI SES) as the difference between the observed NNI and the mean of NNI simulations divided by the standard deviation of the simulations. We used Monte Carlo approach to generate 999 populations of plots location under the condition of a Complete Spatial Randomness (CSR) of the observed number of plots. Then, for each valid simulation we calculated NNI within the extent of the grid cell.

Temporal bias.—Pielou's index (J) is a metric commonly used in ecology to assess how equitable or even the abundance of species is within a specific community or ecosystem (Pielou 1966). In this work, we used Pielou's evenness to estimate the temporal bias of plot data based on the years of different plots were recorded for each grid cell. We computed the metric using the functions provided by the R package *vegan* (version 2.6.6). Pielou's index is calculated as $J = \frac{H}{H_{max}}$ (1), where H is the Shannon-Wiener index and it is calculated as $H = \sum_{i=1}^{N} p_i \ln p_i$ (2).

Traditionally, *N* represents the total number of species and p_i is their relative abundances for each species $i \in \{1, \ldots, N\}$. The maximum value of Shannon's index is expressed as: *Hmax = lnN*. It is the value that indicates an even distribution, which is attained when all species have equal relative abundances. In our study, *N* refers to the total number of years of recording, where *i* is the ith year of recording, and p_i is the proportion of plots in a grid cell being sampled in year *i*. This means that the Pielou's evenness was calculated by taking into account the number of plots per grid cell, instead of the number of individuals, that share the same year of recording. Higher is the value of Pielou's index lower is the temporal bias.

Temporal uncertainty

The information associated with any biodiversity data decays with time. We modelled the temporal decay of the information by applying a negative exponential transformation to our data. The function is defined as $y_{(t)} = e^{-z(t)}$ (3), where y is the temporal precision, i.e., the remaining information associated with a vegetation plot, and t is the difference between the year of the most recent surveyed plot (i.e., 2014) and the date of recording of the data point. Since there is no way of knowing the actual rate of information decay for a vegetation plot, we calculated our results using three different exponents (i.e., z_1= -1, z_2= -1/5, z_3= -1/25) so that the curves decrease with different rates (according to the slope, Appendix: Supplementary Fig. S1). Therefore, for each plot we calculated three values of temporal precision.

Finally, we quantified the temporal uncertainty of the vegetation plot data in a given grid cell as the median value of 1 - temporal precision of each plot. We chose negative exponential functions, as they have four desirable properties, when compared to other linear transformations. First, negative exponentials are consistent with the assumption that the information associated to a vegetation plot can only decrease (or be stable) with time (i.e., is monotonically decreasing), and that this information will never reach zero. This corresponds to the reasonable assumption that having vegetation plot data for an area, no matter how old the data is, will always provide more information than having no data at all. Second, negative exponentials can be used to constrain the amount of remaining information to a 0-1 interval, which is intuitive and easy to communicate. Third, negative exponentials are simple and versatile functions that can assume a range of shapes, including a linear shape for short time intervals. Finally, negative exponentials have often been used to model the decrease of a quantity against time or space. Radioactive decay is the most typical example, but see Xu et al. (2019) for an application on population decrease over time, or Newling (1969) for the decrease in population as a function of the distance from the city center.

However, we also tested the temporal decay of the plot information (i.e., temporal uncertainty) as a linear function of the median value of the differences between the year of the most recent surveyed plot (i.e., 2014) and the year of recording of the ith plot (see Appendix for further details).

Spatial variables of bias

We selected a number of variables (number of plots, human population count, road density, Natura 2000 network, and topographic roughness), which are likely to be related to the facets of bias (taxonomic, spatial, temporal) in sPlotOpen data. We chose these variables because they have already been tested as sources of bias in several studies (Ballesteros-Mejia et al. 2013; Geldmann et al. 2016; Girardello et al. 2019).

Human population count: The human population count per pixel at 0.0083 degrees of spatial resolution for the year 2014 (year of the most recent plots in the database) was obtained from World Pop[1] (Stevens et al. 2015). We calculated the human population count for each grid cell of 39.5 km as the mean value of the human population counts at the plot locations to be consistent with the method applied to calculate the facets of bias. Accordingly, we extracted the values of the variable at 0.0083 degrees for each plot within the grid cell then, we calculated the mean value.

Road density: Road density was employed as a metric to quantify the level of accessibility at the collection sites; road data shapefile for the European network were obtained from the Global Roads Inventory Project (GRIP)[2] (Meijer et al. 2018) and filtered by retaining only highways, primary and secondary roads. The road density was then calculated with a Kernel Density Estimation (KDE) at 1 km of spatial resolution through the *spatstat* package (Baddeley et al. 2016). Kernel density function is frequently employed to produce a continuous, smooth surface that depicts the spatial density of data points. We obtained the road density at 39.5 km by extracting the values from the original raster layer for each plot location then, we calculated the mean of the values included in each grid cell.

Natura 2000 network: We measured the relative number of plots inside the Natura 2000 network to detect if the locations of the records were biased toward Natura 2000 areas. The polygon layer of the Natura 2000 network was obtained from the European Environment Agency.[3] For each grid cell, we calculated the ratio between the number of plots located inside the Natura 2000 area and the total number of plots present in that grid cell, so as to obtain a grid-based measure of the number of plots inside the protected area which accounts also for the records size.

Topographic roughness: It refers to the variation in elevation and the spatial distribution of landform elements. This variable, which measures the topographic heterogeneity, was taken from (Amatulli et al. 2018). We selected the topographic heterogeneity cause it determines the establishment of different habitats and diverse microenvironments that support different species (Stein et al. 2014; Barajas-Barbosa et al. 2020). Therefore, if the sampling is not appropriately distributed across these different habitats, it can underestimate or lose certain species.

The variable at 39.5 km of spatial resolution was obtained by extracting the values from the original rater layer with a spatial resolution of 0.4 degrees for each plot location then, calculating the mean of the values included in each grid cell.

Finally, we used these variables as predictors in three Generalized Additive Models (GAMs), one for each measure of taxonomic, spatial and temporal bias (i.e., completeness of species richness, NNI, and Pielou's evenness). We used the thin plate splines as spline-based technique for each smooth term of GAM. The variables of GAMs were standardized to zero mean and one standard deviation before rescaling to a 0-1 range. We also considered the spatial autocorrelation including the term s(x,y) to the GAM, where s is a smoothing spline and x and y are the longitude and latitude coordinates of the centroid of the grid cell. To control for the varying number of vegetation plots across grid cells, we added sampling effort as an additional explanatory variable to the models. Sampling effort was calculated as the number of plots within each grid cell.

Results

Bias

The taxonomic bias, described by the completeness of the species richness, was not evenly distributed over Europe (Fig. 2, Appendix: Supplementary Fig. S3), following a similar pattern as the number of vegetation plots recorded per grid cell (Appendix: Supplementary Figs. S4, S9). Besides, the spatial distribution of the plots, measured through the Nearest Neighbor Index (NNI), was clustered almost everywhere in Europe (Fig. 2, Appendix: Supplementary Fig. S6). Most grid cells (97.4%) exhibited a clustered spatial pattern. The values of the NNI

[1] https://hub.worldpop.org/geodata/listing?id=64.

[2] https://www.globio.info/download-grip-dataset.

[3] https://www.eea.europa.eu/data-and-maps/data/natura-13, Published: 6 Oct 2022, Temporal coverage: 2021.

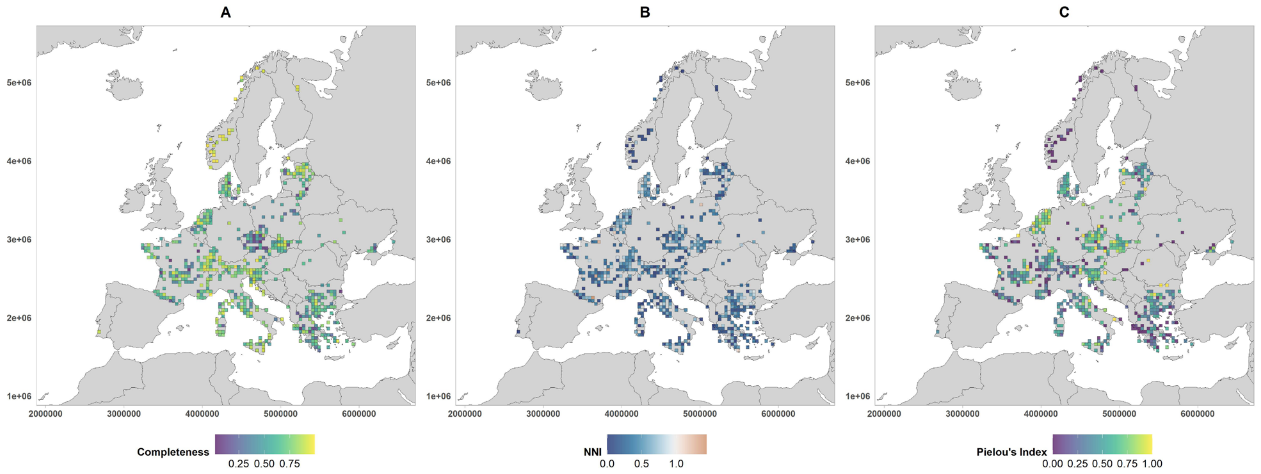

Figure 2: Grid-based map of three facets of bias. The map shows **A** the uneven distribution of the taxonomic bias, **B** the distribution of the vegetation plots through the Nearest Neighbour Index (spatial bias) and, **C** the heterogeneous distribution of the temporal bias. NNI values greater than 1 indicate a random distribution of plots within a grid cell, while values less than 1 indicate a clustered distribution; high completeness of the species richness implies low taxonomic bias; high values of Pielou's index reveals low temporal bias.

confirmed that the effect size was large, pointing out that the magnitude of the deviation from the random expectation was substantial (Appendix: Supplementary Fig. S7).

Furthermore, we observed that the temporal bias, calculated using Pielou's index to estimate the distribution of data across years, followed a different and independent pattern from the taxonomic and spatial bias (Fig. 2, Appendix: Supplementary Fig. S8). However, it highlighted a heterogeneous evenness of plots inventory over time. Indeed, surveys turned out to be evenly distributed (i.e., lower bias) in several countries such as Slovakia, Netherlands and Czech Republic. Overall, the European data in sPlotOpen had high spatial clustering and heterogeneous temporal evenness and completeness of the species richness. Additionally, the prevalence of one type of bias over another varied across geographic areas in Europe, with some countries being characterized by the prevalence of one facet of bias over another (Appendix: Supplementary Fig. S2). The completeness of the species richness (i.e., low taxonomic bias) showed to be preponderant in Norway. An high temporal evenness (i.e., low temporal bias) was observed in some plots in Lithuania and in the Netherlands, while low values of spatial bias were detected in Czech Republic.

