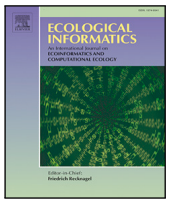
Ecological Informatics 90 (2025) 103289

Contents lists available at ScienceDirect

Ecological Informatics

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



Unravelling Bias: A Sardinian perspective on taxonomic, spatial, and temporal biases in vascular plant biodiversity data from GBIF

Raimondo Melis [a,*,1], Marco Malavasi [a,1], Manuele Bazzichetto [d], Maria Carmela Caria [a], Giovanni Rivieccio [a], Agnese Denaro [a,b], Elisa Marchetto [e], Michela Perrone [d], Martina Livornese [e], Duccio Rocchini [d,e], Simonetta Bagella [a,c]

[a] Department of Chemical, Physical, Mathematical and Natural Sciences, University of Sassari, Via Piandanna 4, Sassari, 07100, Italy
[b] Department of Agricultural Sciences, University of Sassari, Viale Italia 39/a, Sassari, 07100, Italy
[c] Desertification Research Centre, University of Sassari, Via de Nicola, Sassari, 07100, Italy
[d] Department of Spatial Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha-Suchdol, 165 00, Czech Republic
[e] BIOME Lab, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Via Irnerio 42, Bologna, 40126, Italy

ARTICLE INFO

Dataset link: https://github.com/Redkenn/BiasEstimationGBIF

Keywords:
Citizen science
GBIF
Spatial bias
Taxonomic bias
Temporal bias

ABSTRACT

Biodiversity data are expanding rapidly, yet often exhibit significant biases that are rarely mapped or systematically analyzed to understand underlying drivers. This issue is particularly pressing in the era of citizen science, which now contributes a substantial share of biodiversity records. In this study, we assessed taxonomic, temporal, and spatial biases in vascular plant occurrence records from Sardinia, a Mediterranean biodiversity hotspot, using all available occurrence data for the region retrieved from the Global Biodiversity Information Facility (GBIF). The dataset encompasses a range of sources, from structured inventories to citizen science platforms.

Biases were quantified using metrics such as species richness completeness, Pielou's evenness, and the Nearest Neighbor Index (NNI). After mapping these biases, we used Generalized Additive Models (GAMs) to explore their environmental drivers, including road density, the standard deviation of the Normalized Difference Vegetation Index (NDVI), and topographic roughness. Additionally, we evaluated the influence of structured data sources (e.g., Wikiplantbase) versus citizen science platforms (e.g., PlantNet and iNaturalist) on observed bias patterns.

Spatial bias was the most prominent, followed by temporal and taxonomic biases. Road density and NDVI influenced both temporal and taxonomic biases, while topographic roughness affected temporal and spatial biases. Structured data mainly contributed to temporal bias, whereas citizen science data were more associated with spatial bias.

Our findings highlight the importance of addressing biases in biodiversity data, particularly those introduced by citizen science, and provide a replicable framework for improving data quality and biodiversity monitoring at both sampling and interpretation stage.

1. Introduction

Biodiversity is crucial to support multiple ecosystem functions, ultimately providing the ecosystem services essential to sustain human populations (De Groot et al., 2002; Gamfeldt et al., 2008; Cardinale et al., 2012). The change in global biodiversity is occurring at an unprecedented rate becoming one of the most pressing environmental issues of our time. In fact, over one-fifth of all vascular plant species are globally threatened (Willis, 2017) and the associated losses of ecosystem functions and services are expected to cost humanity 7% of the world's gross domestic product by 2050 (Braat et al., 2008). The human appropriation of Earth's natural resources is not only leading to biodiversity loss but also to large alterations in the distribution and abundance of species (Pereira et al., 2012). Within this context, information facilities on plant biodiversity are essential for understanding ecosystem health and making informed conservation decisions.

