ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE DELLA TERRA, DELLA VITA E DELL'AMBIENTE

Ciclo 37

Settore Concorsuale: 05/A1 - BOTANICA

Settore Scientifico Disciplinare: BIO/03 - BOTANICA AMBIENTALE E APPLICATA

ASSESSMENT OF BIAS AND UNCERTAINTY OF SPECIES OCCURRENCE DATA: METRICS AND METHODS

Presentata da: Elisa Marchetto

Coordinatore Dottorato
Barbara Cavalazzi

Supervisore
Duccio Rocchini

Co-supervisore
Enrico Tordoni

Esame finale anno 2025

Abstract

The increasing availability of large quantities of data on species occurrence (i.e., presence and/or presence and absence of species) to support biodiversity studies and conservation actions is not always coupled with the data quality. Indeed, species occurrence data can present forms of bias (i.e., systematic deviation from the true value) and uncertainty (i.e., dispersion of values or lack of knowledge). In the PhD thesis, I evaluated different metrics and methods to address uncertainty and bias of species occurrence data.

In the first chapter, the effect of the sampling method of the presences and absences and the effect of their ratio (i.e., sample prevalence) were tested on different Species Distribution Models: Favourability and Probability-based SDMs. In the second and third chapters, I employed various metrics to evaluate the quality of species occurrence data stored in a case study database, namely the sPlotOpen database. The taxonomic, spatial and temporal bias were measured at European and habitat levels respectively with i) the completeness of the species richness, ii) the Nearest Neighbor Index, iii) the Pielou's index. Besides, the temporal uncertainty—defined as the information decay of the species occurrence over time—was quantified using a negative exponential function.

Among the main results, I found that the sampling methods (i.e., random and stratified sampling) of the species occurrences had no effect on the performance of the Favourability and Probability models. The Favourability model, in contrast, exhibited lower variability and only slightly higher accuracy than Probability in the predictions of species distribution. Moreover, the metrics used to assess the dimensions of bias in species occurrence data proved to be effective, revealing heterogeneous patterns. Additionally, the analysis of temporal uncertainty identified hotspot areas across Europe. Results that highlighted the necessity of assessing data quality prior to its use in biodiversity inferences.

Extended abstract

The collection of species occurrence data has risen considerably in the last decades. Policies and actions, aimed to reduce the intense impact of threats to the conservation status of species and habitats with the support of technological advancements, speeded up the demand for collection effort. However, a larger amount of data is not always coupled with a higher data quality, since many sources of inaccuracies and information gaps can still be hidden. In this PhD thesis, I focused on evaluating the quality of species occurrence data (i.e., presence and/or presence and absence of the species) and testing different methods to address its bias and uncertainty.

Species occurrence data can store knowledge gaps in the actual distribution and in the taxa coverage, as well as imprecision in the records collection, leading to different forms of errors like bias (i.e., systematic deviation from the true value) and uncertainty (i.e., dispersion of values or

lack of knowledge). The data quality was evaluated by measuring and representing bias and uncertainty under two methodological approaches: the first one involved the assessment of the accuracy and the variability of the predictions of species distribution estimated using Species Distribution Models and presences and absences data; the second one dealt with the direct measurement of bias and uncertainty of species occurrences in biodiversity databases.

In the first chapter of the thesis, the assessment of species occurrence data was evaluated in the inference phase. I used virtual species modelling at the European scale for testing the effect of two standardized sampling methods (random and stratified) of presences and absences and different sample prevalence (i.e., ratio of number of presences and absences) in favourability-based and probability-based SDMs calibrated using a fix sample size (i.e., the number of presences and absences) and diverse statistical models (GLM, GAM, RF, BRT). The favourability model derives from the removal of the value of the sample prevalence to the predictions of probability-based SDM.

The sampling methods did not have a great impact on the accuracy of SDMs, although they determined significantly different predictions of species distributions. A significant variation between the predictions of species distribution was also influenced by the ratio of presences and absences being sampled for both sampling methods and all of the statistical models. The probability model showed higher variability between the predicted species distributions calibrated using different sample prevalence. Moreover, the favourability-based SDMs performed slightly better (i.e., higher accuracy) than probability-based SDMs (more than half of the median Continuous Boyce index values were higher) in predicting the species distribution over space. Hence, the favourability model, thanks to the lower variability when the sample prevalence changes, may allow a better improvement in the comparisons between SDMs and a better understanding of the environmental conditions that shape the niche suitability of the species.

In the second chapter of the thesis, the dimensions of bias (taxonomic, spatial and temporal) and the temporal uncertainty were measured on species occurrence data present in biodiversity databases through a new methodological framework aimed to be as reproducible as possible even changing the spatial scale and the ecological level. We used as a case study the vegetation plot records located in Europe of sPlotOpen, an open-access database. The bias was calculated by using common ecological metrics: the completeness of species richness for the taxonomic bias, the Nearest Neighbor Index for the spatial bias and the Pielou's Index for the temporal bias. The temporal uncertainty, defined as the information decay of the species occurrence over time, was measured by applying a negative exponential transformation to the difference between the more recent year of recording among the plots and the date of recording of the data point. Across the grid cells, the completeness of the species richness (taxonomic bias) and the evenness of the sampling years (temporal bias) were heterogeneous, while the distribution of plots was mainly clustered, showing a high spatial bias. This suggests that the sampling of species occurrences was possibly opportunistic and/or not homogeneous and standardized within the grid cells over Europe. The temporal uncertainty highlighted hotspot areas that changed with the exponent being used. Overall, the new methods allowed us to assess the data quality of species occurrence in

biodiversity databases providing a solid framework for a possible correction of information gaps based on the study context or for addressing future resampling campaigns.

In the third chapter of the thesis, I measured the dimensions of bias (taxonomic, spatial, temporal) at the habitat level using the species occurrence data of the sPlotOpen database. The completeness of the species richness, the NNI and the Pielou's Index were calculated for EUNIS level 1 E (i.e., grassland and lands dominated by forbs, mosses and lichens) and G (i.e., woodland, forest and other woodland) and for their sub habitats at level 2 (E1, E2, E3, E4, E5, G1, G2, G3) within grid cells covering European continent with a spatial resolution of 10 km. The patterns of bias were similar for E and G. Both showed low taxonomic bias, high spatial bias and an intermediate level of temporal bias. Also, for level 2 EUNIS, the patterns of bias were similar to those at level 1. However, the spatial and temporal bias exhibited greater differences in the values among the habitat types. The resulting values may have been influenced by biased sampling techniques and procedures, as well as, the geographic distribution of the habitat type. Otherwise, a measure of bias of species occurrences can help, at first instance, to better describe the degree of completeness of the actual description and representation of the habitat state, bearing in mind that the dimensions of the bias of species occurrence data are interconnected and often not independent of each other; evaluating them at the habitat level also adds the environmental dimension to the first three as the habitat is a complex system described also by its climatic and edaphic conditions.

Awareness of the quality of data in biodiversity databases provides perspective on the efforts that still need to be made to ensure more complete monitoring and conservation of species and habitats. Furthermore, in light of this, in the inference phase of the data it is important to test the methods and models used to ensure greater accuracy and precision.

Contents

Introduction

Overview: data on species occurrence

Biotic and abiotic conditions together determine the physical and structural environment, as well as the interactions, behaviours, and biological processes of the organisms within an ecosystem (Schulze et al., 2019). Both living and non-living components play crucial roles in shaping the species population and community dynamics within an ecosystem together with its functions (Maestre et al. 2010). Nowadays, species distributions and habitat integrity suffer spatiotemporal changes under the pressure of different threats which are further exasperated by human activities (Pimm et al. 1995; Latombe et al. 2017; Wiens and Zelinka 2024). Indeed, human activities are known to be responsible for habitat fragmentation, land use change, climate change, biological invasions which, all together, are increasing the rate of biodiversity loss (Tilman et al. 2017). Nowadays, over 45,300 species face the threat of extinction including 26% of mammals, 41% of amphibians, 36% of reef corals, and 34% of conifers (IUCN 2024). Preserving or restoring populations, communities, and ecosystems is critically important for their survival , consequently, global initiatives and policy frameworks (e.g., the Convention on Biological Diversity, the Sustainable Development Goals, Habitat Directives) arise to quantify the conservation status of species and to address specific conservation actions preventing biodiversity loss (Pereira et al. 2013). Many of them support sample collections, monitoring programs and sharing of biodiversity data. Numerous efforts are undertaken to share clear, accessible, harmonized and up-to-date biodiversity information that accurately reflects the species populations across different taxa and regions over time (Pereira et al. 2013; Jetz et al. 2019; Kühl et al. 2020). Indeed, reliable data are essential for obtaining accurate estimates of biodiversity conservation status.

Studies on biodiversity conservation often estimate species diversity, community and species interactions using species occurrence data. Data on species occurrence provide information on the presence of the species, eventually, with their abundances or coverage (Jetz et al. 2019). They can also include the absence of the species when the inventories are carried out in small areas. However, the term absence should be taken with caution because the species could be simply unobserved. Hence, rather than a true absence, it would be better to consider it as a probability of absence which is strictly related to the sampling effort and the accuracy of the sampling design (Lobo, Jiménez-Valverde, and Hortal 2010; Jetz et al. 2019). Absence (or, according to the type of sampling, pseudo-absence or background data) and presence data are often combined with a set of variables that reflect the ecological and geographical conditions to understand species or community potential distribution across space (Elith and Leathwick 2009). The derived predictions are commonly estimated through models denominated Species Distribution Models or Ecological Niche Models. There are two main types of SDMs. Correlative SDMs model the species occurrences as a function of environmental conditions by using different statistical techniques, such as regression-based modeling (e.g., GLM, GAM, MARS), machine learning-based modeling (e.g., RF, GBM, SVM) and envelope modeling (e.g., BIOCLIM), to predict the geographic distribution of the species (Srivastava, Lafond, and Griess 2019). In contrast, mechanistic SDMs

combine spatial habitat features to functional traits (morphology, physiological and behavioural responses) of organisms (Kearney and Porter 2009).

Data on species occurrence is retrieved from a variety of sources like field observations, museum and herbarium collections, scientific literature, along with the ones derived from "citizen science" campaigns (Chytrý 2001; Boakes et al. 2010; Tiago et al. 2017). Even so, species occurrence data can often drag on different facets of bias and uncertainty related to the taxonomic, geographic and temporal dimensions (Meyer, Weigelt, and Kreft 2016; Maldonado et al. 2015) which are generally originated by sampling and non-sampling errors. Sampling errors are often induced by chance or by the sampling designs that do not achieve a proper randomization failing to represent the entire population. In species occurrence data, the sampling errors can be created by opportunistic (i.e., non-probabilistic) sampling designs being used for collecting records (Boakes et al. 2010; Meyer et al. 2015). On the contrary, non-sampling errors can be determined by measurement errors such as researcher misreading and wrong species identification, miscalibrated scale, and a low accuracy of the instruments (e.g., taxonomic misidentification, erroneous or imprecise coordinates). Besides, non-sampling errors can derive from the union and use of heterogeneous, inaccurate and imprecise data being sampled with different goals and protocols and in different time (Boyd et al. 2022). Consequently, these errors stand for information gaps of various nature (Hortal et al. 2015) such as the lack of knowledge about the actual geographic distributions of species (i.e., Wallacean shortfall), the knowledge gap in taxa coverage (i.e., Linnean shortfall). However, although technological advancements over the years, such as remote sensing, camera traps, and GPS devices, have increased the ability to collect data, the potential presence of errors is not always removed (Hofmeester et al. 2019; Feng et al. 2021; Schad and Fischer 2023). Despite new methods are continuously developing also taking advantage of, remote sensing technologies such as Unmanned Aerial Vehicles (UAVs) (Kellenberger et al. 2018, Ferreira et al. 2020) and Terrestrial Laser Scanners (TLS) (Terryn et al. 2020), these intrinsically hold their degree of uncertainty.

Data on species occurrence is often gathered in biodiversity databases, which are structured collections of information of various nature like species distribution and abundance, genetic data, habitat classification, and conservation status. Species occurrence data can be further collected in bigger repositories such as regional or global biodiversity databases of species occurrence and/or co-occurrence (e.g., GBIF, sPlot, iNaturalist, VertNet) where a few of them are also freely accessible (e.g., GBIF, sPlotOpen, BIEN). Indeed, over the years, with the improvement of computational capabilities, these databases have grown in number and size (Feng et al. 2022). However, even the biodiversity databases can hold knowledge gaps stemming from the different sampling protocols of the data types and the combination of multiple databases with diverse sampling designs and projects into a single database (Chytrý et al. 2014; Wüest et al. 2020). For example, the Global Biodiversity Information Facility (GBIF 2024), with more than 3 billion occurrence records, is the largest international network and data infrastructure of free and open access primary biodiversity data. Alongside its extraordinary and undoubtedly value, it exhibits limitations in data quality. Many taxa and regions are still under-sampled with records showing

biased temporal coverage and preferential sampling for specific functional traits (Hughes et al. 2021; Daru and Rodriguez 2023; García-Roselló, González-Dacosta, and Lobo 2023), erroneous or imprecise geographic coordinates (Zizka et al. 2020) and differences in biodiversity patterns between records achieved by observations or from specimen or samples preserved in a natural history collections (Speed et al. 2018; Daru and Rodriguez 2023).

Therefore, given the wide range of possibilities which generate information errors, data on species occurrence can often be biased and uncertain.

Bias and uncertainty in species occurrence data

In statistics, bias alludes to a systematic error or deviation from the true value in a data collection and, it generates an overestimation or underestimation of the outcomes, while uncertainty indicates the dispersion of the values within which the true value is expected to fall (ISO., I., & OIML, B. 1993; Bolker 2008). Bias and uncertainty can be described by accuracy and precision by using different measures. Bias is generally calculated as the expected difference between the estimate and the true value while, uncertainty is often evaluated using several measures such as the variance (the variability of the point estimates around their mean value) or the standard deviation and the confidence interval (upper and lower limit within which the true value is likely to fall) (Bolker 2008). However, uncertainty is also conceived as a lack of knowledge in which the current state cannot be precisely described (Hüllermeier and Waegeman 2021).

Data on species occurrence can mainly show uncertainty in three dimensions: taxonomic, spatial and temporal. The taxonomic uncertainty derives from imprecise or erroneous species names with spelling errors or variants in scientific names, which can broadly depend on changes in taxonomy, and wrong species identification. Generally, longer taxonomic history occurs in regions with lower diversity (e.g., in European temperate forests, Stropp et al. 2022), resulting in a disparity in expertise and taxonomic identification. Therefore, the taxonomic effort is heterogeneous at a global scale and depends on the taxon under consideration (Stropp et al. 2022).

The spatial or geographic uncertainty, instead, reflects the inaccuracy of the data coordinates i.e., the positional error, defined as the difference between the true and recorded location (Gábor et al. 2020; Moudrý et al. 2024). One of the main sources of spatial uncertainty is the historical records held in museums, botanical gardens and other similar institutions (Marcer et al. 2022; Campbell 2024) while the observations derived from field surveys tend to be more precise. However, positional errors can also appear in georeferenced data using global navigation satellite systems (GNSS) where the number and position of satellites are incorrect or where the site characteristics act as a barrier (e.g., canopy density) (Moudrý et al. 2024).

Lastly, a definition of temporal uncertainty is difficult to find in scientific literature. Tessarolo et al. (2017, 2021) defined it as the information decay of the species occurrence over time; in other words, the information becomes more imprecise as time passes from the date the data was recorded. Indeed, biodiversity data is likely to change because species distribution and community

composition are naturally prone to change over time (Wisz et al. 2013) or it is likely to become imprecise because changes in taxonomy or loss of metadata (Tessarolo et al. 2017).

Data on species occurrence can also show bias in the taxonomic, spatial, environmental, and temporal dimensions. The taxonomic bias reflects the sample coverage gap between the observed species pool and the expected (Chao and Jost 2012). The gap is determined by the non detection of the species which is influenced by several factors such as taxonomic inexperience, erroneous sampling, preference for charismatic species, rarity of the species, low sampling effort (Adamo et al. 2021; Cazzolla Gatti et al. 2022). Due to these factors, taxonomic bias is also irregular among different taxa (García-Roselló, González-Dacosta, and Lobo 2023).

A common metric to calculate it is the completeness of the species richness measured as the ratio of the observed species and the expected or rarely as the slope of the species accumulation curve (Chao et al. 2020; Yang, Ma, and Kreft 2013). However, another potential solution relies on the concept of species pool and dark diversity (Pärtel, Szava-Kovats, and Zobel 2011; Ronk, Szava-Kovats, and Pärtel 2015), especially regarding the so-called community completeness (i.e., the proportion of observed diversity from site-specific species pool size) which may also provide hints about the uncertainty around the sampled communities.

A disproportionate sampling in geographic space generates bias in the spatial dimension. Indeed, an heterogeneous sampling effort and an opportunistic sampling of the target species occurrences, which do not represent the entire niche of the species and geographic distribution, can lead to a geographic distortion of the data (Sumner et al. 2019). Similarly, in biodiversity databases, species occurrence data that is over-sampled in some regions rather than being uniformly distributed generates spatial bias (Meyer et al. 2015). The spatial bias is then commonly measured as the number of plots or records per unit area (Meyer, Weigelt, and Kreft 2016; Rocchini et al. 2023) and rarely as an index of data clustering (Boyd et al. 2021; Chesshire et al. 2023).

Generally, a biased sampling in geographic space can also be associated with an environmental bias, i.e., an incomplete representation of the climatic and soil conditions (Speed et al. 2018; Monsarrat, Boshoff, and Kerley 2019) describing the species occurrence. Therefore, a greater sampling effort or an irregular sampling for some geographical areas and bioclimatic regions can create a distortion of the environmental space. For instance, considering global biodiversity databases like sPlot, the temperate zones appear to have greater data coverage than tropical and Mediterranean climates (Sabatini et al. 2021).

The temporal bias, on the other hand, i.e., the uneven temporal coverage of the data (Meyer, Weigelt, and Kreft 2016) is mainly due to opportunistic sampling in time. Indeed, the temporal bias can reflect a non-uniform sampling over the years or an incorrect sampling over time (e.g., seasons, daily cycle) according to the ecology of the species (La Sorte and Somveille 2020; Bowler et al. 2024).

The dimensions of bias typically stem from opportunistic and unstructured sample collections, such as unstandardized sampling or socioeconomic preferences in the choice of sampling locations (Chapman et al. 2024). However, the absence of a common and coordinated monitoring scheme among the parties involved can also result in several information gaps (Kühl et al. 2020). For instance, some sampling campaigns rather than being probabilistic opt only for locations or regions species-rich (Chytrý 2001). Moreover, road accessibility, presence of infrastructures, human densely populated areas, low conflict risk, high research funds, and protected areas often influence the selection of the sampling sites (Mair and Ruete 2016; Meyer, Weigelt, and Kreft 2016; Girardello et al. 2019; Hughes et al. 2021; Zizka, Antonelli, and Silvestro 2021a). Finally, the dimensions of bias of species occurrence data can manifest across various timeframes, at different spatial scales ranging from local to global, and at multiple ecological levels, from individual species to entire realms (Meyer et al. 2015; Hugo and Altwegg 2017; La Sorte and Somveille 2020; Hughes et al. 2021; García-Roselló, González-Dacosta, and Lobo 2023).

Approaches to the assessment of bias and uncertainty

Biased and uncertain species occurrence data can determine under or over-representation of biodiversity patterns (Hughes et al. 2021) but also induce poorly accurate and precise biodiversity estimates and predictions (e.g., Species Distribution Models) (Beck et al. 2014, Moudrý et al. 2024). The presence of bias and uncertainty is then addressed by scientists using mainly three different approaches.

The first approach consists of evaluating them in the inference phase calculating i) the accuracy and precision of estimates, ii) modelling predictions with known sampling errors, and iii) comparing variables affected by sampling errors. However, this approach does not correct the bias and the uncertainty of the data. Accuracy and precision of estimates (i) are tested by calculating statistical measures such as the bias, the variance, the standard deviation, the confidence interval (Bolker 2008) and the performance of the model. For instance, Bazzichetto et al. (2023) measured the bias between the estimated probability of the species distribution and the true probability of the species distribution, the variance of the estimated probabilities of species distribution and the Root Mean Squared Error as a combination of the values of bias and variance. Večeřa et al. (2019), as well as Dyderski et al. (2018), evaluated the effect of sampling bias on species diversity by calculating the species diversity using the species observations resampled with biased sampling strategies (ii). Testolin et al. (2024), instead, tested the possible bias in species richness generated by the sampling effort by comparing the Pearson correlation coefficient between the two variables (iii).

The second approach relies on filtering and data cleaning of the existing database (e.g., García-Roselló et al. (2014), Ronquillo, Stropp, and Hortal (2024)) or on gaps fixing solutions such as subsampling, weighting, imputation techniques (Bowler et al. 2024). This approach can involve removing records according to specific criteria before using them for biodiversity estimates (Maldonado et al. 2015; Führding-Potschkat, Kreft, and Ickert-Bond 2022) with particular attention to avoiding altering the real patterns of the species records (Ronquillo et al. 2023). For instance,

records with high positional error or older than a fixed date are often removed (Večeřa et al., 2019) as well as records duplicates. Despite that, the customization of tests and thresholds for filtering different taxonomic groups or taxa and geographic areas is preferable to automated filtering (Zizka et al. 2020). Actually, each combination of species–ecological question–data set determines the condition under which dealing with gaps correction of species occurrence data (Bowler et al. 2024).

Particular attention is paid to data used in Species Distribution Models in evaluating its effect on model performance (Beck et al. 2014) and in identifying methods to correct information gaps (Phillips et al. 2009; Varela et al. 2014). Information gaps can generally occur when species occurrences are uncertain for high positional error, the sampling does not follow the ecology of the species and/or because of opportunistic (i.e., non-probability) sampling (Moudrý et al. 2024). Indeed, high positional error can mask the actual environmental conditions that describe the species' niche decreasing the performance of the model (Graham et al. 2008; Gábor et al. 2020). Instead, the sampling bias in geographic, temporal and environmental space can distort the predicted distribution of the species (Beck et al. 2014; Cosentino and Maiorano 2021). For instance, geographic bias, created by heterogeneous sampling intensity or distribution, can either alter the environmental characteristics of the species' niche or fail to include them entirely (environmental bias) (Moudrý et al. 2024). There are many techniques that correct geographic and environmental bias of presences and pseudo-absences or background points, for example, by choosing background points with the same bias of the presence data (Phillips et al. 2009), by applying a spatial filter to the presences to reduce the spatial autocorrelation of the sampling effort (Veloz 2009), by uniformly sampling the pseudo-absences in the environmental space (Da Re et al. 2023). However, their effects on both dimensions of bias should be checked before calibrating the model (Varela et al. 2014; Cosentino and Maiorano 2021).

Despite the possibility of applying various correction techniques, it is a good practice to measure bias and uncertainty of species occurrence data depicting them using specific metrics (Boyd et al. 2022), like those of coverage or variance, before carrying out filtering, data cleaning and gaps fixing solutions; as well as, it is recommended to test the correction techniques being used (Boyd, Stewart, and Pescott 2024). However, despite this, bias and uncertainty of data are often not clearly represented. Hughes et al. (2021) described highly biased patterns of marine and terrestrial animal distribution data in GBIF and OBIS databases. According to their study, only 6.74% of the planet has been sampled for animal species with tropical regions poorly represented. Moreover, high elevations and deep ocean areas were largely unexplored and, in most taxonomic groups, more than half of the recorded data denoted less than 2% of all species.

Nevertheless, a third approach still exists to address bias and uncertainty in species occurrence data which relies on performing an initial filtering and data cleaning of data during the creation of the database or a data standardization using tools such as Taxonstand (Cayuela et al. 2012) and CoordinateCleaner (Zizka et al. 2019). For example, the sPlotOpen database (Sabatini et al. 2021), an open-access version of the sPlot database (Bruelheide et al. 2019) was developed by applying a balanced resample over the environmental space to the vegetation plots that were previously

filtered to eliminate lacking coordinates or locations with a position uncertainty greater than three kilometres.

Beyond these three main approaches aiming to address bias and uncertainty in species occurrence data, it is a common practice also to evaluate the effect of several factors, such as road proximity, population density, financial and institutional resources, in determining the different forms of information gaps (Yang, Ma, and Kreft 2014; Meyer et al. 2015; Tiago et al. 2017; Callaghan et al. 2021; Shirey et al. 2021) to increase our insight and awareness. To support the evaluation, different tools are already available. For instance, Zizka, Antonelli, and Silvestro (2021a) provided a tool to facilitate the assessment of the effect of accessibility biases in species occurrence data sets and Zizka et al. (2021b) provided a graphic interface to identify possible relationship between species occurrence and socio-political conditions in time.

However, the availability of large amounts of data is not always coupled with the data quality (Bayraktarov et al. 2019; Wüest et al. 2020). Consequently, the presence of bias and uncertainty should be preventively tested and adequately accounted for or mitigated before in the modelling procedure.

Aims of the thesis

The PhD project aims to provide methodologies to measure and represent the uncertainty and bias of species occurrence data. I addressed bias and uncertainty issues at the inference level (first approach) and at the database level (second approach). The thesis is organized into three distinct chapters.

Chapter 1:

Following the first approach, I tested the effect of sampling methods (i.e., random and stratified) and sample prevalence (i.e., the ratio of presences and absences) of presences and absences in Species Distribution Modeling. Furthermore, I evaluated the effect of Favourability model in Species Distribution Modeling in reducing the variability of predictions which may directly depend on the sampling bias in the geographic and environmental space.

Chapter 2:

Following the second approach, I presented a reproducible methodology to the measure of bias and uncertainty in biodiversity databases providing a detailed workflow of the use of new and common metrics and methods in ecology. The final goal of the study was to increase our awareness of knowledge shortfalls in taxonomic, spatial and temporal dimensions to implement best practices and protocols to fill information gaps and reduce sources of error.

Chapter 3:

Here I measured the dimensions of bias (taxonomic, spatial, temporal) of species occurence data at habitat level across Europe. As a peculiar and complex concept that bring together species interactions and environmental conditions, the assessment of possible forms of information gaps in the habitat type it can be crucial to ensure reliable estimates and conservation actions.

References

Adamo, Martino, Matteo Chialva, Jacopo Calevo, Filippo Bertoni, Kingsley Dixon, and Stefano Mammola. 2021. 'Plant Scientists' Research Attention Is Skewed towards Colourful, Conspicuous and Broadly Distributed Flowers'. *Nature Plants* 7 (5): 574–78. https://doi.org/10.1038/s41477-021-00912-2.

Bayraktarov, Elisa, Glenn Ehmke, James O'Connor, Emma L. Burns, Hoang A. Nguyen, Louise McRae, Hugh P. Possingham, and David B. Lindenmayer. 2019. 'Do Big Unstructured Biodiversity Data Mean More Knowledge?' *Frontiers in Ecology and Evolution* 6 (January):239. https://doi.org/10.3389/fevo.2018.00239.

Bazzichetto, Manuele, Jonathan Lenoir, Daniele Da Re, Enrico Tordoni, Duccio Rocchini, Marco Malavasi, Vojtech Barták, and Marta Gaia Sperandii. 2023. 'Sampling Strategy Matters to Accurately Estimate Response Curves' Parameters in Species Distribution Models'. *Global Ecology and Biogeography* 32 (10): 1717–29. https://doi.org/10.1111/geb.13725.

Beck, Jan, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. 2014. 'Spatial Bias in the GBIF Database and Its Effect on Modeling Species' Geographic Distributions'. *Ecological Informatics* 19:10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Boakes, Elizabeth H., Philip J. K. McGowan, Richard A. Fuller, Ding Chang-qing, Natalie E. Clark, Kim O'Connor, and Georgina M. Mace. 2010. 'Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data'. *PLoS Biology* 8 (6): e1000385. https://doi.org/10.1371/journal.pbio.1000385.

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton University Press. https://doi.org/10.1515/9781400840908.

Boyd, Robin J., Gary D. Powney, Claire Carvell, and Oliver L. Pescott. 2021. 'OccAssess: An R Package for Assessing Potential Biases in Species Occurrence Data'. *Ecology and Evolution* 11 (22): 16177–87. https://doi.org/10.1002/ece3.8299.

Boyd, Robin J., Marcelo A. Aizen, Rodrigo M. Barahona-Segovia, Luis Flores-Prado, Francisco E. Fontúrbel, Tiago M. Francoy, Manuel Lopez-Aliste, et al. 2022. 'Inferring Trends in Pollinator

Distributions across the Neotropics from Publicly Available Data Remains Challenging despite Mobilization Efforts'. Edited by Yoan Fourcade. *Diversity and Distributions* 28 (7): 1404–15. https://doi.org/10.1111/ddi.13551.

Boyd, Robin J., Gavin B. Stewart, and Oliver L. Pescott. 2024. 'Descriptive Inference Using Large, Unrepresentative Nonprobability Samples: An Introduction for Ecologists'. *Ecology* 105 (2): e4214. https://doi.org/10.1002/ecy.4214.

Bowler, Diana E., Robin J. Boyd, Corey T. Callaghan, Robert A. Robinson, Nick J. B. Isaac, and Michael J. O. Pocock. 2024. 'Treating Gaps and Biases in Biodiversity Data as a Missing Data Problem'. *Biological Reviews*, August, brv.13127. https://doi.org/10.1111/brv.13127.

Bruelheide, Helge, Jürgen Dengler, Borja Jiménez-Alfaro, Oliver Purschke, Stephan M. Hennekens, Milan Chytrý, Valério D. Pillar, et al. 2019. 'SPlot – A New Tool for Global Vegetation Analyses'. Edited by Alessandro Chiarucci. *Journal of Vegetation Science* 30 (2): 161–86. https://doi.org/10.1111/jvs.12710.

Callaghan, Corey T., Alistair G. B. Poore, Max Hofmann, Christopher J. Roberts, and Henrique M. Pereira. 2021. 'Large-Bodied Birds Are over-Represented in Unstructured Citizen Science Data'. *Scientific Reports* 11 (1): 19073. https://doi.org/10.1038/s41598-021-98584-7.

Campbell, Peter. 2024. 'Interpreting and Georeferencing the Concept of "Near" in Biodiversity Records'. *Biodiversity Informatics* 18 (May). https://doi.org/10.17161/qe3d5373.

Cayuela, Luis, Íñigo Granzow-de La Cerda, Fabio S. Albuquerque, and Duncan J. Golicher. 2012. 'Taxonstand: An r Package for Species Names Standardisation in Vegetation Databases'. *Methods in Ecology and Evolution* 3 (6): 1078–83. https://doi.org/10.1111/j.2041-210X.2012.00232.x.

Cazzolla Gatti, Roberto, Peter B. Reich, Javier G. P. Gamarra, Tom Crowther, Cang Hui, Albert Morera, Jean-Francois Bastin, et al. 2022. 'The Number of Tree Species on Earth'. *Proceedings of the National Academy of Sciences* 119 (6): e2115329119. https://doi.org/10.1073/pnas.2115329119.

Chapman, Melissa, Benjamin R. Goldstein, Christopher J. Schell, Justin S. Brashares, Neil H. Carter, Diego Ellis-Soto, Hilary Oliva Faxon, et al. 2024. 'Biodiversity Monitoring for a Just Planetary Future'. *Science* 383 (6678): 34–36. https://doi.org/10.1126/science.adh8874.

Chao, Anne, and Lou Jost. 2012. 'Coverage-based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size'. *Ecology* 93 (12): 2533–47. https://doi.org/10.1890/11-1952.1.

Chao, Anne, Yasuhiro Kubota, David Zelený, Chun-Huo Chiu, Ching-Feng Li, Buntarou Kusumoto, Moriaki Yasuhara, et al. 2020. 'Quantifying Sample Completeness and Comparing Diversities among Assemblages'. *Ecological Research* 35 (2): 292–314. https://doi.org/10.1111/1440-1703.12102.

Chesshire, Paige R., Erica E. Fischer, Nicolas J. Dowdy, Terry L. Griswold, Alice C. Hughes, Michael C. Orr, John S. Ascher, et al. 2023. 'Completeness Analysis for over 3000 United States Bee Species Identifies Persistent Data Gap'. *Ecography* 2023 (5): e06584. https://doi.org/10.1111/ecog.06584.

Chytrý, Milan. 2001. 'Phytosociological Data Give Biased Estimates of Species Richness'. *Journal of Vegetation Science* 12 (3): 441–44. https://doi.org/10.1111/j.1654-1103.2001.tb00190.x.

Chytrý, Milan, Lubomír Tichý, Stephan M. Hennekens, and Joop H.J. Schaminée. 2014. 'Assessing Vegetation Change Using Vegetation-plot Databases: A Risky Business'. Edited by Jürgen Dengler. *Applied Vegetation Science* 17 (1): 32–41. https://doi.org/10.1111/avsc.12050.

Cosentino, Francesca, and Luigi Maiorano. 2021. 'Is Geographic Sampling Bias Representative of Environmental Space?' *Ecological Informatics* 64:101369. https://doi.org/10.1016/j.ecoinf.2021.101369.

Da Re, Daniele, Enrico Tordoni, Jonathan Lenoir, Jonas J. Lembrechts, Sophie O. Vanwambeke, Duccio Rocchini, and Manuele Bazzichetto. 2023. 'USE It: Uniformly Sampling Pseudo-absences within the Environmental Space for Applications in Habitat Suitability Models'. *Methods in Ecology and Evolution* 14 (11): 2873–87. https://doi.org/10.1111/2041-210X.14209.

Daru, Barnabas H., and Jordan Rodriguez. 2023. 'Mass Production of Unvouchered Records Fails to Represent Global Biodiversity Patterns'. *Nature Ecology & Evolution* 7 (6): 816–31. https://doi.org/10.1038/s41559-023-02047-3.

Dyderski, Marcin K., Sonia Paź, Lee E. Frelich, and Andrzej M. Jagodziński. 2018. 'How Much Does Climate Change Threaten European Forest Tree Species Distributions?' *Global Change Biology* 24 (3): 1150–63. https://doi.org/10.1111/gcb.13925.

Elith, Jane, and John R. Leathwick. 2009. 'Species Distribution Models: Ecological Explanation and Prediction Across Space and Time'. *Annual Review of Ecology, Evolution, and Systematics* 40 (1): 677–97. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Feng, Tong, Shilin Chen, Zhongke Feng, Chaoyong Shen, and Yi Tian. 2021. 'Effects of Canopy and Multi-Epoch Observations on Single-Point Positioning Errors of a GNSS in Coniferous and Broadleaved Forests'. *Remote Sensing* 13 (12): 2325. https://doi.org/10.3390/rs13122325.

Feng, Xiao, Brian J. Enquist, Daniel S. Park, Brad Boyle, David D. Breshears, Rachael V. Gallagher, Aaron Lien, et al. 2022. 'A Review of the Heterogeneous Landscape of Biodiversity Databases: Opportunities and Challenges for a Synthesized Biodiversity Knowledge Base'. *Global Ecology and Biogeography* 31 (7): 1242–60. https://doi.org/10.1111/geb.13497.

Ferreira, Matheus Pinheiro, Danilo Roberti Alves De Almeida, Daniel De Almeida Papa, Juliano Baldez Silva Minervino, Hudson Franklin Pessoa Veras, Arthur Formighieri, Caio Alexandre Nascimento Santos, Marcio Aurélio Dantas Ferreira, Evandro Orfanó Figueiredo, and Evandro José Linhares Ferreira. 2020. 'Individual Tree Detection and Species Classification of Amazonian Palms Using UAV Images and Deep Learning'. *Forest Ecology and Management* 475:118397. https://doi.org/10.1016/j.foreco.2020.118397.

Führding-Potschkat, Petra, Holger Kreft, and Stefanie M. Ickert-Bond. 2022. 'Influence of Different Data Cleaning Solutions of Point-occurrence Records on Downstream Macroecological Diversity Models'. *Ecology and Evolution* 12 (8): e9168. https://doi.org/10.1002/ece3.9168.

García-Roselló, Emilio, Cástor Guisande, Juergen Heine, Patricia Pelayo-Villamil, Ana Manjarrés-Hernández, Luis González Vilas, Jacinto González-Dacosta, Antonio Vaamonde, and Carlos Granado-Lorencio. 2014. 'Using MODESTR to Download, Import and Clean Species

Distribution Records'. Edited by David Orme. *Methods in Ecology and Evolution* 5 (7): 708–13. https://doi.org/10.1111/2041-210X.12209.

García-Roselló, Emilio, Jacinto González-Dacosta, and Jorge 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. https://doi.org/10.1016/j.biocon.2023.110118.

Gábor, Lukáš, Vítězslav Moudrý, Vincent Lecours, Marco Malavasi, Vojtěch Barták, Michal Fogl, Petra Šímová, Duccio Rocchini, and Tomáš Václavík. 2020. 'The Effect of Positional Error on Fine Scale Species Distribution Models Increases for Specialist Species'. *Ecography* 43 (2): 256–69. https://doi.org/10.1111/ecog.04687.

GBIF: The Global Biodiversity Information Facility (2024) What is GBIF?. Available from https://www.gbif.org/what-is-gbif [13 January 2020]

Girardello, Marco, Anna Chapman, Roger Dennis, Lauri Kaila, Paulo A.V. Borges, and Andrea Santangeli. 2019. 'Gaps in Butterfly Inventory Data: A Global Analysis'. *Biological Conservation* 236:289–95. https://doi.org/10.1016/j.biocon.2019.05.053.