Temporal uncertainty

The different negative exponential functions being used for calculating the temporal uncertainty revealed that different exponents (i.e., z_1= -1, z_2= -1/5, z_3= -1/25) allow for discriminating in different ways the pattern and intensity of the hotspots of temporal uncertainty (Fig. 3). The temporal uncertainty measured using the exponent z_1 = -1 was high across the entire European extent, except for some grid cells primarily distributed in Estonia. The temporal uncertainty calculated with z_2 = -1/5 highlighted new areas with lower uncertainty values, namely the Danish peninsula and Bulgaria. Finally, the temporal uncertainty calculated using z_3 = -1/25 smoothed out the values of temporal uncertainty, making uncertainty hotspots less visible compared to the uncertainties based on the other exponents. This exponent most closely approximated the negative exponential curve to a linear trend. Indeed, its pattern of values was comparable with that obtained by calculating the uncertainty as median difference of the year of recording of the plot with the most recent one (Appendix: Supplementary Fig. S11).

Spatial variables of bias

The Generalized Additive Models showed that most facets of bias are related to the presence of Natura 2000 areas. The regression models of taxonomic, spatial and temporal bias had respectively a deviance explained of 49.5%, 14.8% and 22.2% (Table 1).

Only Natura 2000 network and human population count contributed to influencing the three facets

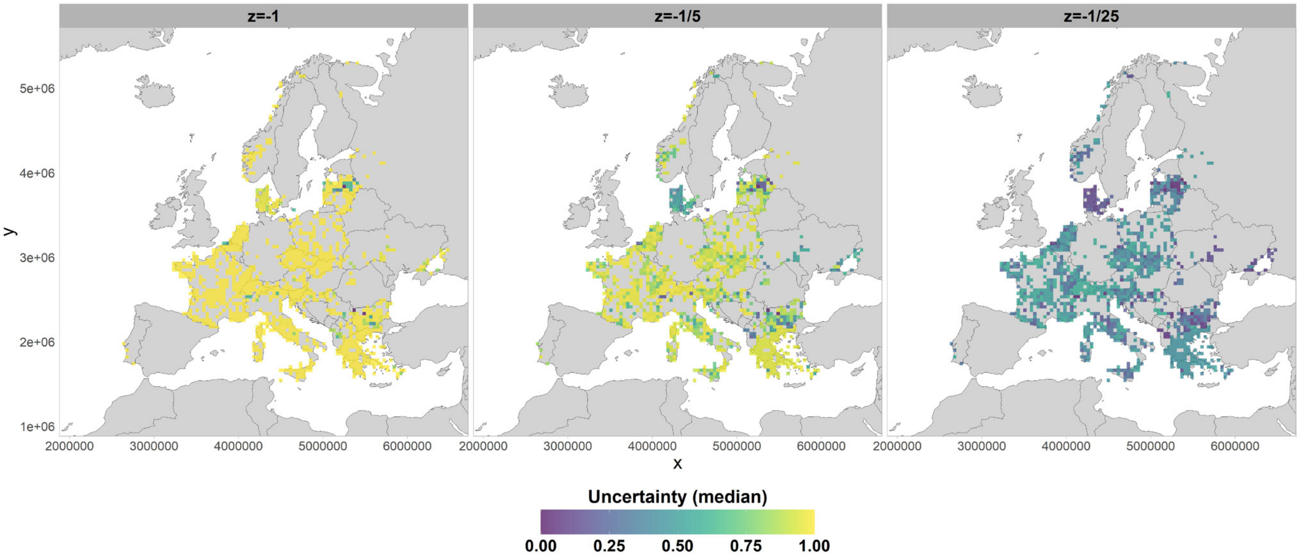

Figure 3: The map shows the median temporal uncertainty of the vegetation plots per grid cell of 39.5 km; the intensity of the temporal uncertainty changes according to the exponents being used setting the exponential negative function (i.e., exponents: z_1= -1, z_2= -1/5 and z_3= -1/25).

Table 1: Terms of quality and fitting process of Generalized Additive Models, as well as, overall significance of explanatory variables. *N2K* refers to relative number of plots in Natura 2000 network, *pop* to human population count, *road* to road density, *rough* to topographic roughness. Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. REML refers to Restricted maximum likelihood, R-squared to coefficient of determination, Deviance expl. to deviance explained.

	Taxonomic bias		Spatial Bias		Temporal Bias	
R-squared	0.461		0.117		0.187	
Deviance expl.	49.5%		14.8%		22.2%	
- REML	-115.35		-89.511		150.18	
	F	p value	F	p value	F	p value
N2K	2.927	**< 0.05 ***	3.807	**< 0.05 ***	4.117	**< 0.01 ****
pop	0.809	0.358	3.547	**< 0.01 ****	1.903	0.118
road	0.081	0.918	1.392	0.239	2.377	0.124
rough	2.088	0.102	0.425	0.687	1.700	0.171

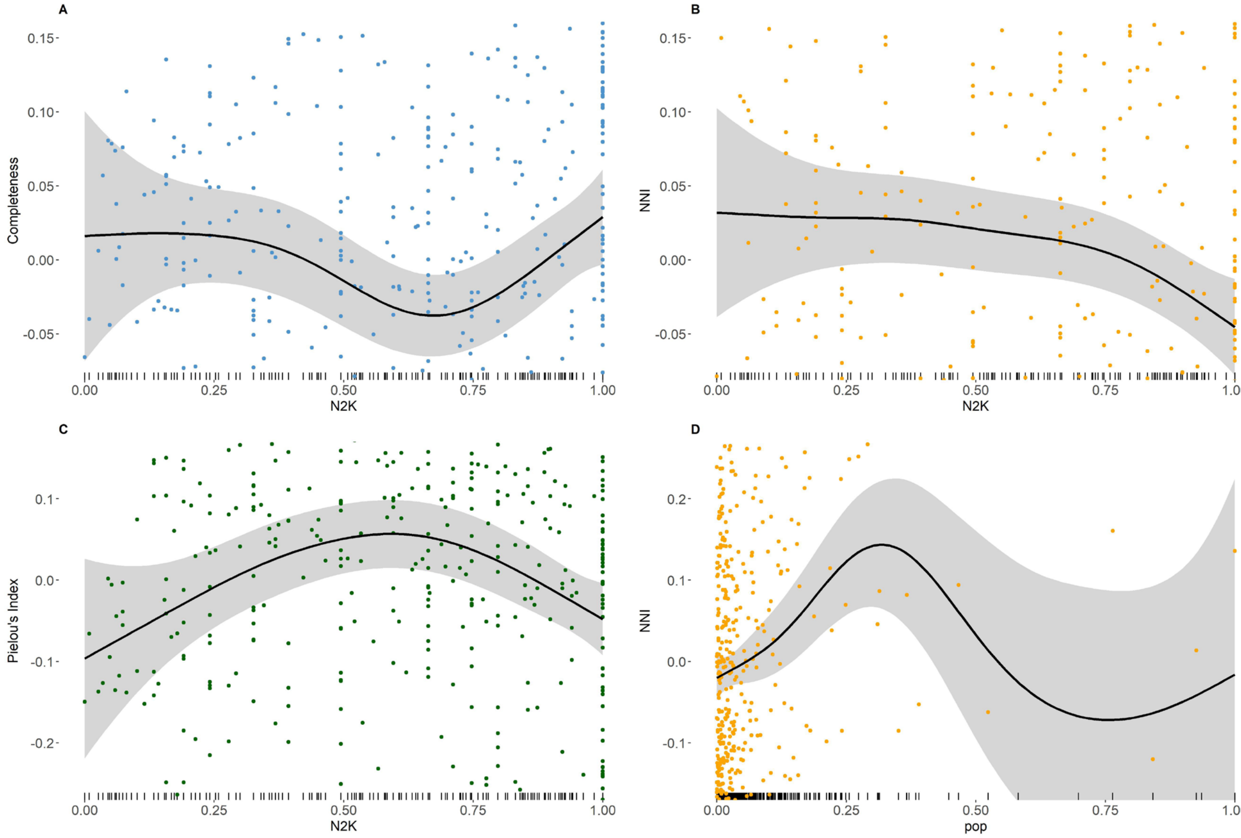

Figure 4: Trends of significative predictors with respect to the response variables of GAMs. *N2K* refers to the relative number of plots inside Natura 2000 network, *pop* refers to the human population count. The plot **A** represents the estimated values of taxonomic bias (i.e., completeness of species richness) at each value of *N2K*, **B** the estimated values of spatial bias (i.e., NNI) at each value of *N2K*, **C** the estimated values of temporal bias (i.e., Pielou's index) at each value of *N2K*, **D** the estimated values of spatial bias at each value of *pop*. The estimated values of the response variable are represented in the y-axis while the observed values of the spatial variable in the x-axis. The "ticks" in the x-axis indicate the distribution of the values. Finally, the line shows the estimated smooth and the point the partial residuals.

of bias (Fig. 4). Specifically, the relative number of plots inside the Natura 2000 network significantly explained the variability in all response variables while human population count was a significant predictor only for spatial bias. Concerning the relative number of plots in Natura 2000 network, lower values were associated with higher completeness of the species richness (lower bias), nevertheless the relationship was not linear (effective degree of freedom (edf) = 3.083); the completeness slightly decreased when the share of plots in Natura 2000 areas increased from about 0.30 to 0.65 and, then increased again. Also the NNI did not follow a complete linear relationship with Natura 2000 protected area (edf = 2.208) showing higher bias (low NNI value) where the share of Natura 2000 areas was higher. Instead, the temporal bias reached its lowest value (highest Pielou's index) when the plots were almost evenly distributed both inside and outside the Natura 2000 network; the degree of non-linearity was low with an edf value of 2.606. Finally, the spatial bias decreased to about 0.30 of the human population count and then increased until it reached almost stability as the covariate increased (edf = 3.830). The control variable sampling effort had a significant effect on the variability of the three biases and the same applied to the term s(x,y) except for the spatial bias. Overall, about 47% of the vegetation plots were inside Natura 2000 protected areas, although this network only accounts for 18% of EU's land area. This showed how vegetation plots were not uniformly distributed inside and outside Natura 2000 areas (Appendix: Supplementary Fig. S10).

Discussion

Biodiversity big data are being increasingly used to understand ecological patterns and monitor biodiversity trends (García-Roselló et al. 2023). Yet, these large collections of opportunistic data come with intrinsic sources of bias, that require careful considerations (Caldwell et al. 2024). Here, we proposed a methodological framework and a set of useful metrics to quantify three different dimensions of bias (taxonomic, spatial, and temporal), as well as the underappreciated dimension of temporal uncertainty in biodiversity data, using vegetation plot data from the open-access database sPlotOpen as an example.