The Global Biodiversity Information Facility (GBIF; http://www.gbif.org) is a key provider of biodiversity data globally (Berendsohn

* Corresponding author.
E-mail address: rmdmelis@gmail.com (R. Melis).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ecoinf.2025.103289
Received 23 January 2025; Received in revised form 17 June 2025; Accepted 18 June 2025
Available online 3 July 2025

et al., 2010). GBIF is a significant data source for researchers, offering millions of occurrence records that have been used in hundreds of different studies, including biodiversity research, conservation, and predictive modeling in various ecosystems and species groups (Rocchini et al., 2011; Fourcade, 2016; Phillips et al., 2016; Colli-Silva et al., 2019). The accessibility and comprehensive nature of GBIF make it a valuable resource to advance ecological and evolutionary studies. GBIF gathers data from a variety of sources, including museum collections, citizen science observations, scientific literature, camera traps and environmental DNA (Ivanova and Shashkov, 2021; Miller, 2022; Robertson et al., 2022). In short, GBIF includes both structured data from systematic scientific surveys and unstructured, opportunistic observations. Structured data are collected following standardized protocols and predefined sampling designs, ensuring consistency and comparability. In contrast, unstructured data – often contributed by citizen scientists or incidental records by researchers – lack formal design and are typically collected without a consistent methodology, yet provide extensive spatial and temporal coverage.

GBIF has faced criticism regarding data and related biases (Meyer et al., 2016b; Rocchini et al., 2023). Bias and uncertainty are terms developed in the statistical literature and refer to the theory of sampling (Walther and Moore, 2005). Bias is the expected difference between the estimate and the true value of the parameter (Bolker, 2008), or, in other words, when the sampling is not representative of the target statistical population. Data bias denotes the existence of systematic inaccuracies within a dataset, that is, consistent, non-random errors arising from structural issues in how data are collected, recorded, or represented. Such bias may result in erroneous predictions when employing those data for analytical purposes or decision-making processes. Uneven sampling across taxonomic groups, geographic areas, or time periods will cause (Walther and Moore, 2005; Fiske et al., 2008; Arceo-Gómez et al., 2018; McLeod et al., 2021), taxonomic, spatial, and temporal bias.

Taxonomic bias is an inherent feature of organismal research, including biodiversity records, with the representation of taxa in the literature and biodiversity databases not reflecting their complete representation of the species pool in nature (Troudet et al., 2017). This bias has been shown to persist throughout history (Monsarrat and Kerley, 2018), resulting in certain taxa receiving more attention than others in biodiversity research (Phaka et al., 2022). For example, data are more abundant for certain taxonomic groups, such as terrestrial plants, mammals, and birds, resulting in a bias towards terrestrial over marine biodiversity (Bellard et al., 2012). This bias can be attributed to various factors, including the preferences of data collectors, both scientists and non-scientists (Phaka et al., 2022), or the difficulties in locating or identifying some organisms. In addition, the bias is reinforced by the fact that terrestrial ecosystems – particularly in terms of plant diversity – harbor a higher percentage of described species compared to marine ecosystems, partly due to the greater accessibility of terrestrial habitats, and the logistical challenges of exploring and sampling marine environments (Díaz and Malhi, 2022).

Similarly, spatial bias arises due to the distribution of sampling effort, such as uneven sampling in different regions (Beck et al., 2014; Moua et al., 2020), and is common in biodiversity data (Oliveira et al., 2016; Gaul et al., 2020; Piccolo et al., 2020) that significantly affects the precision and reliability of ecological studies (Yang et al., 2013). Spatial bias is often due to the lack of a predefined sampling scheme, which leads observers to survey areas based on personal preferences influenced by factors like accessibility, potential observations, or prior knowledge of the study area (Glad et al., 2019). This bias can distort views of biodiversity, biogeography, and species distributions (Cosentino and Maiorano, 2021), where observed patterns may reflect sampling effort rather than actual environmental or demographic causes (Hugo and Altwegg, 2017).

Finally, decisions regarding the temporal allocation of sampling effort can alter perceived results in ecological studies (Banks-Leite et al., 2012; Rosário et al., 2025). It is essential to consider the temporal distribution of the sampling effort, as it can influence the detection of ecological patterns and lead to different results even with similar total effort but varying temporal protocols (Banks-Leite et al., 2012). Furthermore, the impact of temporal sampling bias on environmental models, like dynamic species distribution models, has been acknowledged, emphasizing the need for methods to correct for such biases (El-Gabbas et al., 2021).