Graham, Catherine H, Jane Elith, Robert J Hijmans, Antoine Guisan, A Townsend Peterson, Bette A Loiselle, and The Nceas Predicting Species Distributions Working Group. 2008. 'The Influence of Spatial Errors in Species Occurrence Data Used in Distribution Models'. *Journal of Applied Ecology* 45 (1): 239–47. https://doi.org/10.1111/j.1365-2664.2007.01408.x.

Hofmeester, Tim R., Joris P. G. M. Cromsigt, John Odden, Henrik Andrén, Jonas Kindberg, and John D. C. Linnell. 2019. 'Framing Pictures: A Conceptual Framework to Identify and Correct for Biases in Detection Probability of Camera Traps Enabling Multi-species Comparison'. *Ecology and Evolution* 9 (4): 2320–36. https://doi.org/10.1002/ece3.4878.

Hortal, Joaquín, Francesco De Bello, José Alexandre F. Diniz-Filho, Thomas M. Lewinsohn, Jorge M. Lobo, and Richard J. Ladle. 2015. 'Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity'. *Annual Review of Ecology, Evolution, and Systematics* 46 (1): 523–49. https://doi.org/10.1146/annurev-ecolsys-112414-054400.

Hughes, Alice C., Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, and Huijie Qiao. 2021. 'Sampling Biases Shape Our View of the Natural World'. *Ecography* 44 (9): 1259–69. https://doi.org/10.1111/ecog.05926.

Hugo, Sanet, and Res Altwegg. 2017. 'The Second Southern African Bird Atlas Project: Causes and Consequences of Geographical Sampling Bias'. *Ecology and Evolution* 7 (17): 6839–49. https://doi.org/10.1002/ece3.3228.

Hüllermeier, Eyke, and Willem Waegeman. 2021. 'Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods'. *Machine Learning* 110 (3): 457–506. https://doi.org/10.1007/s10994-021-05946-3.

ISO., I., & OIML, B. 1993. *Guide to the Expression of Uncertainty in Measurement*. Aenor.

IUCN 2024. *The IUCN Red List of Threatened Species. Version 2024-1*. http://www.iucnredlist.org.

Jetz, Walter, Melodie A. McGeoch, Robert Guralnick, Simon Ferrier, Jan Beck, Mark J. Costello, Miguel Fernandez, et al. 2019. 'Essential Biodiversity Variables for Mapping and Monitoring Species Populations'. *Nature Ecology & Evolution* 3 (4): 539–51. https://doi.org/10.1038/s41559-019-0826-1.

Kearney, Michael, and Warren Porter. 2009. 'Mechanistic Niche Modelling: Combining Physiological and Spatial Data to Predict Species' Ranges'. *Ecology Letters* 12 (4): 334–50. https://doi.org/10.1111/j.1461-0248.2008.01277.x.

Kellenberger, Benjamin, Diego Marcos, and Devis Tuia. 2018. 'Detecting Mammals in UAV Images: Best Practices to Address a Substantially Imbalanced Dataset with Deep Learning'. *Remote Sensing of Environment* 216:139–53. https://doi.org/10.1016/j.rse.2018.06.028.

Kühl, Hjalmar S., Diana E. Bowler, Lukas Bösch, Helge Bruelheide, Jens Dauber, David. Eichenberg, Nico Eisenhauer, et al. 2020. 'Effective Biodiversity Monitoring Needs a Culture of Integration'. *One Earth* 3 (4): 462–74. https://doi.org/10.1016/j.oneear.2020.09.010.

La Sorte, Frank A., and Marius Somveille. 2020. 'Survey Completeness of a Global Citizen-science Database of Bird Occurrence'. *Ecography* 43 (1): 34–43. https://doi.org/10.1111/ecog.04632.

Latombe, Guillaume, Petr Pyšek, Jonathan M. Jeschke, Tim M. Blackburn, Sven Bacher, César Capinha, Mark J. Costello, et al. 2017. 'A Vision for Global Monitoring of Biological Invasions'. *Biological Conservation* 213:295–308. https://doi.org/10.1016/j.biocon.2016.06.013.

Lobo, Jorge M., Alberto Jiménez-Valverde, and Joaquín Hortal. 2010. 'The Uncertain Nature of Absences and Their Importance in Species Distribution Modelling'. *Ecography* 33 (1): 103–14. https://doi.org/10.1111/j.1600-0587.2009.06039.x.

Maestre, Fernando T., Matthew A. Bowker, Cristina Escolar, María D. Puche, Santiago Soliveres, Sara Maltez-Mouro, Pablo García-Palacios, Andrea P. Castillo-Monroy, Isabel Martínez, and Adrián Escudero. 2010. 'Do Biotic Interactions Modulate Ecosystem Functioning along Stress Gradients? Insights from Semi-Arid Plant and Biological Soil Crust Communities'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1549): 2057–70. https://doi.org/10.1098/rstb.2010.0016.

Mair, Louise, and Alejandro Ruete. 2016. 'Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa'. Edited by Judi Hewitt. *PLOS ONE* 11 (1): e0147796. https://doi.org/10.1371/journal.pone.0147796.

Maldonado, Carla, Carlos I. Molina, Alexander Zizka, Claes Persson, Charlotte M. Taylor, Joaquina Albán, Eder Chilquillo, Nina Rønsted, and Alexandre Antonelli. 2015. 'Estimating Species Diversity and Distribution in the Era of B Ig D Ata: To What Extent Can We Trust Public Databases?' *Global Ecology and Biogeography* 24 (8): 973–84. https://doi.org/10.1111/geb.12326.

Marcer, Arnald, Arthur D. Chapman, John R. Wieczorek, F. Xavier Picó, Francesc Uribe, John Waller, and Arturo H. Ariño. 2022. 'Uncertainty Matters: Ascertaining Where Specimens in Natural History Collections Come from and Its Implications for Predicting Species Distributions'. *Ecography* 2022 (9): e06025. https://doi.org/10.1111/ecog.06025.

Meyer, Carsten, Holger Kreft, Robert Guralnick, and Walter Jetz. 2015. 'Global Priorities for an Effective Information Basis of Biodiversity Distributions'. *Nature Communications* 6 (1): 8221. https://doi.org/10.1038/ncomms9221.

Meyer, Carsten, Patrick Weigelt, and Holger Kreft. 2016. 'Multidimensional Biases, Gaps and Uncertainties in Global Plant Occurrence Information'. Edited by Janneke Hille Ris Lambers. *Ecology Letters* 19 (8): 992–1006. https://doi.org/10.1111/ele.12624.

Monsarrat, Sophie, Andre F. Boshoff, and Graham I. H. Kerley. 2019. 'Accessibility Maps as a Tool to Predict Sampling Bias in Historical Biodiversity Occurrence Records'. *Ecography* 42 (1): 125–36. https://doi.org/10.1111/ecog.03944.

Moudrý, Vítězslav, Manuele Bazzichetto, Ruben Remelgado, Rodolphe Devillers, Jonathan Lenoir, Rubén G. Mateo, Jonas J. Lembrechts, et al. 2024. 'Optimising Occurrence Data in Species Distribution Models: Sample Size, Positional Uncertainty, and Sampling Bias Matter'. *Ecography*, August, e07294. https://doi.org/10.1111/ecog.07294.

Pärtel, Meelis, Robert Szava-Kovats, and Martin Zobel. 2011. 'Dark Diversity: Shedding Light on Absent Species'. *Trends in Ecology & Evolution* 26 (3): 124–28. https://doi.org/10.1016/j.tree.2010.12.004.

Pereira, H. M., S. Ferrier, M. Walters, G. N. Geller, R. H. G. Jongman, R. J. Scholes, M. W. Bruford, et al. 2013. 'Essential Biodiversity Variables'. *Science* 339 (6117): 277–78. https://doi.org/10.1126/science.1229931.

Phillips, Steven J., Miroslav Dudík, Jane Elith, Catherine H. Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. 2009. 'Sample Selection Bias and Presence-only Distribution Models: Implications for Background and Pseudo-absence Data'. *Ecological Applications* 19 (1): 181–97. https://doi.org/10.1890/07-2153.1.

Pimm, Stuart L., Gareth J. Russell, John L. Gittleman, and Thomas M. Brooks. 1995. 'The Future of Biodiversity'. *Science* 269 (5222): 347–50. https://doi.org/10.1126/science.269.5222.347.

Rocchini, Duccio, Enrico Tordoni, Elisa Marchetto, Matteo Marcantonio, A. Márcia Barbosa, Manuele Bazzichetto, Carl Beierkuhnlein, et al. 2023. 'A Quixotic View of Spatial Bias in Modelling the Distribution of Species and Their Diversity'. *Npj Biodiversity* 2 (1): 10. https://doi.org/10.1038/s44185-023-00014-6.

Ronquillo, Cristina, Juliana Stropp, Nagore G. Medina, and Joaquin Hortal. 2023. 'Exploring the Impact of Data Curation Criteria on the Observed Geographical Distribution of Mosses'. *Ecology and Evolution* 13 (12): e10786. https://doi.org/10.1002/ece3.10786.

Ronquillo, Cristina, Juliana Stropp, and Joaquin Hortal. 2024. 'OCCUR Shiny Application: A User-friendly Guide for Curating Species Occurrence Records'. *Methods in Ecology and Evolution* 15 (5): 816–23. https://doi.org/10.1111/2041-210X.14271.

Ronk, Argo, Robert Szava-Kovats, and Meelis Pärtel. 2015. 'Applying the Dark Diversity Concept to Plants at the European Scale'. *Ecography* 38 (10): 1015–25. https://doi.org/10.1111/ecog.01236.

Sabatini, Francesco Maria, Jonathan Lenoir, Tarek Hattab, Elise Aimee Arnst, Milan Chytrý, Jürgen Dengler, Patrice De Ruffray, et al. 2021. 'SPlotOpen – An Environmentally Balanced, Open-access, Global Dataset of Vegetation Plots'. *Global Ecology and Biogeography* 30 (9): 1740–64. https://doi.org/10.1111/geb.13346.

Schad, Lukas, and Julia Fischer. 2023. 'Opportunities and Risks in the Use of Drones for Studying Animal Behaviour'. *Methods in Ecology and Evolution* 14 (8): 1864–72. https://doi.org/10.1111/2041-210X.13922.

Schulze, Ernst-Detlef, Erwin Beck, Nina Buchmann, Stephan Clemens, Klaus Müller-Hohenstein, and Michael Scherer-Lorenzen. 2019. 'Interactions Between Plants, Plant Communities and the Abiotic and Biotic Environment: With Contributions from C. F. Dormann and H. M. Schaefer'. In *Plant Ecology*, by Ernst-Detlef Schulze, Erwin Beck, Nina Buchmann, Stephan Clemens, Klaus Müller-Hohenstein, and Michael Scherer-Lorenzen, 689–741. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-56233-8_19.

Shirey, Vaughn, Michael W. Belitz, Vijay Barve, and Robert Guralnick. 2021. 'A Complete Inventory of North American Butterfly Occurrence Data: Narrowing Data Gaps, but Increasing Bias'. *Ecography* 44 (4): 537–47. https://doi.org/10.1111/ecog.05396.

Speed, James D. M., Mika Bendiksby, Anders G. Finstad, Kristian Hassel, Anders L. Kolstad, and Tommy Prestø. 2018. 'Contrasting Spatial, Temporal and Environmental Patterns in Observation and Specimen Based Species Occurrence Data'. Edited by Ulrich Joger. *PLOS ONE* 13 (4): e0196417. https://doi.org/10.1371/journal.pone.0196417.

Srivastava, V., V. Lafond, and V. C. Griess. 2019. 'Species Distribution Models (SDM): Applications, Benefits and Challenges in Invasive Species Management.' *CABI Reviews*, April, 1–13. https://doi.org/10.1079/PAVSNNR201914020.

Stropp, Juliana, Richard James Ladle, Thaise Emilio, Thainá Lessa, and Joaquín Hortal. 2022. 'Taxonomic Uncertainty and the Challenge of Estimating Global Species Richness'. *Journal of Biogeography* 49 (9): 1654–56. https://doi.org/10.1111/jbi.14463.

Sumner, Seirian, Peggy Bevan, Adam G. Hart, and Nicholas J.B. Isaac. 2019. 'Mapping Species Distributions in 2 Weeks Using Citizen Science'. Edited by Simon Leather. *Insect Conservation and Diversity* 12 (5): 382–88. https://doi.org/10.1111/icad.12345.

Terryn, Louise, Kim Calders, Mathias Disney, Niall Origo, Yadvinder Malhi, Glenn Newnham, Pasi Raumonen, Markku Å Kerblom, and Hans Verbeeck. 2020. 'Tree Species Classification Using Structural Features Derived from Terrestrial Laser Scanning'. *ISPRS Journal of Photogrammetry and Remote Sensing* 168:170–81. https://doi.org/10.1016/j.isprsjprs.2020.08.009.

Testolin, Riccardo, Fabio Attorre, Vanessa Bruzzaniti, Riccardo Guarino, Borja Jiménez-Alfaro, Michele Lussu, Stefano Martellos, et al. 2024. 'Plant Species Richness Hotspots and Related Drivers across Spatial Scales in Small Mediterranean Islands'. *Journal of Systematics and Evolution* 62 (2): 242–56. https://doi.org/10.1111/jse.13034.

Tessarolo, Geiziane, Richard Ladle, Thiago Rangel, and Joaquin Hortal. 2017. 'Temporal Degradation of Data Limits Biodiversity Research'. *Ecology and Evolution* 7 (17): 6863–70. https://doi.org/10.1002/ece3.3259.

Tessarolo, Geiziane, Richard J. Ladle, Jorge M. Lobo, Thiago Fernando Rangel, and Joaquín Hortal. 2021. 'Using Maps of Biogeographical Ignorance to Reveal the Uncertainty in Distributional Data

Hidden in Species Distribution Models'. *Ecography* 44 (12): 1743–55. https://doi.org/10.1111/ecog.05793.

Tiago, Patrícia, Ana Ceia-Hasse, Tiago A. Marques, César Capinha, and Henrique M. Pereira. 2017. 'Spatial Distribution of Citizen Science Casuistic Observations for Different Taxonomic Groups'. *Scientific Reports* 7 (1): 12832. https://doi.org/10.1038/s41598-017-13130-8.

Tilman, David, Michael Clark, David R. Williams, Kaitlin Kimmel, Stephen Polasky, and Craig Packer. 2017. 'Future Threats to Biodiversity and Pathways to Their Prevention'. *Nature* 546 (7656): 73–81. https://doi.org/10.1038/nature22900.

Varela, Sara, Robert P. Anderson, Raúl García-Valdés, and Federico Fernández-González. 2014. 'Environmental Filters Reduce the Effects of Sampling Bias and Improve Predictions of Ecological Niche Models'. *Ecography* 37 (11): 1084–91. https://doi.org/10.1111/j.1600-0587.2013.00441.x.

Večeřa, Martin, Jan Divíšek, Jonathan Lenoir, Borja Jiménez-Alfaro, Idoia Biurrun, Ilona Knollová, Emiliano Agrillo, et al. 2019. 'Alpha Diversity of Vascular Plants in European Forests'. *Journal of Biogeography* 46 (9): 1919–35. https://doi.org/10.1111/jbi.13624.

Veloz, Samuel D. 2009. 'Spatially Autocorrelated Sampling Falsely Inflates Measures of Accuracy for Presence-only Niche Models'. *Journal of Biogeography* 36 (12): 2290–99. https://doi.org/10.1111/j.1365-2699.2009.02174.x.

Wiens, John J., and Joseph Zelinka. 2024. 'How Many Species Will Earth Lose to Climate Change?' *Global Change Biology* 30 (1): e17125. https://doi.org/10.1111/gcb.17125.

Wisz, Mary Susanne, Julien Pottier, W. Daniel Kissling, Loïc Pellissier, Jonathan Lenoir, Christian F. Damgaard, Carsten F. Dormann, et al. 2013. 'The Role of Biotic Interactions in Shaping Distributions and Realised Assemblages of Species: Implications for Species Distribution Modelling'. *Biological Reviews* 88 (1): 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x.

Wüest, Rafael O., Niklaus E. Zimmermann, Damaris Zurell, Jake M. Alexander, Susanne A. Fritz, Christian Hof, Holger Kreft, et al. 2020. 'Macroecology in the Age of Big Data – Where to Go from Here?' *Journal of Biogeography* 47 (1): 1–12. https://doi.org/10.1111/jbi.13633.

Yang, Wenjing, Keping Ma, and Holger Kreft. 2013. 'Geographical Sampling Bias in a Large Distributional Database and Its Effects on Species Richness–Environment Models'. Edited by W. Daniel Kissling. *Journal of Biogeography* 40 (8): 1415–26. https://doi.org/10.1111/jbi.12108.

Yang, Wenjing, Keping Ma, and Holger Kreft. 2014. 'Environmental and Socio-economic Factors Shaping the Geography of Floristic Collections in C Hina'. *Global Ecology and Biogeography* 23 (11): 1284–92. https://doi.org/10.1111/geb.12225.

Zizka, Alexander, Daniele Silvestro, Tobias Andermann, Josué Azevedo, Camila Duarte Ritter, Daniel Edler, Harith Farooq, et al. 2019. 'CoordinateCleaner : Standardized Cleaning of Occurrence Records from Biological Collection Databases'. Edited by Tiago Quental. *Methods in Ecology and Evolution* 10 (5): 744–51. https://doi.org/10.1111/2041-210X.13152.

Zizka, Alexander, Fernanda Antunes Carvalho, Alice Calvente, Mabel Rocio Baez-Lizarazo, Andressa Cabral, Jéssica Fernanda Ramos Coelho, Matheus Colli-Silva, et al. 2020. 'No One-Size-Fits-All Solution to Clean GBIF'. *PeerJ* 8 (September):e9916. https://doi.org/10.7717/peerj.9916.

Zizka, Alexander, Alexandre Antonelli, and Daniele Silvestro. 2021a. '*Sampbias* , a Method for Quantifying Geographic Sampling Biases in Species Distribution Data'. *Ecography* 44 (1): 25–32. https://doi.org/10.1111/ecog.05102.

Zizka, Alexander, Oskar Rydén, Daniel Edler, Johannes Klein, Allison Perrigo, Daniele Silvestro, Sverker C. Jagers, Staffan I. Lindberg, and Alexandre Antonelli. 2021b. 'B IO -D EM , a Tool to Explore the Relationship between Biodiversity Data Availability and Socio-political Conditions in Time and Space'. *Journal of Biogeography* 48 (11): 2715–26. https://doi.org/10.1111/jbi.14256.

Chapter 1

Testing the effect of sample prevalence and sampling methods on probability- and favourability based SDMs

Elisa Marchetto[1], Daniele Da Re[2], Enrico Tordoni[3], Manuele Bazzichetto[4,5], Piero Zannini[1,7], Simone Celebrin[1], Ludovico Chieffallo[1], Marco Malavasi[5,6], Duccio Rocchini[1,5]

[1] BIOME Lab, Department of Biological, Geological and Environmental Sciences

(BiGeA), Alma Mater Studiorum University of Bologna, Bologna, Italy

[2] George Lemaître Center for Earth and Climate research, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, Belgium

[3]Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

[4]Department of Ecology and Global Change, Desertification Research Centre (CSIC/UV/GV), Valencia, Spain

[5] Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha - Suchdol Czech Republic

[6] Department of Chemistry, Physics, Mathematics and Natural Sciences, University of Sassari, Sassari, Italy

[7] LifeWatch Italy, Lecce, Italy

Published as:

Marchetto, Elisa, Daniele Da Re, Enrico Tordoni, Manuele Bazzichetto, Piero Zannini, Simone Celebrin, Ludovico Chieffallo, Marco Malavasi, and Duccio Rocchini. 2023. 'Testing the Effect of Sample Prevalence and Sampling Methods on Probability- and Favourability-Based SDMs'. *Ecological Modelling* 477:110248. https://doi.org/10.1016/j.ecolmodel.2022.110248.

Abstract

Predicting the occurrence probability of species is intrinsically dependent on the quality of the training dataset and, in particular, on the sample prevalence (i.e., the ratio between presences and absences). Whenever the number of presences and absences is not equal within the training dataset, the predictions deviate towards higher values as the sample prevalence increases and vice versa. As a result, probability models of species occurrence with different sample prevalence cannot be directly compared. The favourability concept was introduced to amend this limitation. Indeed, the favourability – i.e., the variation in the probability of occurrence regardless the sample prevalence – could reduce the degree of uncertainty when comparing species distributions despite different sample prevalences. To test this hypothesis, we simulated 50 virtual species and compared the predictive performance of four probability-based and favourability-based Species Distribution Models (GLM, GAM, RF, BRT) under a set of different prevalence values and sampling strategies (i.e, random and stratified sampling). Favourability-based models performed slightly better than probability-based models in predicting the species distribution over geographic space, confirming also their capability to reduce the variability of the predictions across different degrees of sample prevalence.

Keywords: biodiversity; ecological informatics; spatial bias; spatial ecology; species distribution modelling;

1 Introduction

Correlative Species Distribution Models (SDMs) relate species observations with spatial-explicit environmental variables (e.g., climatic, edaphic, etc.) allowing to (i) possibly infer the relationships between the species and its environment, and (ii) map the habitat suitability of a species across space and time (Guisan and Zimmermann 2000; Guisan and Thuiller 2005; Elith and Leathwick 2009; Guillera-Arroita et al. 2015; Guisan et al., 2017).

Different correlative modelling techniques can be employed depending on the type of the response variable attributed to the species: presence–absence (e.g., Generalized Linear Model (GLM), Generalized Additive Model (GAM), Random Forest (RF), Boosted Regression Trees (BRT)), presence-background (e.g., MaxEnt, ENFA, GARP), and presence-only methods (e.g., Bioclim, Domain) (Sillero et al., 2021). By using presence–absence data, correlative SDMs estimate the occurrence probability of a species given a combination of environmental variables. However, probability-based SDMs estimated with different sample prevalence values suffer from the limitation that they cannot be compared (e.g., by niche overlap (Warren, Glor, and Turelli 2008) or by Stacked Species Distribution Models (D'Amen et al. 2015; Schmitt et al. 2017)) among populations or species and either considering the same species in diverse times without creating any degree of error in the outputs. To overcome these limitations Real, Barbosa, and Vargas (2006) introduced the concept of favourability. They used Laplace's definition of probability (marquis de Laplace 1840), which is defined as the ratio of the number of favourable cases to the whole number of possible cases, to modify the 'ordinary' probability of species response and derive the favourability of species response. Favourability can be then calculated as follows:

$$F = \frac{\frac{P}{(1-P)}}{\frac{n1}{n0} + \frac{P}{(1-P)}} \quad (1)$$

being P the probability and $n1$ and $n0$ the respective number of presences and absences sampled where the ratio is defined as sample prevalence.

Strictly speaking, the occurrence probability of the species depends on both the predictors and the sample prevalence, whereas the species favourability is determined by correcting the estimated probabilities for the sample prevalence value, regardless of the statistical model used (Acevedo and Real 2012). Therefore, favourability can be a suitable approach to compare SDMs calibrated for species with unequal proportions of presences and absences within the sample (Real, Barbosa, and Vargas 2006).

However, despite this achievement, the species response curves estimated by SDMs are still conditioned by the collected data (i.e., presence samples, presence/absence samples, pseudo-absences, background points) used for the model calibration. Indeed, different survey strategies may influence the accuracy and the quality of predictions (Hirzel and Guisan 2002; Thibaud et al. 2014; Bazzichetto et al. 2022). Therefore, an efficient sampling method is crucial for avoiding spatial heterogeneity in the sampling intensity (e.g., incomplete sampling and over-sampling) of species occurrences and pseudo-absences/background points (Inman et al, 2021).

Accordingly, virtual species, i.e., simulated entities with known species-environment relationships, can represent a proper approach for testing new methodologies and practices in species distribution modelling before applying them to real data (Schweiger et al. 2016; Meynard, Leroy, and Kaplan 2019). Indeed, virtual species modelling promises to be a suitable approach for understanding the effect of sample prevalence and sampling method on probability- and favourability-based models, allowing to a priori known the species-environment relationships and to simulate multiple species.

In this study, we created 50 virtual species to test the effects of sample prevalence and sampling method on Species Distribution Models fitted by applying four modelling techniques (i.e., Generalized Linear Models, Generalized Additive Models, Random Forest and Boosted Regression Trees). Especially, we evaluated (i) the effect of sample prevalence and sampling method on model performances of probability- based and favourability-based SDMs; we tested (ii) the tendency of the favourability to maintain unchanged the prediction values across different degrees of sample prevalence in juxtaposition with the probability outcomes; finally, we investigated (iii) the impact of the sampling method on the probability-based and favourability-based SDMs.

2 Materials and methods

We generated 50 virtual species from bioclimatic variables. For each virtual species, we calibrated four modelling techniques (GLM, GAM, RF, BRT) using 1000 presence–absence points collected according to different sample prevalences (i.e., 0.2, 0.4, 0.5, 0.6, 0.8) and sampling method (random vs stratified). After having estimated the probability-based SDMs, we calculated the favourability-based SDMs. For each SDM we carried out different model evaluations (Coefficient of Variation, AUC, Continuous Boyce Index) and statistical tests (predictions' levels of dispersion, Kruskal–Wallis rank sum test, Dunn's test) Fig. 1.

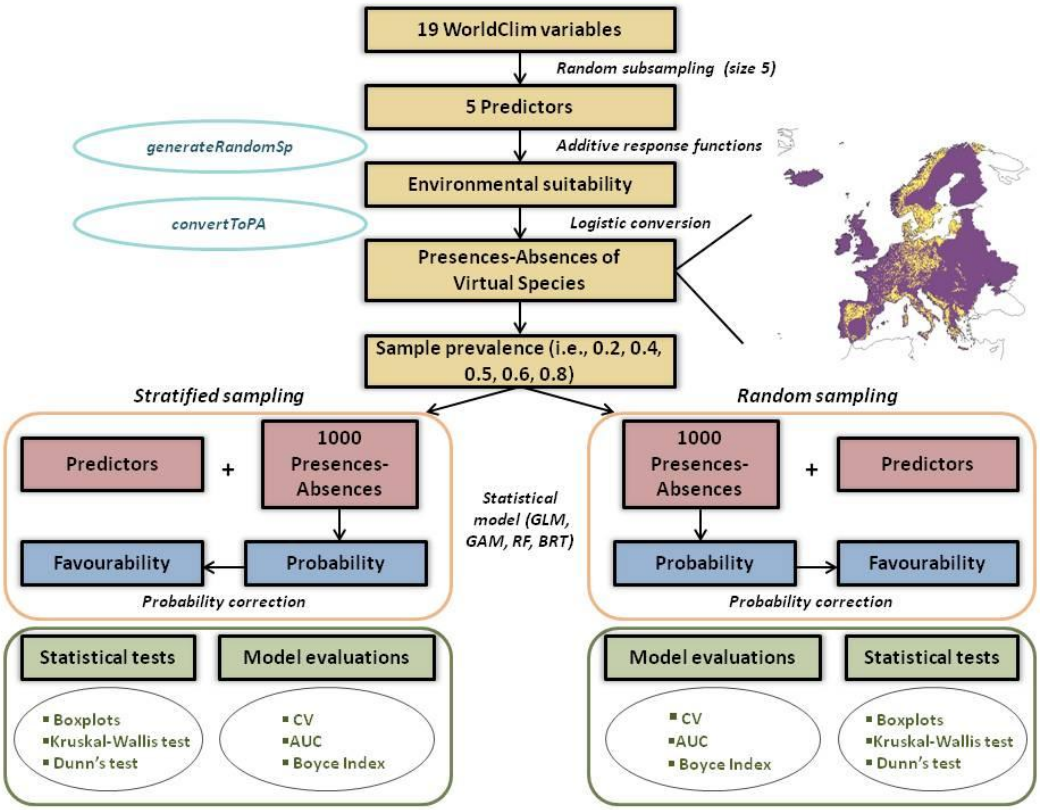


Fig. 1. Workflow methodology for a single virtual species. First, we generated the virtual species (yellow boxes), then we fitted the statistical models in accordance with the sampling method and the sample prevalence being used to derive probability-based and favourability-based SDMs (orange square). Finally, we evaluated the models and tested the predictions with different statistical analyses (green square).

2.1 Generating virtual species

In order to compare favourability-based and probability-based SDMs we used virtual species that were created by the virtualspecies R package (Leroy et al. 2016). We derived a virtual species using a subset of the WorldClim Bioclimatic variables at the European extent. We used the generateRandomSp function to create the environmental suitability for the virtual species distribution which was generated from a random subsampling (5 replicates) of the 19 bioclimatic variables (https://www.worldclim.org/data/bioclim.html) with 10 arc-minutes of spatial

resolution. The environmental suitability was calculated using an additive approach to the response functions of each bioclimatic variable, where the possible types of response function implemented are ‘‘gaussian’’, ‘‘linear’’, ‘‘logistic’’ and ‘‘quadratic’’. The obtained environmental suitability was then rescaled between 0 and 1 (i.e., range of possible probability values of the virtual species distribution). We used the *convertToPA* function to convert the raster layer reporting the environmental suitability into a probability of occurrence; the weighted probability of occurrence was then used to sample the presence or absence in each cell. We transformed the environmental suitability with a logistic conversion setting α and β parameters that determine the shape of the logistic curve (Meynard and Kaplan 2013). β controls the inflexion point and α drives the ‘slope’ of the curve, the latter was set equal to -0.05 such that the function detects an appropriate conversion by testing different values of β ; the species prevalence, i.e., the proportion of sites occupied by the species (Meynard and Kaplan 2012), was fixed at 0.2.

2.2 Sampling methods

We sampled 1000 presence–absence points for each virtual species (Wisz et al. 2008; van Proosdij et al. 2016), according to the different sample prevalence (i.e., 0.2, 0.4, 0.5, 0.6 and 0.8), using two different sampling methods: a random sampling and a stratified sampling. The random approach (*sampleOccurrences* function) consisted in randomly selecting the coordinates of presence and absence points across the study area, which makes all points equally likely to be sampled. The stratified approach collected presences–absences points by overlapping a grid of 0.3 degree of spatial resolution across the geographic area. Afterwards, if any binary pixel value (1 or 0) belonging to each polygon was equal to 1, then all of them were set as presence (1) otherwise to absence (0). Finally, in accordance with the sample prevalence, we randomly sampled 1000 presence–absence points with coordinates respectively associated with the centroids of the spatial polygons.

2.3 Models settings

For each virtual species, we estimated probability-based SDMs using four different modelling techniques which were trained relying on two sampling methods and 5 sample prevalences. We used the following modelling techniques available in different R packages: GLM, GAM, RF and BRT. The generalized linear models were generated with the R functions provided by *FuzzySim* package (Barbosa 2015), the generalized additive models with *mgcv* package (Wood 2017), the random forest regressions with *ranger* package (Wright and Ziegler 2017) and the boosted regression trees with *dismo* package (Hijmans et al. 2022). GLM algorithms were set using the default parameters of *multGLM* function avoiding a selected removal of variables (step=FALSE and trim=FALSE); GAM algorithms were set by *gam* function using the default parameters of thin plate regression splines (smooth term s and smooth class bs=‘‘tp’’); RF algorithms were set using the default parameters of *ranger* function providing as variable importance mode the variance of the responses; BRT algorithms were set using *gbm.step* function assigning tree.complexity=5, bag.fraction=0.75, learning.rate=0.005. Finally, to convert probability predicted values to favourability we employed equation (1).

Hence, we obtained 4000 SDMs as follows: 50 virtual species x 5 sample prevalence values x 2 sampling methods (random vs stratified) x 4 modelling techniques x 2 strategies (favourability vs probability).

2.4 Models evaluation

We estimated the model performances of probability-based and favourability-based SDMs of 50 virtual species with the Continuous Boyce Index, a presence-only based analysis focused on model predictions that removes the dependence on the Presence/Absence ratio. Especially, it measures how much model predictions differ from a random distribution of the observed presences (Boyce et al. 2002; Hirzel et al. 2006).

Besides, the accuracy of the 50 virtual species' probability SDMs under different degrees of sample prevalence were estimated by calculating the Area Under the Curve (AUC) of the receiver operator characteristic (ROC).

Furthermore, for a single virtual species, we evaluated the variability of the predictions (i.e, the variability of the pixels) which was calculated as Coefficient of Variation (CV) of the probability and favourability predictions according to the change of sample prevalence. We also calculated the difference between the coefficients of variations of probability and favourability (i.e., CV probability - CV favourability) for each statistical model. The analysis was performed on multiple species in order to verify the consistency.

2.5 Statistical tests

The levels of dispersion of the predictions of 50 virtual species (for each statistical model) were compared by calculating the lower quartile qn (0.25) and the upper quartile qn (0.75). In addition, we carried out a Kruskal–Wallis rank sum test (Kruskal and Wallis 1952) for testing the evenness of SDMs across different sample prevalence degrees. The test was performed on favourability-based and probability-based predicted values of 50 virtual species for each sampling method and each statistical model comparing the sample prevalence groups. Eventually, we evaluated the effect of the sampling design on the favourability and the probability predicted values of 50 virtual species for each sample prevalence performing a posthoc pairwise comparisons using Dunn's test (Dunn 1964). The pairwise comparisons were carried out on a subsample of the favourability-based and the probability-based distribution values.

3 Results

3.1 Models performance evaluation

For more than half of the sample prevalences, the favourability model had slightly higher median Continuous Boyce index values (i.e., better performances) than the probability model for all of the statistical models and for both sampling methods, except for RF trained using a random sampling of presences and absences. Especially, GLM had higher performances using the favourability-based approach for both the sampling methods and for all of the sample prevalences. Overall, the

sampling method (i.e., random and stratified) did not have a great impact on the model performances Fig. 2.

Furthermore, for all probability-based SDMs over the set of sample prevalence, calibrated using both the random sampling method and the stratified sampling method, the model performances, estimated with the Area Under the Curve (AUC) of the receiver operator characteristic (ROC), had a good accuracy ranging between 0.80 and 0.95 (Appendix: Fig. S6–S9).

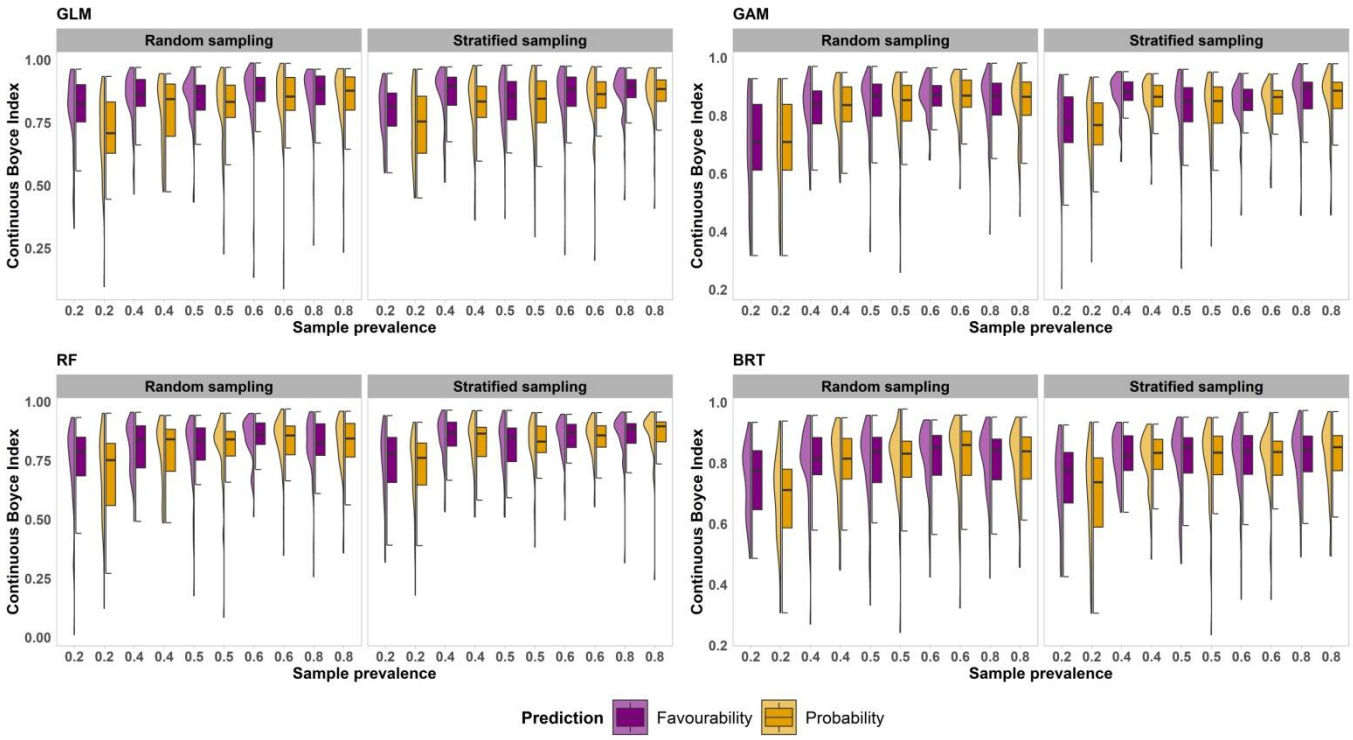


Fig. 2. Distribution between first and third quartiles of continuous Boyce index values of favourability-based and probability-based SDMs of 50 virtual species. The graphs show the distribution values of Continuous Boyce indices estimated by applying Generalized Linear Models, Generalized Additive Models, Random Forest and Boosted Regression Trees using random and stratified sampling methods.