We found that the completeness of the species richness estimates varied across grid cells in Europe, and vegetation plot data varied both in terms of their level of spatial clustering, and their level of temporal unevenenness at the European extent. In addition, the prevalence of one dimension of bias over the others also exhibited a non-uniform distribution, highlighting the presence of several hotspots of bias.

In sPlotOpen, the taxonomic bias varied unevenly across Europe and accordingly to the plot size (Appendix: Supplementary Fig. S5). As expected, we found that the observed species richness is significantly influenced by the sample size (Chao and Jost 2012), with high completeness occurring in grid cells with a high number of plots. However, the sampling completeness still presents some limitations. In particular, the Species Accumulation Curve assumes that there is no spatial and temporal autocorrelation between the species occurrences (Gotelli and Colwell 2001; Yang et al. 2013), and the values of the completeness of the species richness do not represent the degree of sampling of different habitat types (Lobo et al. 2018). Regardless, to address this constraint, a measure of the dark diversity, i.e., the species that are potentially present in a given community but have not yet been detected, can provide a more complete representation of the taxonomic sampling bias by associating it with the value of sampling completeness (Carmona and Pärtel 2021). Despite these limitations, the use of sampling completeness is particularly common. Its use appears in several applications such as for calculating the taxonomic gaps of species records at both multi (La Sorte and Somveille 2020) and single-taxa level (Chesshire et al. 2023), and in assessing the efficacy of a sampling method (Pelayo-Villamil et al. 2018). Here, we provided an example of how sampling completeness can be employed to depict the distribution of the taxonomic information gaps, based on the taxonomic coverage of the vascular plants, at the continental scale.

As far as we know, the use of the Nearest Neighbor Index to assess the spatial bias of raw data is not widespread (e.g., (Geldmann et al. 2016; Oliveira et al. 2016; Hughes et al. 2021; Rocchini et al. 2023)). In sPlotOpen, we observed a high spatial bias, where most of the grid cells had a clustered distribution of plots. Consequently, a high spatial bias in data collection can alter the current representation of community composition and environmental conditions, as well as the potential distribution of a species (Michalcová et al. 2011; Bazzichetto et al. 2023). However, the high clustering we found may depend on the environmental-based resampling of sPlotOpen and possibly on the further filtering we applied to the database which, may have promoted the process of concentration of the plots in a restricted area. Furthermore, the NNI displayed values different from random expectations, suggesting a clustered pattern which, can have been determined by multiple factors, such as the sampling within the network of protected areas.

Although the sampling effort is the most commonly used method to represent the spatial bias of raw data, recent studies (Sumner et al. 2019; Boyd et al. 2021) have proposed the NNI as a suitable index to measure and represent it. In this regard, combining the NNI with the sampling effort can complement our understanding of spatial bias in its possible facets.

Here, we also represented the temporal bias, calculated using Pielou's index. In sPlotOpen, the temporal bias follows a heterogeneous distribution across Europe; high values (i.e., low bias) indicate a more uniform distribution of data across years. However, most of the studies tested the effect of irregular collection over time of raw data in ecological modelling or indices. Examples are the temporal variation of the inventory completeness (Stropp et al. 2016; Ronquillo et al. 2020), the temporal change in species occupancy (Powney et al. 2019; Outhwaite et al. 2020), the temporal coverage of the species records (Meyer et al. 2016; Daru and Rodriguez 2023), or the temporal variation of Species Distribution Models due to biased sampling of species records under land-use change (Bowler et al. 2022). Here, we propose a new method to quantify temporal bias using a common metric employed in ecology, i.e., the Pielou's index, focusing on the distribution of the year of recording of the plots data rather than determining the impact of an uneven sampling over time of the species records.

In our study, we tested three metrics commonly used in ecology to measure the bias of raw data at different dimensions. Nevertheless, many other approaches exist to assess gaps and biases in biodiversity data and one does not exclude the others. Some methods use directly raw data to evaluate the errors, others use predictions or estimations. For instance, Ruete (2015) proposed an ignorance score representing the sampling effort of raw data; Oliver et al. (2021) developed indicators of biodiversity data coverage and sampling effectiveness; Moura and Jetz (2021) analyzed one aspect of taxonomic and geographic knowledge gaps by modelling species discovery probability. Eventually, it is even possible to face biases in raw data by a pre-processing procedure through their standardization and filtering to improve the accuracy of the inferences (Ronquillo et al. 2023).

In this study, we also provide a measure of the temporal uncertainty. To account for the wide uncertainties in the process of temporal decay, we quantified temporal precision using different negative exponential curves. With the method proposed, it is possible to appreciate different patterns of temporal uncertainty based on the exponents used. As lower z-values are used, the rate of decay of information increases. This allows us to identify areas where temporal uncertainty is always low and the information contained is consistently more precise. On the other side, it is possible to notice how areas that appeared to be more precise with higher z-values (e.g., -1/25) become highly uncertain with lower exponents. However, temporal precision is likely to decrease with different rates across different regions and vegetation types, due to many possible drivers of changes, such as anthropogenic pressures, climatic changes, or successional trajectories. This means that using the same function to model information decay across large areas is just an approximation since different contexts might be subjected to different drivers and intensities of change. In future research, it would be interesting to relate the rate of biodiversity information decay to rates of habitat loss and species assemblage turnover (Jandt et al. 2022a,b).

Only a few studies paid attention to the temporal uncertainty of raw data (Meyer et al. 2016; Tessarolo et al. 2021; D’Antraccoli et al. 2022). For instance, when creating a map of ignorance (Rocchini et al. 2011) for species distribution models, Tessarolo et al. (2021) calculated the temporal decay of the information provided by each occurrence record through a kernel Gaussian function that increases the uncertainty for the increment in years since the last recording date. To our knowledge, no study has modelled temporal uncertainty using negative exponential functions. However, future research should investigate how to calibrate the most appropriate set of decay functions to model information loss across regions and vegetation types rather than arbitrarily choosing the exponent.

It is most likely that the biases and uncertainties of the vegetation plots we found in sPlotOpen reflect those of European Vegetation Archive (EVA) (Chytrý et al. 2016); in fact, the integration into EVA database is necessary before European data can be contributed to sPlot. EVA is an archive of multiple databases, and has continued accumulating, compared to the version sPlotOpen was built upon. Although many of the gaps in geographic coverage and representation of specific vegetation types might have been filled in the meantime (Chytrý et al. 2014; Sporbert et al. 2019), it is likely that some aspects of spatial, taxonomic or temporal bias remain. The resulting biases inevitably stem from errors embedded in individual contributing databases as well as challenges related to integrating data from databases with different objectives and adhering to diverse national and regional rules for structuring them.

The relative number of plots inside the Natura 2000 network and the human population count play a role in determining some facets of bias. Ballesteros-Mejia et al. (2013); Girardello et al. (2019) showed how the sampling collection in protected areas increases the completeness of the species richness, as well as, Ricci et al. (2024) demonstrated the effectiveness of Natura 2000 protected area in increasing the species diversity. Furthermore, we found that, as the number of plots inside the Natura 2000 network increases, the distribution of the plots is more clustered (i.e., higher spatial bias). Regarding the temporal evenness of the record collection, we found a non-linear relationship with the number of plots inside the Natura 2000 network, with data collection being more even in time where plots are located both inside and outside the network of protected areas (Fig. 4). In any case, the initial removal of vegetation plots in sPlotOpen to maximize the representation of the environmental space may have altered the current representation of bias dimensions from that of the original sPlot database and their subsequent relationship with the spatial variables we considered. Nevertheless, our outcomes show the strength of the presence of protected areas in shaping the three facets of bias and in influencing the sampling location of the vegetation plots (Boakes et al. 2010). However, the role played by each spatial variable is limited by its release year, which does not reflect the entire temporal period covered by the plots considered in the analysis.

It is crucial to note that in many studies the taxonomic and spatial bias of biodiversity databases correlates with human population density and road density (Ballesteros-Mejia et al. 2013; Geldmann et al. 2016; Mair and Ruete 2016). This was partially observed in our models. In fact, only the spatial bias was significantly influenced by the human population count. This can probably depend on the initial environmental-based resampling of sPlotOpen or by the possible masking effect that the sampling effort had on the other spatial variables in explaining the variability of the models. Eventually, it is likely that human population count and presence of roads are better predictors of the spatial bias in sampling effort across grid cells, rather than predicting the level of clustering within cells (Geldmann et al. 2016; Mair and Ruete 2016; Oliveira et al. 2016).

Different facets of bias and uncertainty can be present in biodiversity databases because of many natural and anthropogenic factors that influence the choice of collecting data in a specific place and at a specific time. Not accounting for these sources of errors in biodiversity data could create knowledge shortfalls and hinder our capacity to monitor real trends in biodiversity and consequently develop effective conservation strategies. It is, therefore, necessary to take into consideration the different facets of bias and uncertainty in biodiversity data by incorporating a routine to check for their presence. Here, we proposed and tested a methodological framework that can be reproduced and applied at different spatial scales (local, ecoregions, biomes, global) and for other databases such as vegetation plots, or simple occurrence data, as those contained in GBIF (GBIF, 2020).

We argue that our framework can be useful for quantifying, making visible, and possibly addressing different sources of bias and uncertainty transparently both when creating a new biodiversity database, and when highlighting priorities for gap-filling in existing ones. For instance, it can be helpful to point out where more actions to fix gaps and sources of errors could be allocated and to provide guidance to data users on how to avoid falling into potential pitfalls and drawing biased inferences.

Declaration of conflict of interest

The authors have declared no competing interests exist.

Data availability statement

The data that support the findings of this study are openly available in Zenodo.[4]

Acknowledgments

F.M.S. gratefully acknowledges financial support from the Italian Ministry of University and Research, within the Rita-Levi Montalcini 2019 program. R.T., R.C.G., A.C., J.I., D.S. and D.R. were supported by the European Union—NextGenerationEU, under the National Recovery and Resilience Plan (NRRP), project title "National Biodiversity Future Center -NBFC" (project code CN_00000033) CUP J33C22001190001. D.R. was also partially funded by the Horizon Europe project B3-Biodiversity Building Blocks for policy (grant agreement 101059592). D.R., V.M., and P.S. were partially funded by the Horizon Europe project EarthBridge (grant agreement 101079310). Views and opinions expressed are, however, those of the authors only, and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Author Contributions

E.M. and M.L. equally contributed at developing the idea and performed the formal analyses. They also wrote the first version of the manuscript and headed the review editing. F.M.S., E.T., D.R. provided substantial input on the conceptualization and the analytical and methodological framework. J.L., D.D.R., R.T. provided methodological revision and gave considerable suggestions on writing – review and editing. G.B., R.C.G., A.C., G.M.F., L.G., Q.G., J.I., M.M., V.M., D.S., P.S., P.Z. contributed to writing – review and editing.