Therefore, recognizing and rectifying biases in biodiversity databases is a crucial step towards enhancing the accuracy and utility of these valuable resources for biodiversity research and conservation efforts. For our study, we explored the vascular plants of Sardinia in the GBIF database to map the related sampling effort and the resulting taxonomic, temporal, and spatial bias. Citizen science and non-citizen science data were aggregated, with the relative contribution of each source within GBIF then used as predictor variables. In particular, we estimated the relationship between the aforementioned biases and environmental predictors such as road density, standard deviation of the Normalized Difference Vegetation Index (NDVI), and topographic roughness. Data sources instead included Wikiplantbase (Peruzzi et al., 2017), PlantNet (Li et al., 2022a), and iNaturalist (Nugent, 2018). Wikiplantbase mainly represents structured scientific surveys primarily involving expert collaboration (Peruzzi et al., 2017), while PlantNet and iNaturalist, mainly deriving from citizen science initiatives, contain unstructured, opportunistic observations (Li et al., 2022b; Nugent, 2018). However, The above variety of data sources is believed to impact the sampling effort, thus potentially influencing the biases accordingly. We consider that knowledge and mapping of such biases (and their drivers) are crucial to improving the GBIF database use by explicitly considering sampling effort to guarantee an effective conservation of biodiversity.

Table 1
Description of vascular plant data sources within GBIF.

Project	Percentage (%)	Time interval	Source
Wikiplantbase	43.19	1950–2020	Bibliography and herbarium specimens
PlantNet	38.42	2009–2022	Citizen science
iNaturalist	7.84	1994–2023	Citizen science
Others	10.55	1952–2023	Various unverified institutional datasets

2. Material and methods

2.1. Data preparation

From GBIF (accessed in November 2023), we downloaded 110,826 records of vascular plant occurrence from the island of Sardinia (Italy), documented over 74 years (1950 to 2023). These records are part of different projects: Wikiplantbase (Peruzzi et al., 2017), PlantNet, iNaturalist (Li et al., 2022a; Nugent, 2018). The data-sets differ in both time interval and origin (Table 1).

We worked with two data-sets derived from the initial 110,826 occurrence records. The first data-set (105,169 records) retained duplicate observations (defined as plant records with identical geographic coordinates but sampled in different years) to estimate the temporal bias of the sampling effort. The second dataset was cleaned to support spatial and taxonomic analyses. Data filtering involved: (a) removing records with low geographic precision (i.e., less than two decimal places in longitude or latitude) or misplaced at sea; (b) excluding duplicate records with identical geographic coordinates; and (c) deleting records with missing data. After these procedures, the cleaned dataset comprised 92,342 observations.

2.2. Bias calculation

We divided the study area into 307 10×10 km^2 grid cells, and then calculated the biases for each grid cell. This procedure led to some

of the grid cells, namely those located at the edges of the study area, having a lower area than those located in the inner areas of Sardinia. The process of estimating bias was thoroughly described in Marchetto et al. (2024) and Backstrom et al. (2025).

2.2.1. Taxonomic bias

To quantify taxonomic bias, we first mapped the 92,342 plant occurrence records into their respective grid cells. As a bias metric, we used species richness completeness, defined as the ratio between the observed species richness (S_{obs}) in a sample and the estimated true species richness (S), which includes both observed and undetected species (Chao et al., 2020):

$$ {}^{q}\mathrm{C} = \frac{S_{obs}}{S} \tag{1} $$

where q is the order of diversity of the Hill number (for theoretical details, see Hill, 1973). When $q = 0$, Eq. (1) turns out to be:

$$ {}^{0}\mathrm{C} = \frac{S_{obs}}{\hat{S}_{Chao1}} \tag{2} $$

which is equivalent to replacing true species richness in the formula by the Chao1 species richness estimator (Chao, 1984). We computed the sample completeness using the R package function *iNEXT* (Chao et al., 2014; Hsieh et al., 2016), by specifying the order of diversity of the Hill number $q = 0$ and the number of equally-spaced knots $k = 5$. With the condition $q = 0$, all species are counted equally regardless of their relative abundances. Input data for each grid cell (assemblage) include species abundances in an empirical sample of n individuals. The index ranges from 0 to 1. High completeness values correspond to low taxonomic bias, whereas low values reflect high taxonomic bias.