3.2 Effect of the sample prevalence on the predictions

The favourability distribution values of 50 virtual species' predictions were steadier across the degrees of sample prevalence than the probability distribution values (Appendix: Fig. S2–S5). Nevertheless, the Kruskal–Wallis test proved that there were significant variations in both probability and favourability predicted values as the sample prevalence changes for both sampling methods and for all of the statistical models (Appendix: Tables S1–S2).

Besides, the variability of the predictions for a single species – i.e., the pixels variability – calculated as Coefficient of Variation across the sample prevalence values, showed higher stability (i.e., lower CV) for the favourability-based SDM both for the random sampling and for the

stratified sampling Figs. 3 and 4. Moreover, the difference between the pixels variability of the probability predictions and the pixels variability of the favourability predictions confirmed that the favourability SDM generates higher pixels stability as the sample prevalence changes. However, the generalized linear model showed a larger decrease in the pixels variability once the sample prevalence was removed from the probability predicted values Fig. 5.

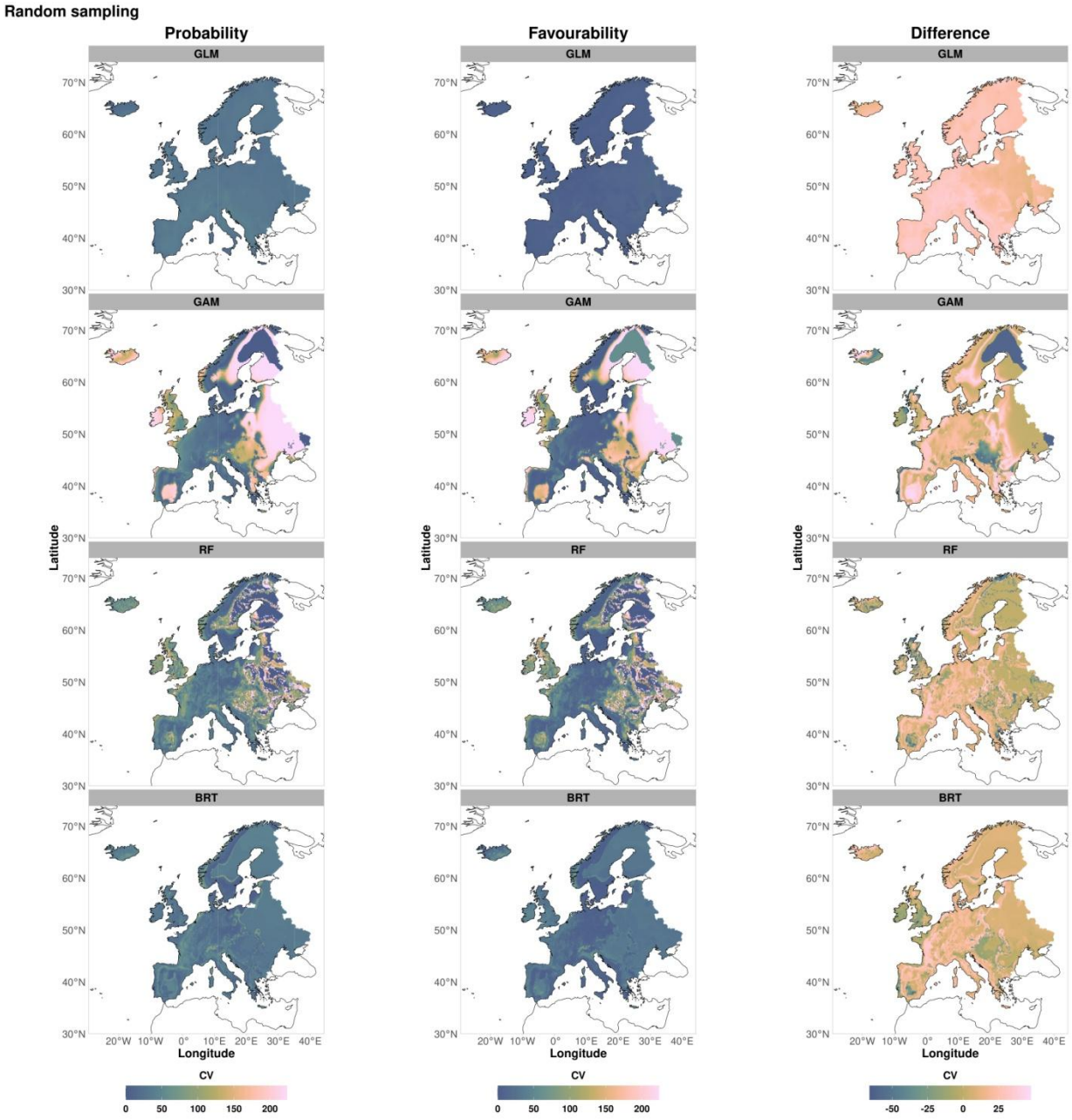


Fig. 3. Predictions variability, i.e., pixels variability, of a single virtual species calculated as Coefficient of Variation (CV) of favourability and probability SDMs related to a random sampling of presence–absence points. The left column shows the probability-based coefficients of variation for each statistical model (GLM, GAM, RF, BRT), the central column the favourability-based coefficients of variation, and the right column the difference values between the probability-based CV and the favourability-based CV for each statistical model.

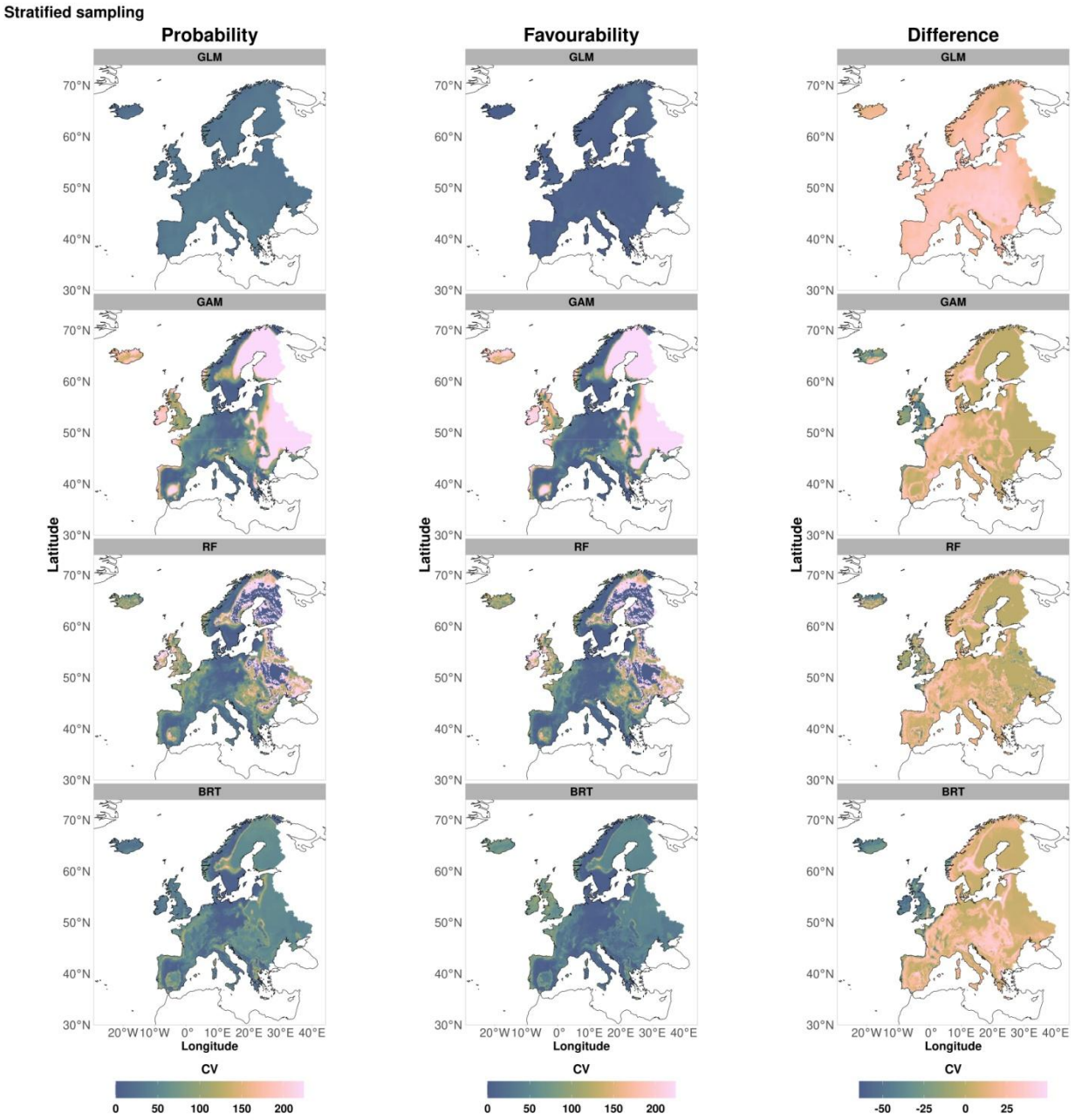


Fig. 4. Predictions variability, i.e. pixels variability, of a single virtual species calculated as Coefficient of Variation (CV) of favourability and probability SDMs related to a stratified sampling of presence–absence points. The left column shows the probability-based coefficients of variation for each statistical model (GLM, GAM, RF, BRT), the central column the favourability-based coefficients of variation, and the right column the difference values between the probability-based CV and the favourability-based CV for each statistical model.

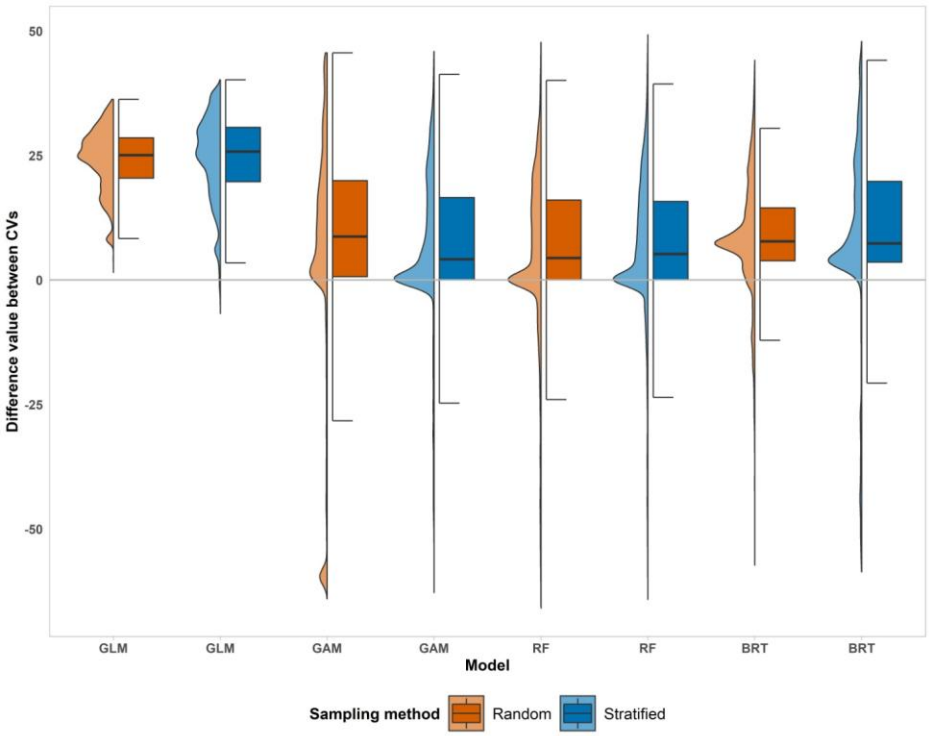


Fig. 5. Distribution values between first and third quartiles of the difference between probability-based and favourability-based coefficients of variation for each statistical model (i.e., GLM, GAM, RF, BRT) and sampling method (i.e., random and stratified sampling) for a single virtual species.

3.3 Effect of the sampling method on the predictions

Although the sampling methods did not determine a great difference in the range of the predicted values, the random sampling showed a lower range in comparison to SDMs estimated using the stratified sampling (Appendix: Fig. S2–S5). Besides, the Dunn's test proved that the sampling designs generated significantly different species predictions for all probability and favourability outcomes at the spatial scale (Appendix: Tables S3–S6).

4 Discussion

In this study we tested to which extent the favourability-based and the probability-based SDMs are affected by sample prevalence and sampling method.

Concerning models' performance, the Continuous Boyce index did not show a great difference in the performance efficiency between favourability and probability models. This behaviour could depend on the fact that the models have been calibrated with the default parameters in order to be extended to 50 different species. Indeed, several authors showed that the model parametrization has an impact on SDMs output (e.g., Fourcade 2021). However, for more than half of sample prevalences we considered, median Continuous Boyce Index values were slightly higher for favourability-based SDMs than for probability-based SDMs. Besides, although van Proosdij et al. (2016) and Tessarolo et al. (2021) report a linear relationship between the model performance

and the sample prevalence, our outcomes of AUCs indicate an independence of the accuracy of predictive models with respect to prevalence values (Guo et al. 2015).

Concerning the spatial variability of predictions of a single virtual species, the pixels variability across the degrees of sample prevalence was lower for the favourability-based SDMs than for the probability-based. By comparing the distribution values of 50 virtual species' predictions this pattern was also retained; favourability-based predictions showed steadier values across different sample prevalences than the probability-based predictions, although they retain a certain degree of variability. Indeed, the Kruskal Wallis rank sum tests highlighted a difference in the values of both probability and favourability predictions across the different degrees of sample prevalence. Hence, favourability-based SDMs do not maintain unchanged the prediction values across different sample prevalence values (Real, Barbosa, and Vargas 2006; Acevedo et al., 2010; Romero et al. 2019), since they are created with a posteriori removal of the sample prevalence after statistical model calibration, so that the correction is made on the probability predictions. Consequently, the favourability model does not lose any information about sample prevalence and, therefore, about species or species-environment interactions, since the statistical model is still dependent on the prevalence value. Furthermore, it allows obtaining more effective comparisons among SDMs of different species or populations and time scales as a consequence of a lower pixels variability. This makes the favourability an extremely powerful tool to broaden our understanding of ecological trends such as the ecological niches pattern between species (Pulido-Pastor et al. 2021), the environmental factors that favour the spread of an invasive species (Romero et al., 2014; Baquero et al., 2021) or an epidemiological vector (Aliaga-Samanez et al. 2021), the areal shift range under land and climate changes (Muñoz et al, 2005; Chamorro, Real, and Muñoz 2020).

However, it is of paramount importance to point out that the favourability-based SDMs are dependent on the extent of analysis being chosen (VanDerWal et al. 2009), as well as on the spatial resolution of the predictors (Sillero and Barbosa 2021), and it is unequivocally associated with the environmental features of the study area (Barbosa et al. 2009). On the other hand, the possible errors deriving from the modeling technique being chosen (Rocchini et al. 2017) can invalidate the overall performance and make the favourability SDMs matchless (Elith and Graham 2009). Moreover, although the benefits of favorability are promising, we cannot exclude the uncertainty determined by biased sample collections both in occurrence (Rocchini et al. 2011) and in background data (Phillips et al. 2009; Grimmett, Whitsed, and Horta 2020) which can affect the results of the modelling process (Leitão, Moreira, and Osborne 2011; Beck et al. 2014). Indeed, misleading or unstandardized sampling schemes can result in the so-called Wallacean shortfalls (Lomolino 2004; Hortal et al. 2015). For instance, biased sampling effort, as a consequence of survey preferences in proximity to roads, centres of research, infrastructures, or protected areas, may cause incomplete and distorted presences-absences samples (Oliveira et al. 2016; Ronquillo et al., 2020). Consequently, for those modelling procedures that ignore the sampling effort bias, the local density of occurrences of a species may be over- or under-estimated over space (Rocchini et al. 2017, 2019).

According to our results, the sampling method of presences and absences does not have a decisive impact on the predictions variability of favourabiliy-based and probability-based SDMs across the degrees of sample prevalences. However, the random sampling determined a greater uniformity around a narrower range of values (Appendix: Fig. S2–S5). Besides, Dunn's test confirmed that the sampling methods generate different prediction values at the spatial scale.

Broadly speaking, the sampling strategy we applied did not affect model performances. Indeed, the influence of the sampling design often depends on the intensity of bias of the samples used to train the model (e.g., location bias, geographical bias and so on) (Syfert et al. 2013) but, in our case, there was no source of bias in the sampling that could have considerably affected the accuracy. It has been shown as some sampling methods can actually reduce the effect of the bias and increase the model performance (Fourcade et al. 2014). However, Tessarolo et al. (2014) stated that the sampling method is not the most important factor affecting SDMs performance, even though they also concluded that the design may become more and more important as the spatial extent of the species' geographical distribution increases. However, the question of what effect the sampling method would have on the model calibration using presences and pseudo-absences rather than presences and absences remains still open (Barbet-Massin et al. 2012).

5 Conclusion

Favourability might provide an important contribution to map species distribution, especially for the fact that training datasets are often biased, as in the case of rare species of important conservation value. Indeed, favourability-based SDM, although it does not maintain unchanged the predictions values across different sample prevalences, proved to be effective in reducing their variability across different prevalence degrees compared to probability-based SDM. Besides, favourability model showed high model performances for all of the modelling techniques being applied. Therefore, according to our results, favourability-based SDM can definitely improve knowledge in community and population dynamics and provide useful tools for biogeography conservation allowing to achieve more effective comparisons among species distributions in space and their possible shifts over time. Nevertheless, being aware that the favourability is not independent of the uncertainty related to the sampling effort, the extent and the resolution of analysis, in a future study, these elements should also be considered. In our study, the sampling methods, i.e., random and stratified, revealed that they have a great impact neither in the variability of the predictions across the set of sample prevalence values nor in the performance of the models if no source of bias is present in the sampling. However, having proved the advantages of favourability-based SDM with virtual species, future studies on real species distribution models can be definitively promising in testing the real empirical power of favourability-based approach.

Declaration of conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

Data will be made available on request.

Acknowledgements

We are grateful to the handling editor and two anonymous reviewers for precious suggestions which helped us improving a previous version of the manuscript. We are also thankful to Arianna Ferrara for her graphics recommendations. Duccio Rocchini has received funding from the Project SHOWCASE (SHOWCASing synergies between agriculture, biodiversity and ecosystems services to help farmers capitalizing on native biodiversity) within the European Union Horizon 2020 Researcher and Innovation Programme under grant agreement No 862480. Piero Zannini has been supported by LifeWatch Italy through the project LifeWatchPLUS (CIR-01_00028).

Author Contributions

Elisa Marchetto: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. Daniele Da Re: Conceptualization, Methodology, Writing – review & editing. Enrico Tordoni: Conceptualization, Methodology, Writing – review & editing. Manuele Bazzichetto: Writing – review & editing. Piero Zannini: Writing – review & editing. Simone Celebrin: Conceptualization, Methodology. Ludovico Chieffallo: Writing – review & editing. Marco Malavasi: Writing – review & editing. Duccio Rocchini: Conceptualization, Methodology, Writing – review & editing, Project administration.

References

Acevedo, Pelayo, Alastair I. Ward, Raimundo Real, and Graham C. Smith. 2010. 'Assessing Biogeographical Relationships of Ecologically Related Species Using Favourability Functions: A Case Study on British Deer'. *Diversity and Distributions* 16 (4): 515–28. https://doi.org/10.1111/j.1472-4642.2010.00662.x.

Acevedo, Pelayo, and Raimundo Real. 2012. 'Favourability: Concept, Distinctive Characteristics and Potential Usefulness'. *Naturwissenschaften* 99 (7): 515–22. https://doi.org/10.1007/s00114-012-0926-0.

Aliaga-Samanez, Alisa, Marina Cobos-Mayo, Raimundo Real, Marina Segura, David Romero, Julia E. Fa, and Jesús Olivero. 2021. 'Worldwide Dynamic Biogeography of Zoonotic and Anthroponotic Dengue'. Edited by Michael R. Holbrook. *PLOS Neglected Tropical Diseases* 15 (6): e0009496. https://doi.org/10.1371/journal.pntd.0009496.

Baquero, Rocío A., A. Márcia Barbosa, Daniel Ayllón, Carlos Guerra, Enrique Sánchez, Miguel B. Araújo, and Graciela G. Nicola. 2021. 'Potential Distributions of Invasive Vertebrates in the Iberian Peninsula under Projected Changes in Climate Extreme Events'. Edited by Zhixin Zhang. *Diversity and Distributions* 27 (11): 2262–76. https://doi.org/10.1111/ddi.13401.

Barbet-Massin, Morgane, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. 2012. 'Selecting Pseudo-absences for Species Distribution Models: How, Where and How Many?' *Methods in Ecology and Evolution* 3 (2): 327–38. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Barbosa, A. Márcia, Raimundo Real, and J. Mario Vargas. 2009. 'Transferability of Environmental Favourability Models in Geographic Space: The Case of the Iberian Desman (Galemys Pyrenaicus) in Portugal and Spain'. *Ecological Modelling* 220 (5): 747–54. https://doi.org/10.1016/j.ecolmodel.2008.12.004.

Barbosa, A. Márcia. 2015. 'FuzzySim: Applying Fuzzy Logic to Binary Similarity Indices in Ecology'. Edited by Robert B. O'Hara. *Methods in Ecology and Evolution* 6 (7): 853–58. https://doi.org/10.1111/2041-210X.12372.

Bazzichetto, Manuele, Jonathan Lenoir, Daniele Da Re, Enrico Tordoni, Duccio Rocchini, Marco Malavasi, Vojtech Barták, and Marta Sperandii. 2022. 'Effect of Sampling Strategies on the Response Curves Estimated by Plant Species Distribution Models'. https://doi.org/10.32942/OSF.IO/RHYS3.

Beck, Jan, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. 2014. 'Spatial Bias in the GBIF Database and Its Effect on Modeling Species' Geographic Distributions'. *Ecological Informatics* 19:10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Boyce, Mark S, Pierre R Vernier, Scott E Nielsen, and Fiona K.A Schmiegelow. 2002. 'Evaluating Resource Selection Functions'. *Ecological Modelling* 157 (2–3): 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4.

Chamorro, Darío, Raimundo Real, and Antonio-Román Muñoz. 2020. 'Fuzzy Sets Allow Gaging the Extent and Rate of Species Range Shift Due to Climate Change'. *Scientific Reports* 10 (1): 16272. https://doi.org/10.1038/s41598-020-73509-y.

D'Amen, Manuela, Anne Dubuis, Rui F. Fernandes, Julien Pottier, Loïc Pellissier, and Antoine Guisan. 2015. 'Using Species Richness and Functional Traits Predictions to Constrain Assemblage Predictions from Stacked Species Distribution Models'. *Journal of Biogeography* 42 (7): 1255–66. https://doi.org/10.1111/jbi.12485.

Dunn, Olive Jean. 1964. 'Multiple Comparisons Using Rank Sums'. *Technometrics* 6 (3): 241–52. https://doi.org/10.1080/00401706.1964.10490181.

Elith, Jane, and Catherine H. Graham. 2009. 'Do They? How Do They? WHY Do They Differ? On Finding Reasons for Differing Performances of Species Distribution Models'. *Ecography* 32 (1): 66–77. https://doi.org/10.1111/j.1600-0587.2008.05505.x.

Elith, Jane, and John R. Leathwick. 2009. 'Species Distribution Models: Ecological Explanation and Prediction Across Space and Time'. *Annual Review of Ecology, Evolution, and Systematics* 40 (1): 677–97. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Fourcade, Yoan, Jan O. Engler, Dennis Rödder, and Jean Secondi. 2014. 'Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias'. Edited by John F. Valentine. *PLoS ONE* 9 (5): e97122. https://doi.org/10.1371/journal.pone.0097122.

Fourcade, Yoan. 2021. 'Fine-Tuning Niche Models Matters in Invasion Ecology. A Lesson from the Land Planarian Obama Nungara.' *Ecological Modelling* 457:109686. https://doi.org/10.1016/j.ecolmodel.2021.109686.

Grimmett, Liam, Rachel Whitsed, and Ana Horta. 2020. 'Presence-Only Species Distribution Models Are Sensitive to Sample Prevalence: Evaluating Models Using Spatial Prediction Stability and Accuracy Metrics'. *Ecological Modelling* 431:109194. https://doi.org/10.1016/j.ecolmodel.2020.109194.

Guisan, Antoine, and Niklaus E. Zimmermann. 2000. 'Predictive Habitat Distribution Models in Ecology'. *Ecological Modelling* 135 (2): 147–86. https://doi.org/10.1016/S0304-3800(00)00354-9.

Guisan, Antoine, and Wilfried Thuiller. 2005. 'Predicting Species Distribution: Offering More than Simple Habitat Models'. *Ecology Letters* 8 (9): 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x.

Guo, Chuanbo, Sovan Lek, Shaowen Ye, Wei Li, Jiashou Liu, and Zhongjie Li. 2015. 'Uncertainty in Ensemble Modelling of Large-Scale Species Distribution: Effects from Species Characteristics and Model Techniques'. *Ecological Modelling* 306:67–75. https://doi.org/10.1016/j.ecolmodel.2014.08.002.

Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2022. Dismo: Species distribution modeling. R package version 1.3-8.

Hirzel, Alexandre, and Antoine Guisan. 2002. 'Which Is the Optimal Sampling Strategy for Habitat Suitability Modelling'. *Ecological Modelling* 157 (2–3): 331–41. https://doi.org/10.1016/S0304-3800(02)00203-X.

Hirzel, Alexandre H., Gwenaëlle Le Lay, Véronique Helfer, Christophe Randin, and Antoine Guisan. 2006. 'Evaluating the Ability of Habitat Suitability Models to Predict Species Presences'. *Ecological Modelling* 199 (2): 142–52. https://doi.org/10.1016/j.ecolmodel.2006.05.017.

Hortal, Joaquín, Francesco De Bello, José Alexandre F. Diniz-Filho, Thomas M. Lewinsohn, Jorge M. Lobo, and Richard J. Ladle. 2015. 'Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity'. *Annual Review of Ecology, Evolution, and Systematics* 46 (1): 523–49. https://doi.org/10.1146/annurev-ecolsys-112414-054400.

Inman, Richard, Janet Franklin, Todd Esque, and Kenneth Nussear. 2021. 'Comparing Sample Bias Correction Methods for Species Distribution Modeling Using Virtual Species'. *Ecosphere* 12 (3): e03422. https://doi.org/10.1002/ecs2.3422.

Kruskal, William H., and W. Allen Wallis. 1952. 'Use of Ranks in One-Criterion Variance Analysis'. *Journal of the American Statistical Association* 47 (260): 583–621. https://doi.org/10.1080/01621459.1952.10483441.

Leitão, Pedro J., Francisco Moreira, and Patrick E. Osborne. 2011. 'Effects of Geographical Data Sampling Bias on Habitat Models of Species Distributions: A Case Study with Steppe Birds in Southern Portugal'. *International Journal of Geographical Information Science* 25 (3): 439–54. https://doi.org/10.1080/13658816.2010.531020.

Leroy, Boris, Christine N. Meynard, Céline Bellard, and Franck Courchamp. 2016. 'Virtualspecies, an R Package to Generate Virtual Species Distributions'. *Ecography* 39 (6): 599–607. https://doi.org/10.1111/ecog.01388.

Lomolino, Mark V. 'Conservation biogeography.' *Frontiers of Biogeography: new directions in the geography of nature* 293 (2004).

marquis de Laplace, P.S., 1840. Essai Philosophique sur Les Probabilités. Bachelier.

Meynard, Christine N., and David M. Kaplan. 2012. 'The Effect of a Gradual Response to the Environment on Species Distribution Modeling Performance'. *Ecography* 35 (6): 499–509. https://doi.org/10.1111/j.1600-0587.2011.07157.x.

Meynard, Christine N., and David M. Kaplan. 2013. 'Using Virtual Species to Study Species Distributions and Model Performance'. Edited by Miles Silman. *Journal of Biogeography* 40 (1): 1–8. https://doi.org/10.1111/jbi.12006.

Meynard, Christine N., Boris Leroy, and David M. Kaplan. 2019. 'Testing Methods in Species Distribution Modelling Using Virtual Species: What Have We Learnt and What Are We Missing?' *Ecography* 42 (12): 2021–36. https://doi.org/10.1111/ecog.04385.

Muñoz, A. Román, Raimundo Real, A. Márcia Barbosa, and J. Mario Vargas. 2005. 'Modelling the Distribution of Bonelli's Eagle in Spain: Implications for Conservation Planning'. *Diversity and Distributions* 11 (6): 477–86. https://doi.org/10.1111/j.1366-9516.2005.00188.x.

Oliveira, Ubirajara, Adriano Pereira Paglia, Antonio D. Brescovit, Claudio J. B. De Carvalho, Daniel Paiva Silva, Daniella T. Rezende, Felipe Sá Fortes Leite, et al. 2016. 'The Strong Influence of Collection Bias on Biodiversity Knowledge Shortfalls of B Razilian Terrestrial Biodiversity'. Edited by

Jeremy VanDerWal. *Diversity and Distributions* 22 (12): 1232–44. https://doi.org/10.1111/ddi.12489.

Phillips, Steven J., Miroslav Dudík, Jane Elith, Catherine H. Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. 2009. 'Sample Selection Bias and Presence-only Distribution Models: Implications for Background and Pseudo-absence Data'. *Ecological Applications* 19 (1): 181–97. https://doi.org/10.1890/07-2153.1.

Pulido-Pastor, Antonio, Ana Luz Márquez, José Carlos Guerrero, Enrique García-Barros, and Raimundo Real. 2021. 'Metapopulation Patterns of Iberian Butterflies Revealed by Fuzzy Logic'. *Insects* 12 (5): 392. https://doi.org/10.3390/insects12050392.

Real, Raimundo, A. Márcia Barbosa, and J. Mario Vargas. 2006. 'Obtaining Environmental Favourability Functions from Logistic Regression'. *Environmental and Ecological Statistics* 13 (2): 237–45. https://doi.org/10.1007/s10651-005-0003-3.

Rocchini, Duccio, Joaquín Hortal, Szabolcs Lengyel, Jorge M. Lobo, Alberto Jiménez-Valverde, Carlo Ricotta, Giovanni Bacaro, and Alessandro Chiarucci. 2011. 'Accounting for Uncertainty When Mapping Species Distributions: The Need for Maps of Ignorance'. *Progress in Physical Geography: Earth and Environment* 35 (2): 211–26. https://doi.org/10.1177/0309133311399491.

Rocchini, Duccio, Carol X. Garzon-Lopez, Matteo Marcantonio, Valerio Amici, Giovanni Bacaro, Lucy Bastin, Neil Brummitt, et al. 2017. 'Anticipating Species Distributions: Handling Sampling Effort Bias under a Bayesian Framework'. *Science of The Total Environment* 584–585:282–90. https://doi.org/10.1016/j.scitotenv.2016.12.038.

Rocchini, Duccio, Matteo Marcantonio, George Arhonditsis, Alessandro Lo Cacciato, Heidi C. Hauffe, and Kate S. He. 2019. 'Cartogramming Uncertainty in Species Distribution Models: A Bayesian Approach'. *Ecological Complexity* 38:146–55. https://doi.org/10.1016/j.ecocom.2019.04.002.

Romero, David, José C. Báez, Francisco Ferri-Yáñez, Jesús J. Bellido, and Raimundo Real. 2014. 'Modelling Favourability for Invasive Species Encroachment to Identify Areas of Native Species Vulnerability'. *The Scientific World Journal* 2014:1–9. https://doi.org/10.1155/2014/519710.

Romero, David, Jesús Olivero, Raimundo Real, and José Carlos Guerrero. 2019. 'Applying Fuzzy Logic to Assess the Biogeographical Risk of Dengue in South America'. *Parasites & Vectors* 12 (1): 428. https://doi.org/10.1186/s13071-019-3691-5.

Ronquillo, Cristina, Fernanda Alves-Martins, Vicente Mazimpaka, Thadeu Sobral-Souza, Bruno Vilela-Silva, Nagore G. Medina, and Joaquín Hortal. 2020. 'Assessing Spatial and Temporal Biases and Gaps in the Publicly Available Distributional Information of Iberian Mosses'. *Biodiversity Data Journal* 8 (September):e53474. https://doi.org/10.3897/BDJ.8.e53474.

Schmitt, Sylvain, Robin Pouteau, Dimitri Justeau, Florian De Boissieu, and Philippe Birnbaum. 2017. 'SSDM : An R Package to Predict Distribution of Species Richness and Composition Based on Stacked Species Distribution Models'. Edited by Nick Golding. *Methods in Ecology and Evolution* 8 (12): 1795–1803. https://doi.org/10.1111/2041-210X.12841.

Schweiger, Andreas H., Severin D. H. Irl, Manuel J. Steinbauer, Jürgen Dengler, and Carl Beierkuhnlein. 2016. 'Optimizing Sampling Approaches along Ecological Gradients'. Edited by Matthew Schofield. *Methods in Ecology and Evolution* 7 (4): 463–71. https://doi.org/10.1111/2041-210X.12495.

Sillero, Neftalí, Salvador Arenas-Castro, Urtzi Enriquez-Urzelai, Cândida Gomes Vale, Diana Sousa-Guedes, Fernando Martínez-Freiría, Raimundo Real, and A.Márcia Barbosa. 2021. 'Want to Model a Species Niche? A Step-by-Step Guideline on Correlative Ecological Niche Modelling'. *Ecological Modelling* 456:109671. https://doi.org/10.1016/j.ecolmodel.2021.109671.

Sillero, Neftalí, and A. Márcia Barbosa. 2021. 'Common Mistakes in Ecological Niche Models'. *International Journal of Geographical Information Science* 35 (2): 213–26. https://doi.org/10.1080/13658816.2020.1798968.

Syfert, Mindy M., Matthew J. Smith, and David A. Coomes. 2013. 'The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models'. Edited by David L. Roberts. *PLoS ONE* 8 (2): e55158. https://doi.org/10.1371/journal.pone.0055158.

Tessarolo, Geiziane, Thiago F. Rangel, Miguel B. Araújo, and Joaquín Hortal. 2014. 'Uncertainty Associated with Survey Design in Species Distribution Models'. Edited by Linda Beaumont. *Diversity and Distributions* 20 (11): 1258–69. https://doi.org/10.1111/ddi.12236.

Tessarolo, Geiziane, Jorge M. Lobo, Thiago Fernando Rangel, and Joaquín Hortal. 2021. 'High Uncertainty in the Effects of Data Characteristics on the Performance of Species Distribution Models'. *Ecological Indicators* 121:107147. https://doi.org/10.1016/j.ecolind.2020.107147.

Thibaud, Emeric, Blaise Petitpierre, Olivier Broennimann, Anthony C. Davison, and Antoine Guisan. 2014. 'Measuring the Relative Effect of Factors Affecting Species Distribution Model Predictions'. Edited by Robert B. OHara. *Methods in Ecology and Evolution* 5 (9): 947–55. https://doi.org/10.1111/2041-210X.12203.

VanDerWal, Jeremy, Luke P. Shoo, Catherine Graham, and Stephen E. Williams. 2009. 'Selecting Pseudo-Absence Data for Presence-Only Distribution Modeling: How Far Should You Stray from What You Know?' *Ecological Modelling* 220 (4): 589–94. https://doi.org/10.1016/j.ecolmodel.2008.11.010.

Van Proosdij, André S. J., Marc S. M. Sosef, Jan J. Wieringa, and Niels Raes. 2016. 'Minimum Required Number of Specimen Records to Develop Accurate Species Distribution Models'. *Ecography* 39 (6): 542–52. https://doi.org/10.1111/ecog.01509.

Warren, Dan L., Richard E. Glor, and Michael Turelli. 2008. 'ENVIRONMENTAL NICHE EQUIVALENCY VERSUS CONSERVATISM: QUANTITATIVE APPROACHES TO NICHE EVOLUTION'. *Evolution* 62 (11): 2868–83. https://doi.org/10.1111/j.1558-5646.2008.00482.x.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, A. Guisan, and NCEAS Predicting Species Distributions Working Group. 2008. 'Effects of Sample Size on the Performance of Species Distribution Models'. *Diversity and Distributions* 14 (5): 763–73. https://doi.org/10.1111/j.1472-4642.2008.00482.x.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279.

Wright, Marvin N., and Andreas Ziegler. 2017. '**Ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*'. *Journal of Statistical Software* 77 (1). https://doi.org/10.18637/jss.v077.i01.

Appendix

Complete code with virtual data for probability- and favourability-based SDMs:
https://github.com/elisamarchetto/Favourability-Probability

Tables

Table S1: p-values and chi-squared values of Kruskal-Wallis rank sum test. The test was carried out on the favourability-based and probability-based predictions of 50 virtual species estimated using a random sampling of presence/absence points comparing the sample prevalence groups.