Literature Cited

Amatulli, G., S. Domisch, M.-N. Tuanmu, B. Parmentier, A. Ranipeta, J. Malczyk, and W. Jetz. 2018. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci Data 5:180040.

Baddeley, A., E. Rubak, and R. Turner. 2016. Spatial point patterns: methodology and applications with R. CRC Press, Boca Raton London New York.

Ball-Damerow, J. E., L. Brenskelle, N. Barve, P. S. Soltis, P. Sierwald, R. Bieler, R. LaFrance, A. H. Ariño, and R. P. Guralnick. 2019. Research applications of primary biodiversity

[4] https://zenodo.org/doi/10.5281/zenodo.12179384.

databases in the digital age. PLoS ONE 14:e0215794.

Ballesteros-Mejia, L., I. J. Kitching, W. Jetz, P. Nagel, and J. Beck. 2013. Mapping the biodiversity of tropical insects: species richness and inventory completeness of A frican sphingid moths. Global Ecology and Biogeography 22:586–595.

Barajas-Barbosa, M. P., P. Weigelt, M. K. Borregaard, G. Keppel, and H. Kreft. 2020. Environmental heterogeneity dynamics drive plant diversity on oceanic islands. Journal of Biogeography 47:2248–2260.

Bazzichetto, M., J. Lenoir, D. Da Re, E. Tordoni, D. Rocchini, M. Malavasi, V. Barták, and M. G. Sperandii. 2023. Sampling strategy matters to accurately estimate response curves' parameters in species distribution models. Global Ecol Biogeogr 32:1717–1729.

Bertrand, R., J. Lenoir, C. Piedallu, G. Riofrío-Dillon, P. De Ruffray, C. Vidal, J.-C. Pierrat, and J.-C. Gégout. 2011. Changes in plant community composition lag behind climate warming in lowland forests. Nature 479:517–520.

Boakes, E. H., P. J. K. McGowan, R. A. Fuller, D. Chang-qing, N. E. Clark, K. O'Connor, and G. M. Mace. 2010. Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. PLoS Biol 8:e1000385.

Boitani, L., L. Maiorano, D. Baisero, A. Falcucci, P. Visconti, and C. Rondinini. 2011. What spatial data do we need to develop global mammal conservation strategies? Phil. Trans. R. Soc. B 366:2623–2632.

Bowler, D. E., C. T. Callaghan, N. Bhandari, K. Henle, M. Benjamin Barth, C. Koppitz, R. Klenke, M. Winter, F. Jansen, H. Bruelheide, and A. Bonn. 2022. Temporal trends in the spatial bias of species occurrence records. Ecography 2022:e06219.

Boyd, R. J., G. D. Powney, C. Carvell, and O. L. Pescott. 2021. occAssess: An R package for assessing potential biases in species occurrence data. Ecology and Evolution 11:16177–16187.

Bruelheide, H., J. Dengler, B. Jiménez-Alfaro, O. Purschke, S. M. Hennekens, M. Chytrý, V. D. Pillar, F. Jansen, J. Kattge, B. Sandel, I. Aubin, I. Biurrun, R. Field, S. Haider, U. Jandt, J. Lenoir, R. K. Peet, G. Peyre, F. M. Sabatini, M. Schmidt, F. Schrodt, M. Winter, S. Aćić, E. Agrillo, M. Alvarez, D. Ambarlı, P. Angelini, I. Apostolova, M. A. S. Arfin Khan, E. Arnst, F. Attorre, C. Baraloto, M. Beckmann, C. Berg, Y. Bergeron, E. Bergmeier, A. D. Bjorkman, V. Bondareva, P. Borchardt, Z. Botta-Dukát, B. Boyle, A. Breen, H. Brisse, C. Byun, M. R. Cabido, L. Casella, L. Cayuela, T. Černý, V. Chepinoga, J. Csiky, M. Curran, R. Ćušterevska, Z. Dajić Stevanović, E. De Bie, P. De Ruffray, M. De Sanctis, P. Dimopoulos, S. Dressler, R. Ejrnæs, M. A. E. M. El-Sheikh, B. Enquist, J. Ewald, J. Fagúndez, M. Finckh, X. Font, E. Forey, G. Fotiadis, I. García-Mijangos, A. L. De Gasper, V. Golub, A. G. Gutierrez, M. Z. Hatim, T. He, P. Higuchi, D. Holubová, N. Hölzel, J. Homeier, A. Indreica, D. Işık Gürsoy, S. Jansen, J. Janssen, B. Jedrzejek, M. Jiroušek, N. Jürgens, Z. Kącki, A. Kavgacı, E. Kearsley, M. Kessler, I. Knollová, V. Kolomiychuk, A. Korolyuk, M. Kozhevnikova, Ł. Kozub, D. Krstonošić, H. Kühl, I. Kühn, A. Kuzemko, F. Küzmič, F. Landucci, M. T. Lee, A. Levesley, C. Li, H. Liu, G. Lopez-Gonzalez, T. Lysenko, A. Macanović, P. Mahdavi, P. Manning, C. Marcenò, V. Martynenko, M. Mencuccini, V. Minden, J. E. Moeslund, M. Moretti, J. V. Müller, J. Munzinger, Ü. Niinemets, M. Nobis, J. Noroozi, A. Nowak, V. Onyshchenko, G. E. Overbeck, W. A. Ozinga, A. Pauchard, H. Pedashenko, J. Peñuelas, A. Pérez-Haase, T. Peterka, P. Petřík, O. L. Phillips, V. Prokhorov, V. Rašomavičius, R. Revermann, J. Rodwell, E. Ruprecht, S. Rūsiņa, C. Samimi, J. H. J. Schaminée, U. Schmiedel, J. Šibík, U. Šilc, Ž. Škvorc, A. Smyth, T. Sop, D. Sopotlieva, B. Sparrow, Z. Stančić, J. Svenning, G. Swacha, Z. Tang, I. Tsiripidis, P. D. Turtureanu, E. Uğurlu, D. Uogintas, M. Valachovič, K. A. Vanselow, Y. Vashenyak, K. Vassilev, E. Vélez-Martin, R. Venanzoni, A. C. Vibrans, C. Violle, R. Virtanen, H. Von Wehrden, V. Wagner, D. A. Walker, D. Wana, E. Weiher, K. Wesche, T. Whitfeld, W. Willner, S. Wiser, T. Wohlgemuth, S. Yamalov, G. Zizka, and A. Zverev. 2019. sPlot – A new tool for global vegetation analyses. J Vegetation Science 30:161–186.

Caldwell, I. R., J.-P. A. Hobbs, B. W. Bowen, P. F. Cowman, J. D. DiBattista, J. L. Whitney, P. A. Ahti, R. Belderok, S. Canfield, R. R. Coleman, M. Iacchei, E. C. Johnston, I. Knapp, E. M. Nalley, T. M. Staeudle, and Á. J. Láruson. 2024. Global trends and biases in biodiversity conservation research. Cell Reports Sustainability 1:100082.

Carmona, C. P., and M. Pärtel. 2021. Estimating probabilistic site-specific species pools and dark diversity from co-occurrence data. Global Ecol. Biogeogr. 30:316–326.

Chao, A., and L. Jost. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93:2533–2547.

Chao, A., Y. Kubota, D. Zelený, C. Chiu, C. Li, B. Kusumoto, M. Yasuhara, S. Thorn, C. Wei, M. J. Costello, and R. K. Colwell. 2020. Quantifying sample completeness and comparing diversities among assemblages. Ecological Research 35:292–314.

Chesshire, P. R., E. E. Fischer, N. J. Dowdy, T. L. Griswold, A. C. Hughes, M. C. Orr, J. S. Ascher, L. M. Guzman, K. J. Hung, N. S. Cobb, and L. M. McCabe. 2023. Completeness analysis for over 3000 United States bee species identifies persistent data gap. Ecography 2023:e06584.

Chiarucci, A., R. M. Di Biase, L. Fattorini, M. Marcheselli, and C. Pisani. 2018. Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness. Ann. Appl. Stat. 12.

Chytrý, M., S. M. Hennekens, B. Jiménez-Alfaro, I. Knollová, J. Dengler, F. Jansen, F. Landucci, J. H. J. Schaminée, S. Aćić, E. Agrillo, D. Ambarlı, P. Angelini, I. Apostolova, F. Attorre, C. Berg, E. Bergmeier, I. Biurrun, Z. Botta-Dukát, H. Brisse, J. A. Campos, L. Carlón, A. Čarni, L. Casella, J. Csiky, R. Ćušterevska, Z. Dajić Stevanović, J. Danihelka, E. De Bie, P. De Ruffray, M. De Sanctis, W. B. Dickoré, P. Dimopoulos, D. Dubyna, T. Dziuba, R. Ejrnæs, N. Ermakov, J. Ewald, G. Fanelli, F. Fernández-González, Ú.

FitzPatrick, X. Font, I. García-Mijangos, R. G. Gavilán, V. Golub, R. Guarino, R. Haveman, A. Indreica, D. Işık Gürsoy, U. Jandt, J. A. M. Janssen, M. Jiroušek, Z. Kącki, A. Kavgacı, M. Kleikamp, V. Kolomiychuk, M. Krstivojević Ćuk, D. Krstonošić, A. Kuzemko, J. Lenoir, T. Lysenko, C. Marcenò, V. Martynenko, D. Michalcová, J. E. Moeslund, V. Onyshchenko, H. Pedashenko, A. Pérez-Haase, T. Peterka, V. Prokhorov, V. Rašomavičius, M. P. Rodríguez-Rojo, J. S. Rodwell, T. Rogova, E. Ruprecht, S. Rūsiņa, G. Seidler, J. Šibík, U. Šilc, Ž. Škvorc, D. Sopotlieva, Z. Stančić, J. Svenning, G. Swacha, I. Tsiripidis, P. D. Turtureanu, E. Uğurlu, D. Uogintas, M. Valachovič, Y. Vashenyak, K. Vassilev, R. Venanzoni, R. Virtanen, L. Weekes, W. Willner, T. Wohlgemuth, and S. Yamalov. 2016. European Vegetation Archive (EVA): an integrated database of European vegetation plots. Applied Vegetation Science 19:173–180.