2.2.2. Temporal bias

To quantify the temporal bias, we first mapped the 105,169 plant occurrence records (i.e., records with duplicates) in the respective grid cells. As a bias metric, we used the Shannon–Wiener index (H), which is widely used in ecology to quantify the diversity of species in a community (Shannon, 1948). The index was calculated as:

$$ H = -\sum_{i=1}^{S} p_i \ln p_i \tag{3} $$

where p_i is the relative abundance of species i and S represents the total number of species in the sample. We used Pielou's evenness (J), a measure derived from the Shannon–Wiener index to quantify the degree of uniformity in the distribution of individuals across the various species comprising a given community (Pielou, 1966). The index was calculated as the ratio of the Shannon–Wiener index (H) to its theoretical maximum value (H_{max}):

$$ J = \frac{H}{H_{max}} = \frac{H}{\ln S} \tag{4} $$

We adapted Eqs. (3) and (4) to quantify the temporal bias in the sampling effort by making the following modifications: (1) S (total number of species) → Y (total number of years of recording), where $Y = 74$ years in this study; (2) p_i (species relative abundance) → r_i (proportion of records in year i), calculated as:

$$ r_i = \frac{n_i}{N} \tag{5} $$

where n_i is the number of records in year i and N is the total number of records across all years. Thus, the modified equations were:

$$ J' = \frac{H'}{H'_{max}} = \frac{H'}{\ln Y} \tag{6} $$

$$ H' = -\sum_{i=1}^{Y} r_i \ln r_i \tag{7} $$

In this context, J' approaching 1 indicates a uniform distribution of sampling effort across years, suggesting minimal temporal bias. Conversely, low J' values indicate that sampling effort is concentrated in a few years, revealing a strong temporal bias. Pielou's evenness values were mapped to the grid, providing a spatial representation of temporal bias across Sardinia. The metric was calculated using the R package *vegan* (Oksanen et al., 2007).

2.2.3. Spatial bias

To quantify spatial bias, we first mapped the 92,342 plant occurrence records in the respective grid cells. We used the Nearest Neighbor Index (NNI) to determine the degree of clustering of plant occurrence records in a cell (Zhao et al., 2023). We computed NNI using the R package function *clarkevansCalc* (Clark and Evans, 1954) with the Cumulative Distribution Function method *cdf*, which applies edge corrections. The Clark-Evans criterion (R) was calculated as follows:

$$ R = \frac{\bar{r}A}{\bar{r}E} \tag{8} $$

where $\bar{r}A$ is the actual mean nearest neighbor distance and $\bar{r}E$ is the expected mean nearest neighbor distance at random distribution of objects. If the Clark-Evans criterion equals 1, the spatial distribution type is random (chaotic). If it exceeds 1, the distribution is even (dispersed). If the Clark-Evans criterion is less than 1, then the distribution is aggregated (clustered).

2.2.4. Trivariate map

To visualize the statistical (or geographic) relationship among temporal, taxonomic, and spatial bias, we used a trivariate map, a thematic representation displaying three variables simultaneously. The bias metrics (Completeness, J Index, and NNI) were transformed to a common scale before visualization. Since the indices were already within the 0–1 range, no rescaling was applied. However, for generalizability, if any index were to exceed 1, it could be transformed using the following min–max normalization:

$$ I_{scaled} = \frac{I - I_{min}}{I_{max} - I_{min}} \tag{9} $$

where I is the original bias index, and I_{min} and I_{max} are its observed minimum and maximum values in the dataset. The variables selected for the trivariate map were the completeness of the species richness for the taxonomic bias (Completeness), Pielou's evenness for the temporal bias (J Index), and the nearest neighbor index (NNI) for the spatial bias. The map was generated using the R package *tricolore* (Schöley, 2018). The final visualization of the map highlights grid cells with lower bias values.