RANDOM SAMPLING

MODEL	TEST	FAVOURABILITY	PROBABILITY
GLM	chi-squared	631.912 (p value <0.001)	396401.700 (p value <0.001)
GAM	chi-squared	225.600 (p value <0.001)	115436.000 (p value <0.001)
RF	chi-squared	2588.089 (p value <0.001)	144298.500 (p value <0.001)
BRT	chi-squared	29368.390 (p value <0.001)	238044.100 (p value <0.001)

Table S2: p-values and chi-squared values of Kruskal-Wallis rank sum test. The test was carried out on the favourability-based and probability-based predictions of 50 virtual species estimated using a stratified sampling of presence/absence points comparing the sample prevalence groups.

STRATIFIED SAMPLING

MODEL	TEST	FAVOURABILITY	PROBABILITY
GLM	chi-squared	433.493 (p-value <0.001)	367401.200 (p-value <0.001)
GAM	chi-squared	275.230 (p-value <0.001)	118533.600 (p-value <0.001)
RF	chi-squared	3201.369 (p-value <0.001)	146264.600 (p-value <0.001)
BRT	chi-squared	13729.390 (p-value	217202.200 (p-value

		<0.001)	<0.001)

Table S3: p-values and Z values of Dunn's test computed for GLM. According to the pairwise comparisons, "F" indicates favourability, "P" probability, "R" random sampling and "S" stratified sampling, while the decimal number is referred to the sample prevalence.

GLM

comparison	Z test statistic	adjusted p value
F0.2R - F0.2S	- 10.501	3.829 x 10−24
F0.4R - F0.4S	- 10.020	5.604 x 10−22
F0.5R - F0.5S	- 10.521	3.098 x 10−24
F0.6R - F0.6S	- 10.714	3.930 x 10−25
F0.8R - F0.8S	- 10.457	6.156 x 10−24
P0.2R - P0.2S	- 10.641	8.655 x 10−25
P0.4R - P0.4S	- 9.948	1.156 x 10−21
P0.5R - P0.5S	- 10.213	7.806 x 10−23
P0.6R - P0.6S	- 10.344	1.998 x 10−23
P0.8R - P0.8S	- 9.652	2.169 x 10−20

Table S4: p-values and Z values of Dunn's test computed for GAM. According to the pairwise comparisons, "F" indicates favourability, "P" probability, "R" random sampling and "S" stratified sampling, while the decimal number is referred to the sample prevalence.

GAM

comparison	Z test statistic	adjusted p value
F0.2R - F0.2S	-16.169	3.769 x 10−57
F0.4R - F0.4S	-15.604	3.063 x 10−53
F0.5R - F0.5S	-16.311	3.693 x 10−58
F0.6R - F0.6S	-16.194	2.506 x 10−57
F0.8R - F0.8S	15.612	2.727 x 10−53
P0.2R - P0.2S	-15.656	1.355 x 10−53
P0.4R - P0.4S	-15.521	1.131 x 10−52
P0.5R - P0.5S	-16.224	1.540 x 10−57
P0.6R - P0.6S	-16.088	1.385 x 10−56
P0.8R - P0.8S	-15.493	1.742 x 10−52

Table S5: p-values and Z values of Dunn's test computed for RF. According to the pairwise comparisons, "F" indicates favourability, "P" probability, "R" random sampling and "S" stratified sampling, while the decimal number is referred to the sample prevalence.

RF

comparison	Z test statistic	adjusted p value
F0.2R - F0.2S	-12.601	9.383 x 10−35

F0.4R - F0.4S	-12.752	1.361 x 10−35
F0.5R - F0.5S	-13.432	1.761 x 10−39
F0.6R - F0.6S	-13.980	9.343 x 10−43
F0.8R - F0.8S	-13.467	1.100 x 10−39
P0.2R - P0.2S	-12.494	3.602 x 10−34
P0.4R - P0.4S	-12.696	2.799 x 10−35
P0.5R - P0.5S	-13.345	5.704 x 10−39
P0.6R - P0.6S	-13.911	2.431 x 10−42
P0.8R - P0.8S	-13.332	6.744 x 10−39

Table S6: p-values and Z values of Dunn's test computed for BRT. According to the pairwise comparisons, "F" indicates favourability, "P" probability, "R" random sampling and "S" stratified sampling, while the decimal number is referred to the sample prevalence.

BRT

comparison	Z test statistic	adjusted p value
F0.2R - F0.2S	-11.616	1.536 x 10−29
F0.4R - F0.4S	-13.220	3.011 x 10−38
F0.5R - F0.5S	-13.814	9.383 x 10−42
F0.6R - F0.6S	-14.413	1.934 x 10−45
F0.8R - F0.8S	-14.823	4.711 x 10−48
P0.2R - P0.2S	-12.824	5.410 x 10−36
P0.4R - P0.4S	-13.281	1.341 x 10−38
P0.5R - P0.5S	-13.380	3.572 x 10−39
P0.6R - P0.6S	-13.596	1.907 x 10−40
P0.8R - P0.8S	-13.208	3.555 x 10−38

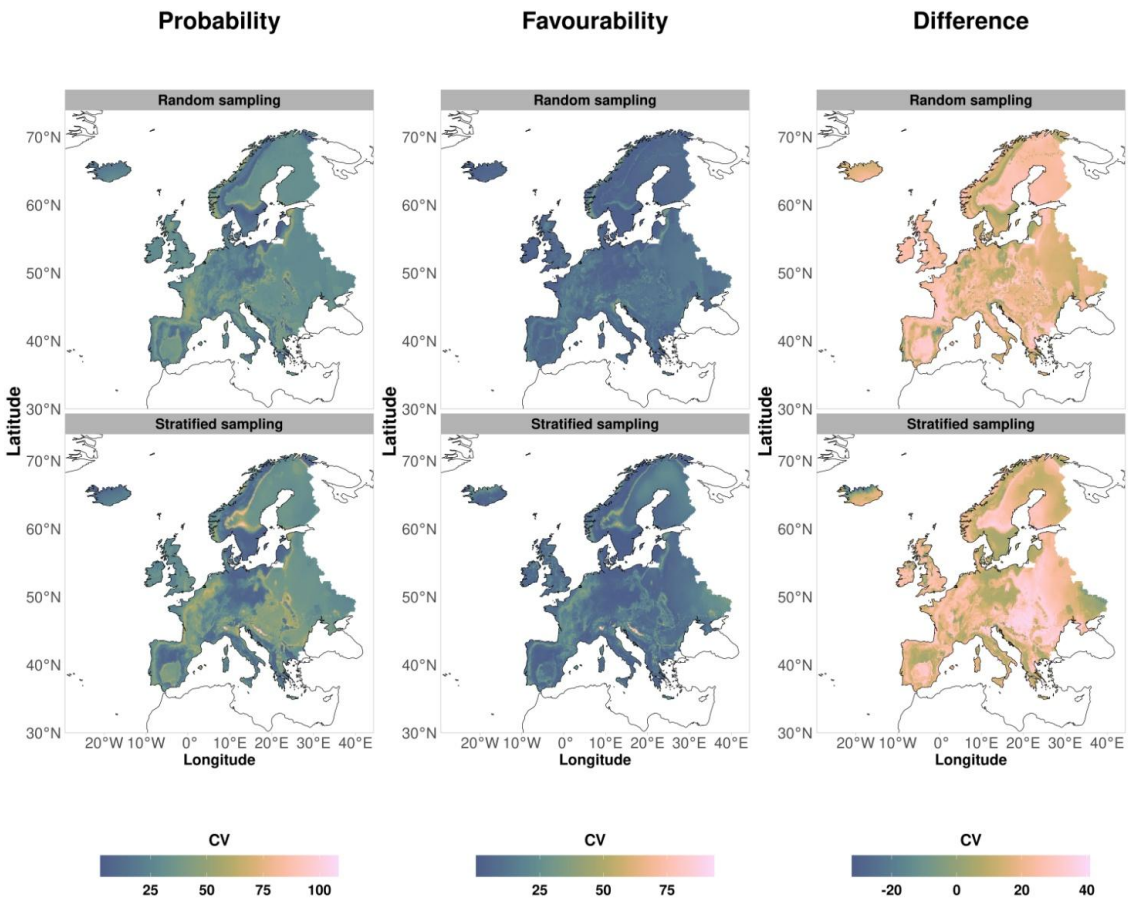


Figure S1: Pixels variabilities of mean probability and mean favourability predictions. The first column on the left shows the pixels variability of the mean probability of species occurrence calculated for GLM, RF, GAM and BRT, the central column, the pixels variability of the mean favourability of species occurrence calculated for GLM, RF, GAM and BRT and the right column the difference between the pixels variabilities of mean probability and mean favourability outcomes.

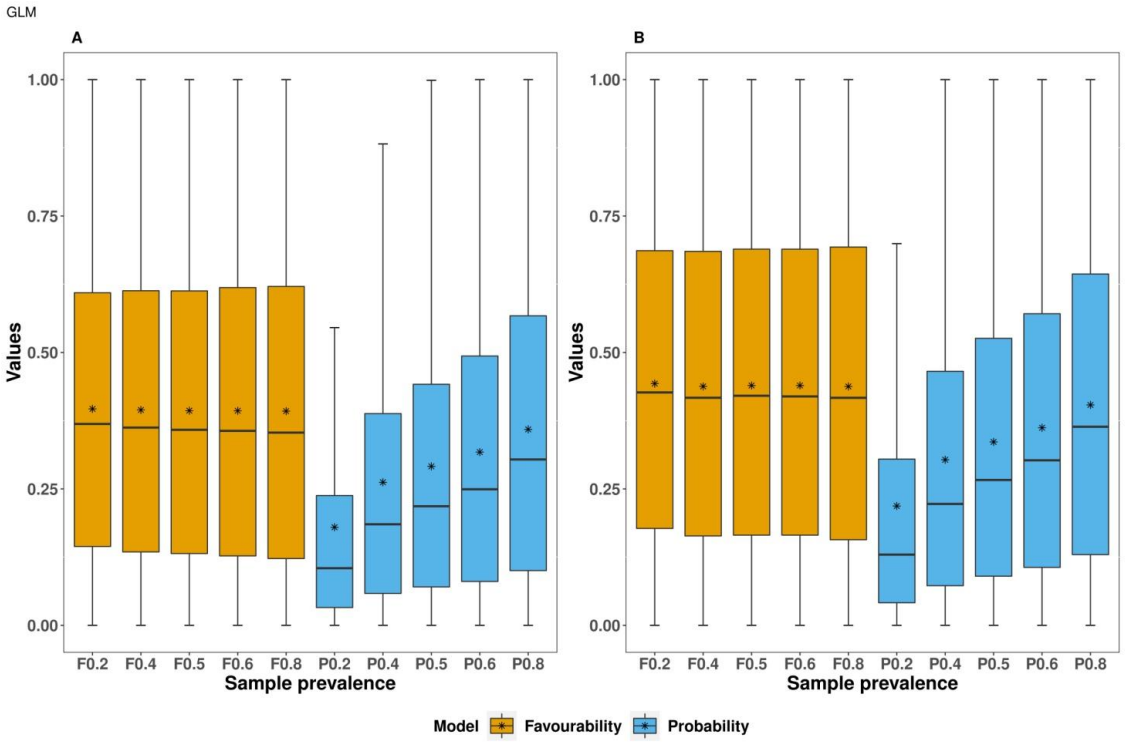


Figure S2: Distribution values of 50 virtual species between first and third quartiles within the range of prevalence values of the favourability and the probability predictions estimated with GLM. (A) favourability-based and probability-based species distribution values related to a random sampling of presence-absence points, (B) favourability-based and probability-based species distribution values related to a stratified sampling of presence-absence points. The mean value is represented with * symbol.

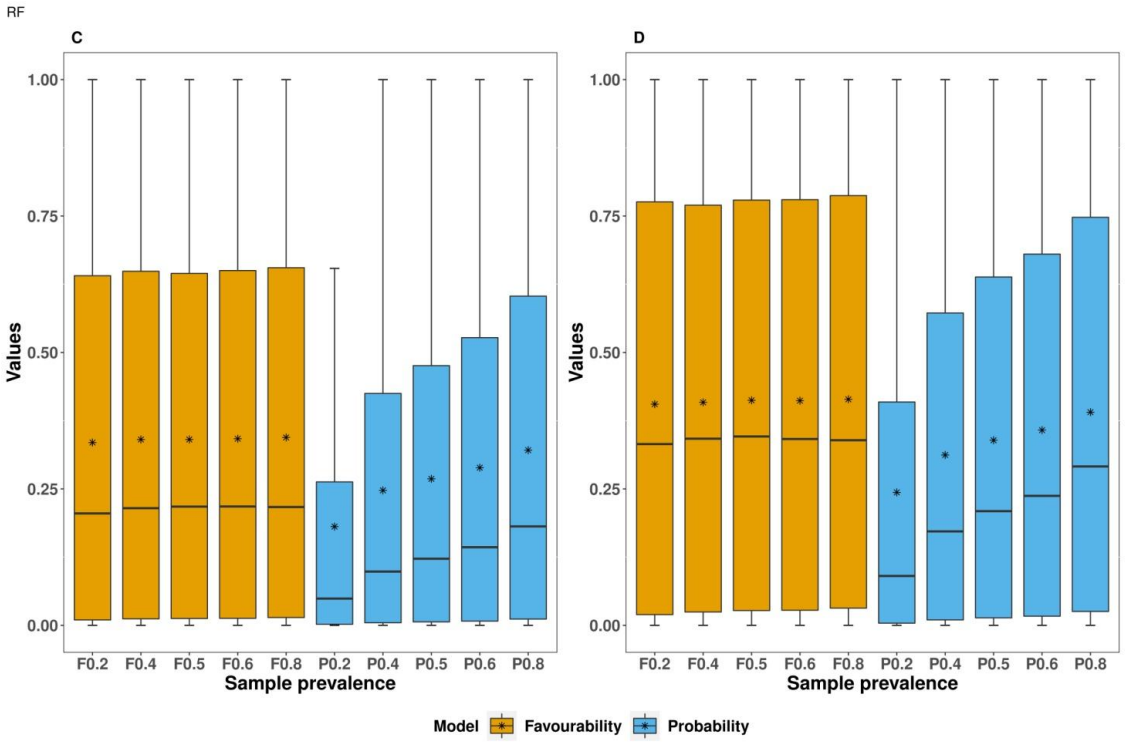


Figure S3: Distribution values of 50 virtual species between first and third quartiles within the range of prevalence values of the favourability and the probability predictions estimated with RF. (C) favourability-based and probability-based species distribution values related to a random sampling of presence-absence points, (D) favourability-based

and probability-based species distribution values related to a stratified sampling of presence-absence points. The mean value is represented with * symbol.

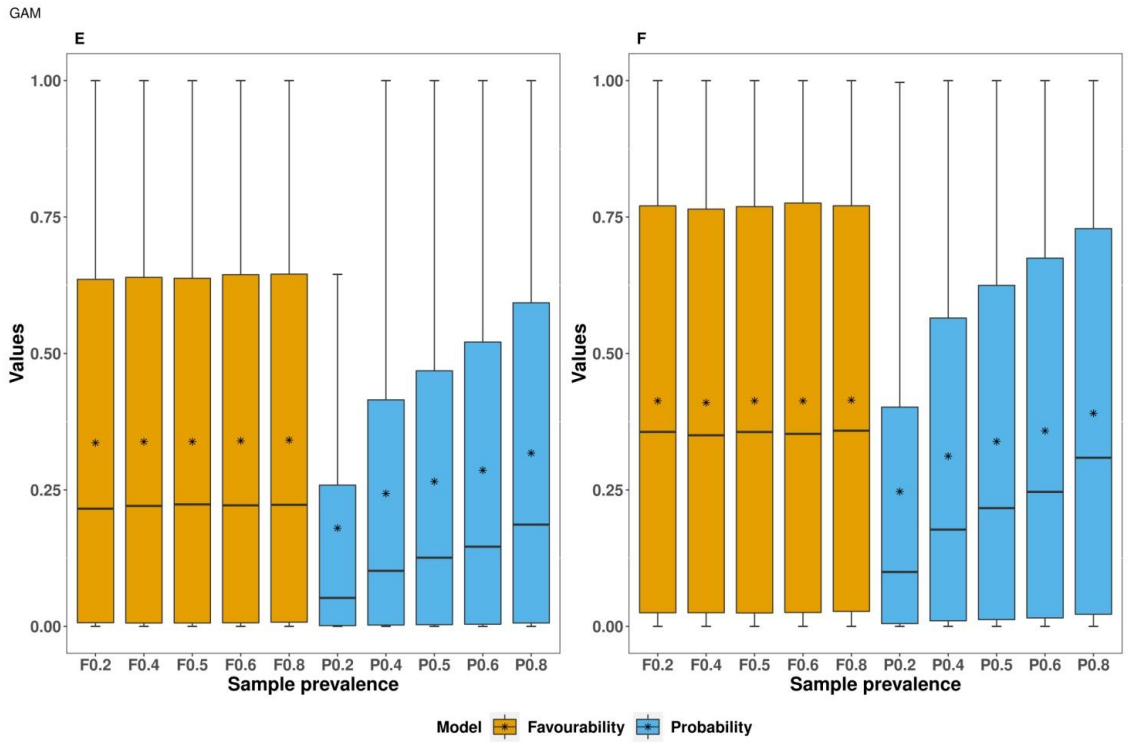


Figure S4: Distribution values of 50 virtual species between first and third quartiles within the range of prevalence values of the favourability and the probability predictions estimated with GAM. (E) favourability-based and probability-based species distribution values related to a random sampling of presence-absence points, (F) favourability-based and probability-based species distribution values related to a stratified sampling of presence-absence points. The mean value is represented with * symbol.

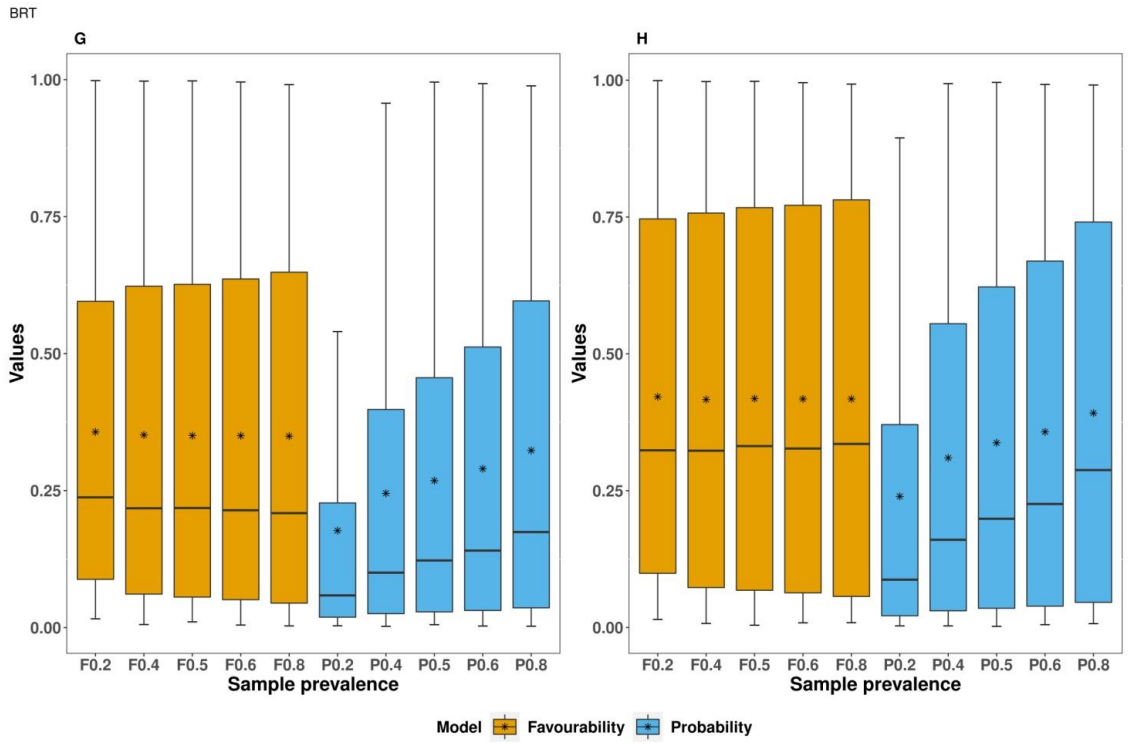


Figure S5: Distribution values of 50 virtual species between first and third quartiles within the range of prevalence values of the favourability and the probability predictions estimated with BRT. (G) favourability-based and probability-based species distribution values related to a random sampling of presence-absence points, (H) favourability-based and probability-based species distribution values related to a stratified sampling of presence-absence points. The mean value is represented with * symbol.

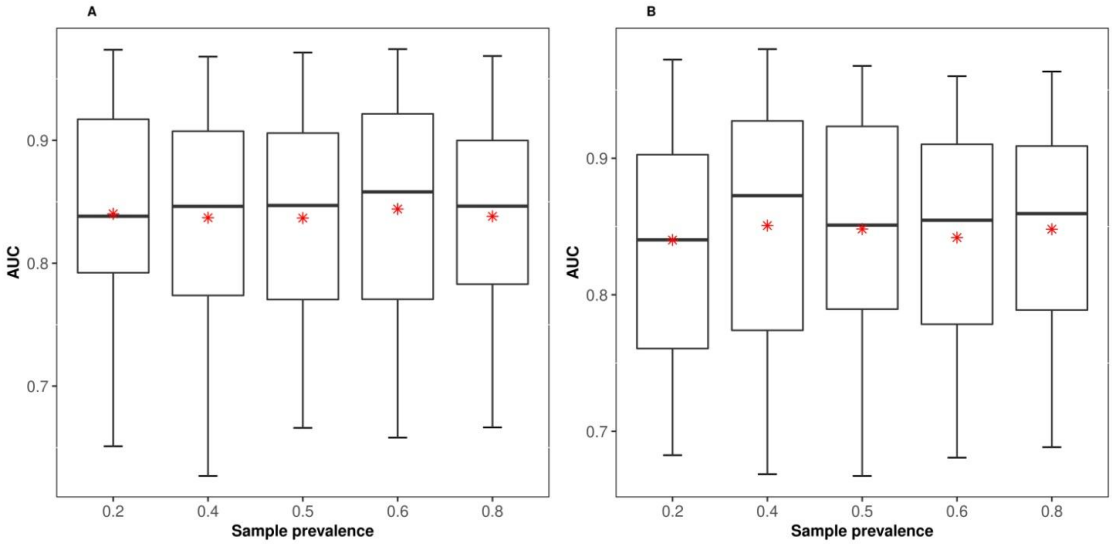


Figure S6: Distribution values between first and third quartiles of AUCs related to GLM within the range of sample prevalence. The first distribution (A) is referred to the random sampling method, the second distribution (B) is referred to the stratified sampling method. The mean value is represented with * symbol.

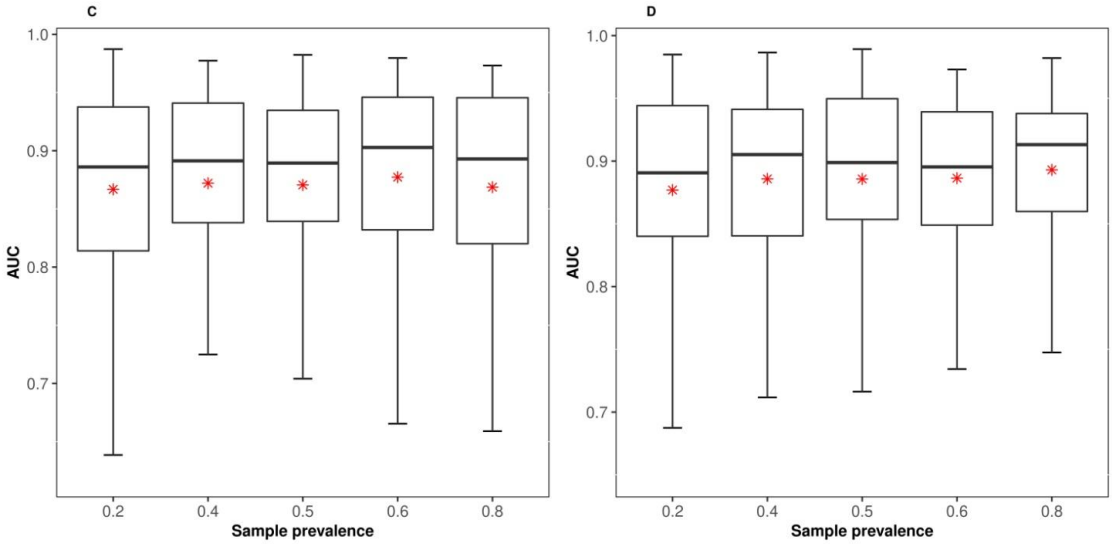


Figure S7: Distribution values between first and third quartiles of AUCs related to RF within the range of sample prevalence. The first distribution (C) is referred to the random sampling method, the second distribution (D) is referred to the stratified sampling method. The mean value is represented with * symbol.

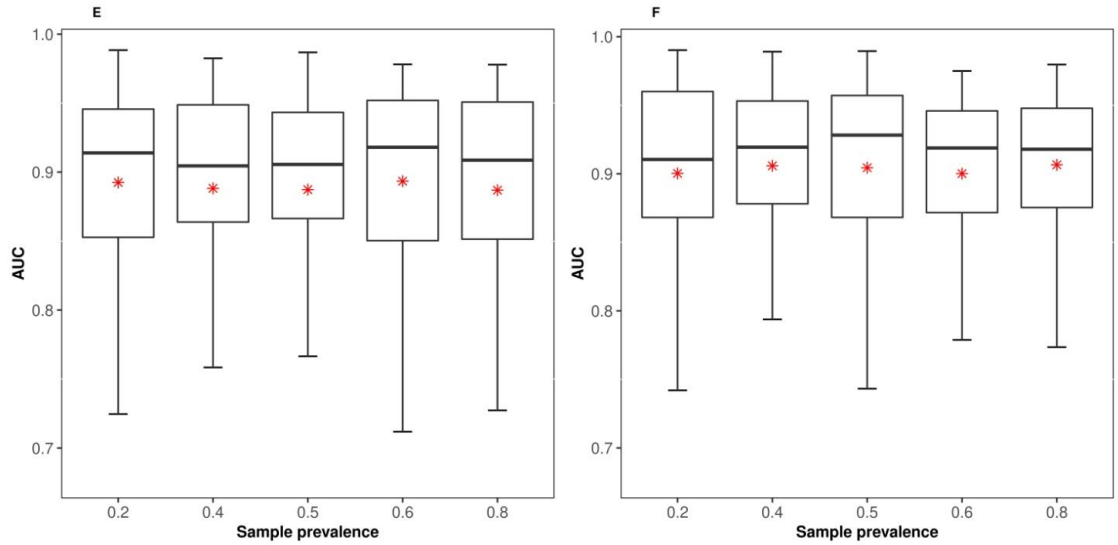


Figure S8: Distribution values between first and third quartiles of AUCs related to GAM within the range of sample prevalence. The first distribution (E) is referred to the random sampling method, the second distribution (F) is referred to the stratified sampling method. The mean value is represented with * symbol.

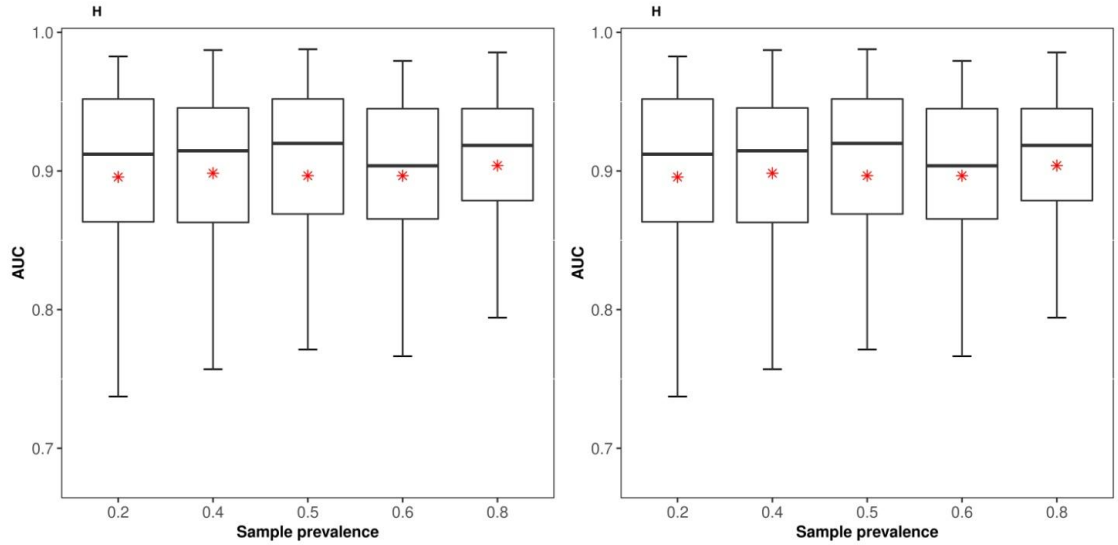


Figure S9: Distribution values between first and third quartiles of AUCs related to BRT within the range of sample prevalence. The first distribution (G) is referred to the random sampling method, the second distribution (H) is referred to the stratified sampling method. The mean value is represented with * symbol.

REAL SPECIES SCENARIO

We analyzed the tendency of favourability to reduce the variability of the predictions across different degrees of sample prevalence in juxtaposition with the suitability outcomes for two real species that share part of their geographic distribution (Fig. S10). Since the real absences were not provided we could infer only the habitat suitability and we could not estimate the actual probability of occurence. However, the study focused on the calculation of the coefficients of variations of *Picea abies* and *Fagus sylvatica*.

The species occurrences in Italy of *Picea abies* and *Fagus sylvatica* were selected from EU-Forest (Mauri et al., 2017), a dataset of European tree species distribution

(http://dx.doi.org/10.6084/m9.figshare.c.3288407) that blends forest plot surveys from National Forest Inventories on an INSPIRE-compliant 1km × 1km grid. We randomly sampled 1000 presence and pseudo absence points in order to fulfill the sequence of sample prevalence values (i.e., 0.2, 0.4, 0.5, 0.6, 0.8) and we randomly subsampled 5 replicates of the 19 bioclimatic variables with 10 arc-minutes of spatial resolution which were also used as predictors. Suitability and favorability models for each prevalence values were estimated using the generalized linear model. The favourability predictions variability of *Picea abies* and *Fagus sylvatica* proved to be more stable across the degrees of sample prevalence than the suitability outcomes (Fig. S11), (Fig. S12) demostrating that comparisons of species distributions of real species are more reliable using the favourability models.

References

Mauri, A., Strona, G., San-Miguel-Ayanz, J (2017). EU-Forest, a high resolution tree occurrence dataset for Europe. Sci Data 4, 160123. https://doi.org/10.1038/sdata.2016.123

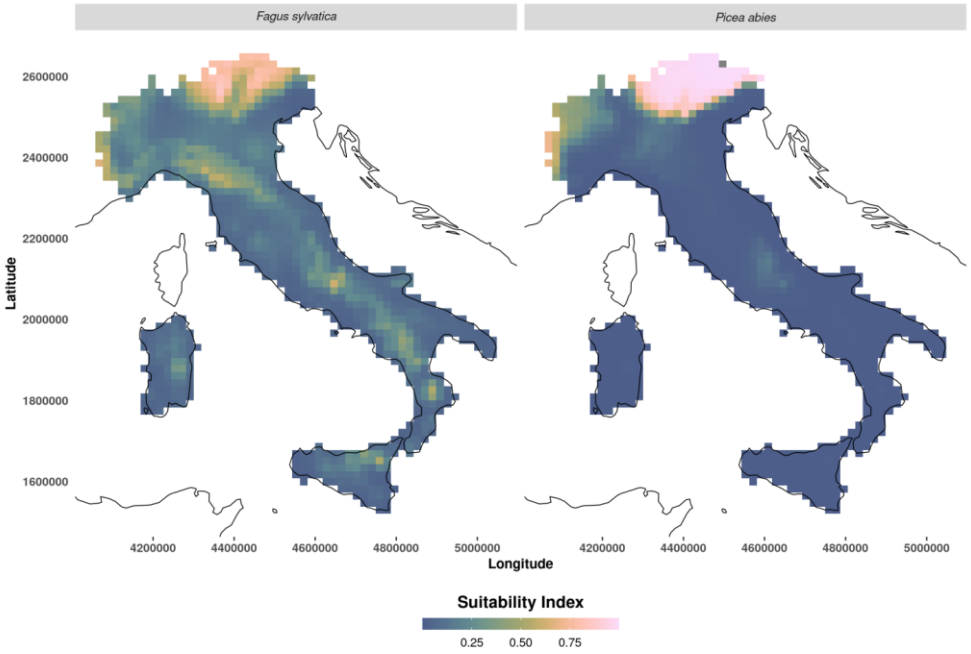


Figure S10: Habitat suitability of *Fagus sylvatica* and *Picea abies*. The first map on the left shows the suitability of *Fagus sylvatica*'s occurrence estimated using a random sampling of presences and pseudo-absences (sample prevalence value: 0.2), the second on the right represents the suitability of *Picea abies*'s occurrence estimated using a random sampling of presences and pseudo-absences (sample prevalence value: 0.2).

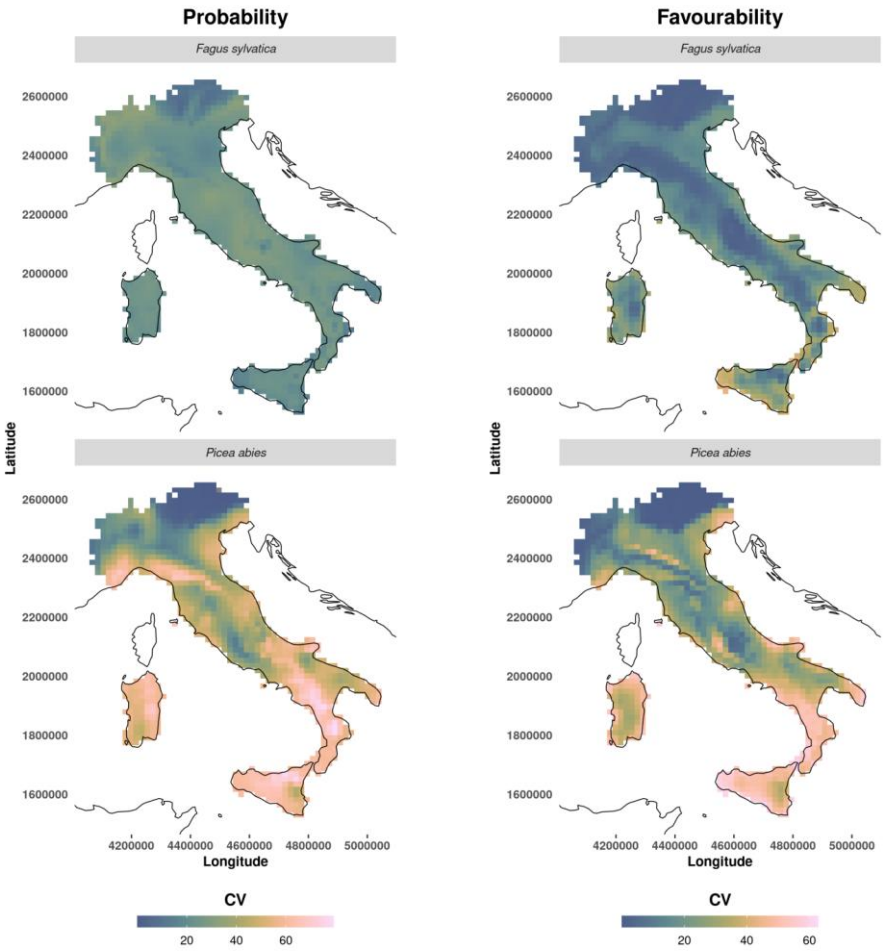


Figure S11: Predictions variability of *Fagus sylvatica* and *Picea abies*. The first column on the left represents the suitability-based predictions variability, calculated as Coefficient of Variation (CV) across the sample prevalence values of *Fagus sylvatica* and *Picea abies*, while, the second column shows the favourability-based predictions variability.

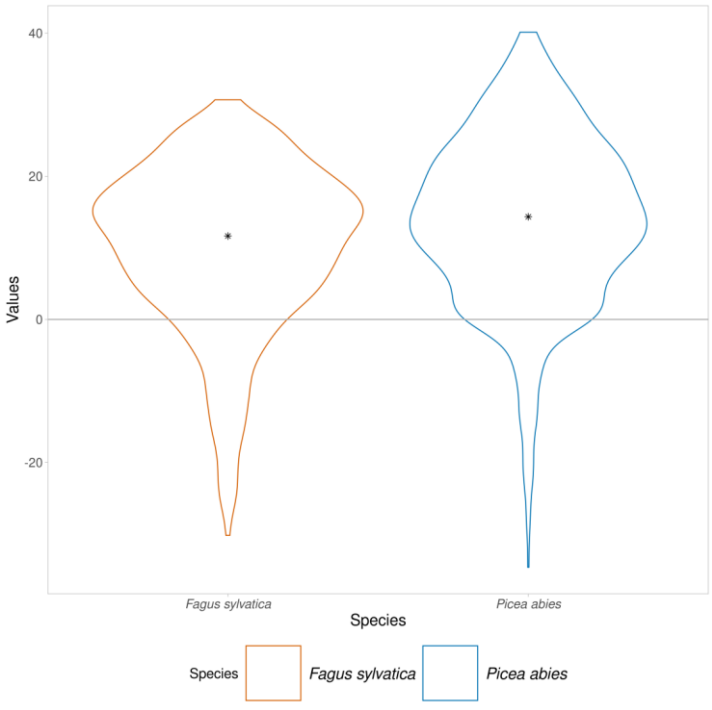


Figure S12: Distribution values of the difference between suitability-based coefficients of variation and favourability-based coefficients of variation for *Fagus sylvatica* and *Picea abies*. The median value is represented with * symbol.