Chytrý, M., L. Tichý, S. M. Hennekens, and J. H. J. Schaminée. 2014. Assessing vegetation change using vegetation-plot databases: a risky business. Applied Vegetation Science 17:32–41.

Clark, P. J., and F. C. Evans. 1954. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. Ecology 35:445–453.

Colli-Silva, M., M. Reginato, A. Cabral, R. C. Forzza, J. R. Pirani, and T. N. D. C. Vasconcelos. 2020. Evaluating shortfalls and spatial accuracy of biodiversity documentation in the Atlantic Forest, the most diverse and threatened Brazilian phytogeographic domain. TAXON 69:567–577.

D'Antraccoli, M., G. Bedini, and L. Peruzzi. 2022. Maps of relative floristic ignorance and virtual floristic lists: An R package to incorporate uncertainty in mapping and analysing biodiversity data. Ecological Informatics 67:101512.

Daru, B. H., and J. Rodriguez. 2023. Mass production of unvouchered records fails to represent global biodiversity patterns. Nat Ecol Evol 7:816–831.

Di Marco, M., T. D. Harwood, A. J. Hoskins, C. Ware, S. L. L. Hill, and S. Ferrier. 2019. Projecting impacts of global climate and land-use scenarios on plant biodiversity using compositional-turnover modelling. Global Change Biology 25:2763–2778.

Enquist, B. J., R. Condit, R. K. Peet, M. Schildhauer, and B. M. Thiers. 2016. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. PeerJ Preprints.

Gábor, L., V. Moudrý, V. Lecours, M. Malavasi, V. Barták, M. Fogl, P. Šímová, D. Rocchini, and T. Václavík. 2020. The effect of positional error on fine scale species distribution models increases for specialist species. Ecography 43:256–269.

García-Roselló, E., J. González-Dacosta, and J. M. Lobo. 2023. The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently. Biological Conservation 283:110118.

GBIF: The Global Biodiversity Information Facility (year) What is GBIF?. Available from https://www.gbif.org/what-is-gbif [13 January 2020]

Geldmann, J., J. Heilmann-Clausen, T. E. Holm, I. Levinsky, B. Markussen, K. Olsen, C. Rahbek, and A. P. Tøttrup. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Diversity and Distributions 22:1139–1149.

Girardello, M., A. Chapman, R. Dennis, L. Kaila, P. A. V. Borges, and A. Santangeli. 2019. Gaps in butterfly inventory data: A global analysis. Biological Conservation 236:289–295.

Gotelli, N. J., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4:379–391.

Hortal, J., F. De Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle. 2015. Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. Annu. Rev. Ecol. Evol. Syst. 46:523–549.

Hsieh, T. C., K. H. Ma, and A. Chao. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (H ill numbers). Methods Ecol Evol 7:1451–1456.

Hughes, A. C., M. C. Orr, K. Ma, M. J. Costello, J. Waller, P. Provoost, Q. Yang, C. Zhu, and H. Qiao. 2021. Sampling biases shape our view of the natural world. Ecography 44:1259–1269.

IPBES. 2019. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. [object Object].

Jandt, U., H. Bruelheide, C. Berg, M. Bernhardt-Römermann, V. Blüml, F. Bode, J. Dengler, M. Diekmann, H. Dierschke, I. Doerfler, U. Döring, S. Dullinger, W. Härdtle, S. Haider, T. Heinken, P. Horchler, F. Jansen, T. Kudernatsch, G. Kuhn, M. Lindner, S. Matesanz, K. Metze, S. Meyer, F. Müller, N. Müller, T. Naaf, C. Peppler-Lisbach, P. Poschlod, C. Roscher, G. Rosenthal, S. B. Rumpf, W. Schmidt, J. Schrautzer, A. Schwabe, P. Schwartze, T. Sperle, N. Stanik, H.-G. Stroh, C. Storm, W. Voigt, A. Von Heßberg, G. Von Oheimb, E.-R. Wagner, U. Wegener, K. Wesche, B. Wittig, and M. Wulf. 2022a. ReSurveyGermany: Vegetation-plot time-series over the past hundred years in Germany. Sci Data 9:631.

Jandt, U., H. Bruelheide, F. Jansen, A. Bonn, V. Grescho, R. A. Klenke, F. M. Sabatini, M. Bernhardt-Römermann, V. Blüml, J. Dengler, M. Diekmann, I. Doerfler, U. Döring, S. Dullinger, S. Haider, T. Heinken, P. Horchler, G. Kuhn, M. Lindner, K. Metze, N. Müller, T. Naaf, C. Peppler-Lisbach, P. Poschlod, C. Roscher, G. Rosenthal, S. B. Rumpf, W. Schmidt, J. Schrautzer, A. Schwabe, P. Schwartze, T. Sperle, N. Stanik, C. Storm, W. Voigt, U. Wegener, K. Wesche, B. Wittig, and M. Wulf. 2022b. More losses than gains during one century of plant biodiversity change in Germany. Nature 611:512–518.

La Sorte, F. A., and M. Somveille. 2020. Survey completeness of a global citizen-science database of bird occurrence. Ecography 43:34–43.

Lobo, J. M., J. Hortal, J. L. Yela, A. Millán, D. Sánchez-Fernández, E. García-Roselló, J. González-Dacosta, J. Heine, L. González-Vilas, and C. Guisande. 2018. KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. Ecological Indicators 91:241–248.

Mair, L., and A. Ruete. 2016. Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa. PLoS ONE 11:e0147796.

Meijer, J. R., M. A. J. Huijbregts, K. C. G. J. Schotten, and A. M. Schipper. 2018. Global patterns of current and future road infrastructure. Environ. Res. Lett. 13:064006.

Meyer, C., H. Kreft, R. Guralnick, and W. Jetz. 2015. Global priorities for an effective information basis of biodiversity distributions. Nat Commun 6:8221.

Meyer, C., P. Weigelt, and H. Kreft. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology Letters 19:992–1006.

Michalcová, D., S. Lvončík, M. Chytrý, and O. Hájek. 2011. Bias in vegetation databases? A comparison of stratified-random and preferential sampling: Stratified-random and preferential sampling. Journal of Vegetation Science 22:281–291.

Monsarrat, S., A. F. Boshoff, and G. I. H. Kerley. 2019. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. Ecography 42:125–136.

Moura, M. R., and W. Jetz. 2021. Shortfalls and opportunities in terrestrial vertebrate species discovery. Nat Ecol Evol 5:631–639.

Newbold, T., L. N. Hudson, S. L. L. Hill, S. Contu, I. Lysenko, R. A. Senior, L. Börger, D. J. Bennett, A. Choimes, B. Collen, J. Day, A. De Palma, S. Díaz, S. Echeverria-Londoño, M. J. Edgar, A. Feldman, M. Garon, M. L. K. Harrison, T. Alhusseini, D. J. Ingram, Y. Itescu, J. Kattge, V. Kemp, L. Kirkpatrick, M. Kleyer, D. L. P. Correia, C. D. Martin, S. Meiri, M. Novosolov, Y. Pan, H. R. P. Phillips, D. W. Purves, A. Robinson, J. Simpson, S. L. Tuck, E. Weiher, H. J. White, R. M. Ewers, G. M. Mace, J. P. W. Scharlemann, and A. Purvis. 2015. Global effects of land use on local terrestrial biodiversity. Nature 520:45–50.

Newling, B. E. 1969. The Spatial Variation of Urban Population Densities. Geographical Review 59:242.

Oliveira, U., A. P. Paglia, A. D. Brescovit, C. J. B. De Carvalho, D. P. Silva, D. T. Rezende, F. S. F. Leite, J. A. N. Batista, J. P. P. P. Barbosa, J. R. Stehmann, J. S. Ascher, M. F. De Vasconcelos, P. De Marco, P. Löwenberg-Neto, P. G. Dias, V. G. Ferro, and A. J. Santos. 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of B razilian terrestrial biodiversity. Diversity and Distributions 22:1232–1244.

Oliver, R. Y., C. Meyer, A. Ranipeta, K. Winner, and W. Jetz. 2021. Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. PLoS Biol 19:e3001336.

Outhwaite, C. L., R. D. Gregory, R. E. Chandler, B. Collen, and N. J. B. Isaac. 2020. Complex long-term biodiversity change among invertebrates, bryophytes and lichens. Nat Ecol Evol 4:384–392.

Pelayo-Villamil, P., C. Guisande, A. Manjarrés-Hernández, L. F. Jiménez, C. Granado-Lorencio, E. García-Roselló, J. González-Dacosta, J. Heine, L. González-Vilas, and J. M. Lobo. 2018. Completeness of national freshwater fish species inventories around the world. Biodivers Conserv 27:3807–3817.

Pielou, E. C. 1966. The measurement of diversity in different types of biological collections. Journal of Theoretical Biology 13:131–144.

Powney, G. D., C. Carvell, M. Edwards, R. K. A. Morris, H. E. Roy, B. A. Woodcock, and N. J. B. Isaac. 2019. Widespread losses of pollinating insects in Britain. Nat Commun 10:1018.

Ricci, L., M. Di Musciano, F. M. Sabatini, A. Chiarucci, P. Zannini, R. C. Gatti, C. Beierkuhnlein, A. Walentowitz, A. Lawrence, A. R. Frattaroli, and S. Hoffmann. 2024. A multitaxonomic assessment of Natura 2000 effectiveness across European biogeographic regions. Conservation Biology 38:e14212.

Rocchini, D., J. Hortal, S. Lengyel, J. M. Lobo, A. Jiménez-Valverde, C. Ricotta, G. Bacaro, and A. Chiarucci. 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Progress in Physical Geography: Earth and Environment 35:211–226.

Rocchini, D., E. Tordoni, E. Marchetto, M. Marcantonio, A. M. Barbosa, M. Bazzichetto, C. Beierkuhnlein, E. Castelnuovo, R. C. Gatti, A. Chiarucci, L. Chieffallo, D. Da Re, M. Di Musciano, G. M. Foody, L. Gabor, C. X. Garzon-Lopez, A. Guisan, T. Hattab, J. Hortal, W. E. Kunin, F. Jordán, J. Lenoir, S. Mirri, V. Moudrý, B. Naimi, J. Nowosad, F. M. Sabatini, A. H. Schweiger, P. Šímová, G. Tessarolo, P. Zannini, and M. Malavasi. 2023. A quixotic view of spatial bias in modelling the distribution of species and their diversity. npj biodivers 2:10.

Ronquillo, C., F. Alves-Martins, V. Mazimpaka, T. Sobral-Souza, B. Vilela-Silva, N. G. Medina, and J. Hortal. 2020. Assessing spatial and temporal biases and gaps in the publicly available distributional information of Iberian mosses. BDJ 8:e53474.