2.3. Explanatory variables

We calculated the explanatory variables within the same 10 × 10 km^2 grid cells used to assess the biases. Seven explanatory variables (Table 2) were selected to describe the spatial pattern of the different biases. Road density, standard deviation of the Normalized Difference Vegetation Index (NDVI), and mean topographic roughness (which are landscape or environmental features); and four are proxy of the sampling effort associated with the original sources of the plant occurrence records: Wikiplantbase, PlantNet, iNaturalist, Others.

We used road density (Fig. 1A) as a proxy of cell-specific (average) accessibility. We calculated Road density as the ratio of the length (m) of the total road network in the cell to the area (m^2) of the cell. Road network data of the study area was obtained from the Sardinia Geoportal (https://www.sardegnageoportale.it, scale of 1:10000).

The Normalized Difference Vegetation Index (NDVI) provides a measure of vegetation reflectance that is widely used to quantify vegetation greenness (Tucker, 1979). The standard deviation of the NDVI has often been used as a continuous measure of the dispersion (variation) of NDVI values in a given area since it explains a reasonable portion of the variability in situ diversity data, therefore serving as a proxy of plant diversity within each cell (Perrone et al., 2023). The

Table 2
List of variables used to model the relationship between the three biases (taxonomic, temporal, and spatial) and the explanatory variables. All variables range between 0 and 1.

Explanatory variables	
Acronym	Description
RD	Road density
sdNDVI	Standard deviation of the Normalized Difference Vegetation Index
MR	Mean topographic roughness
TotalSE	Total sampling effort
WikiW	Wikiplantbase weight
PlantW	PlantNet weight
iNaturW	iNaturalist weight
OthersW	Weight of others sources
Lon	Longitude
Lat	Latitude
Response variables	
Acronym	Description
Completeness	Completeness of the species richness for the taxonomic bias
J Index	Pielou's evenness for the temporal bias
NNI	Nearest Neighbor Index for the spatial bias

NDVI data used in this study were derived from a 'greenest-pixel' composite image, constructed using the USGS Landsat 8 atmospherically corrected Surface Reflectance Red (B4) and Near-infrared (B5) bands at a 30 m spatial resolution. This composite was generated in Google Earth Engine (https://earthengine.google.com, Gorelick et al., 2017) by processing images acquired over the study area during the growing seasons (April to September) from 2013 to 2016, selecting the pixels with the highest NDVI values from overlapping images. Subsequently, the standard deviation of the NDVI was calculated for each grid cell (Fig. 1B).

Topographic roughness is an essential geomorphological variable that can be quantified using digital elevation models (DEMs; https://spacedata.copernicus.eu) (Wilson, 2012). Topographic roughness is a measure of terrain variability, calculated as the difference in elevation between the highest and lowest points within a given cell and the eight neighboring cells. We used the worldwide DEM coverage (raster of 30 m spatial resolution) to obtain the DEM coverage of Sardinia, which was required to determine the topographic roughness (in degrees) of the entire research region. The roughness values were computed using the *terrain()* function from the *raster* package in R (Hijmans et al., 2015). This approach ensures that roughness captures local variations in topography, reflecting the degree of terrain heterogeneity. The mean topographic roughness was calculated for each grid cell (Fig. 1C) by aggregating the roughness values within the cell. Any grid cells lacking data were assigned a roughness value of zero, and the resulting values were normalized between 0 and 1.

All plant occurrence records (105,169) were mapped in the respective grid cells as a proxy for the total sampling effort, and the same procedure was performed for each data source (Wikiplantbase, PlantNet, Others, and iNaturalist). The "others" source included a variety of potentially unreliable sources, such as various institutional data or unverified public contributions. Despite their heterogeneous origins, these sources were treated as a homogeneous group due to the absence of metadata that would allow for more precise classification. The sampling effort for each data source was then divided by the total sampling effort (Fig. 1D) to obtain their contribution (weight) on a scale ranging from 0 to 1 (Fig. 1E, F, G, and H).