Chapter 2

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], Jonathan Lenoir[5], Riccardo Testolin[1], Giovanni Bacaro[6], Roberto Cazzolla Gatti[1], Alessandro Chiarucci[1], Giles M. Foody[7], Lukáš Gábor[8,9], Quentin Groom[10], Jacopo Iaria[1], Marco Malavasi[11], Vítězslav Moudrý[8], Diletta Santovito[1], Petra Šímová [8], Piero Zannini[1], and Duccio Rocchini[1,8]

[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 Agricolture Food Environment, Via Edmund Mach, 1, 38098 San Michele all'Adige, Italy

[5]UMR CNRS 7058 Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), Université de Picardie Jules Verne, 1 rue des Louvels, 80037 Amiens, France7

[6]University of Trieste, Department of Life Sciences, Via L. Giorgieri 10, 34127 Trieste, Italy

[7]University of Nottingham, School of Geography, University Park, Nottingham NG7 2RD, UK

[8]Czech University of Life Sciences Prague, Faculty of Environmental Sciences, Department of Spatial Sciences, Kamýcka 129, 16500 Praha - Suchdol, Czech Republic

[9]Yale University, New Haven, CT 06520, United States

[10]Meise Botanic Garden, Nieuwelaan 38, 1860 Meise, Belgium

[11]University of Sassari, Department of Chemistry, Physics, Mathematics and Natural Sciences, Via Vienna 2, 07100 Sassari, Italy

*Equal contribution

Published as:

Marchetto, Elisa, Martina Livornese, Francesco Maria Sabatini, Enrico Tordoni, Daniele Da Re, Jonathan Lenoir, Riccardo Testolin, et al. 2024. 'Addressing Multiple Facets of Bias and Uncertainty in Continental Scale Biodiversity Databases'. *Biodiversity Informatics* 18 (September). https://doi.org/10.17161/bi.v18i.21810.

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, roads 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.

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

1 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, Malamud, and Francis 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, González-Dacosta, and Lobo 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, Weigelt, and Kreft 2016). While assessments of the limitations posed by the use of biodiversity databases do exist (Ronquillo et al. 2020; Monsarrat, Boshoff, and Kerley 2019; Colli-Silva et al. 2020), most studies focus on one dimension at the time, commonly spatial or taxonomic (but see Meyer, Weigelt, and Kreft (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, González-Dacosta, and Lobo 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, Ma, and Kreft 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, Ma, and Kreft 2014). The spatial distortion of the data resulting from these factors might yield inaccurate modelling outputs, especially when modelling species distribution (Rocchini et al. 2023; Bazzichetto 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.

2 Material and Methods

2.1 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.

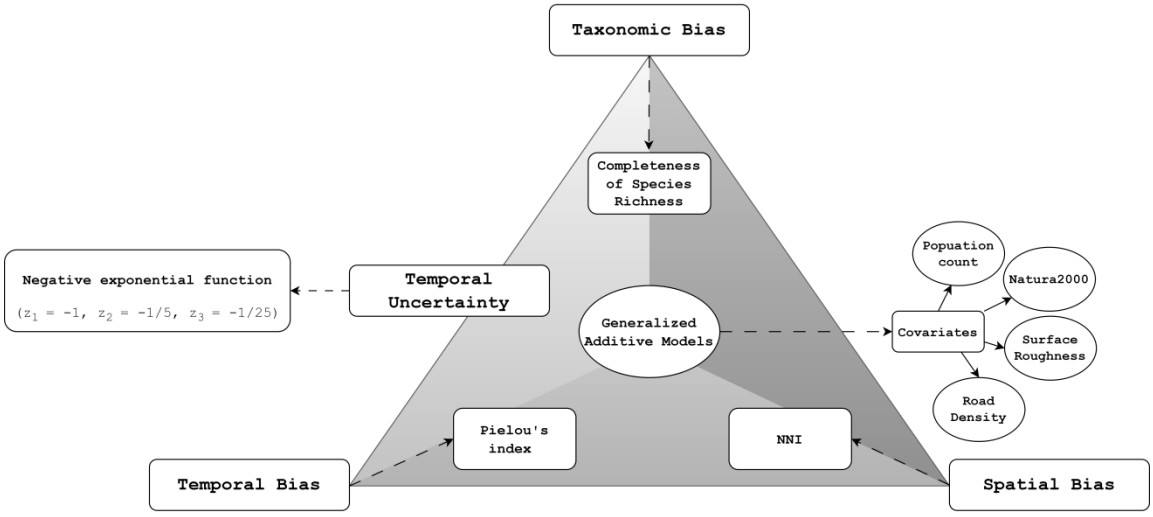


Figure 1: Methodological workflow to assess the presence of taxonomic, spatial and temporal shortfalls in biodiversity databases. The assessment of bias in raw data involves the following measurements: sampling completeness for taxonomic bias, Nearest Neighbour Index for spatial bias, Pielou's index 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.

2.2 Bias

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

2.2.1 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, Ma, and Chao 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.

2.2.2 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, Rubak, and Turner 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.

2.2.3 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 follow:

$$J = \frac{H}{Hmax} \quad (1)$$

Where H is the Shannon-Wiener index and it is calculated as:

$$H = -\sum_{i=1}^{N} pi \ln pi \quad (2)$$

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.

2.3 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 follows:

$$y_{(t)} = e^{-z(t)} \quad (3)$$

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: 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: Supplementary methods for further details).

2.4 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 (https://hub. worldpop.org/geodata/listing?id=64 (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) (https://www.globio.info/download-grip-dataset) (Meijer J.R. 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, Rubak, and Turner 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 website (https://www.eea.europa.eu/data-and-maps/data/natura-13, Published: 6 Oct 2022, Temporal coverage: 2021). 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, Gerstner, and Kreft 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.

3 Results

3.1 Bias

The taxonomic bias, described by the completeness of the species richness, was not evenly distributed over Europe (Fig. 2, Appendix: Fig. 3), following a similar pattern as the number of vegetation plots recorded per grid cell (Appendix: Fig. S4, Fig. S9). Besides, the spatial distribution of the plots, measured through the Nearest Neighbor Index (NNI), was clustered almost everywhere in Europe (Fig. 2, Appendix: Fig. S6). Most grid cells (97.4%) exhibited a clustered spatial pattern. The values of the NNI SES confirmed that the effect size was large, pointing out that the magnitude of the deviation from the random expectation was substantial (Appendix: 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: 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: 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.

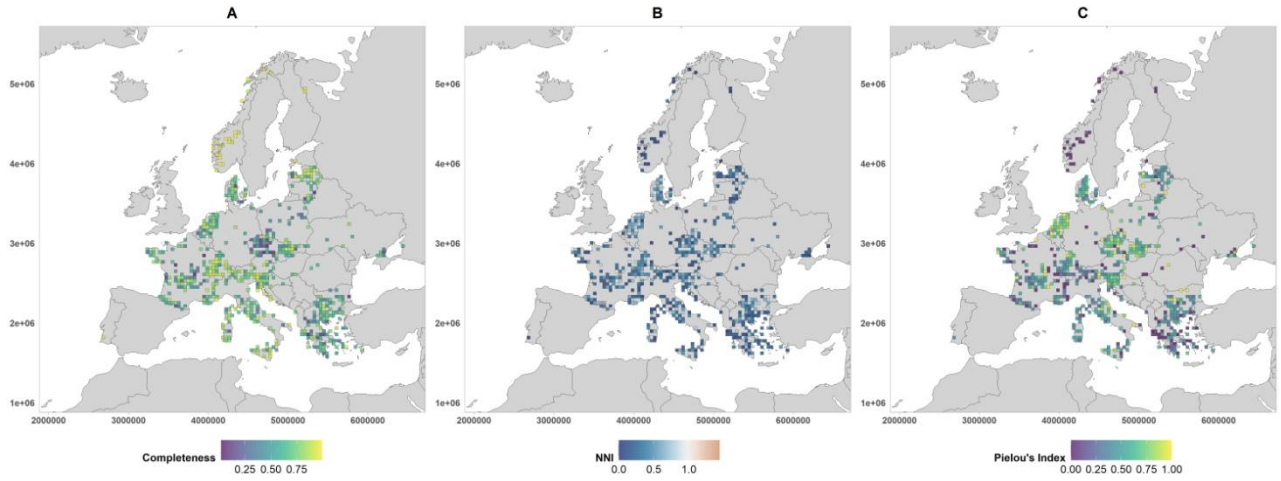


Figure 2: Grid-based map of three facets of bias. The map shows **A** the uneven distribution of the taxonomic bias, **B** the distribution of the vegetation plots through the Nearest Neighbour Index (spatial bias) and, **C** the heterogeneous distribution of the temporal bias. NNI values greater than 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.

3.2 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: Fig. S11).

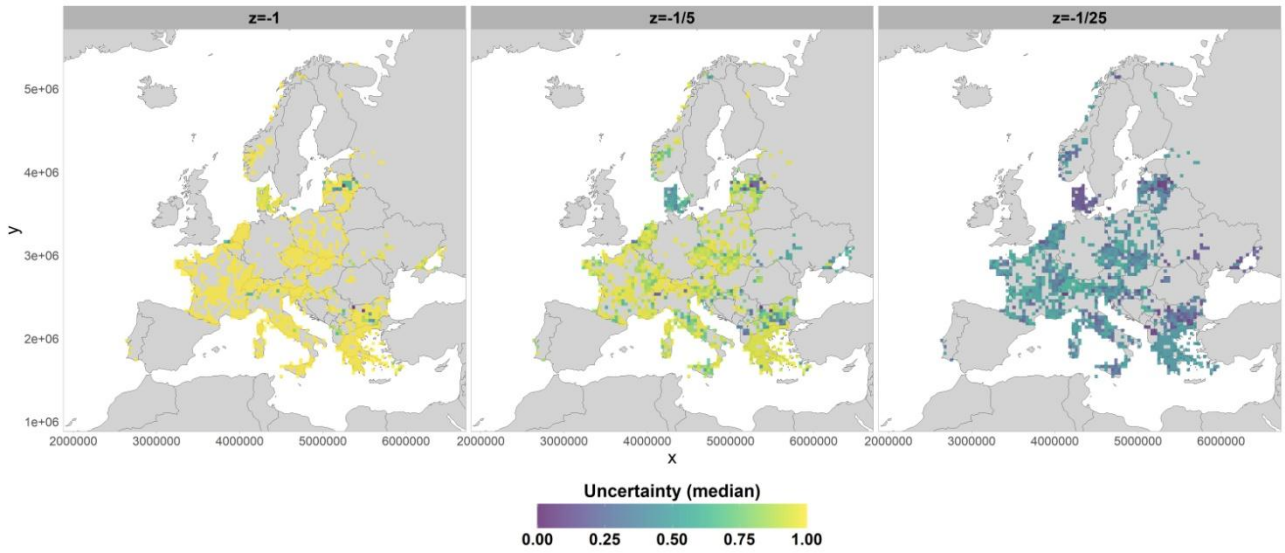


Figure 3: The map shows the median temporal uncertainty of the vegetation plots per grid cell of 39.5 km; the intensity of the temporal uncertainty changes according to the exponents being used setting the exponential negative function (i.e., exponents: z_1= -1, z_2= -1/5 and z_3= -1/25).

3.3 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 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: Fig. S10).

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

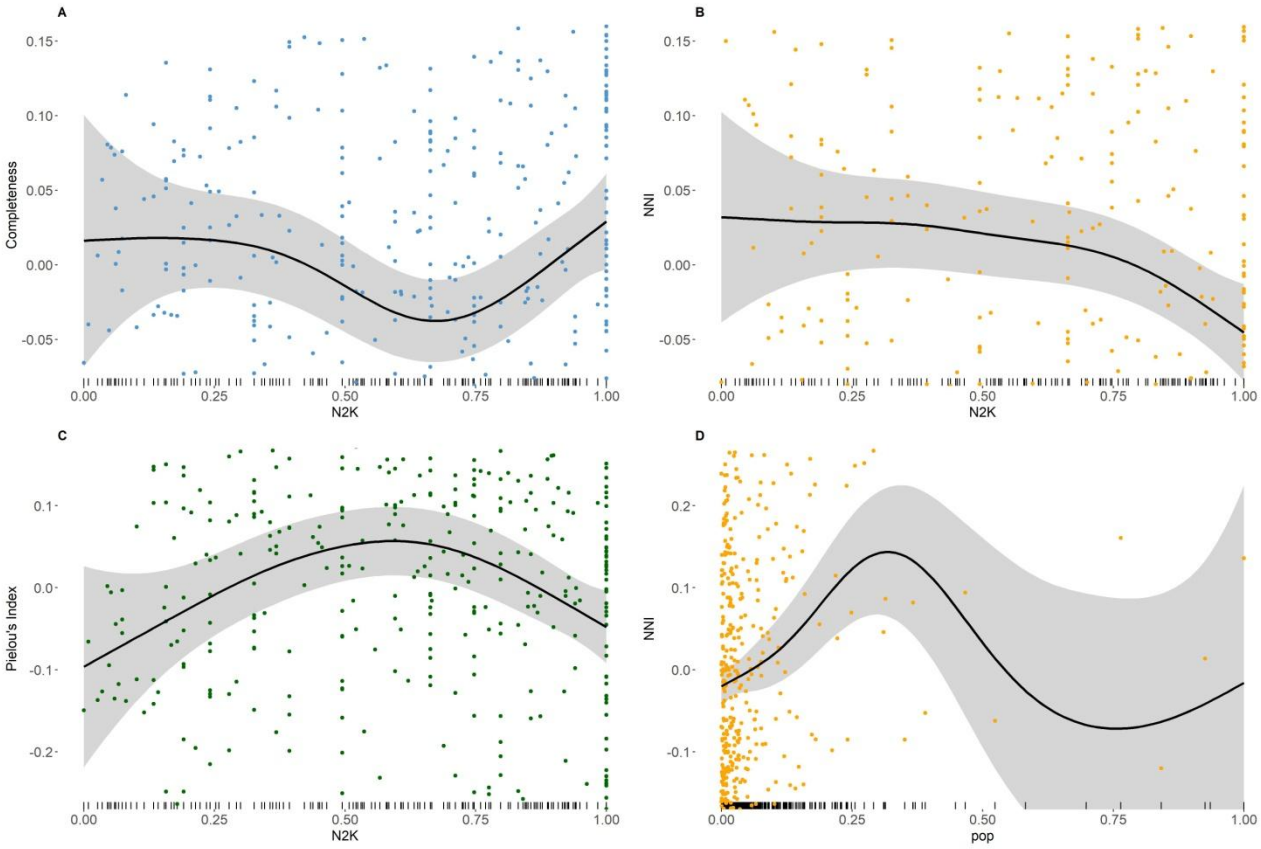


Figure 4: Trends of significative predictors with respect to the response variables of GAMs. *N2K* refers to the relative number of plots inside Natura 2000 network, *pop* refers to the human population count. The plot **A**) represents the estimated values of taxonomic bias (i.e., completeness of species 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.

4 Discussion

Biodiversity big data are being increasingly used to understand ecological patterns and monitor biodiversity trends (García-Roselló, González-Dacosta, and Lobo 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: 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, Ma, and Kreft 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 SES 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 an 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, Weigelt, and Kreft 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, Weigelt, and Kreft 2016; Tessarolo et al. 2021; D'Antraccoli, Bedini, and Peruzzi 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 2024).

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 at DOI https://zenodo.org/doi/10.5281/zenodo.12179384.

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.

References

Amatulli, Giuseppe, Sami Domisch, Mao-Ning Tuanmu, Benoit Parmentier, Ajay Ranipeta, Jeremy Malczyk, and Walter Jetz. 2018. 'A Suite of Global, Cross-Scale Topographic Variables for Environmental and Biodiversity Modeling'. *Scientific Data* 5 (1): 180040. https://doi.org/10.1038/sdata.2018.40.

Baddeley, Adrian, Ege Rubak, and Rolf Turner. 2016. *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics Series. Boca Raton London New York: CRC Press.

Ball-Damerow, Joan E., Laura Brenskelle, Narayani Barve, Pamela S. Soltis, Petra Sierwald, Rüdiger Bieler, Raphael LaFrance, Arturo H. Ariño, and Robert P. Guralnick. 2019. 'Research Applications of Primary Biodiversity Databases in the Digital Age'. Edited by Daniel De Paiva Silva. *PLOS ONE* 14 (9): e0215794. https://doi.org/10.1371/journal.pone.0215794.

Ballesteros-Mejia, Liliana, Ian J. Kitching, Walter Jetz, Peter Nagel, and Jan Beck. 2013. 'Mapping the Biodiversity of Tropical Insects: Species Richness and Inventory Completeness of A Frican Sphingid Moths'. *Global Ecology and Biogeography* 22 (5): 586–95. https://doi.org/10.1111/geb.12039.

Barajas-Barbosa, Martha Paola, Patrick Weigelt, Michael Krabbe Borregaard, Gunnar Keppel, and Holger Kreft. 2020. 'Environmental Heterogeneity Dynamics Drive Plant Diversity on Oceanic Islands'. *Journal of Biogeography* 47 (10): 2248–60. https://doi.org/10.1111/jbi.13925.

Bazzichetto, Manuele, Jonathan Lenoir, Daniele Da Re, Enrico Tordoni, Duccio Rocchini, Marco Malavasi, Vojtech Barták, and Marta Gaia Sperandii. 2023. 'Sampling Strategy Matters to Accurately Estimate Response Curves' Parameters in Species Distribution Models'. *Global Ecology and Biogeography* 32 (10): 1717–29. https://doi.org/10.1111/geb.13725.

Bertrand, Romain, Jonathan Lenoir, Christian Piedallu, Gabriela Riofrío-Dillon, Patrice De Ruffray, Claude Vidal, Jean-Claude Pierrat, and Jean-Claude Gégout. 2011. 'Changes in Plant Community Composition Lag behind Climate Warming in Lowland Forests'. *Nature* 479 (7374): 517–20. https://doi.org/10.1038/nature10548.

Boakes, Elizabeth H., Philip J. K. McGowan, Richard A. Fuller, Ding Chang-qing, Natalie E. Clark, Kim O'Connor, and Georgina M. Mace. 2010. 'Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data'. *PLoS Biology* 8 (6): e1000385. https://doi.org/10.1371/journal.pbio.1000385.

Boitani, Luigi, Luigi Maiorano, Daniele Baisero, Alessandra Falcucci, Piero Visconti, and Carlo Rondinini. 2011. 'What Spatial Data Do We Need to Develop Global Mammal Conservation Strategies?' *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1578): 2623–32. https://doi.org/10.1098/rstb.2011.0117.

Bowler, Diana E., Corey T. Callaghan, Netra Bhandari, Klaus Henle, M. Benjamin Barth, Christian Koppitz, Reinhard Klenke, et al. 2022. 'Temporal Trends in the Spatial Bias of Species Occurrence Records'. *Ecography* 2022 (8): e06219. https://doi.org/10.1111/ecog.06219.

Boyd, Robin J., Gary D. Powney, Claire Carvell, and Oliver L. Pescott. 2021. 'OccAssess: An R Package for Assessing Potential Biases in Species Occurrence Data'. *Ecology and Evolution* 11 (22): 16177–87. https://doi.org/10.1002/ece3.8299.

Bruelheide, Helge, Jürgen Dengler, Borja Jiménez-Alfaro, Oliver Purschke, Stephan M. Hennekens, Milan Chytrý, Valério D. Pillar, et al. 2019. 'SPlot – A New Tool for Global Vegetation Analyses'. Edited by Alessandro Chiarucci. *Journal of Vegetation Science* 30 (2): 161–86. https://doi.org/10.1111/jvs.12710.

Caldwell, Iain R., Jean-Paul A. Hobbs, Brian W. Bowen, Peter F. Cowman, Joseph D. DiBattista, Jon L. Whitney, Pauliina A. Ahti, et al. 2024. 'Global Trends and Biases in Biodiversity Conservation Research'. *Cell Reports Sustainability*, 100082. https://doi.org/10.1016/j.crsus.2024.100082.

Carmona, Carlos P., and Meelis Pärtel. 2021. 'Estimating Probabilistic Site-specific Species Pools and Dark Diversity from Co-occurrence Data'. Edited by Petr Keil. *Global Ecology and Biogeography* 30 (1): 316–26. https://doi.org/10.1111/geb.13203.

Chao, Anne, and Lou Jost. 2012. 'Coverage-based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size'. *Ecology* 93 (12): 2533–47. https://doi.org/10.1890/11-1952.1.

Chesshire, Paige R., Erica E. Fischer, Nicolas J. Dowdy, Terry L. Griswold, Alice C. Hughes, Michael C. Orr, John S. Ascher, et al. 2023. 'Completeness Analysis for over 3000 United States Bee Species Identifies Persistent Data Gap'. *Ecography* 2023 (5): e06584. https://doi.org/10.1111/ecog.06584.

Chiarucci, Alessandro, Rosa Maria Di Biase, Lorenzo Fattorini, Marzia Marcheselli, and Caterina Pisani. 2018. 'Joining the Incompatible: Exploiting Purposive Lists for the Sample-Based Estimation of Species Richness'. *The Annals of Applied Statistics* 12 (3). https://doi.org/10.1214/17-AOAS1126.

Chytrý, Milan, Lubomír Tichý, Stephan M. Hennekens, and Joop H.J. Schaminée. 2014. 'Assessing Vegetation Change Using Vegetation-plot Databases: A Risky Business'. Edited by Jürgen Dengler. *Applied Vegetation Science* 17 (1): 32–41. https://doi.org/10.1111/avsc.12050.

Chytrý, Milan, Stephan M. Hennekens, Borja Jiménez-Alfaro, Ilona Knollová, Jürgen Dengler, Florian Jansen, Flavia Landucci, et al. 2016. 'European Vegetation Archive (EVA): An Integrated Database of European Vegetation Plots'. Edited by Meelis Pärtel. *Applied Vegetation Science* 19 (1): 173–80. https://doi.org/10.1111/avsc.12191.

Clark, Philip J., and Francis C. Evans. 1954. 'Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations'. *Ecology* 35 (4): 445–53. https://doi.org/10.2307/1931034.

Colli-Silva, Matheus, Marcelo Reginato, Andressa Cabral, Rafaela Campostrini Forzza, José Rubens Pirani, and Thais N. Da 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 (3): 567–77. https://doi.org/10.1002/tax.12239.

D'Antraccoli, Marco, Gianni Bedini, and Lorenzo 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. https://doi.org/10.1016/j.ecoinf.2021.101512.

Daru, Barnabas H., and Jordan Rodriguez. 2023. 'Mass Production of Unvouchered Records Fails to Represent Global Biodiversity Patterns'. *Nature Ecology & Evolution* 7 (6): 816–31. https://doi.org/10.1038/s41559-023-02047-3.

Di Marco, Moreno, Tom D. Harwood, Andrew J. Hoskins, Chris Ware, Samantha L. L. Hill, and Simon Ferrier. 2019. 'Projecting Impacts of Global Climate and Land-use Scenarios on Plant Biodiversity Using Compositional-turnover Modelling'. *Global Change Biology* 25 (8): 2763–78. https://doi.org/10.1111/gcb.14663.

Enquist, Brian J, Rick Condit, Robert K Peet, Mark Schildhauer, and Barbara M. Thiers. 2016. 'Cyberinfrastructure for an Integrated Botanical Information Network to Investigate the Ecological Impacts of Global Climate Change on Plant Biodiversity'. Preprint. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.2615v2.

Gábor, Lukáš, Vítězslav Moudrý, Vincent Lecours, Marco Malavasi, Vojtěch Barták, Michal Fogl, Petra Šímová, Duccio Rocchini, and Tomáš Václavík. 2020. 'The Effect of Positional Error on Fine Scale Species Distribution Models Increases for Specialist Species'. *Ecography* 43 (2): 256–69. https://doi.org/10.1111/ecog.04687.

García-Roselló, Emilio, Jacinto González-Dacosta, and Jorge 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. https://doi.org/10.1016/j.biocon.2023.110118.

GBIF: The Global Biodiversity Information Facility (2024) What is GBIF?. Available from https://www.gbif.org/what-is-gbif [13 January 2020]

Geldmann, Jonas, Jacob Heilmann-Clausen, Thomas E. Holm, Irina Levinsky, Bo Markussen, Kent Olsen, Carsten Rahbek, and Anders P. Tøttrup. 2016. 'What Determines Spatial Bias in Citizen Science? Exploring Four Recording Schemes with Different Proficiency Requirements'. Edited by Brian Leung. *Diversity and Distributions* 22 (11): 1139–49. https://doi.org/10.1111/ddi.12477.

Girardello, Marco, Anna Chapman, Roger Dennis, Lauri Kaila, Paulo A.V. Borges, and Andrea Santangeli. 2019. 'Gaps in Butterfly Inventory Data: A Global Analysis'. *Biological Conservation* 236:289–95. https://doi.org/10.1016/j.biocon.2019.05.053.

Gotelli, Nicholas J., and Robert K. Colwell. 2001. 'Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness'. *Ecology Letters* 4 (4): 379–91. https://doi.org/10.1046/j.1461-0248.2001.00230.x.

Hortal, Joaquín, Francesco De Bello, José Alexandre F. Diniz-Filho, Thomas M. Lewinsohn, Jorge M. Lobo, and Richard J. Ladle. 2015. 'Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity'. *Annual Review of Ecology, Evolution, and Systematics* 46 (1): 523–49. https://doi.org/10.1146/annurev-ecolsys-112414-054400.

Hsieh, T. C., K. H. Ma, and Anne Chao. 2016. 'INEXT: An R Package for Rarefaction and Extrapolation of Species Diversity (H Ill Numbers)'. Edited by Greg McInerny. *Methods in Ecology and Evolution* 7 (12): 1451–56. https://doi.org/10.1111/2041-210X.12613.

Hughes, Alice C., Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, and Huijie Qiao. 2021. 'Sampling Biases Shape Our View of the Natural World'. *Ecography* 44 (9): 1259–69. https://doi.org/10.1111/ecog.05926.

IPBES. 2019. 'Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services'. [object Object]. https://doi.org/10.5281/ZENODO.3831673.

Jandt, Ute, Helge Bruelheide, Christian Berg, Markus Bernhardt-Römermann, Volker Blüml, Frank Bode, Jürgen Dengler, et al. 2022a. 'ReSurveyGermany: Vegetation-Plot Time-Series over the Past Hundred Years in Germany'. *Scientific Data* 9 (1): 631. https://doi.org/10.1038/s41597-022-01688-6.

Jandt, Ute, Helge Bruelheide, Florian Jansen, Aletta Bonn, Volker Grescho, Reinhard A. Klenke, Francesco Maria Sabatini, et al. 2022b. 'More Losses than Gains during One Century of Plant Biodiversity Change in Germany'. *Nature* 611 (7936): 512–18. https://doi.org/10.1038/s41586-022-05320-w.

La Sorte, Frank A., and Marius Somveille. 2020. 'Survey Completeness of a Global Citizen-science Database of Bird Occurrence'. *Ecography* 43 (1): 34–43. https://doi.org/10.1111/ecog.04632.

Lobo, Jorge M., Joaquín Hortal, José Luís Yela, Andrés Millán, David Sánchez-Fernández, Emilio García-Roselló, Jacinto González-Dacosta, Juergen Heine, Luís González-Vilas, and Castor Guisande. 2018. 'KnowBR: An Application to Map the Geographical Variation of Survey Effort and Identify Well-Surveyed Areas from Biodiversity Databases'. *Ecological Indicators* 91:241–48. https://doi.org/10.1016/j.ecolind.2018.03.077.

Mair, Louise, and Alejandro Ruete. 2016. 'Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa'. Edited by Judi Hewitt. *PLOS ONE* 11 (1): e0147796. https://doi.org/10.1371/journal.pone.0147796.

Meijer, Johan R, Mark A J Huijbregts, Kees C G J Schotten, and Aafke M Schipper. 2018. 'Global Patterns of Current and Future Road Infrastructure'. *Environmental Research Letters* 13 (6): 064006. https://doi.org/10.1088/1748-9326/aabd42.

Meyer, Carsten, Holger Kreft, Robert Guralnick, and Walter Jetz. 2015. 'Global Priorities for an Effective Information Basis of Biodiversity Distributions'. *Nature Communications* 6 (1): 8221. https://doi.org/10.1038/ncomms9221.

Meyer, Carsten, Patrick Weigelt, and Holger Kreft. 2016. 'Multidimensional Biases, Gaps and Uncertainties in Global Plant Occurrence Information'. Edited by Janneke Hille Ris Lambers. *Ecology Letters* 19 (8): 992–1006. https://doi.org/10.1111/ele.12624.

Michalcová, Dana, Samuel Lvončík, Milan Chytrý, and Ondřej Hájek. 2011. 'Bias in Vegetation Databases? A Comparison of Stratified-Random and Preferential Sampling: Stratified-Random and Preferential Sampling'. *Journal of Vegetation Science* 22 (2): 281–91. https://doi.org/10.1111/j.1654-1103.2010.01249.x.

Monsarrat, Sophie, Andre F. Boshoff, and Graham I. H. Kerley. 2019. 'Accessibility Maps as a Tool to Predict Sampling Bias in Historical Biodiversity Occurrence Records'. *Ecography* 42 (1): 125–36. https://doi.org/10.1111/ecog.03944.

Moura, Mario R., and Walter Jetz. 2021. 'Shortfalls and Opportunities in Terrestrial Vertebrate Species Discovery'. *Nature Ecology & Evolution* 5 (5): 631–39. https://doi.org/10.1038/s41559-021-01411-5.

Newbold, Tim, Lawrence N. Hudson, Samantha L. L. Hill, Sara Contu, Igor Lysenko, Rebecca A. Senior, Luca Börger, et al. 2015. 'Global Effects of Land Use on Local Terrestrial Biodiversity'. *Nature* 520 (7545): 45–50. https://doi.org/10.1038/nature14324.

Newling, Bruce E. 1969. 'The Spatial Variation of Urban Population Densities'. *Geographical Review* 59 (2): 242. https://doi.org/10.2307/213456.

Oliveira, Ubirajara, Adriano Pereira Paglia, Antonio D. Brescovit, Claudio J. B. De Carvalho, Daniel Paiva Silva, Daniella T. Rezende, Felipe Sá Fortes Leite, et al. 2016. 'The Strong Influence of Collection Bias on Biodiversity Knowledge Shortfalls of B Razilian Terrestrial Biodiversity'. Edited by Jeremy VanDerWal. *Diversity and Distributions* 22 (12): 1232–44. https://doi.org/10.1111/ddi.12489.

Oliver, Ruth Y., Carsten Meyer, Ajay Ranipeta, Kevin Winner, and Walter Jetz. 2021. 'Global and National Trends, Gaps, and Opportunities in Documenting and Monitoring Species Distributions'. Edited by Craig Moritz. *PLOS Biology* 19 (8): e3001336. https://doi.org/10.1371/journal.pbio.3001336.

Outhwaite, Charlotte L., Richard D. Gregory, Richard E. Chandler, Ben Collen, and Nick J. B. Isaac. 2020. 'Complex Long-Term Biodiversity Change among Invertebrates, Bryophytes and Lichens'. *Nature Ecology & Evolution* 4 (3): 384–92. https://doi.org/10.1038/s41559-020-1111-z.

Pelayo-Villamil, Patricia, Cástor Guisande, Ana Manjarrés-Hernández, Luz Fernanda Jiménez, Carlos Granado-Lorencio, Emilio García-Roselló, Jacinto González-Dacosta, Juergen Heine, Luis González-Vilas, and Jorge M. Lobo. 2018. 'Completeness of National Freshwater Fish Species Inventories around the World'. *Biodiversity and Conservation* 27 (14): 3807–17. https://doi.org/10.1007/s10531-018-1630-y.

Pielou, E.C. 1966. 'The Measurement of Diversity in Different Types of Biological Collections'. *Journal of Theoretical Biology* 13:131–44. https://doi.org/10.1016/0022-5193(66)90013-0.

Powney, Gary D., Claire Carvell, Mike Edwards, Roger K. A. Morris, Helen E. Roy, Ben A. Woodcock, and Nick J. B. Isaac. 2019. 'Widespread Losses of Pollinating Insects in Britain'. *Nature Communications* 10 (1): 1018. https://doi.org/10.1038/s41467-019-08974-9.

Ricci, Lorenzo, Michele Di Musciano, Francesco Maria Sabatini, Alessandro Chiarucci, Piero Zannini, Roberto Cazzolla Gatti, Carl Beierkuhnlein, et al. 2024. 'A Multitaxonomic Assessment of Natura 2000 Effectiveness across European Biogeographic Regions'. *Conservation Biology*, February, e14212. https://doi.org/10.1111/cobi.14212.

Rocchini, Duccio, Joaquín Hortal, Szabolcs Lengyel, Jorge M. Lobo, Alberto Jiménez-Valverde, Carlo Ricotta, Giovanni Bacaro, and Alessandro Chiarucci. 2011. 'Accounting for Uncertainty When Mapping Species Distributions: The Need for Maps of Ignorance'. *Progress in Physical Geography: Earth and Environment* 35 (2): 211–26. https://doi.org/10.1177/0309133311399491.

Rocchini, Duccio, Enrico Tordoni, Elisa Marchetto, Matteo Marcantonio, A. Márcia Barbosa, Manuele Bazzichetto, Carl Beierkuhnlein, et al. 2023. 'A Quixotic View of Spatial Bias in Modelling the Distribution of Species and Their Diversity'. *Npj Biodiversity* 2 (1): 10. https://doi.org/10.1038/s44185-023-00014-6.

Ronquillo, Cristina, Fernanda Alves-Martins, Vicente Mazimpaka, Thadeu Sobral-Souza, Bruno Vilela-Silva, Nagore G. Medina, and Joaquín Hortal. 2020. 'Assessing Spatial and Temporal Biases and Gaps in the Publicly Available Distributional Information of Iberian Mosses'. *Biodiversity Data Journal* 8 (September):e53474. https://doi.org/10.3897/BDJ.8.e53474.

Ronquillo, Cristina, Juliana Stropp, Nagore G. Medina, and Joaquin Hortal. 2023. 'Exploring the Impact of Data Curation Criteria on the Observed Geographical Distribution of Mosses'. *Ecology and Evolution* 13 (12): e10786. https://doi.org/10.1002/ece3.10786.

Ruete, Alejandro. 2015. 'Displaying Bias in Sampling Effort of Data Accessed from Biodiversity Databases Using Ignorance Maps'. *Biodiversity Data Journal* 3 (July):e5361. https://doi.org/10.3897/BDJ.3.e5361.

Sabatini, Francesco Maria, Jonathan Lenoir, Helge 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. https://doi.org/10.25829/IDIV.3474-40-3292.

Sabatini, Francesco Maria, Jonathan Lenoir, Tarek Hattab, Elise Aimee Arnst, Milan Chytrý, Jürgen Dengler, Patrice De Ruffray, et al. 2021b. 'SPlotOpen – An Environmentally Balanced, Open-access, Global Dataset of Vegetation Plots'. *Global Ecology and Biogeography* 30 (9): 1740–64. https://doi.org/10.1111/geb.13346.

Sporbert, Maria, Helge Bruelheide, Gunnar Seidler, Petr Keil, Ute Jandt, Gunnar Austrheim, Idoia Biurrun, et al. 2019. 'Assessing Sampling Coverage of Species Distribution in Biodiversity Databases'. Edited by Duccio Rocchini. *Journal of Vegetation Science* 30 (4): 620–32. https://doi.org/10.1111/jvs.12763.

Stein, Anke, Katharina Gerstner, and Holger Kreft. 2014. 'Environmental Heterogeneity as a Universal Driver of Species Richness across Taxa, Biomes and Spatial Scales'. Edited by Hector Arita. *Ecology Letters* 17 (7): 866–80. https://doi.org/10.1111/ele.12277.

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. 2015. 'Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data'. *PLOS ONE* 10 (2): e0107042. https://doi.org/10.1371/journal.pone.0107042.

Stropp, Juliana, Richard J. Ladle, Ana C. M. Malhado, Joaquín Hortal, Julien Gaffuri, William H. Temperley, Jon Olav Skøien, and Philippe Mayaux. 2016. 'Mapping Ignorance: 300 Years of Collecting Flowering Plants in Africa'. *Global Ecology and Biogeography* 25 (9): 1085–96. https://doi.org/10.1111/geb.12468.