Ronquillo, C., J. Stropp, N. G. Medina, and J. Hortal. 2023. Exploring the impact of data curation criteria on the observed geographical distribution of mosses. Ecology and Evolution 13:e10786.

Ruete, A. 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. BDJ 3:e5361.

Sabatini, F. M., J. Lenoir, H. Bruelheide, and the sPlot Consortium. 2021a. sPlotOpen – An environmentally-balanced, open-access, global dataset of vegetation plots. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig.

Sabatini, F. M., J. Lenoir, T. Hattab, E. A. Arnst, M. Chytrý, J.

Dengler, P. De Ruffray, S. M. Hennekens, U. Jandt, F. Jansen, B. Jiménez-Alfaro, J. Kattge, A. Levesley, V. D. Pillar, O. Purschke, B. Sandel, F. Sultana, T. Aavik, S. Aćić, A. T. R. Acosta, E. Agrillo, M. Alvarez, I. Apostolova, M. A. S. Arfin Khan, L. Arroyo, F. Attorre, I. Aubin, A. Banerjee, M. Bauters, Y. Bergeron, E. Bergmeier, I. Biurrun, A. D. Bjorkman, G. Bonari, V. Bondareva, J. Brunet, A. Čarni, L. Casella, L. Cayuela, T. Černý, V. Chepinoga, J. Csiky, R. Ćušterevska, E. De Bie, A. L. De Gasper, M. De Sanctis, P. Dimopoulos, J. Dolezal, T. Dziuba, M. A. E. M. El-Sheikh, B. Enquist, J. Ewald, F. Fazayeli, R. Field, M. Finckh, S. Gachet, A. Galán-de-Mera, E. Garbolino, H. Gholizadeh, M. Giorgis, V. Golub, I. G. Alsos, J. Grytnes, G. R. Guerin, A. G. Gutiérrez, S. Haider, M. Z. Hatim, B. Hérault, G. Hinojos Mendoza, N. Hölzel, J. Homeier, W. Hubau, A. Indreica, J. A. M. Janssen, B. Jedrzejek, A. Jentsch, N. Jürgens, Z. Kącki, J. Kapfer, D. N. Karger, A. Kavgacı, E. Kearsley, M. Kessler, L. Khanina, T. Killeen, A. Korolyuk, H. Kreft, H. S. Kühl, A. Kuzemko, F. Landucci, A. Lengyel, F. Lens, D. V. Lingner, H. Liu, T. Lysenko, M. D. Mahecha, C. Marcenò, V. Martynenko, J. E. Moeslund, A. Monteagudo Mendoza, L. Mucina, J. V. Müller, J. Munzinger, A. Naqinezhad, J. Noroozi, A. Nowak, V. Onyshchenko, G. E. Overbeck, M. Pärtel, A. Pauchard, R. K. Peet, J. Peñuelas, A. Pérez-Haase, T. Peterka, P. Petřík, G. Peyre, O. L. Phillips, V. Prokhorov, V. Rašomavičius, R. Revermann, G. Rivas-Torres, J. S. Rodwell, E. Ruprecht, S. Rūsiņa, C. Samimi, M. Schmidt, F. Schrodt, H. Shan, P. Shirokikh, J. Šibík, U. Šilc, P. Sklenář, Ž. Škvorc, B. Sparrow, M. G. Sperandii, Z. Stančić, J. Svenning, Z. Tang, C. Q. Tang, I. Tsiripidis, K. A. Vanselow, R. Vásquez Martínez, K. Vassilev, E. Vélez-Martin, R. Venanzoni, A. C. Vibrans, C. Violle, R. Virtanen, H. Von Wehrden, V. Wagner, D. A. Walker, D. M. Waller, H. Wang, K. Wesche, T. J. S. Whitfeld, W. Willner, S. K. Wiser, T. Wohlgemuth, S. Yamalov, M. Zobel, and H. Bruelheide. 2021b. sPlotOpen – An environmentally balanced, open-access, global dataset of vegetation plots. Global Ecol. Biogeogr. 30:1740–1764.

Sporbert, M., H. Bruelheide, G. Seidler, P. Keil, U. Jandt, G. Austrheim, I. Biurrun, J. A. Campos, A. Čarni, M. Chytrý, J. Csiky, E. De Bie, J. Dengler, V. Golub, J. Grytnes, A. Indreica, F. Jansen, M. Jiroušek, J. Lenoir, M. Luoto, C. Marcenò, J. E. Moeslund, A. Pérez-Haase, S. Rūsiņa, V. Vandvik, K. Vassilev, and E. Welk. 2019. Assessing sampling coverage of species distribution in biodiversity databases. J Vegetation Science 30:620–632.

Stein, A., K. Gerstner, and H. Kreft. 2014. Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. Ecology Letters 17:866–880.

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem. 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS ONE 10:e0107042.

Stropp, J., R. J. Ladle, T. Emilio, T. Lessa, and J. Hortal. 2022. Taxonomic uncertainty and the challenge of estimating global species richness. Journal of Biogeography 49:1654–1656.

Stropp, J., R. J. Ladle, A. C. M. Malhado, J. Hortal, J. Gaffuri, W. H. Temperley, J. Olav Skøien, and P. Mayaux. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. Global Ecol. Biogeogr. 25:1085–1096.

Sumner, S., P. Bevan, A. G. Hart, and N. J. B. Isaac. 2019. Mapping species distributions in 2 weeks using citizen science. Insect Conserv Diversity 12:382–388.

Tessarolo, G., R. J. Ladle, J. M. Lobo, T. F. Rangel, and J. Hortal. 2021. Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models. Ecography 44:1743–1755.

Tessarolo, G., R. Ladle, T. Rangel, and J. Hortal. 2017. Temporal degradation of data limits biodiversity research. Ecology and Evolution 7:6863–6870.

Tessarolo, G., T. F. Rangel, M. B. Araújo, and J. Hortal. 2014. Uncertainty associated with survey design in Species Distribution Models. Diversity and Distributions 20:1258–1269.

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. Taxonomic bias in biodiversity data and societal preferences. Sci Rep 7:9132.

Turbelin, A. J., B. D. Malamud, and R. A. Francis. 2017. Mapping the global state of invasive alien species: patterns of invasion and policy responses. Global Ecol. Biogeogr. 26:78–92.

Walther, B. A., and J. L. Moore. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28:815–829.

Wüest, R. O., N. E. Zimmermann, D. Zurell, J. M. Alexander, S. A. Fritz, C. Hof, H. Kreft, S. Normand, J. S. Cabral, E. Szekely, W. Thuiller, M. Wikelski, and D. N. Karger. 2020. Macroecology in the age of Big Data – Where to go from here? Journal of Biogeography 47:1–12.

Xu, G., L. Jiao, M. Yuan, T. Dong, B. Zhang, and C. Du. 2019. How does urban population density decline over time? An exponential model for Chinese cities with international comparisons. Landscape and Urban Planning 183:59–67.

Yang, W., K. Ma, and H. Kreft. 2014. Environmental and socio-economic factors shaping the geography of floristic collections in C hina. Global Ecology and Biogeography 23:1284–1292.

Yang, W., K. Ma, and H. Kreft. 2013. Geographical sampling bias in a large distributional database and its effects on species richness–environment models. Journal of Biogeography 40:1415–1426.

APPENDIX

Supplementary Methods

Trivariate map

We represented in a trivariate map, which is a graphic representation that shows the relationship between three variables at once, the three dimensions of bias. We used the functions provided by "tricolore" R package for creating the map. The variables selected for the trivariate map were the Nearest Neighbour Index, the completeness of the species richness and the Pielou's evenness. The variables were first standardized to have a mean of zero and a standard deviation of one, then rescaled to a 0-1 range and subsequently mapped over our study area. We also removed all grid cells that had missing values for at least one facet of bias.

The trivariate map highlighted those area where the prevalence of one type of bias prevail to the others. The grid cells with different colours from those of vertices (e.g., brown) tend to be more and more influenced uniformly by the three dimensions of bias as the colour approaches the center of the triangle.

Facets of bias

Single grid-based map of each metric of bias with unstandardized grids number.

Temporal uncertainty

Here we represented the temporal uncertainty by assessing the difference between the most recent year in the database (2014) and the year of each record. The higher the difference, the more uncertainty we have. The uncertainty per grid cell was calculated as the median value of the differences between the year of the most recent surveyed plot (i.e., 2014) and the date of recording of the ith plot within the grid cell.

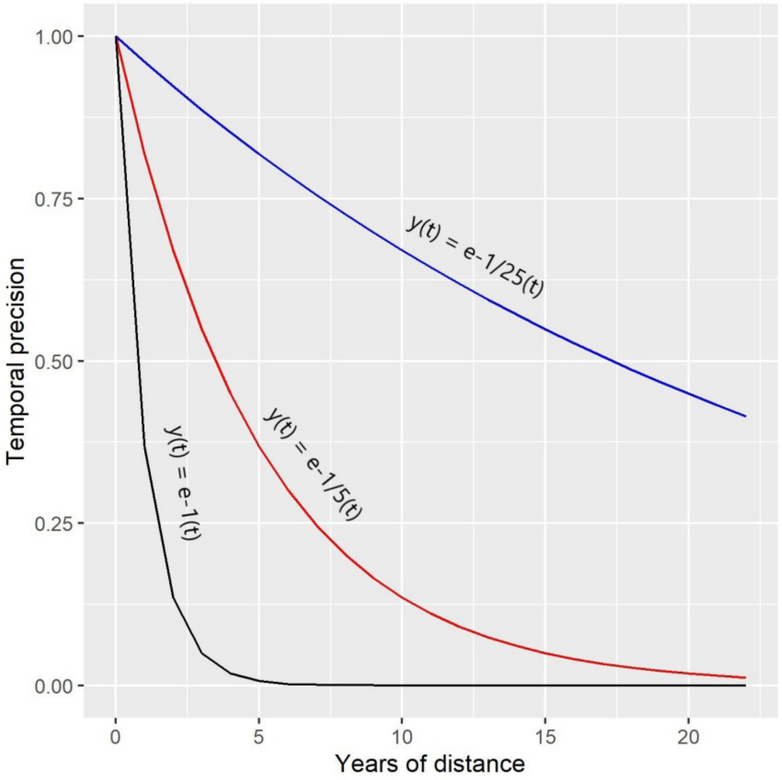

Figure S1: Negative exponential curves fitting using three different exponents, i.e., z_1=-1, z_2=-1/5 and z_3=-1/25.