2.4. Generalized additive models (GAMs)

To assess the relationship between the three biases (taxonomic, temporal, and spatial) and the explanatory variables, we used Generalized Additive Models (GAMs; Wood and Augustin, 2002). GAMs embody versatile modeling tools that combine parametric and nonparametric methodologies, thereby enabling the integration of 'almost linear' patterns within regression models (Wood, 2017). The smoothing function *s*() of the GAM model used for our analysis is the default, i.e., the thin plate regression spline. We incorporated spatial autocorrelation into the model using the term *s*(Lon, Lat), where Lon and Lat are the longitude and latitude coordinates of the grid cell centers. Neglecting spatial autocorrelation might lead to overfitting and an overoptimistic perception of predictive capacity (Gazis and Greinert, 2021). All GAMs were developed using the *mgcv* package (Wood and Wood, 2015). For each bias metric, we compared five candidate models with Akaike's information criterion (AIC; Akaike, 1973). The first model includes a combination of environmental variables (RD, sdNDVI, MR), while the other five models incorporate these environmental variables along with total sampling effort (TotalSE) and data sources from bibliographic and herbarium records, citizen science, and other unspecified sources (WikiW, PlantW, iNaturW, OthersW). All combinations of variables showed no collinearity. The model with the lowest AIC score was regarded as the best candidate model. For the best models, we showed the statistical significance of the predictors and the partial effects on the response variables.

3. Results

3.1. Taxonomic, spatial and temporal biases

The taxonomic bias map of the vascular plants (Fig. 2A) showed a heterogeneous distribution in the study area, with a mean value of completeness of the species richness of 0.54. The highest values of taxonomic bias were identified in the grid cells of the north-central area of Sardinia, while the lowest mainly along the coast (Fig. 2A). The temporal bias map (Fig. 2B), on the other hand, showed a mean J Index value of 0.46. The variability of temporal bias was less pronounced among cells; however, some areas exhibited a more even distribution over time compared to others. The spatial bias map (Fig. 2C) described by the NNI showed that all grid cells exhibited a clustering of the distribution of plant records, with a mean value of 0.33 and a maximum value of 0.86, i.e., no evidence of a random distribution (NNI = 1).

The ternary plot showed that the majority of grid cells were distributed near the center of the triangle, indicating a relatively balanced contribution of the three bias components (Fig. 3, right panel). In particular, 38 grid cells had low temporal bias, and 22 had low taxonomic bias. Only 5 grid cells had the spatial bias lower than the taxonomic and temporal bias.

3.2. Generalized additive models

The model with the lowest AIC for taxonomic bias (Table 3) was the second one, which included TotalSE (AIC = −251.03, adj. $R^2 = 0.54$). In this best-fitting model, RD ($p < 0.001$), sdNDVI ($p = 0.043$), TotalSE ($p < 0.001$), and spatial autocorrelation ($p < 0.001$) were all statistically significant predictors of completeness.

For the temporal bias (Table 3), the model with the lowest AIC value was the third with WikiW (AIC = −646.6940 and adj. $R^2 = 0.59$). Based on the best temporal bias model, we found that RD ($p < 0.001$), sdNDVI ($p = 0.003$), MR (< 0.001), WikiW ($p < 0.001$), and spatial autocorrelation ($p < 0.001$) were statistically significant for the response variable (J Index).

Last, for the spatial bias (Table 3), the best model was the fourth with PlantW (AIC = −510.2732 and adj. $R^2 = 0.65$). MR ($p = 0.030$), PlantW ($p < 0.001$), and spatial autocorrelation ($p = 0.006$) were all statistically significant predictors for the response variable (NNI).