Stropp, Juliana, Richard James Ladle, Thaise Emilio, Thainá Lessa, and Joaquín Hortal. 2022. 'Taxonomic Uncertainty and the Challenge of Estimating Global Species Richness'. *Journal of Biogeography* 49 (9): 1654–56. https://doi.org/10.1111/jbi.14463.

Sumner, Seirian, Peggy Bevan, Adam G. Hart, and Nicholas J.B. Isaac. 2019. 'Mapping Species Distributions in 2 Weeks Using Citizen Science'. Edited by Simon Leather. *Insect Conservation and Diversity* 12 (5): 382–88. https://doi.org/10.1111/icad.12345.

Tessarolo, Geiziane, Thiago F. Rangel, Miguel B. Araújo, and Joaquín Hortal. 2014. 'Uncertainty Associated with Survey Design in Species Distribution Models'. Edited by Linda Beaumont. *Diversity and Distributions* 20 (11): 1258–69. https://doi.org/10.1111/ddi.12236.

Tessarolo, Geiziane, Richard Ladle, Thiago Rangel, and Joaquin Hortal. 2017. 'Temporal Degradation of Data Limits Biodiversity Research'. *Ecology and Evolution* 7 (17): 6863–70. https://doi.org/10.1002/ece3.3259.

Tessarolo, Geiziane, Richard J. Ladle, Jorge M. Lobo, Thiago Fernando Rangel, and Joaquín Hortal. 2021. 'Using Maps of Biogeographical Ignorance to Reveal the Uncertainty in Distributional Data Hidden in Species Distribution Models'. *Ecography* 44 (12): 1743–55. https://doi.org/10.1111/ecog.05793.

Troudet, Julien, Philippe Grandcolas, Amandine Blin, Régine Vignes-Lebbe, and Frédéric Legendre. 2017. 'Taxonomic Bias in Biodiversity Data and Societal Preferences'. *Scientific Reports* 7 (1): 9132. https://doi.org/10.1038/s41598-017-09084-6.

Turbelin, Anna J., Bruce D. Malamud, and Robert A. Francis. 2017. 'Mapping the Global State of Invasive Alien Species: Patterns of Invasion and Policy Responses'. *Global Ecology and Biogeography* 26 (1): 78–92. https://doi.org/10.1111/geb.12517.

Walther, Bruno A., and Joslin 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 (6): 815–29. https://doi.org/10.1111/j.2005.0906-7590.04112.x.

Wüest, Rafael O., Niklaus E. Zimmermann, Damaris Zurell, Jake M. Alexander, Susanne A. Fritz, Christian Hof, Holger Kreft, et al. 2020. 'Macroecology in the Age of Big Data – Where to Go from Here?' *Journal of Biogeography* 47 (1): 1–12. https://doi.org/10.1111/jbi.13633.

Xu, Gang, Limin Jiao, Man Yuan, Ting Dong, Boen Zhang, and Chunmeng 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. https://doi.org/10.1016/j.landurbplan.2018.11.005.

Yang, Wenjing, Keping Ma, and Holger Kreft. 2013. 'Geographical Sampling Bias in a Large Distributional Database and Its Effects on Species Richness–Environment Models'. Edited by W. Daniel Kissling. *Journal of Biogeography* 40 (8): 1415–26. https://doi.org/10.1111/jbi.12108.

Yang, Wenjing, Keping Ma, and Holger Kreft. 2014. 'Environmental and Socio-economic Factors Shaping the Geography of Floristic Collections in C Hina'. *Global Ecology and Biogeography* 23 (11): 1284–92. https://doi.org/10.1111/geb.12225.

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.

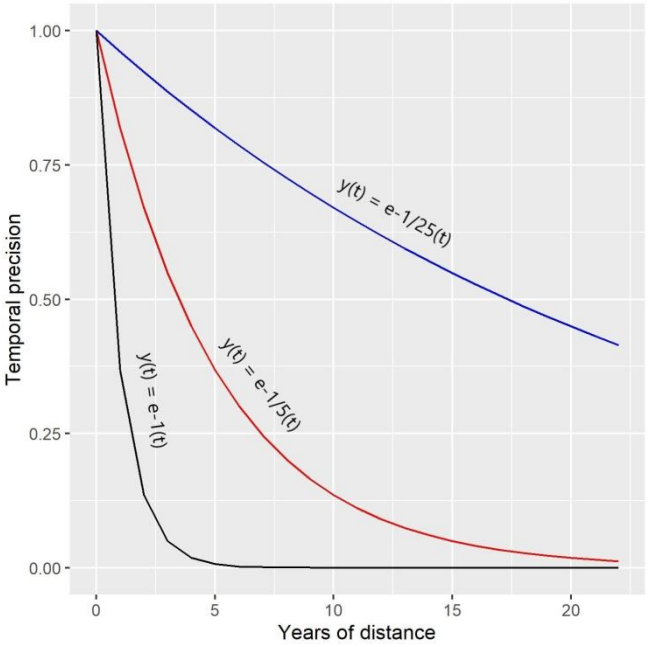


Figure S1: Negative exponential curves fitting using three different exponents, i.e., z_1=-1, z_2=-1/5 and z_3=-1/25.

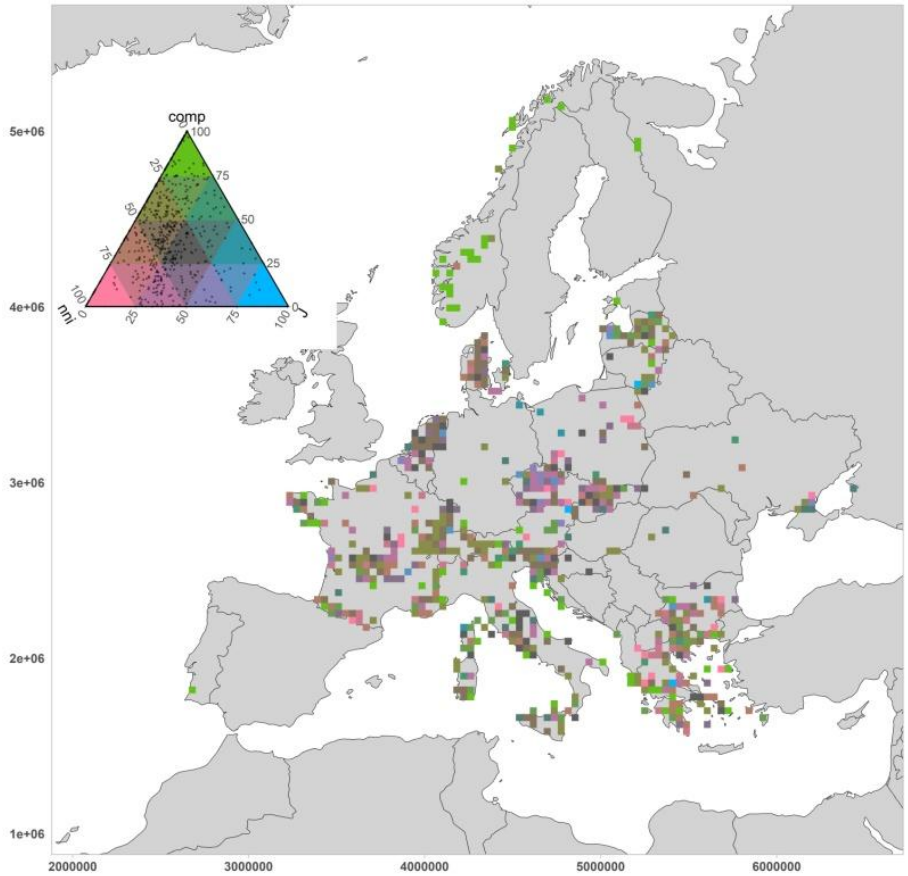


Figure S2: Grid-based trivariate map of taxonomic bias (i.e., completeness of species richness, abbrv. legend comp), spatial bias (i.e., NNI, abbrv. legend nni) and temporal bias (i.e., Pielou's evenness, abbrv. legend J). Each grid cell has a spatial resolution of 39.5 km. The highest sampling completeness 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.

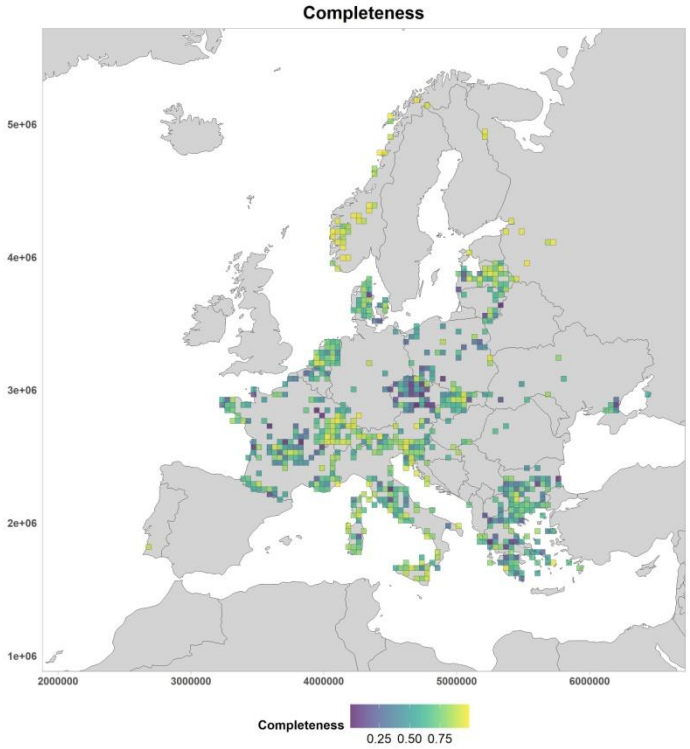


Figure S3: Completeness of the species richness per grid cell of 39.5 km.

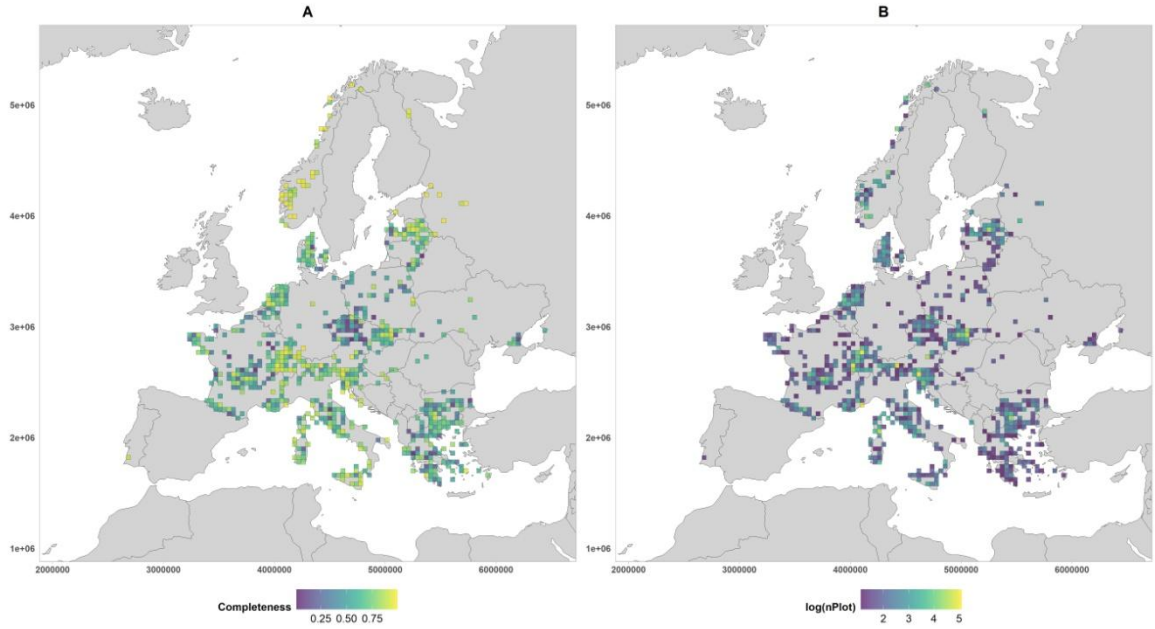


Figure S4: A) Completeness of the species richness and B) logarithm to base 10 of the number of plots per grid cell of 39.5 km with standardized number of grid cells.

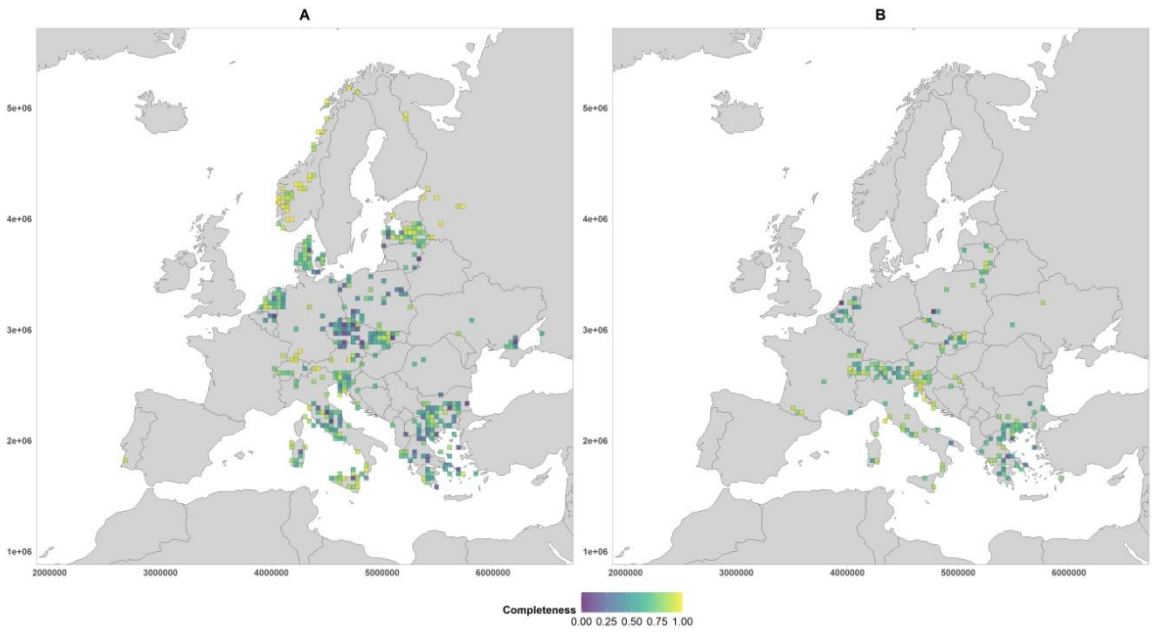
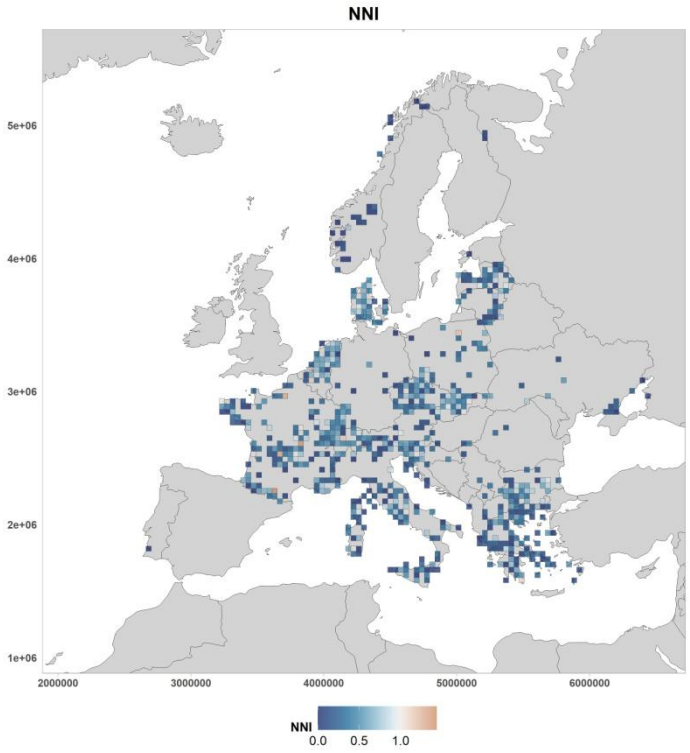


Figure S5: A) Completeness of the species richness per grid cell of 39.5km including only the vegetation plots with area less than or equal to 150 m^2. B) Completeness of the species richness for plots with an area greater than 150 m^2. The area size was determined by relying on Sabatini et al. 2022[1]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

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.

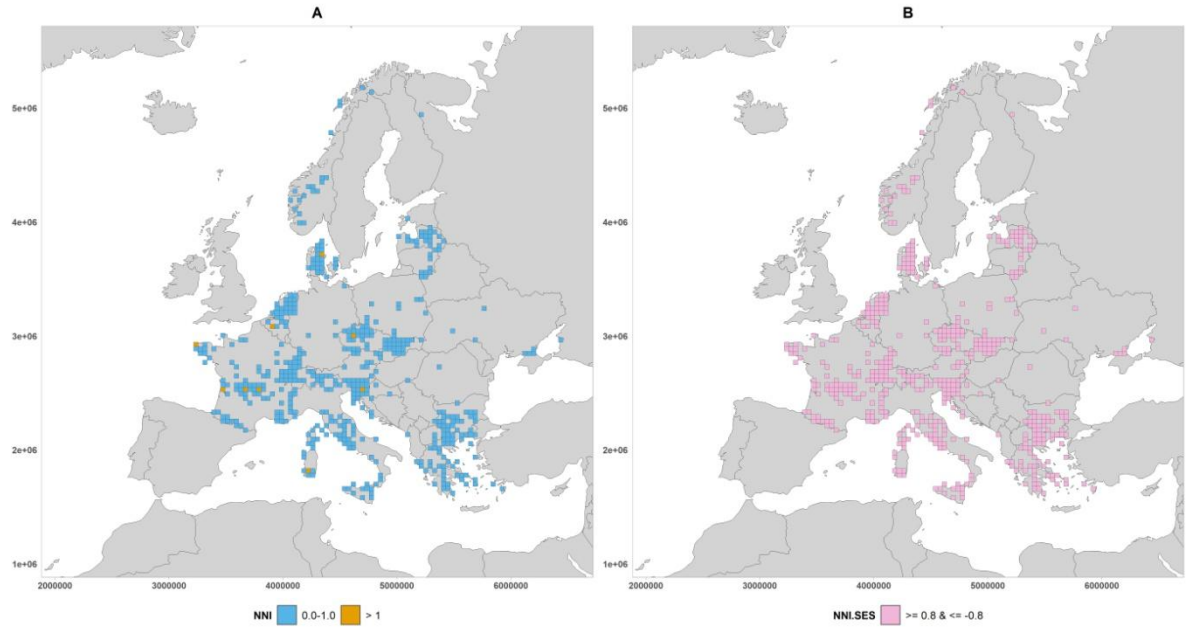


Figure S7: Spatial distribution of the vegetation plots per grid cell. A represents the map of NNI and B represents the map of NNI with a standardized effect size. NNI values greater than 1 indicate a random distribution of plots within a grid cell, while values less than 1 indicate a clustered distribution.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

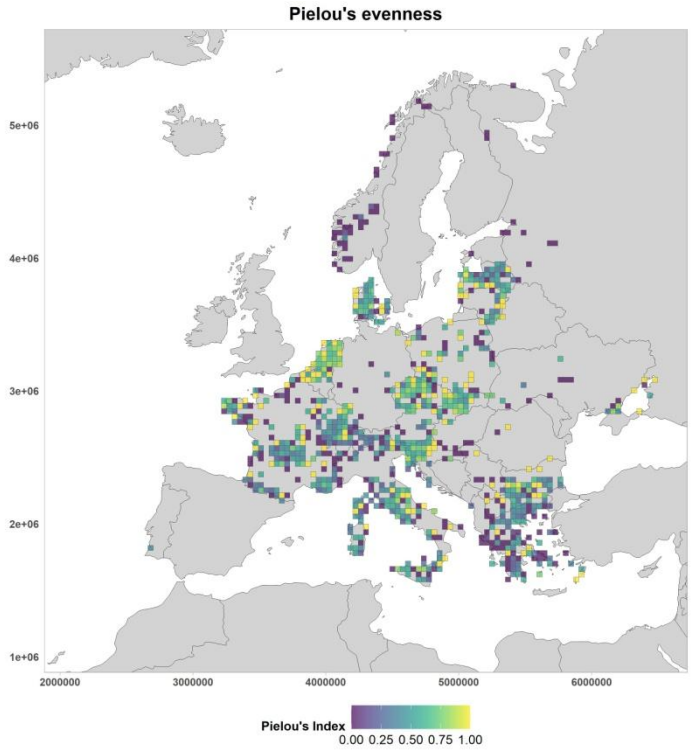


Figure S8: Pielou's Index per grid cell of 39.5 km.

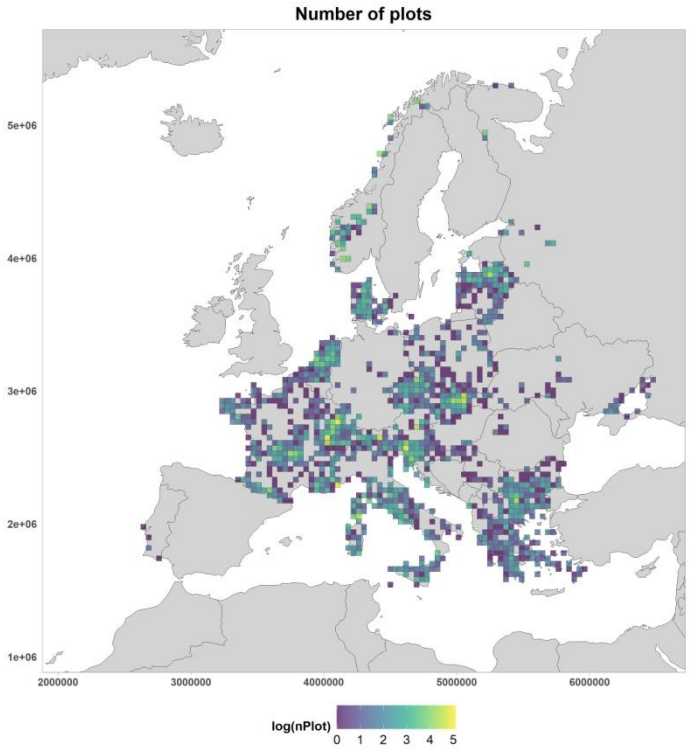


Figure S9: Logarithm to base 10 of the number of plots per grid cell of 39.5 km.

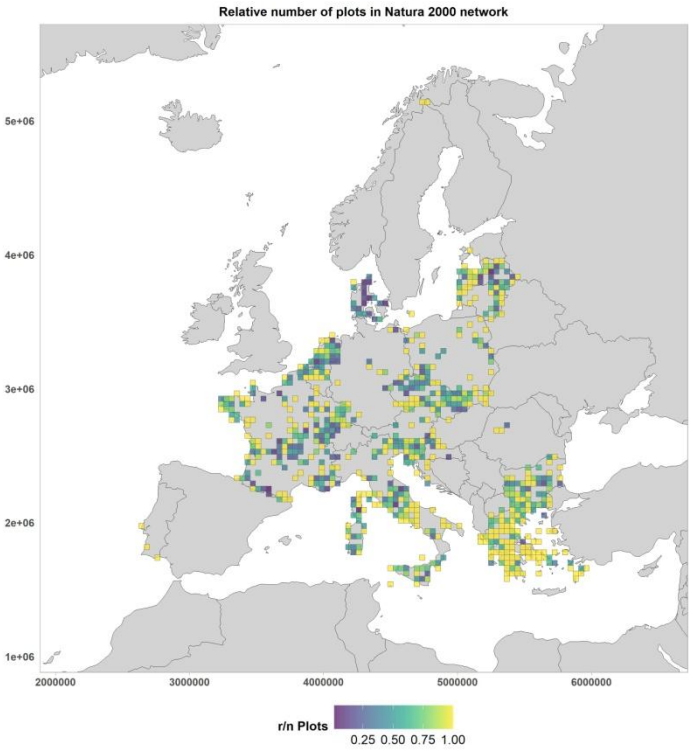


Figure S10: Map of the relative number of plots in the Natura 2000 network per grid cell.

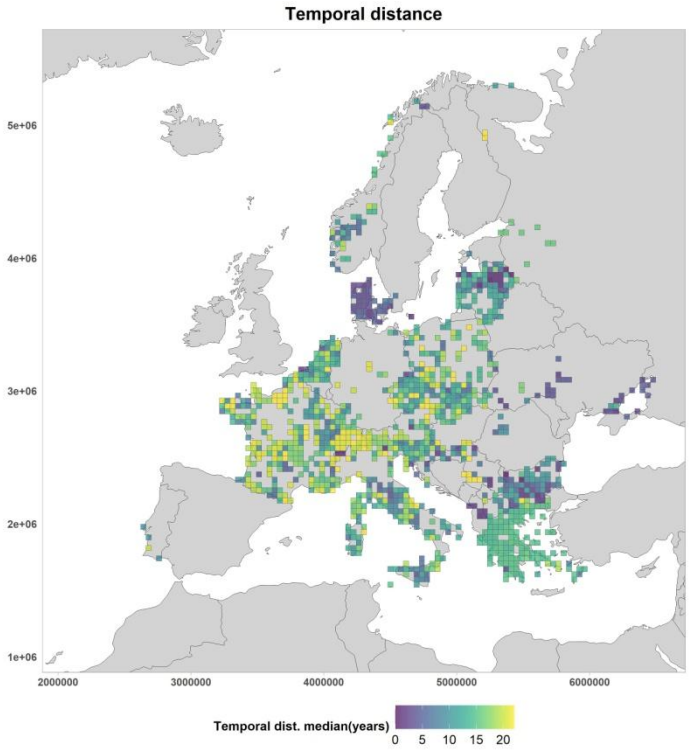


Figure S11: Median of the temporal distance between the most recent year (i.e., 2014) and the year of each record per grid cell of 39.5 km.

Chapter 3

The dimensions of habitat bias: a case study using big data

Elisa Marchetto[1], Enrico Tordoni[2], Duccio Rocchini[1,3]

[1]Alma Mater Studiorum - University of Bologna, Department of Biological, Geological and Environmental Sciences, via Irnerio 42, 40126 Bologna, Italy

[2]University of Tartu, Institute of Ecology and Earth Science, J. Liivi 2, 50409 Tartu, Estonia

[3]Czech University of Life Sciences Prague, Faculty of Environmental Sciences, Department of Spatial Sciences, Kamýcka 129, 16500 Praha - Suchdol, Czech Republic

Not submitted in scientific journal

Abstract

Habitats are essential to the survival of organisms, playing a fundamental role in biodiversity conservation. Monitoring habitat changes and loss is thus of critical importance, however, sampled data may inherently present bias and precision errors. A measure of the data quality may ensure a correct representation of the habitat state and, if required, it help address information gaps corrections. In this study, we propose three metrics (completeness of the species richness, Nearest Neighbor Index, Pielou's Index) for the assessment respectively of taxonomic, spatial, and temporal bias of species occurrence data at habitat level. The study was applied to the European plant records of sPlotOpen database representing two EUNIS habitat types, i.e., E (i.e., grassland and lands dominated by forbs, mosses and lichens) and G (i.e., woodland, forest and other woodland), observed in 3,642 vegetation plots across Europe.

For both habitats at EUNIS level 1 and 2, we found a generally low taxonomic bias, high spatial bias, and moderate values of temporal bias for both EUNIS level 1 and 2. An exception occurred with the temporal bias, where broadleaved evergreen woodland (G2) had particularly low values of Pielou's Index (i.e, high bias). Nonetheless, there was a greater variation in values among the habitat types at level 2 for the spatial and temporal bias suggesting that the number of plots, the habitat distribution and a possible opportunistic sampling can play a key role in shaping the dimensions of bias. Assessing data quality is crucial for addressing information gap-filling and deriving the most accurate representation and estimates for the conservation of a complex system like that of habitat.

Keywords: bias; data quality; habitat type; sampling data; species occurrence

1 Introduction

The quantity of biodiversity observations increased in the last decades (Wüest et al. 2020), and similarly their integration in large databases (e.g., GBIF, EVA, sPlot). Also, the quality of the observations improved significantly. Nevertheless, several sources of errors are still present (Hughes et al. 2021). Typically, opportunistic (i.e., non-probability) sample collections— such as non-standardized sampling or sampling driven by socioeconomic preferences (e.g., road accessibility, protected area, physical barriers) in the sampling locality — led to biased species observations (Yang, Ma, and Kreft 2014; Lessa et al. 2019). Moreover, social and political inequity, such as armed conflicts, the presence of infrastructures and urban centers, and the availability of research funds, shapes biodiversity data disparities in space and time (Zizka et al. 2021; Maitner et al. 2023; Chapman et al. 2024) generating inevitably gaps and bias in biodiversity information. Likewise, the union of biodiversity observations in big data repositories, raised from multiple databases with different projects and sampling methods (Beck et al. 2014; Chytrý et al. 2014), produces knowledge shortfalls which affect the accuracy and reliability of the ecological estimates (Daru and Rodriguez 2023; Johnson et al. 2024).

The bias is one of the possible forms of error in biodiversity data (Marchetto et al. 2024). It denotes a systematic error that deviates the observed values from the true value (Bolker 2008). In

this regard, the sampling design or even the union of biodiversity observations in bigger data inventories can generate several facets of bias (Garcia-Rosello et al. 2023a) of which the most common are the taxonomic bias, the spatial bias and the temporal bias (Marchetto et al. 2024). The taxonomic bias reflects the discrepancy between the observed pool of species and the expected, the spatial bias the irregular distribution of the samples (sample units like plots or, species occurences) in the geographic space and the temporal bias represents the inequality of the sampling across time.

Preference in specific taxa (Troudet et al. 2017) or the occurrence of undetected species (Moura and Jetz 2021; Lessa et al. 2024a) distorts the taxonomic dimension of the data. Typically, sample completeness—which calculates the taxonomic coverage of the gathered data within a certain area—is used to quantify the taxonomic bias. Traditionally, it is calculated as the ratio of the observed versus the expected species richness (Cazzolla Gatti et al. 2022) or rarely as the slope of Species Accumulation Curves (SACs) (Yang, Ma, and Kreft 2013). In the spatial dimension, an uneven distribution of the plots or species occurences (i.e., spatial bias) is frequently caused by sampling preferences such as for nature protected areas, accessible roads or unequal funding (Girardello et al. 2019) and, it also determined by an heterogeneous sampling effort over space of the sample units. In this context, the sampling effort (Ruete 2015; Geldmann et al. 2016), commonly measured as the number of plots or species occurrences per unit area, is the most investigated metric to evaluate the spatial bias. Variations in the strength of sampling efforts exist among data sets and taxonomic groups, as well as, in terms of time period and geographic regions (Stropp et al. 2016; Hughes et al. 2021). Nevertheless, in some studies, the Nearest Neighbour Index (Boyd et al. 2021, 2022) has been proposed to evaluate the possible biased distribution of data rather than its quantity or intensity in space. Finally, in the temporal dimension, if species-specific monitoring is missed or there are temporal gaps in data information, the identification of changes in time, such as the community loss and turnover (Graco-Roza et al. 2022; Jandt et al. 2022), can be biased.

Biased species occurrences are generally caused by opportunistic and disorganized survey efforts, similarly, inaccurate resurveys or irregular temporal monitoring might further skew species observations (Lobo et al. 2007). Hence, any inferred variables, indices, and models — such as Species Distribution Modeling (Baker et al. 2022; Baker, Maclean, and Gaston 2024) and diversity metrics (Maldonado et al. 2015; Ronquillo et al. 2023) — can be impacted by biased species occurrences.

In biodiversity databases, the three dimensions of the bias can appear at different temporal ranges, spatial levels, from local to global scale, and ecological levels, from species to realms (Meyer et al. 2015; Hugo and Altwegg 2017; La Sorte and Somveille 2020; Hughes et al. 2021; García-Roselló, González-Dacosta, and Lobo 2023b). At the habitat level, forms of bias can occur and their characterization is particularly distinctive given the habitat's peculiar ecological concept and classification. Habitat is defined, according to EUNIS habitat classification, as "a place where plants or animals normally live, characterized primarily by its physical features (topography, plant or animal physiognomy, soil characteristics, climate, water quality etc.) and secondarily by the

species of plants and animals that live there" (Davies, Moss, and Hill 2004). Habitat is then a species-specific concept (Hall, Krausman, and Morrison 1997). Given its unique characteristic, it allows only peculiar species to inhabit and adapt under specific resources and environmental conditions (Pardini and Püttker 2017). Habitat loss and fragmentation are some of the greatest threats to biodiversity (Haddad et al. 2015), raising the attention for their particular importance for conservation biology. The most important European initiative for habitat conservation was established by the EU in 1992, including the habitat types in Directive 92/43/EEC. The Habitat Directives envisages the conservation of habitats of community interest and the integration of the sites hosting the habitat types listed in Annex I in protected areas of Natura 2000 network. In Europe, the main habitat classification method is the EUNIS (European Nature Information System) Habitat Classification, developed by the European Environment Agency (EEA) in collaboration with the European Topic Centre on Biological Diversity (ETC/BD) in the 1990s and early 2000s (Davies and Moss 1999; Davies, Moss, and Hill 2004). The classification and identification are based on vegetation types in terms of species composition and vegetation structure and on the abiotic environment distinguishing the geographic location.

Nevertheless, programs aimed at the characterization of habitats and their monitoring can present forms of bias in the information collected. For instance, some habitats are more attractive than others, leading to disparities in sample coverage and possibly showing irregularities in the survey over time. Indeed, according to Geldmann et al. (2016), natural habitats received higher sampling intensity than human-modified habitats, as well as terrestrial habitats and aquatic habitats showed different sampling efforts (Rocha-Ortega, Rodriguez, and Córdoba-Aguilar 2021). However, even the characteristics of the landscape can influence people's sampling making some areas more accessible than others.

Regardless of the conditions under which data are collected, from a survey collection or released from biodiversity databases (big data), an assessment of data quality is recommended before any ecological inference (Boyd et al. 2022). Therefore, maps of ignorance or awareness (Ruete 2015; Lessa et al. 2024b) of the sampled data are crucial to prevent or correct as much as possible potential errors in biodiversity estimates and in conservation actions.

Leveraging the European plot records in sPlotOpen (Sabatini et al. 2021a), an open-access subset of the sPlot, we aimed to identify ecological metrics to measure and represent the bias at habitat level and within habitat types of the same EUNIS level in three dimensions (taxonomic, spatial and temporal). Here, we hypothesize that the values of the dimensions of bias change between habitat types and among sub habitat types. Finally, we aimed to define a methodological framework (Fig. 1) for measuring the dimensions of habitat bias that can be reproducible to different data sources such as inventories and local, regional or global databases.

2 Material and Methods

For the measurement of the dimensions of bias, we replicated the same methodology proposed by Marchetto et al. (2024) which uses three common metrics to calculate them: the completeness of the species richness, the NNI, and the Pielou’s Index.

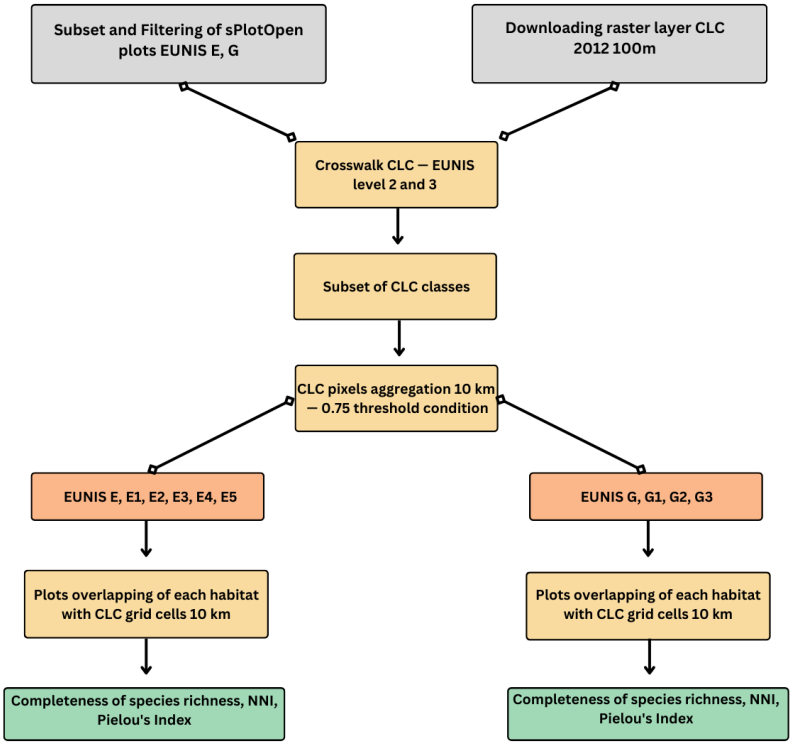


Figure 1: Methodological workflow which provides a first phase of data preparation and a second phase involving the calculation of the metrics for each habitat type.

2.1 Data preparation

We extracted the data records from sPlotOpen (March 2023 version 2.0, Sabatini et al., 2021b) in Europe with coordinates uncertainty lower than 250 m and with information about the plot size area. We also projected the vegetation plot coordinates to the LAEA Europe coordinates system (ETRS89-extended, EPSG:3035) to mitigate area distortions at higher latitudes.