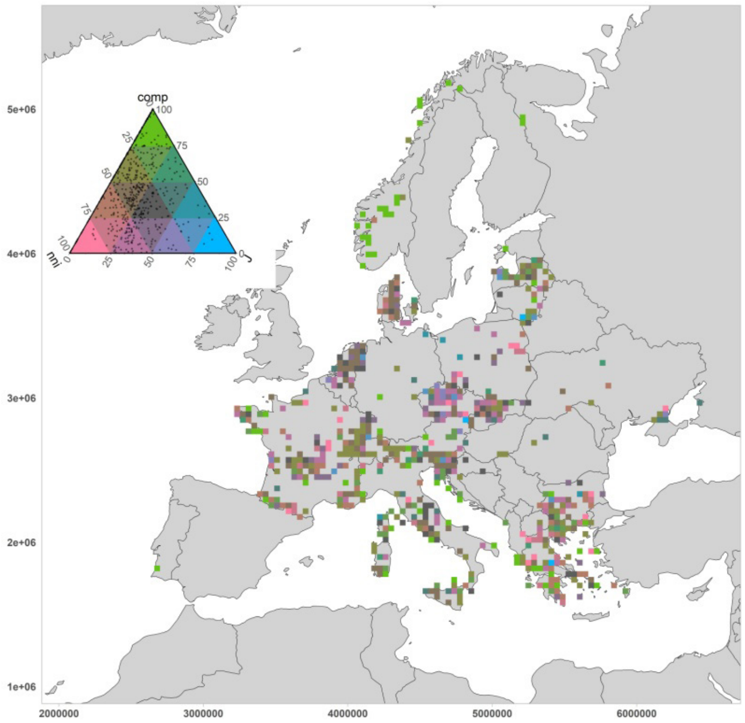

Figure S2: Grid-based trivariate map of taxonomic bias (i.e., completeness of species richness, abbrv. legend comp), spatial bias (i.e., NNI, abbrv. legend nni) and temporal bias (i.e., Pielou's evenness, abbrv. legend J). Each grid cell has a spatial resolution of 39.5 km. The highest sampling completeness is represented by light green color (low taxonomic bias), the highest temporal evenness by light blue color (low temporal bias), the highest uniform distribution of the plots (low spatial bias) by pink color.

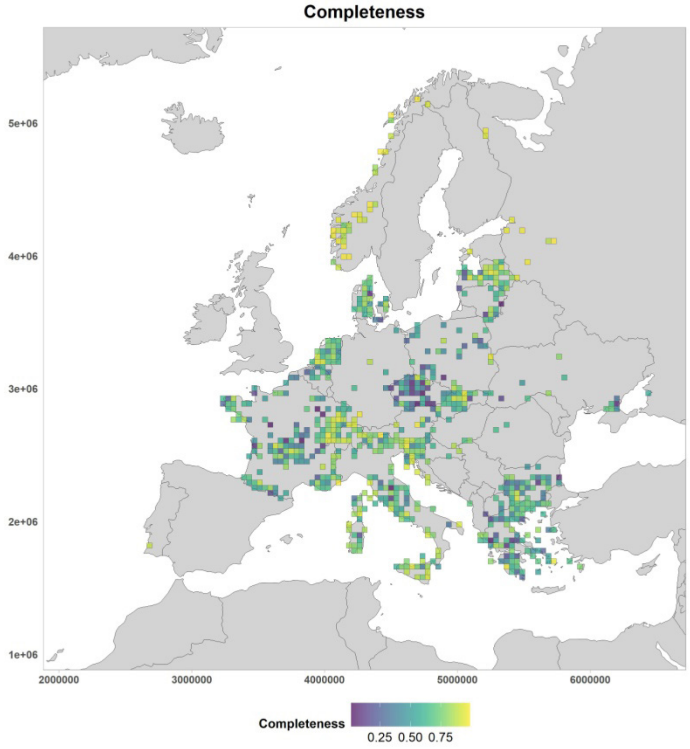

Figure S3: Completeness of the species richness per grid cell of 39.5 km.

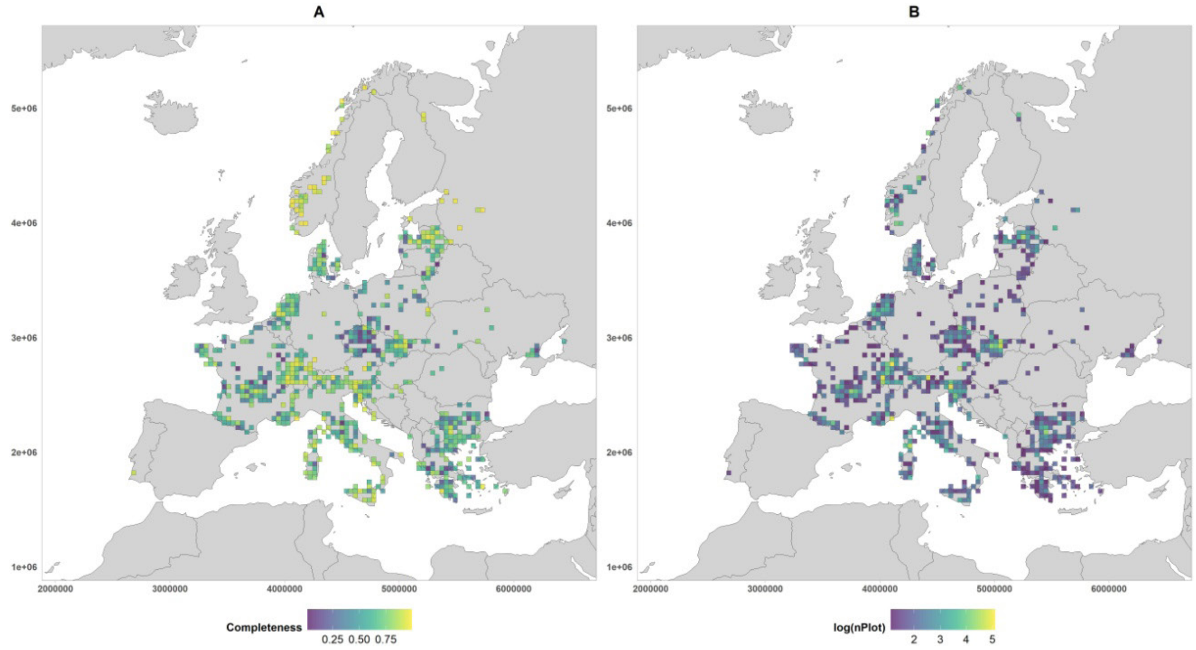

Figure S4: A) Completeness of the species richness and B) logarithm to base 10 of the number of plots per grid cell of 39.5 km with standardized number of grid cells.

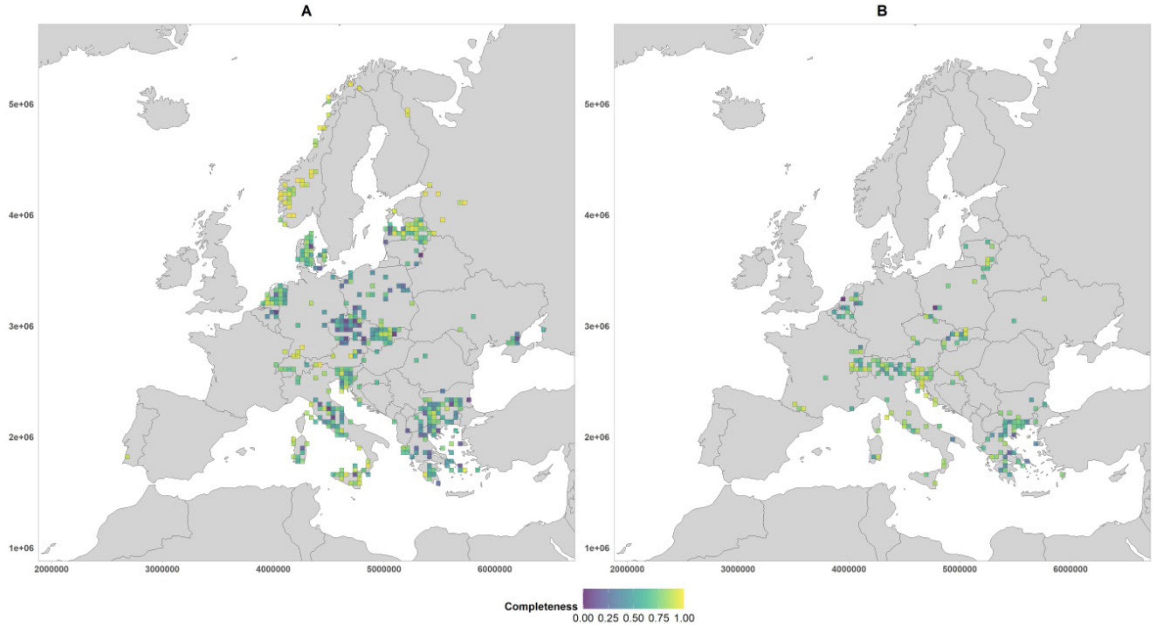

Figure S5: A) Completeness of the species richness per grid cell of 39.5km including only the vegetation plots with area less than or equal to 150 m^2. B) Completeness of the species richness for plots with an area greater than 150 m^2. The area size was determined by relying on Sabatini et al. 2022[1] and to have a comparable number of plots belonging to the two categories.

[1] Sabatini, F. M., Jiménez-Alfaro, B., Jandt, U., Chytrý, M., Field, R., Kessler, M., ... & Bruelheide, H. (2022). Global patterns of vascular plant alpha diversity. *Nature Communications*, *13*(1), 4683.

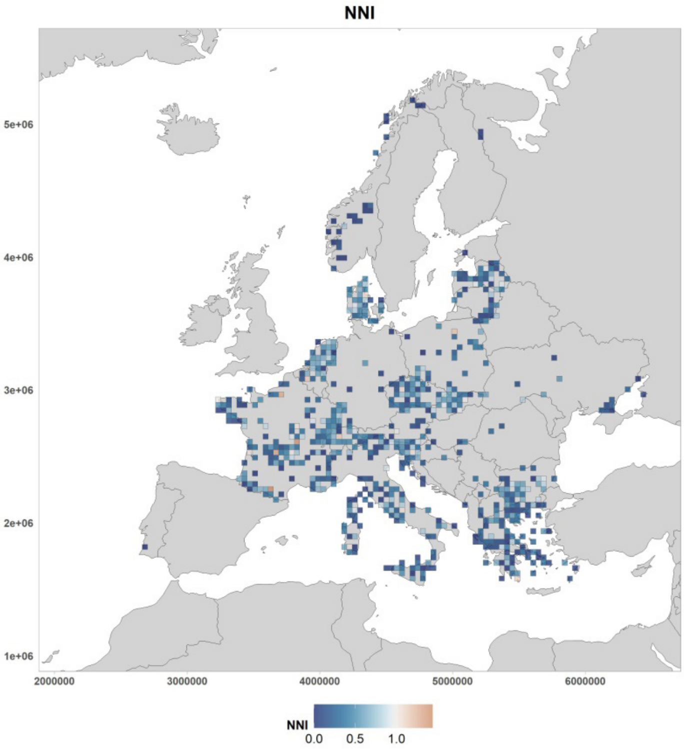

Figure S6: NNI per grid cell of 39.5 km. NNI values greater than 1 indicate a random distribution of plots within a grid cell, while values less than 1 indicate a clustered distribution.

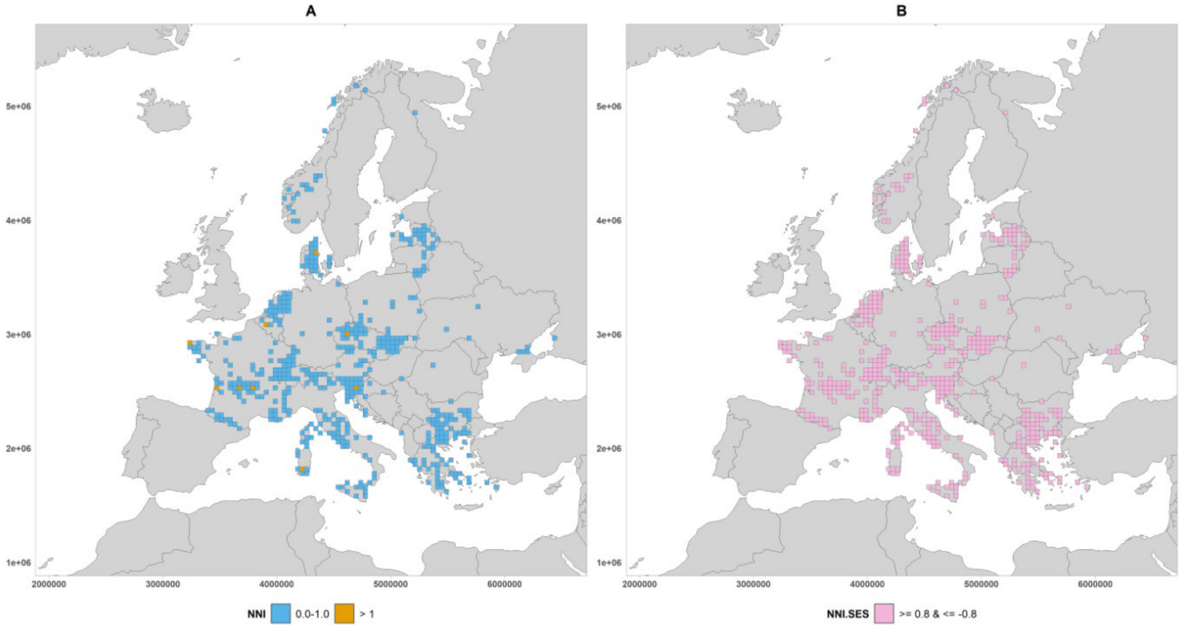

Figure S7: Spatial distribution of the vegetation plots per grid cell. A represents the map of NNI and B represents the map of NNI with a standardized effect size. NNI values greater than 1 indicate a random distribution of plots within a grid cell, while values less than 1 indicate a clustered distribution. There is no value of NNI with a standardized effect size between – 0.8 and 0.8, meaning that the effect size is large.

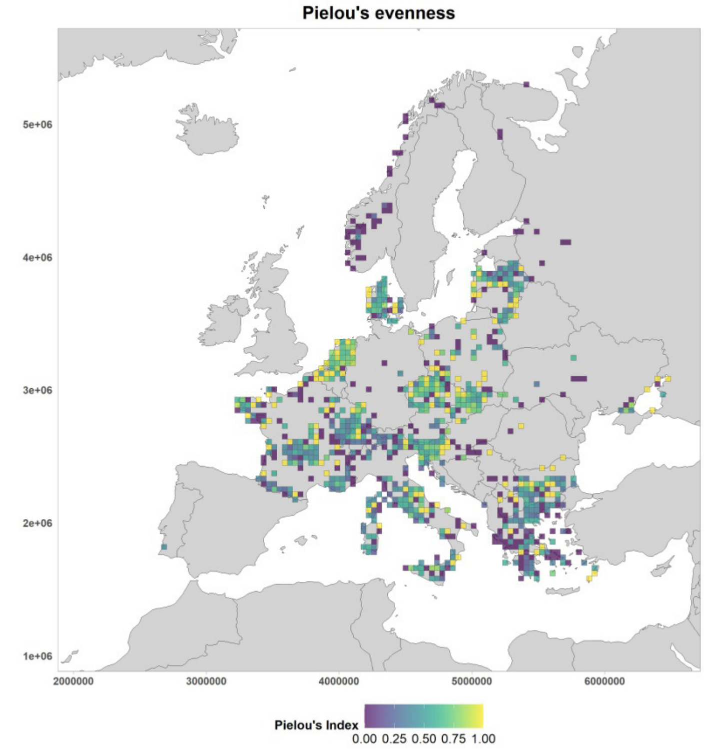

Figure S8: Pielou’s Index per grid cell of 39.5 km.

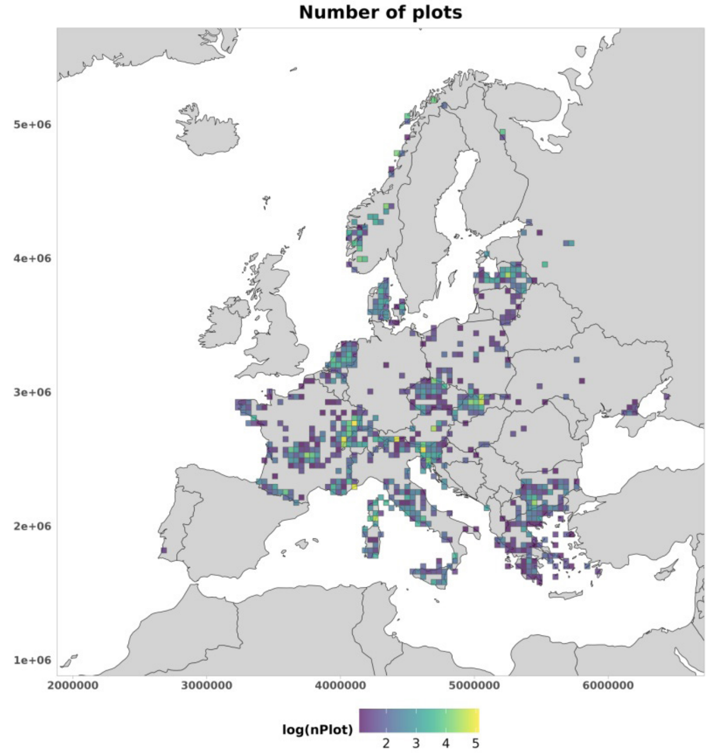

Figure S9: Logarithm to base 10 of the number of plots per grid cell of 39.5 km.

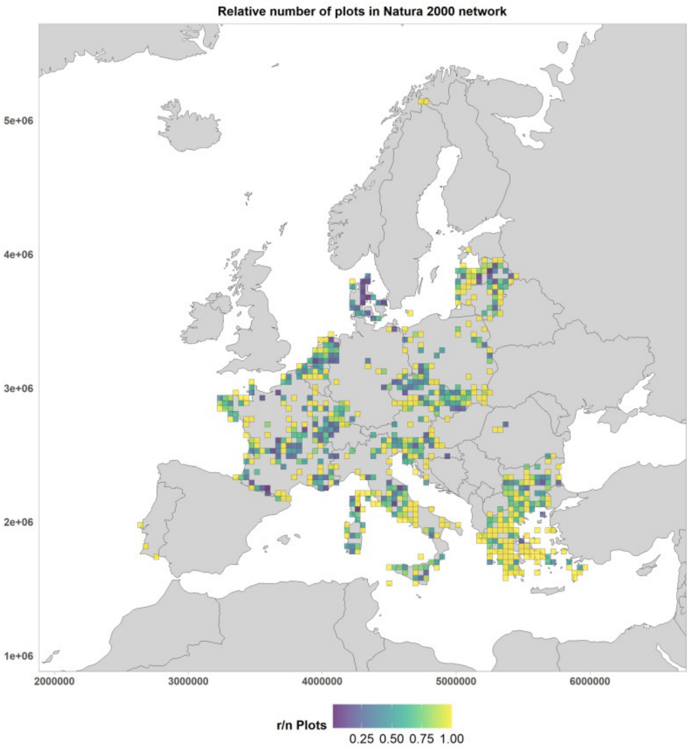

Figure S10: Map of the relative number of plots in the Natura 2000 network per grid cell.

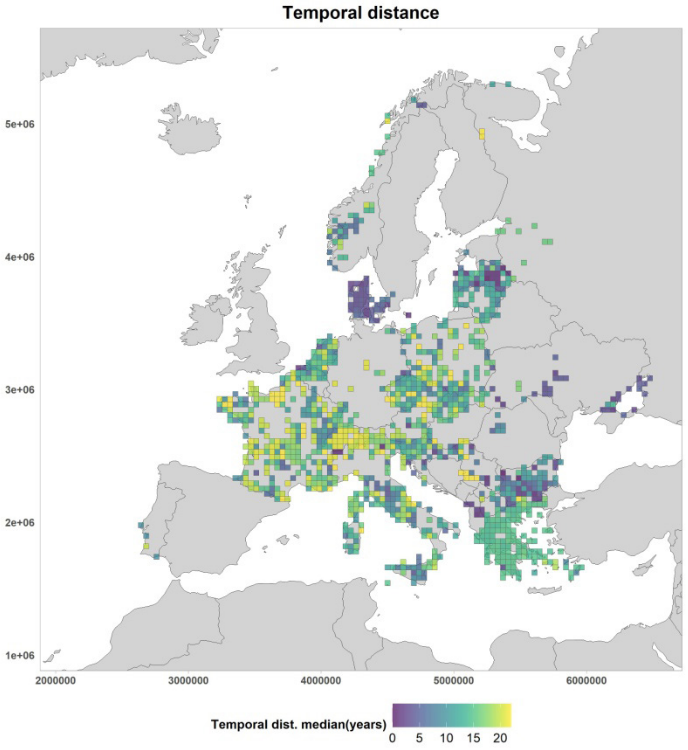

Figure S11: Median of the temporal distance between the most recent year (i.e., 2014) and the year of each record per grid cell of 39.5 km.