The taxonomic, spatial and temporal biases were measured at EUNIS (European Nature Information System) habitat types of levels 1 and 2. We performed the analyses for the habitat types E (i.e., grassland and lands dominated by forbs, mosses and lichens) and G (i.e., woodland, forest and other woodland), classified according to EUNIS-ESy (Chytrý et al. 2020). Since not all habitat types included in sPlotOpen were equally distributed within the European continent, we selected only the ones that have a wide distribution and at least 2500 plots in the database according to our filtering criteria. In particular, the dimensions of bias were measured for the following habitat types: E, G (EUNIS level 1) and E1, E2, E3, E4, E5, G1, G2, G3 (EUNIS level 2). We only retained the plots located within the land cover type (Corine Land Cover) matching with the EUNIS habitat type at level 1. We overlaid the vegetation plots with the grid cells of the land cover type aggregated at 10 km of spatial resolution and we assigned each vegetation plot to its corresponding grid cell. To identify the CLC classes that match with the level 1 habitats, we followed the crosswalk between EUNIS habitats Classification and Corine Land Cover defined by

European Topic Centre on Biological Diversity (https://www.eea.europa.eu/data-and-maps/data/eunis-habitat-classification-1/documentation/eunis-clc.pdf). Specifically, the crosswalk matches the EUNIS code at level 2 and 3 with CLC at level 3. The Corine Land Cover was downloaded from Copernicus in the raster format for the year 2012 to be as much as possible consistent with the last year of recording of the database (i.e., 2014). We opted for a raster file with a resolution of 100 meters which represented the best trade-off between accuracy and computation effort. To ensure optimal computational speed, a sufficient number of plots per grid cell and a great number of pixels covered by the land cover, the spatial resolution of the CLC was aggregated to 10 kilometres. We aggregated the classes fixing the condition, as a threshold, that the new pixel became an NA value if at least 75\% of the pixels were NA values. For testing the effect of the threshold, we replicated the measure of the dimensions of bias using different values of thresholds (i.e., 0.65 and 0.85) for the aggregation of the pixels of the CLC (results in Appendix, Fig. S1 – S4). The data filtering decreased the plot records from 94,951 (total number of plots in sPlotOpen) to 3,642 (sum of the number of plots in E and G for threshold value 0.75).

According to the value of the threshold, the number of grid cells, in which the dimensions of the bias were calculated, changed for each habitat type as a consequence of their different geographic distribution and number of plots. In addition, we decided to retain the values only of those grid cells shared by all of the dimensions of bias (Table 1).

Table 1: Number of grid cells (0.75 threshold condition) per habitat type (E, G, E1, E2, E3, E4, E5, G1, G2, G3) in which the dimensions of bias were calculated.

Habitat	Name	N. of grid cells
E	Grassland and lands dominated by forbs, mosses and lichens	58
G	Woodland, forest and other woodland	186
E1	Dry grasslands	5
E2	Mesic grasslands	25
E3	Seasonally wet and wet grasslands	5
E4	Alpine and subalpine grasslands	19
E5	Woodland fringes and clearings and tall forb stands	7
G1	Broadleaved deciduous woodland	98
G2	Broadleaved evergreen woodland	8
G3	Coniferous woodland	78

2.2 Dimensions of bias

The dimensions of bias were calculated using different metrics (completeness of the species richness, NNI, Pielou's Index) for each grid cell of 10 km of spatial resolution as explained below.

2.2.1 Taxonomic bias

Based on the taxonomic coverage of vascular plants in sPlotOpen, we depicted the taxonomic bias as the completeness of species richness. Sample completeness was gauged as the ratio of the observed species richness to the true richness (observed plus undetected) (Chao et al., 2020). The sample completeness was calculated via the R package *iNEXT* (version 3.0.1) using Hill number q = 0, the vegetation plots as sampling units, and the observed incidence frequencies for the species records.

We set "incidence" as datatype, and for EUNIS level 1 and 2, we designated *k*, as the number of equally spaced knots (sample sizes), equal to 5. Finally, we excluded from the input data all grid cells that contained two or fewer vegetation plots. Essentially, the higher the value of the completeness of the species richness, the lower the taxonomic bias.

2.2.2 Spatial bias

We evaluated the extent of spatial bias of plot records by analyzing the spatial arrangement of plot locations using the Nearest Neighbor Index (NNI) (Clark and Evans 1954). This index is determined as the ratio of the observed average distance between each plot and its nearest neighbor to the expected average distance in a random distribution with an equivalent number of plots. Values less than 1 indicate a clustered pattern, values close to 1 imply a random distribution, and values greater than 1 suggest overdispersion or a more organized, systematic arrangement. A clustered distribution of plots corresponds to high spatial bias while a random or systematic distribution denotes low spatial bias. We used the R package *spatstat* (version 3.0.8) and the package *spatstat.explore* (version 3.2.7) to calculate the NNI by means of *clarkevans.test* function (Baddeley, Rubak, and Turner 2016), which assesses whether the plots demonstrate clustering or random distribution. The NNI was calculated for each EUNIS level 1 habitat type selected (i.e., E, G) and within each habitat type at EUNIS level 2. We considered as a geographic area for the index computation each Corine Land Cover (CLC) grid cell at 10 km of spatial resolution. Hence, for the analysis, only the plots that overlaid the CLC grid cells were selected for the measure of the NNI. The grid cells (spatstat windows of the analysis) were obtained by vectorizing the CLC pixels of the raster object at 10 km of spatial resolution.

2.2.3 Temporal bias

In this study, we used the Pielou's index (Pielou 1966) to estimate the temporal bias of plot records based on the years of recording of the vegetation plots. Pielou's index (J) commonly assesses the evenness of the abundance of the species within a community. It was calculated using the functions provided by the R package *vegan* (version 2.6.6) and it is defined as follows:

$$J = \frac{H}{Hmax} \quad (1)$$

Where H is the Shannon-Wiener index and it is calculated as:

$$H = -\sum_{i=1}^{N} pi \ln pi \quad (2)$$

N represents the total number of species, p_i their relative abundances for every species $_i \in \{1, \ldots, N\}$ and Hmax = lnN the maximum value of Shannon's index. Here, we refer to *N* as the total number of years of recording, where *i* is the ith year of recording, and *pi* is the proportion of plots (belonging to the habitat type) in a grid cell being sampled in the year *i*. So, high values of Pielou's Index indicated low temporal bias and vice versa.

In other words, the temporal bias was calculated for each habitat as the distribution of the year of recording of the plots within each grid cell.

2.3 Statistical analyses

We measured for each habitat type and dimension of bias the mean, the median and, the range between the minimum and the maximum value. We also calculated the Standard Error of the median value of 1000 nonparametric bootstrap replications with replacement. To test if there was any statistical difference between E and G, we performed a Wilcoxon rank sum test for each dimension of bias, then for the habitats at EUNIS level 2, we carried out a Kruskal-Wallis test. In the case of a significant Kruskal-Wallis test, we evaluated pairwise comparisons across habitats using Dunn's test with Holm's correction of the p-value (R package *FSA* version 0.9.5).

3 Results

The completeness of the species richness showed high values (i.e., low taxonomic bias) for both E and G habitat types, with a median value of 0.789 for G and 0.772 for E (Appendix: Table S1). On the contrary, the NNI was low (i.e., high spatial bias) with a median value of 0.382 for G and 0.352 for E (Appendix: Table S2). Finally, the temporal bias at level 1 EUNIS was medium. The habitat type E showed a slightly higher value (0.523) of Pielou's Index (i.e., lower temporal bias) than G (0.445) (Fig. 1 and Appendix: Table S3). With respect to the entire spectrum of values that the metrics can cover, the temporal bias was the one that covered the whole range of possible values (range from 0 to 1). For the level 2 EUNIS, the habitats showed a taxonomic bias with all the median values above 0.75, except for E1 (0.703). Whereas, the spatial bias was generally high, with a clustered distribution of plots, and a median below 0.5, except for E3 (0.716). The median values of Pielou's index were heterogeneous. E3 showed greater temporal evenness, with low bias and a median value of 0.750, while G2 showed higher temporal bias with a median value of 0 (Fig. 2). Even for level 2 EUNIS habitats, the distribution of values was less widespread for taxonomic

bias while the temporal bias showed a generally higher dispersion of values (Appendix: Tables S1-S2-S3).
We found no significant difference between E and G for all dimensions of bias (Appendix: Table S4), on the contrary, the habitat types at level 2 EUNIS showed significant differences for the spatial and the temporal bias (Appendix: Table S5). For the spatial bias, only the pairwise combination E3 – E5 exhibited significant variation. Otherwise, the temporal bias showed significant differences for E3 – G1 and E3 – G2 combinations (Appendix: Tables S7-S8).

The patterns of values for EUNIS level 1 of E and G were the same irrespective of the value of threshold set for the aggregation of the pixels of CLC, except for the median of the spatial bias obtained with 0.65 of threshold which was slightly higher for E than G (Fig. 1, Figs. S1, S2). In contrast, the distribution of values for level 2 EUNIS was different changing the thresholds (Fig. 2, Figs. S3, S4), especially for the spatial and temporal bias (Appendix: Tables S8, S9, S10). Peculiarities in the median values of some habitat types exist. In E1, the median value of the temporal bias was 0 for 0.65 of the threshold, while the median value was 0.406 for 0.75 of the threshold and 0.579 for 0.85 of the threshold. In E3, the median value of the spatial bias for 0.65 of the threshold was 0.860, which was also higher than the median value of the taxonomic bias. Whereas the median value of the spatial bias for 0.75 of threshold was 0.716, and for 0.85 of threshold was 0.335. Finally, the median value of the temporal bias of G2 was very low for all cases: 0 for 0.75 of the threshold, 0.154 for 0.65 of the threshold and 0.154 for 0.85 of the threshold.

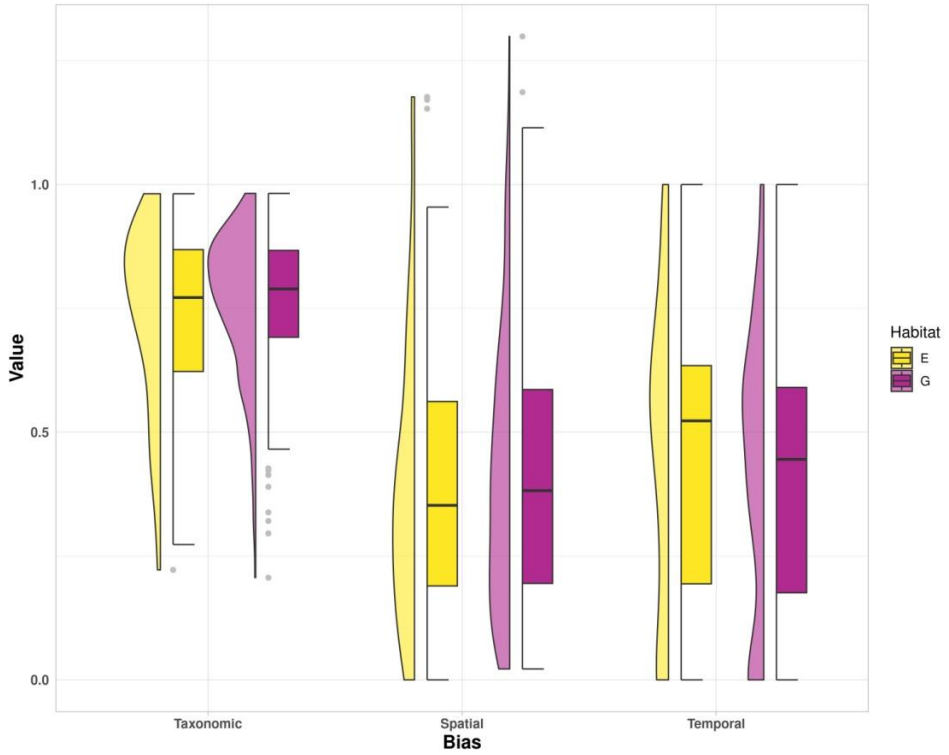


Figure 1: Distribution values of the dimensions of bias (taxonomic, spatial, temporal) at EUNIS level 1 for E (grassland and lands dominated by forbs, mosses and lichens) and G (woodland, forest and other woodland).

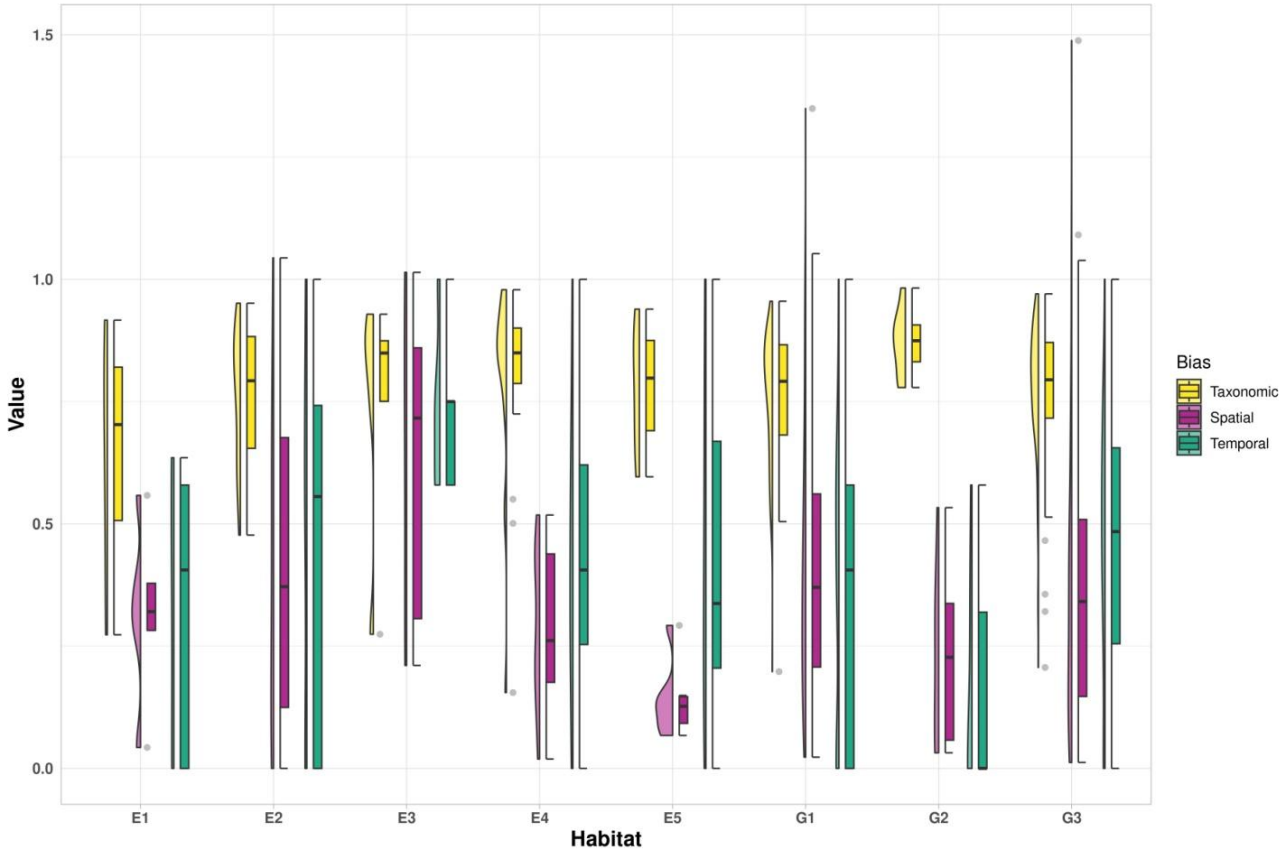


Figure 2: Distribution of values of the dimensions of bias (taxonomic, spatial, temporal) at EUNIS level 2 for E (i.e., E1,E2, E3, E4, E5) and G (i.e., G1, G2, G3).

4 Discussion

We measured the dimensions of bias (i.e., taxonomic, spatial and temporal) at the habitat level of sPlotOpen records to point out which habitat might suffer the highest level of inaccuracy across Europe. We observed a generally low level of taxonomic bias, a high level of spatial bias, and moderate levels of temporal bias for habitat types E and G at levels 1 and 2 EUNIS. The pattern of values at level 2 EUNIS changed with different values of threshold and showed variations among habitats especially for the spatial and the temporal bias. No significant variations were highlighted between E and G, demonstrating that the intensity of the bias was the same regardless of the habitat type. Nevertheless, few habitats at level 2 showed significant differences between them, as in the case of E3 – E5 for the spatial bias and E3 – G1, E3 – G2 for the temporal bias. These results may rely on the low number of recorded plots, according to our filtering criteria, over Europe representing the habitats E3 (seasonally wet and wet grasslands), E5 (woodland fringes and clearings and tall forb stands), and G2 (broadleaved evergreen woodland). Indeed, many are the conditions that could have influenced these outcomes: the number of plots describing each habitat type, their geographic distribution, a non-probability sampling, and the aggregation set-up of CLC which excluded different grid cells according to the threshold value (as happened for 0.65 of threshold). Differences may also occur when changing the spatial resolution of the grid cells, however, particular attention should be paid to the choice of the resolution because the grid cell

must cover the corresponding land cover as much as possible, especially for the calculation of the spatial bias. Overall, higher effort is needed to sample species occurrence data at spatial and temporal scales in a way that the sampling is as uniform, complete and standardized as possible taking into account the specific condition needed to fulfil the study purposes.

In this study, the taxonomic bias at EUNIS level 1 was low (i.e., higher completeness of the species richness) for both E and G habitat types. Also, the habitats at EUNIS level 2 showed low bias. Nevertheless, differences can exist between habitats. Indeed, the completeness of the species richness can be driven by several natural and anthropogenic drivers. High species richness does not always reflect high sample completeness because many species may still be undetected (Cazzolla Gatti et al. 2022), and their potential for discovery can depend on the sampling effort (Button and Borzée 2024). Moreover, some species can be more difficult to detect because of their inaccessibility or rarity. However, the tendency to sample more charismatic species can be another factor which leads to ignoring some species rather than others (Troudet et al. 2017; Adamo et al. 2021; Callaghan et al. 2021). On overall, the knowledge of species distribution can be biased by road accessibility (Hughes et al. 2021), human population density (Mair and Ruete 2016), the amount of financial resources (Meyer, Weigelt, and Kreft 2016) and the presence of protected area (Girardello et al. 2019). For instance, Chanachai et al. (2024) showed that some ecoregions had low forest sample completeness associated with low protected land area and natural habitat, as well as, low sample completeness per sampling units where prevailed low forest integrity. Heterogeneity in our actual knowledge of taxonomic coverage may result in misrepresentations of the species diversity and occurrences and unsuccessful conservation actions (Cazzolla Gatti et al. 2022).

Eventually, higher accuracy on the outcomes of the completeness of species richness may be obtained if the metric is calculated for vegetation plots of equal size area and by using plots with the same sampling year, as the completeness changes over time (Hortal et al. 2008; La Sorte and Somveille 2020). However, we should always consider the potential presence of taxonomic bias in the sampling procedures by prioritizing unbiased sampling, particularly in regions with unique climatic conditions where endemic and rare species are more likely to occur (Enquist et al. 2019; Sandel et al. 2020).

The spatial bias was high (i.e., low NNI) for both E and G habitat types at level 1 and for the habitats at EUNIS level 2, meaning that the sampling in the geographic space had a clustered distribution for all of them. Several factors may have determined this arrangement of the plots, such as the environmental resampling of the plots of sPlot, which may affect the geographic coverage (Sabatini et al. 2021a), the combination of several databases in one with different designs and research purposes, the opportunistic data collections with socio-economic preferences. Another factor that can have shaped the spatial bias is the original biased distribution of the vegetation plots of sPlot gathered from the EVA database (version Chytrý et al. 2014).

Furthermore, the geographic distribution of the habitats may have influenced the NNI outcomes. Indeed, although grassland habitats are widely distributed around Europe, we observed that many of them have a peculiar location and some of them have a restricted distribution. Just to cite a

few, E4.4b is distinctive of the high mountains of the Balkan and Apennines, E2.4 is a unique Iberian summer pasture, and E5.2c is a Macaronesian habitat in the Canary Islands catheterized by perennial herbaceous communities. The same goes for the forest and woodlands which have some habitats of restricted distribution, such as Macaronesian laurophyllous woodland (G 2.3), sub endemic *Alnus cordata* woodland (G 1.Ba), Mediterranean montane Cedrus woodland (G 3.4d). These specific locations and restricted distributions of the habitats may have determined a clustered sampling of the vegetation plots. In any case, since there are no georeferenced polygons of EUNIS habitats at the second and third levels for the European extent, we could not evaluate how much the values of spatial bias were determined by the distribution of the habitats themselves. However, we believe that providing free accessible EUNIS habitat polygons or vector data on a small spatial scale (e.g., region or country) may address this issue.

Finally, particular attention must be paid to data derived from preferential sampling. Many studies demonstrated that a preferential sampling of plots rather than a probability sampling can distort estimates (Botta-Dukát et al. 2007; Michalcová et al. 2011; Alessi et al. 2023). Chytrý (2001) showed that preferential sampling is more prone to sample species-rich locations. Especially, this sampling method is particularly common in historical surveys (Chytrý 2001; Reddy and Dávalos 2003; Monsarrat, Boshoff, and Kerley 2019).

In this study, the temporal bias had medium values for EUNIS level 1 E and G meaning that the habitats surveys were more intense in some years rather than others being the relative abundances of the years not uniform. The EUNIS habitats at level 2 showed different values of Pielou's Index with a distribution of values that in most of the cases covered all the possible values of the index (range from 0 to 1), denoting that the sampling was very irregular in the temporal dimension across the grid cells in Europe. Peculiar is the case of G2 with a median value of the Index equal to zero, circumstance that happens when the year of recording of the plots inside the grid cell is equal. Indeed, if a habitat has a high Pielou's Index (low temporal bias), the data information could have a greater coverage at the taxonomic level; nevertheless, this assumption is not sufficient to confirm the reliability of the data since the data might still suffer from possible taxonomic and spatial bias. The bias dimensions are not independent of each other (Meyer, Weigelt, and Kreft 2016), in fact, one dimension of bias can influence the others and vice-versa with a specific direction and intensity (Fisher-Phelps et al. 2017). In our view, direction and intensity should be evaluated depending on the purpose of the study (Zizka et al. 2020).

The habitat, as a complex ecological identity (Yapp 1922), can involve changes in the interactions and co-presence of multiple species in space and time but also in the physical and environmental conditions. Hence, if data characterizing a habitat type shows bias in one or more dimensions (taxonomic, spatial and temporal), intrinsically, it can implicate a bias in the environmental conditions (edaphic and climatic). A few studies measured the environmental dimension of bias of collected data at the species level only. Oliveira et al. (2016); Monsarrat et al. (2019) measured the environmental bias as the effect of uneven sampling of the species observations.

Given the peculiar characteristics of the habitat and its importance for conservation biology, the extent of sampling bias should not be underestimated, even more so considering that habitats are

prone to habitat fragmentation and loss. Approximately half of grassland habitats are threatened by agricultural intensification or abandonment, natural succession, urbanization, and forestry activities (Habel et al. 2013; Schils et al. 2022). Forest habitats, instead, are threatened by sylviculture, natural hazards, climate change, alien species invasion (Janssen et al. 2016; Dyderski et al. 2018). These threats contribute to habitats change in their coverage and distribution and to the re-assemblage of the local pool of species (Lindborg et al. 2012; Riibak et al. 2020; Pazúr et al. 2024). For instance, sampling of species occurrence data that does not reflect the condition of the habitat prior to and after habitat alteration can influence biodiversity trends and produce biased estimates (Zhang et al. 2021). Similarly, sampling bias of the species occurrences alongside with their environmental conditions can affect the accuracy of Species Distribution Models in predicting the habitat suitability of the species (Stolar and Nielsen 2015; Bardon et al. 2021) providing distort reccommendations to biodiversity conservation and invasive species distribution restrictions.

As a consequence of these changes, the quality of data and how data is sampled need to be taken into account when dealing with habitat-based studies and conservation actions. In fact, it applies the principle that the data collected or taken from databases should be as free as possible from errors and information gaps.

Data availability statement

The data that support the findings of this study are openly available in Zenodo at DOI https://doi.org/10.5281/zenodo.14840330

References

Adamo, Martino, Matteo Chialva, Jacopo Calevo, Filippo Bertoni, Kingsley Dixon, and Stefano Mammola. 2021. 'Plant Scientists' Research Attention Is Skewed towards Colourful, Conspicuous and Broadly Distributed Flowers'. *Nature Plants* 7 (5): 574–78. https://doi.org/10.1038/s41477-021-00912-2.

Alessi, Nicola, Gianmaria Bonari, Piero Zannini, Borja Jiménez-Alfaro, Emiliano Agrillo, Fabio Attorre, Roberto Canullo, et al. 2023. 'Probabilistic and Preferential Sampling Approaches Offer Integrated Perspectives of Italian Forest Diversity'. *Journal of Vegetation Science* 34 (1): e13175. https://doi.org/10.1111/jvs.13175.

Baddeley, Adrian, Ege Rubak, and Rolf Turner. 2016. *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics Series. Boca Raton London New York: CRC Press.

Baker, David J., Ilya M. D. Maclean, Martin Goodall, and Kevin J. Gaston. 2022. 'Correlations between Spatial Sampling Biases and Environmental Niches Affect Species Distribution Models'.

Edited by Pedro Peres-Neto. *Global Ecology and Biogeography* 31 (6): 1038–50. https://doi.org/10.1111/geb.13491.

Baker, David J., Ilya M. D. Maclean, and Kevin J. Gaston. 2024. 'Effective Strategies for Correcting Spatial Sampling Bias in Species Distribution Models without Independent Test Data'. *Diversity and Distributions* 30 (3): e13802. https://doi.org/10.1111/ddi.13802.

Bardon, L. R., B. A. Ward, S. Dutkiewicz, and B. B. Cael. 2021. 'Testing the Skill of a Species Distribution Model Using a 21st Century Virtual Ecosystem'. *Geophysical Research Letters* 48 (22): e2021GL093455. https://doi.org/10.1029/2021GL093455.

Beck, Jan, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. 2014. 'Spatial Bias in the GBIF Database and Its Effect on Modeling Species' Geographic Distributions'. *Ecological Informatics* 19:10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton University Press. https://doi.org/10.1515/9781400840908.

Botta-Dukát, Zoltán, Edit Kovács-Láng, Tamás Rédei, Miklós Kertész, and János Garadnai. 2007. 'Statistical and Biological Consequences of Preferential Sampling in Phytosociology: Theoretical Considerations and a Case Study'. *Folia Geobotanica* 42 (2): 141–52. https://doi.org/10.1007/BF02893880.

Boyd, Robin J., Gary D. Powney, Claire Carvell, and Oliver L. Pescott. 2021. 'OccAssess: An R Package for Assessing Potential Biases in Species Occurrence Data'. *Ecology and Evolution* 11 (22): 16177–87. https://doi.org/10.1002/ece3.8299.

Boyd, Robin J., Marcelo A. Aizen, Rodrigo M. Barahona-Segovia, Luis Flores-Prado, Francisco E. Fontúrbel, Tiago M. Francoy, Manuel Lopez-Aliste, et al. 2022. 'Inferring Trends in Pollinator Distributions across the Neotropics from Publicly Available Data Remains Challenging despite Mobilization Efforts'. Edited by Yoan Fourcade. *Diversity and Distributions* 28 (7): 1404–15. https://doi.org/10.1111/ddi.13551.

Button, Sky, and Amaël Borzée. 2024. 'Estimates of the Number of Undescribed Species Should Account for Sampling Effort'. *Nature Ecology & Evolution* 8 (4): 637–40. https://doi.org/10.1038/s41559-023-02312-5.

Callaghan, Corey T., Alistair G. B. Poore, Max Hofmann, Christopher J. Roberts, and Henrique M. Pereira. 2021. 'Large-Bodied Birds Are over-Represented in Unstructured Citizen Science Data'. *Scientific Reports* 11 (1): 19073. https://doi.org/10.1038/s41598-021-98584-7.

Cazzolla Gatti, Roberto, Peter B. Reich, Javier G. P. Gamarra, Tom Crowther, Cang Hui, Albert Morera, Jean-Francois Bastin, et al. 2022. 'The Number of Tree Species on Earth'. *Proceedings of the National Academy of Sciences* 119 (6): e2115329119. https://doi.org/10.1073/pnas.2115329119.

Chanachai, Jariya, Ernest F. Asamoah, Joseph M. Maina, Peter D. Wilson, David A. Nipperess, Manuel Esperon-Rodriguez, and Linda J. Beaumont. 2024. 'What Remains to Be Discovered: A Global Assessment of Tree Species Inventory Completeness'. *Diversity and Distributions* 30 (7): e13862. https://doi.org/10.1111/ddi.13862.

Chao, Anne, Yasuhiro Kubota, David Zelený, Chun-Huo Chiu, Ching-Feng Li, Buntarou Kusumoto, Moriaki Yasuhara, et al. 2020. 'Quantifying Sample Completeness and Comparing Diversities among Assemblages'. *Ecological Research* 35 (2): 292–314. https://doi.org/10.1111/1440-1703.12102.

Chapman, Melissa, Benjamin R. Goldstein, Christopher J. Schell, Justin S. Brashares, Neil H. Carter, Diego Ellis-Soto, Hilary Oliva Faxon, et al. 2024. 'Biodiversity Monitoring for a Just Planetary Future'. *Science* 383 (6678): 34–36. https://doi.org/10.1126/science.adh8874.

Chytrý, Milan. 2001. 'Phytosociological Data Give Biased Estimates of Species Richness'. *Journal of Vegetation Science* 12 (3): 441–44. https://doi.org/10.1111/j.1654-1103.2001.tb00190.x.

Chytrý, Milan, Lubomír Tichý, Stephan M. Hennekens, and Joop H.J. Schaminée. 2014. 'Assessing Vegetation Change Using Vegetation-plot Databases: A Risky Business'. Edited by Jürgen Dengler. *Applied Vegetation Science* 17 (1): 32–41. https://doi.org/10.1111/avsc.12050.

Chytrý, Milan, Lubomír Tichý, Stephan M. Hennekens, Ilona Knollová, John A. M. Janssen, John S. Rodwell, Tomáš Peterka, et al. 2020. 'EUNIS Habitat Classification: Expert System, Characteristic Species Combinations and Distribution Maps of European Habitats'. Edited by Sebastian Schmidtlein. *Applied Vegetation Science* 23 (4): 648–75. https://doi.org/10.1111/avsc.12519.

Clark, Philip J., and Francis C. Evans. 1954. 'Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations'. *Ecology* 35 (4): 445–53. https://doi.org/10.2307/1931034.

Daru, Barnabas H., and Jordan Rodriguez. 2023. 'Mass Production of Unvouchered Records Fails to Represent Global Biodiversity Patterns'. *Nature Ecology & Evolution* 7 (6): 816–31. https://doi.org/10.1038/s41559-023-02047-3.

Davies, Cynthia E., and Dorian Moss. 1999. 'EUNIS habitat classification. Final report to the European Topic Centre on Nature Conservation'. *European Environment Agency*, 256.

Davies, Cynthia E., Dorian Moss, and Mark O. Hill. 2004. 'EUNIS habitat classification revised 2004'. *Report to: European environment agency-European topic centre on nature protection and biodiversity*, 127-143.

Dyderski, Marcin K., Sonia Paź, Lee E. Frelich, and Andrzej M. Jagodziński. 2018. 'How Much Does Climate Change Threaten European Forest Tree Species Distributions?' *Global Change Biology* 24 (3): 1150–63. https://doi.org/10.1111/gcb.13925.

Enquist, Brian J., Xiao Feng, Brad Boyle, Brian Maitner, Erica A. Newman, Peter Møller Jørgensen, Patrick R. Roehrdanz, et al. 2019. 'The Commonness of Rarity: Global and Future Distribution of Rarity across Land Plants'. *Science Advances* 5 (11): eaaz0414. https://doi.org/10.1126/sciadv.aaz0414.

Fisher-Phelps, Marina, Guofeng Cao, Rebecca M. Wilson, and Tigga Kingston. 2017. 'Protecting Bias: Across Time and Ecology, Open-Source Bat Locality Data Are Heavily Biased by Distance to Protected Area'. *Ecological Informatics* 40:22–34. https://doi.org/10.1016/j.ecoinf.2017.05.003.

Garcia-Rosello, Emilio, Jacinto Gonzalez-Dacosta, Cástor Guisande, and Jorge M. Lobo. 2023a. 'GBIF Falls Short of Providing a Representative Picture of the Global Distribution of Insects'. *Systematic Entomology* 48 (4): 489–97. https://doi.org/10.1111/syen.12589.

García-Roselló, Emilio, Jacinto González-Dacosta, and Jorge M. Lobo. 2023b. '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. https://doi.org/10.1016/j.biocon.2023.110118.

Geldmann, Jonas, Jacob Heilmann-Clausen, Thomas E. Holm, Irina Levinsky, Bo Markussen, Kent Olsen, Carsten Rahbek, and Anders P. Tøttrup. 2016. 'What Determines Spatial Bias in Citizen Science? Exploring Four Recording Schemes with Different Proficiency Requirements'. Edited by Brian Leung. *Diversity and Distributions* 22 (11): 1139–49. https://doi.org/10.1111/ddi.12477.

Girardello, Marco, Anna Chapman, Roger Dennis, Lauri Kaila, Paulo A. V. Borges, and Andrea Santangeli. 2019. 'Gaps in Butterfly Inventory Data: A Global Analysis'. *Biological Conservation* 236 (August):289–95. https://doi.org/10.1016/j.biocon.2019.05.053.

Graco-Roza, Caio, Sonja Aarnio, Nerea Abrego, Alicia T. R. Acosta, Janne Alahuhta, Jan Altman, Claudia Angiolini, et al. 2022. 'Distance Decay 2.0 – A Global Synthesis of Taxonomic and Functional Turnover in Ecological Communities'. *Global Ecology and Biogeography* 31 (7): 1399–1421. https://doi.org/10.1111/geb.13513.

Habel, Jan Christian, Jürgen Dengler, Monika Janišová, Péter Török, Camilla Wellstein, and Michal Wiezik. 2013. 'European Grassland Ecosystems: Threatened Hotspots of Biodiversity'. *Biodiversity and Conservation* 22 (10): 2131–38. https://doi.org/10.1007/s10531-013-0537-x.

Haddad, Nick M., Lars A. Brudvig, Jean Clobert, Kendi F. Davies, Andrew Gonzalez, Robert D. Holt, Thomas E. Lovejoy, et al. 2015. 'Habitat Fragmentation and Its Lasting Impact on Earth's Ecosystems'. *Science Advances* 1 (2): e1500052. https://doi.org/10.1126/sciadv.1500052.

Hall, Linnea S., Paul R. Krausman, and Michael L. Morrison. 1997. 'The Habitat Concept and a Plea for Standard Terminology'. *Wildlife Society Bulletin (1973-2006)* 25 (1): 173–82. https://www.jstor.org/stable/3783301.

Hortal, Joaquín, Alberto Jiménez-Valverde, José F. Gómez, Jorge M. Lobo, and Andrés Baselga. 2008. 'Historical Bias in Biodiversity Inventories Affects the Observed Environmental Niche of the Species'. *Oikos* 117 (6): 847–58. https://doi.org/10.1111/j.0030-1299.2008.16434.x.

Hughes, Alice C., Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, and Huijie Qiao. 2021. 'Sampling Biases Shape Our View of the Natural World'. *Ecography* 44 (9): 1259–69. https://doi.org/10.1111/ecog.05926.

Hugo, Sanet, and Res Altwegg. 2017. 'The Second Southern African Bird Atlas Project: Causes and Consequences of Geographical Sampling Bias'. *Ecology and Evolution* 7 (17): 6839–49. https://doi.org/10.1002/ece3.3228.

Jandt, Ute, Helge Bruelheide, Florian Jansen, Aletta Bonn, Volker Grescho, Reinhard A. Klenke, Francesco Maria Sabatini, et al. 2022. 'More Losses than Gains during One Century of Plant Biodiversity Change in Germany'. *Nature* 611 (7936): 512–18. https://doi.org/10.1038/s41586-022-05320-w.

Janssen, J. a. M., J. S. Rodwell, M. García Criado, S. Gubbay, T. Haynes, A. Nieto, N. Sanders, et al. 2016. *European Red List of Habitats Part 2. Terrestrial and freshwater habitats*. England. https://doi.org/10.2779/091372.

Johnson, T. F., A. P. Beckerman, D. Z. Childs, T. J. Webb, K. L. Evans, C. A. Griffiths, P. Capdevila, et al. 2024. 'Revealing Uncertainty in the Status of Biodiversity Change'. *Nature* 628 (8009): 788–94. https://doi.org/10.1038/s41586-024-07236-z.

La Sorte, Frank A., and Marius Somveille. 2020. 'Survey Completeness of a Global Citizen-science Database of Bird Occurrence'. *Ecography* 43 (1): 34–43. https://doi.org/10.1111/ecog.04632.

Lessa, Thainá, Janisson W. Dos Santos, Ricardo A. Correia, Richard J. Ladle, and Ana C. M. Malhado. 2019. 'Known Unknowns: Filling the Gaps in Scientific Knowledge Production in the Caatinga'. Edited by Tzai-Hung Wen. *PLOS ONE* 14 (7): e0219359. https://doi.org/10.1371/journal.pone.0219359.

Lessa, Thainá, Juliana Stropp, Joaquín Hortal, and Richard J. Ladle. 2024a. 'How Taxonomic Change Influences Forecasts of the Linnean Shortfall (and What We Can Do about It)?' *Journal of Biogeography* 51 (8): 1365–73. https://doi.org/10.1111/jbi.14829.

Lessa, Thainá, Fernanda Alves-Martins, Javier Martinez-Arribas, Ricardo A. Correia, John Mendelsohn, Ezequiel Chimbioputo Fabiano, Simon T. Angombe, Ana C.M. Malhado, and Richard J. Ladle. 2024b. 'Quantifying Spatial Ignorance in the Effort to Collect Terrestrial Fauna in Namibia, Africa'. *Ecological Indicators* 158:111490. https://doi.org/10.1016/j.ecolind.2023.111490.

Lindborg, Regina, Aveliina Helm, Riccardo Bommarco, Risto K. Heikkinen, Ingolf Kühn, Juha Pykälä, and Meelis Pärtel. 2012. 'Effect of Habitat Area and Isolation on Plant Trait Distribution in European Forests and Grasslands'. *Ecography* 35 (4): 356–63. https://doi.org/10.1111/j.1600-0587.2011.07286.x.

Lobo, Jorge M., Andrés Baselga, Joaquín Hortal, Alberto Jiménez-Valverde, and Jose F. Gómez. 2007. 'How Does the Knowledge about the Spatial Distribution of Iberian Dung Beetle Species Accumulate over Time?' *Diversity and Distributions* 13 (6): 772–80. https://doi.org/10.1111/j.1472-4642.2007.00383.x.

Mair, Louise, and Alejandro Ruete. 2016. 'Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa'. Edited by Judi Hewitt. *PLOS ONE* 11 (1): e0147796. https://doi.org/10.1371/journal.pone.0147796.

Maitner, Brian, Rachael Gallagher, Jens-Christian Svenning, Melanie Tietje, Elizabeth H. Wenk, and Wolf L. Eiserhardt. 2023. 'A Global Assessment of the Raunkiæran Shortfall in Plants: Geographic Biases in Our Knowledge of Plant Traits'. *New Phytologist* 240 (4): 1345–54. https://doi.org/10.1111/nph.18999.

Maldonado, Carla, Carlos I. Molina, Alexander Zizka, Claes Persson, Charlotte M. Taylor, Joaquina Albán, Eder Chilquillo, Nina Rønsted, and Alexandre Antonelli. 2015. 'Estimating Species Diversity and Distribution in the Era of B Ig D Ata: To What Extent Can We Trust Public Databases?' *Global Ecology and Biogeography* 24 (8): 973–84. https://doi.org/10.1111/geb.12326.

Marchetto, Elisa, Martina Livornese, Francesco Maria Sabatini, Enrico Tordoni, Daniele Da Re, Jonathan Lenoir, Riccardo Testolin, et al. 2024. 'Addressing Multiple Facets of Bias and Uncertainty

in Continental Scale Biodiversity Databases'. *Biodiversity Informatics* 18 (September). https://doi.org/10.17161/bi.v18i.21810.

Meyer, Carsten, Holger Kreft, Robert Guralnick, and Walter Jetz. 2015. 'Global Priorities for an Effective Information Basis of Biodiversity Distributions'. *Nature Communications* 6 (1): 8221. https://doi.org/10.1038/ncomms9221.

Meyer, Carsten, Patrick Weigelt, and Holger Kreft. 2016. 'Multidimensional Biases, Gaps and Uncertainties in Global Plant Occurrence Information'. Edited by Janneke Hille Ris Lambers. *Ecology Letters* 19 (8): 992–1006. https://doi.org/10.1111/ele.12624.

Michalcová, Dana, Samuel Lvončík, Milan Chytrý, and Ondřej Hájek. 2011. 'Bias in Vegetation Databases? A Comparison of Stratified-Random and Preferential Sampling: Stratified-Random and Preferential Sampling'. *Journal of Vegetation Science* 22 (2): 281–91. https://doi.org/10.1111/j.1654-1103.2010.01249.x.

Monsarrat, Sophie, Andre F. Boshoff, and Graham I. H. Kerley. 2019. 'Accessibility Maps as a Tool to Predict Sampling Bias in Historical Biodiversity Occurrence Records'. *Ecography* 42 (1): 125–36. https://doi.org/10.1111/ecog.03944.

Moura, Mario R., and Walter Jetz. 2021. 'Shortfalls and Opportunities in Terrestrial Vertebrate Species Discovery'. *Nature Ecology & Evolution* 5 (5): 631–39. https://doi.org/10.1038/s41559-021-01411-5.

Oliveira, Ubirajara, Adriano Pereira Paglia, Antonio D. Brescovit, Claudio J. B. De Carvalho, Daniel Paiva Silva, Daniella T. Rezende, Felipe Sá Fortes Leite, et al. 2016. 'The Strong Influence of Collection Bias on Biodiversity Knowledge Shortfalls of B Razilian Terrestrial Biodiversity'. Edited by Jeremy VanDerWal. *Diversity and Distributions* 22 (12): 1232–44. https://doi.org/10.1111/ddi.12489.

Pazúr, Robert, Jozef Nováček, Matthias Bürgi, Monika Kopecká, Juraj Lieskovský, Zuzana Pazúrová, and Ján Feranec. 2024. 'Changes in Grassland Cover in Europe from 1990 to 2018: Trajectories and Spatial Patterns'. *Regional Environmental Change* 24 (2): 51. https://doi.org/10.1007/s10113-024-02197-5.

Pielou, E.C. 1966. 'The Measurement of Diversity in Different Types of Biological Collections'. *Journal of Theoretical Biology* 13:131–44. https://doi.org/10.1016/0022-5193(66)90013-0.

Prendergast, J. R., R. M. Quinn, J. H. Lawton, B. C. Eversham, and D. W. Gibbons. 1993. 'Rare Species, the Coincidence of Diversity Hotspots and Conservation Strategies'. *Nature* 365 (6444): 335–37. https://doi.org/10.1038/365335a0.

Pardini, R., E. Nichols, and T. Püttker. 2018. 'Biodiversity Response to Habitat Loss and Fragmentation'. In *Encyclopedia of the Anthropocene*, 229–39. Elsevier. https://doi.org/10.1016/B978-0-12-809665-9.09824-4.

Reddy, Sushma, and Liliana M. Dávalos. 2003. 'Geographical Sampling Bias and Its Implications for Conservation Priorities in Africa'. *Journal of Biogeography* 30 (11): 1719–27. https://doi.org/10.1046/j.1365-2699.2003.00946.x.

Riibak, Kersti, Jonathan A. Bennett, Ene Kook, Ülle Reier, Riin Tamme, C. Guillermo Bueno, and Meelis Pärtel. 2020. 'Drivers of Plant Community Completeness Differ at Regional and Landscape Scales'. *Agriculture, Ecosystems & Environment* 301:107004. https://doi.org/10.1016/j.agee.2020.107004.

Rocha-Ortega, Maya, Pilar Rodriguez, and Alex Córdoba-Aguilar. 2021. 'Geographical, Temporal and Taxonomic Biases in Insect GBIF Data on Biodiversity and Extinction'. *Ecological Entomology* 46 (4): 718–28. https://doi.org/10.1111/een.13027.

Ronquillo, Cristina, Juliana Stropp, Nagore G. Medina, and Joaquin Hortal. 2023. 'Exploring the Impact of Data Curation Criteria on the Observed Geographical Distribution of Mosses'. *Ecology and Evolution* 13 (12): e10786. https://doi.org/10.1002/ece3.10786.

Ruete, Alejandro. 2015. 'Displaying Bias in Sampling Effort of Data Accessed from Biodiversity Databases Using Ignorance Maps'. *Biodiversity Data Journal* 3 (July):e5361. https://doi.org/10.3897/BDJ.3.e5361.

Sabatini, Francesco Maria, Jonathan Lenoir, Tarek Hattab, Elise Aimee Arnst, Milan Chytrý, Jürgen Dengler, Patrice De Ruffray, et al. 2021. 'SPlotOpen – An Environmentally Balanced, Open-access, Global Dataset of Vegetation Plots'. *Global Ecology and Biogeography* 30 (9): 1740–64. https://doi.org/10.1111/geb.13346.

Sabatini, Francesco Maria, Jonathan Lenoir, Helge Bruelheide, and the sPlot Consortium. 2021b. 'SPlotOpen – An Environmentally-Balanced, Open-Access, Global Dataset of Vegetation Plots'. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. https://doi.org/10.25829/IDIV.3474-40-3292.

Sandel, Brody, Patrick Weigelt, Holger Kreft, Gunnar Keppel, Masha T. Van Der Sande, Sam Levin, Stephen Smith, Dylan Craven, and Tiffany M. Knight. 2020. 'Current Climate, Isolation and History Drive Global Patterns of Tree Phylogenetic Endemism'. Edited by Ruth Kelly. *Global Ecology and Biogeography* 29 (1): 4–15. https://doi.org/10.1111/geb.13001.

Schils, René L.M., Conny Bufe, Caroline M. Rhymer, Richard M. Francksen, Valentin H. Klaus, Mohamed Abdalla, Filippo Milazzo, et al. 2022. 'Permanent Grasslands in Europe: Land Use Change and Intensification Decrease Their Multifunctionality'. *Agriculture, Ecosystems & Environment* 330:107891. https://doi.org/10.1016/j.agee.2022.107891.

Stolar, Jessica, and Scott E. Nielsen. 2015. 'Accounting for Spatially Biased Sampling Effort in Presence-only Species Distribution Modelling'. Edited by Janet Franklin. *Diversity and Distributions* 21 (5): 595–608. https://doi.org/10.1111/ddi.12279.

Stropp, Juliana, Richard J. Ladle, Ana C. M. Malhado, Joaquín Hortal, Julien Gaffuri, William H. Temperley, Jon Olav Skøien, and Philippe Mayaux. 2016. 'Mapping Ignorance: 300 Years of Collecting Flowering Plants in Africa'. *Global Ecology and Biogeography* 25 (9): 1085–96. https://doi.org/10.1111/geb.12468.

Troudet, Julien, Philippe Grandcolas, Amandine Blin, Régine Vignes-Lebbe, and Frédéric Legendre. 2017. 'Taxonomic Bias in Biodiversity Data and Societal Preferences'. *Scientific Reports* 7 (1): 9132. https://doi.org/10.1038/s41598-017-09084-6.

Zhang, Wenyuan, Ben C. Sheldon, Richard Grenyer, and Kevin J. Gaston. 2021. 'Habitat Change and Biased Sampling Influence Estimation of Diversity Trends'. *Current Biology* 31 (16): 3656-3662.e3. https://doi.org/10.1016/j.cub.2021.05.066.

Zizka, Alexander, Oskar Rydén, Daniel Edler, Johannes Klein, Allison Perrigo, Daniele Silvestro, Sverker C. Jagers, Staffan I. Lindberg, and Alexandre Antonelli. 2021. 'B IO -D EM , a Tool to Explore the Relationship between Biodiversity Data Availability and Socio-political Conditions in Time and Space'. *Journal of Biogeography* 48 (11): 2715–26. https://doi.org/10.1111/jbi.14256.

Zizka, Alexander, Fernanda Antunes Carvalho, Alice Calvente, Mabel Rocio Baez-Lizarazo, Andressa Cabral, Jéssica Fernanda Ramos Coelho, Matheus Colli-Silva, et al. 2020. 'No One-Size-Fits-All Solution to Clean GBIF'. *PeerJ* 8 (September):e9916. https://doi.org/10.7717/peerj.9916.

Wüest, Rafael O., Niklaus E. Zimmermann, Damaris Zurell, Jake M. Alexander, Susanne A. Fritz, Christian Hof, Holger Kreft, et al. 2020. 'Macroecology in the Age of Big Data – Where to Go from Here?' *Journal of Biogeography* 47 (1): 1–12. https://doi.org/10.1111/jbi.13633.

Yang, Wenjing, Keping Ma, and Holger Kreft. 2013. 'Geographical Sampling Bias in a Large Distributional Database and Its Effects on Species Richness–Environment Models'. Edited by W. Daniel Kissling. *Journal of Biogeography* 40 (8): 1415–26. https://doi.org/10.1111/jbi.12108.

Yang, Wenjing, Keping Ma, and Holger Kreft. 2014. 'Environmental and Socio-economic Factors Shaping the Geography of Floristic Collections in C Hina'. *Global Ecology and Biogeography* 23 (11): 1284–92. https://doi.org/10.1111/geb.12225.

Yapp, R. H. 1922. 'The Concept of Habitat'. *The Journal of Ecology* 10 (1): 1. https://doi.org/10.2307/2255427.

Appendix

Table S1: Statistical measures of the mean, the median, the bootstrap estimate of the standard error, and the range for each habitat type for the taxonomic bias.

Habitat	Mean	Median	SE	Range
E	0.732	0.772	0.031	0.222 - 0.982
G	0.761	0.789	0.009	0.206 - 0.982
E1	0.644	0.703	0.157	0.273 - 0.917
E2	0.763	0.793	0.041	0.477 - 0.951
E3	0.735	0.849	0.130	0.274 - 0.929
E4	0.796	0.850	0.025	0.155 - 0.979
E5	0.781	0.798	0.074	0.596 - 0.939
G1	0.761	0.791	0.015	0.198 - 0.955
G2	0.872	0.874	0.030	0.779 - 0.982
G3	0.772	0.795	0.017	0.206 - 0.970

Table S2: Statistical measures of the mean, the median, the bootstrap the estimate of standard error, and the range for each habitat type for the spatial bias.

Habitat	Mean	Median	SE	Range
E	0.399	0.352	0.044	0 - 1.177
G	0.415	0.382	0.025	0.022 - 1.299
E1	0.316	0.321	0.095	0.043 - 0.558
E2	0.399	0.372	0.115	0 - 1.044
E3	0.621	0.716	0.246	0.211 - 1.014
E4	0.287	0.261	0.078	0.019 - 0.518
E5	0.141	0.127	0.031	0.067 - 0.292

G1	0.411	0.370	0.033	0.023 - 1.349
G2	0.225	0.227	0.101	0.032 - 0.533
G3	0.363	0.341	0.046	0.012 - 1.488

Table S3: Statistical measures of the mean, the median, the bootstrap the estimate of standard error, and the range for each habitat type for the temporal bias.

Habitat	**Mean**	**Median**	**SE**	**Range**
E	0.452	0.523	0.053	0 - 1
G	0.419	0.445	0.038	0 - 1
E1	0.324	0.406	0.237	0 - 0.635
E2	0.466	0.556	0.115	0 - 1
E3	0.732	0.750	0.101	0.579 - 1
E4	0.419	0.406	0.091	0 - 1
E5	0.441	0.337	0.189	0 - 1
G1	0.372	0.406	0.027	0 - 1
G2	0.155	0	0.123	0 - 0.579
G3	0.451	0.484	0.048	0 - 1

Table S4: Outcomes of Wilcoxon rank sum tests of the habitat types E and G EUNIS level 1 for each dimension of bias.

Bias	**W statistic**	**p-value**
Taxonomic	5405.5	0.572
Spatial	5366.5	0.519
Temporal	606	0.439

Table S5: Outcomes of Kruskal-Wallis tests of the habitat types EUNIS 2 for each dimension of bias.

Bias	**chi-squared**	**p- value**
Taxonomic	10.700	0.15

Spatial	17.322	**< 0.05**
Temporal	22.3375	**< 0.05**

Table S6: Outcomes of Dunn's test of the habitat types EUNIS 2 for spatial bias.

Comparison	**Z statistic**	**p-value adjusted**
E1 - E2	-0.266	1.000
E1 - E3	-1.564	1.000
E1 - E4	0.266	1.000
E1 - E5	1.425	1.000
E1 - G1	-0.541	1.000
E1 - G2	0.804	1.000
E1 - G3	-0.081	0.936
E2 - E3	-1.941	1.000
E2 - E4	0.867	1.000
E2 - E5	2.185	0.664
E2 - G1	-0.526	1.000
E2 - G2	1.449	1.000
E2 - G3	0.405	1.000
E3 - E4	2.535	0.281
E3 - E5	3.330	**0.024**
E3 - G1	1.833	1.000
E3 - G2	2.772	0.150

E3 - G3	2.440	0.353
E4 - E5	1.558	1.000
E4 - G1	-1.523	1.000
E4 - G2	0.770	1.000
E4 - G3	-0.668	1.000
E5 - G1	-2.642	0.214
E5 - G2	-0.750	1.000
E5 - G3	-2.125	0.739
G1 - G2	1.921	1.000
G1 - G3	1.391	1.000
G2 - G3	-1.334	1.000

Table S7: Outcomes of Dunn's test of the habitat types EUNIS 2 for temporal bias.

Comparison	**Z statistic**	**p-value adjusted**
E1 - E2	-0.890	1.000
E1 - E3	-2.471	0.296
E2 - E3	-2.452	0.298
E1 - E4	-0.517	1.000
E2 - E4	0.578	1.000
E3 - E4	2.799	0.128
E1 - E5	-0.534	1.000
E2 - E5	0.288	1.000

E3 - E5	2.111	0.695
E4 - E5	-0.120	0.905
E1 - G1	-0.253	1.000
E2 - G1	1.428	1.000
E3 - G1	3.728	**0.005**
E4 - G1	0.574	1.000
E5 - G1	0.503	1.000
E1 - G2	1.070	1.000
E2 - G2	2.574	0.231
E3 - G2	4.138	**0.001**
E4 - G2	2.063	0.743
E5 - G2	1.782	1.000
G1 - G2	1.973	0.873
E1 - G3	-0.833	1.000
E2 - G3	0.225	1.000
E3 - G3	2.886	0.102
E4 - G3	-0.486	1.000
E5 - G3	-0.181	1.000
G1 - G3	-1.768	1.000
G2 - G3	-2.677	0.178

Effect of threshold condition

The dimensions of bias (taxonomic, spatial, temporal) were calculated per grid cells of 10 km of spatial resolution. Each grid cell was determined by aggregating the CLC raster layer at 100m of

spatial resolution, applying a threshold (i.e., 0.65, 0.65, 0.85) for the number of NA values to which the pixels were classified as NA.

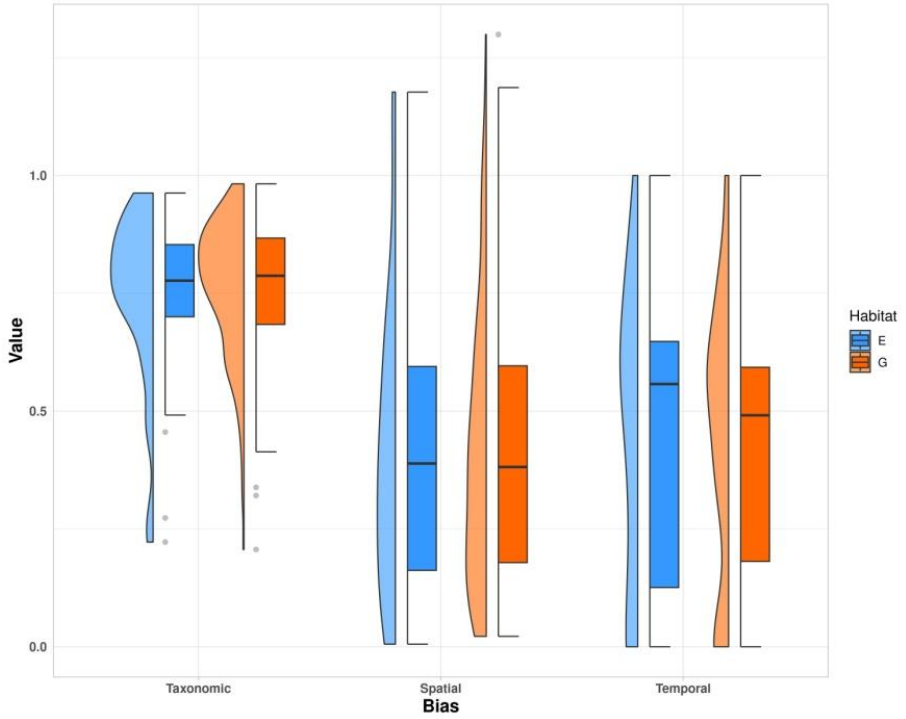


Figure S1: Distribution values of the dimensions of bias (taxonomic, spatial, temporal) at EUNIS level 1 for E and G per grid cells aggregated with a threshold of 0.65.

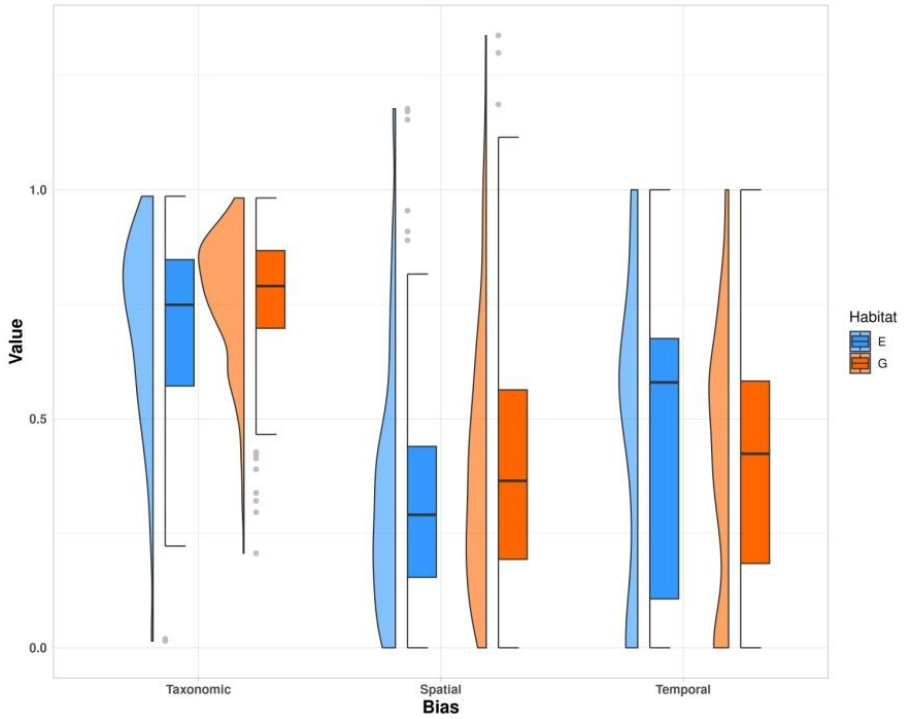


Figure S2: Distribution values of the dimensions of bias (taxonomic, spatial, temporal) at EUNIS level 1 for E and G per grid cells aggregated with a threshold of 0.85.

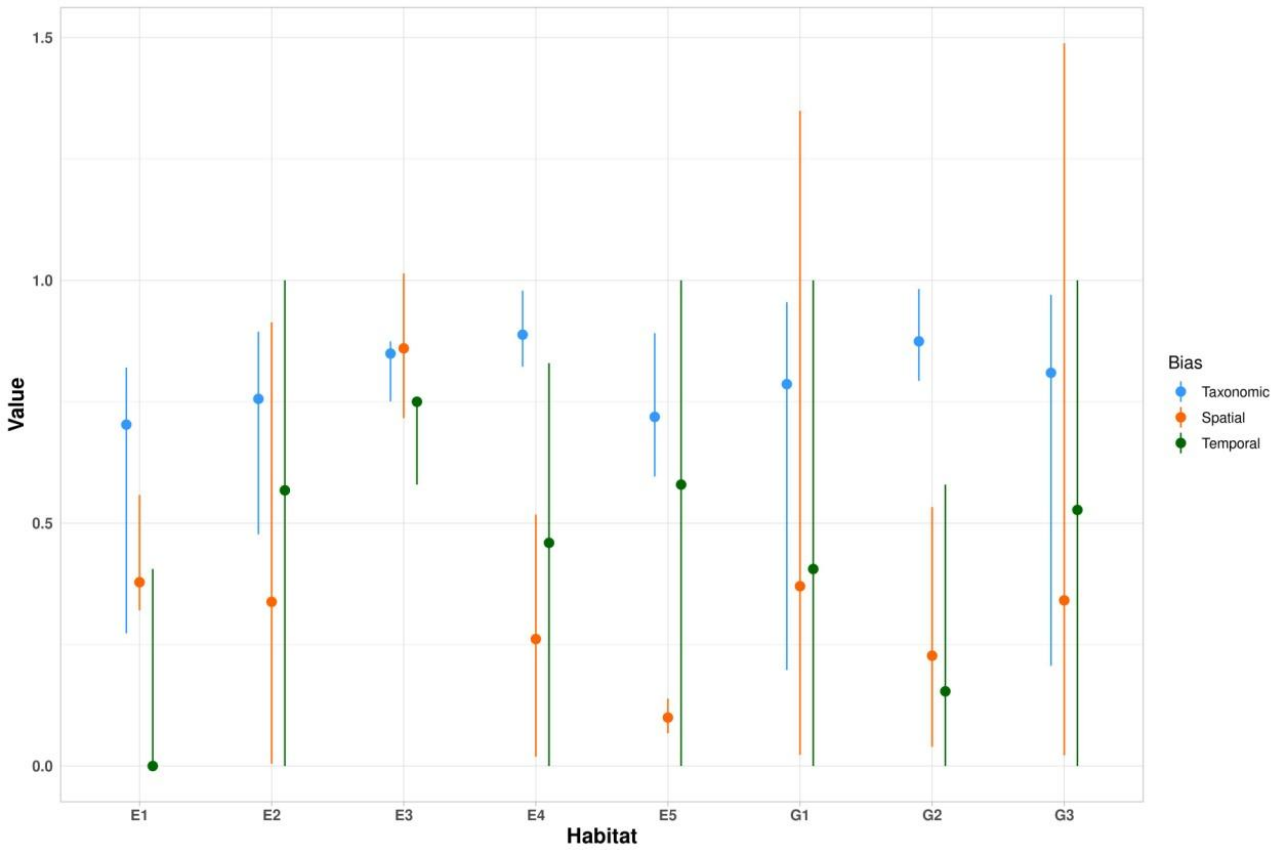


Figure S3: Median (dot) and range (max-min) between minimum and maximum values (line) of the dimensions of bias (taxonomic, spatial, temporal) at EUNIS level 2 for E (i.e., E1, E2, E3, E4, E5) and G (i.e., G1, G2, G3) per grid cells aggregated with a threshold of 0.65.

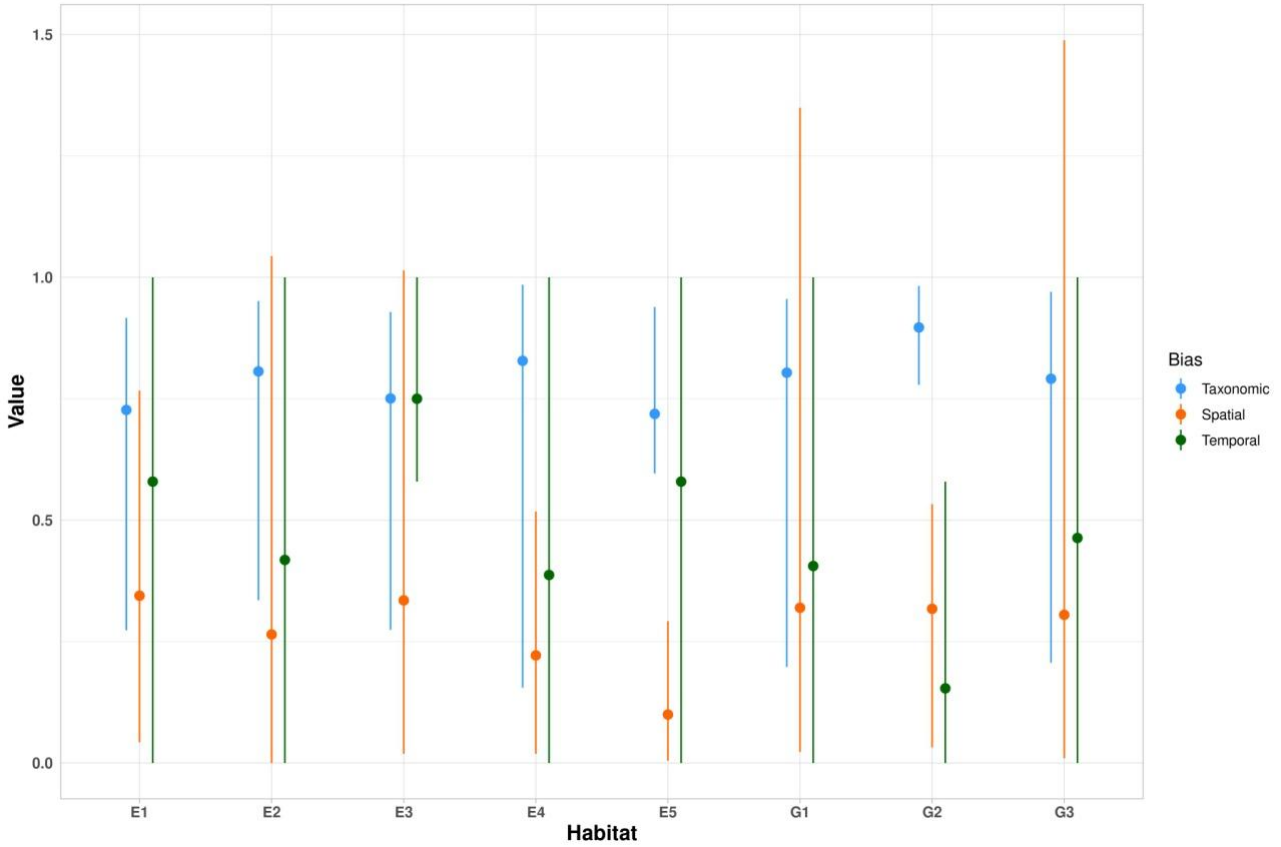


Figure S4: Median (dot) and range (max-min) between minimum and maximum values (line) of the dimensions of bias (taxonomic, spatial, temporal) at EUNIS level 2 for E (i.e., E1, E2, E3, E4, E5) and G (i.e., G1, G2, G3) per grid cells aggregated with a threshold of 0.85.

Table S8: Median values of the taxonomic bias per grid cells aggregated with 0.65, 0.75, 0.85 of thresholds.

Habitat	Median 0.65	Median 0.75	Median 0.85
E	0.777	0.772	0.749
G	0.787	0.789	0.790
E1	0.703	0.703	0.727
E2	0.756	0.793	0.806
E3	0.849	0.849	0.751
E4	0.888	0.850	0.828
E5	0.719	0.798	0.719
G1	0.786	0.791	0.804
G2	0.874	0.874	0.897
G3	0.810	0.795	0.791

Table S9: Median values of the spatial bias per grid cells aggregated with 0.65, 0.75, 0.85 of thresholds.

Habitat	Median 0.65	Median 0.75	Median 0.85
E	0.389	0.352	0.291
G	0.381	0.382	0.364
E1	0.378	0.321	0.344
E2	0.338	0.372	0.265
E3	0.860	0.716	0.335
E4	0.261	0.261	0.222
E5	0.100	0.127	0.100
G1	0.370	0.370	0.320
G2	0.227	0.227	0.318
G3	0.341	0.341	0.305

Table S10: Median values of the temporal bias per grid cells aggregated with 0.65, 0.75, 0.85 of thresholds.

Habitat	Median 0.65	Median 0.75	Median 0.85
E	0.557	0.523	0.579
G	0.491	0.445	0.424
E1	0	0.406	0.579
E2	0.568	0.556	0.418
E3	0.750	0.750	0.75
E4	0.460	0.406	0.387
E5	0.579	0.337	0.579
G1	0.406	0.406	0.406
G2	0.154	0	0.154
G3	0.527	0.484	0.463

Conclusion

In my PhD thesis, I addressed the issue of bias and uncertainty of species occurrence data by using different methodological approaches.

In Chapter 1, I tested the effect of two sampling methods (i.e., random and stratified) and different sample prevalences in Favourability and Probability-based Species Distribution Models. In this study, we highlighted that the standardized sampling methods used did not considerably affect the accuracy of the models although the predictions of the species distribution changed with the spatial scale. Possibly, a biased sampling method might have a greater impact on the performances of the SDMs. Although the effect of sampling bias was not tested in this study, the sampling bias may have different effects on the predictions of favorability and probability model. Indeed, a disproportionate sampling effort for some geographic areas or environmental conditions can fail to determine the actual niche of the species. Favorability and probability predictions can change significantly when the sampling bias leads to a variation in the sample prevalence. A test of the impact of the sampling bias under different conditions on the performances of favourability-based and probability-based models would provide valuable insights. However, even without testing the effect of the sampling bias, we found that the Favourability showed high accuracy in the prediction values and slightly higher performance of the models with respect to probability models (more than half of the median Continuous Boyce index values were higher). Then, it exhibited lower variability of the predicted species distributions when varying the ratio of presences and absences (i.e., sample prevalence) while keeping the sample size fixed (i.e., number of presences and absences being sampled). In future studies, it would be interesting to compare the outcomes by setting different sample size values. However, the property of lower variability of the favourability model has the potential to get more precise comparisons between SDMs and to better detect the environmental conditions that favour the presence of the species.

In Chapter 2, we provided a method to evaluate the bias and uncertainty of species occurrence data in biodiversity databases. We aimed to measure and represent them to raise awareness of possible knowledge gaps in taxonomic, spatial and temporal dimensions when using this data or for correcting them by using methods appropriately calibrated to the study context. We proposed three common metrics to assess bias: i) the completeness of the species richness, ii) the Nearest Neighbor Index, iii) the Pielou's index, and iv) a new easy-apply method for measuring the temporal uncertainty, which relies on the negative exponential function. Certainly, a strength of this study is that the codes are freely accessible making this methodological framework reproducible for biodiversity databases currently available regardless of the spatial scale and for different ecological levels. Future studies might combine the proposed metrics for measuring bias and uncertainty with tests to identify the most suitable methods to correct them by contextualizing them with differente data use (e.g., type of study, ecological level) and the ecology of the species. In fact, the species ecology, for example for mobile and immobile species (generally plants vs animals), can lead to different patterns of bias depending on how the species occurrence data is sampled in the spatial and temporal dimensions. In this context, the metrics of bias were also applied to the species occurrence data to identify possible taxonomic, spatial and temporal

bias at the habitat level (Chapter 3). The advantage of assessing the three dimensions of bias at the habitat level is that it inherently encompasses the environmental bias being the habitat a complex concept that also includes the climatic, edaphic and physical conditions. Likewise, getting reliable information on data gaps can help address resampling campaigns or data fixing actions for data completion and ensure more reliable recommendations for biodiversity conservation.

In this thesis, I also dealt with the possible role played by different spatial variables (presence of Natura 2000 network, human population count, road density, topographic roughness) in shaping the dimensions of bias (taxonomic, spatial, temporal) of species distribution data in sPlotOpen founding out that Natura 2000 network had a significant impact on the sampling bias in the three dimensions. Future studies may test further variables to obtain an even more complete representation of the factors influencing the sampling bias to address correct design campaigns and monitoring programs. It would also be interesting to test whether the data from before the Natura 2000 establishment shows variations on their dependence with the spatial variables.

Understanding the quality of data in biodiversity databases helps address efforts to guarantee more thorough monitoring and conservation of species and their habitats. Additionally, given this, testing the models and techniques employed throughout the data inference phase is crucial to assure increased accuracy and precision.

Acknowledgements

I'd like to thank my family for always believing in me, for always supporting me in my studies, and for always encouraging me towards a path of enrichment. I thank my parents for the love they gave me and the trust in filling my life with experiences. I thank my brother for always believing in our bond.

I want to take this opportunity to thank profoundly Prof. Duccio Rocchini for his precious teachings and for guiding me through the challenging and joyful periods of my PhD studies. I thank him for believing in my abilities and trusting me. A special mention to the convivial moments spent together at the various conferences and during the climbing days.

I would like to express my sincere gratitude to Dr. Enrico Tordoni for being by my side throughout my three years of PhD and for teaching me to conduct thorough scientific studies critically and meticulously. I thank him for his perseverance, dedication and availability.

I thank all the collaborators and colleagues for sharing their knowledge and research projects with me. Special thanks go to my supervisors Prof. Kim Calders and Prof. Vítězslav Moudrý, and colleagues of the two periods abroad for welcoming me into their labs and sharing their knowledge with me leaving good memories of the exchange periods.

I want to sincerely thank the colleagues of lab BIOME with whom I shared moments of support, fun and growth inside and outside the University. I thank them for all the precious scientific and non-scientific chats which taught me a lot. A special thanks to Martina Livornese, the adventure companion of the most difficult project of the PhD thesis.

A heartfelt thank you to all friends who contributed to making me the person I am and who have made me feel at home. I thank Elena, Marco, Daniel, Emanuele, Francesca, Emma, Giulia, Pasquale, Giulio, Morgana, Giorgia, Antonin, Samuel, Anna, Francesco, and the team Black Mamba for being part of my life.