The partial effects of predictors on the response variable suggested that an increase in RD resulted in a decrease in taxonomic and temporal bias, both following non-linear trends (Fig. 4A and E), with most RD values of 0–0.30. The taxonomic bias decreased linearly with increasing

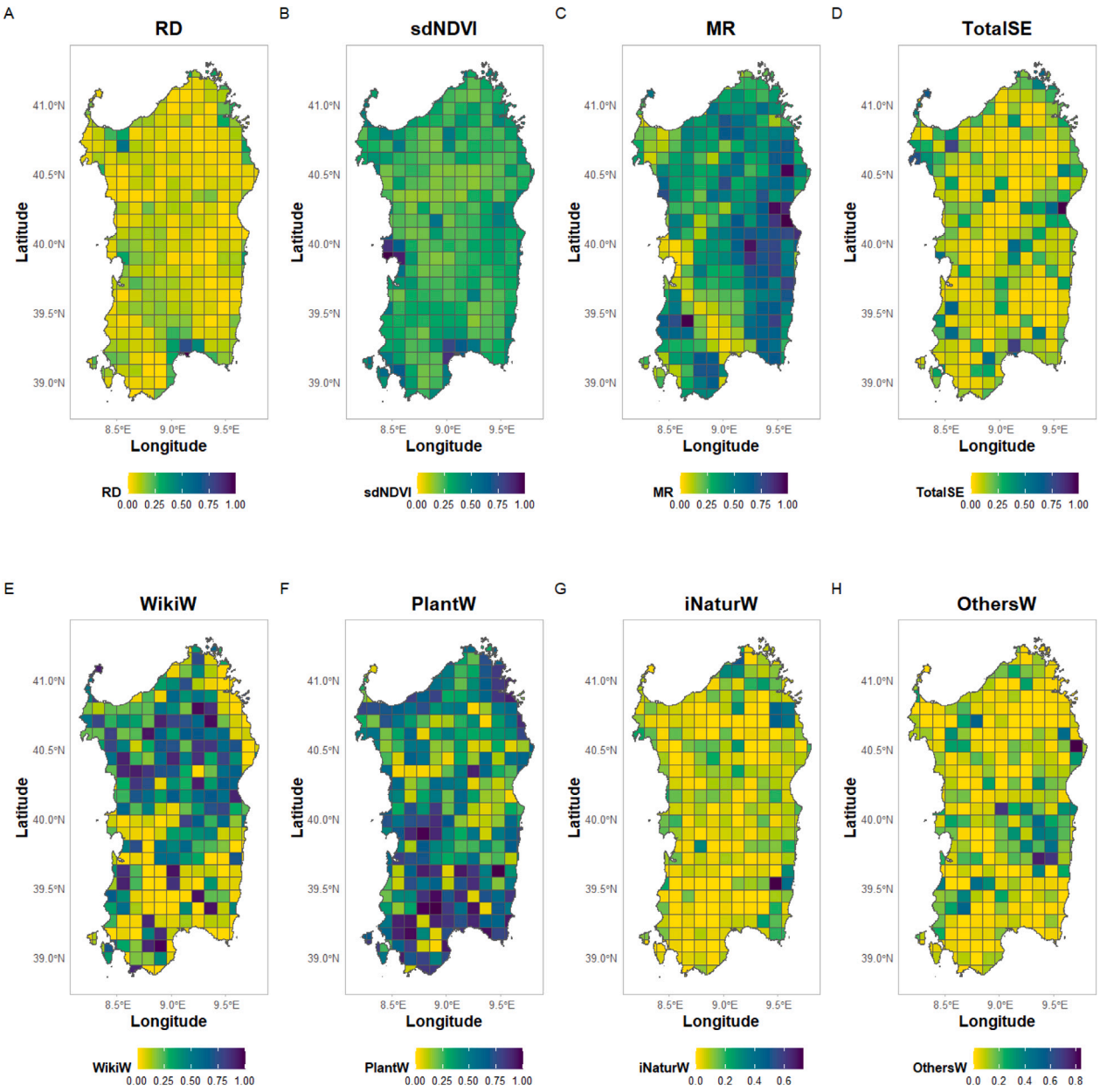


Fig. 1. Explanatory variables map of Sardinia. (A) road density (RD), (B) standard deviation of the Normalized Difference Vegetation Index (sdNDVI), (C) mean topographic roughness (MR), (D) total sampling effort (TotalSE), (E) Wikiplantbase weight (WikiW), (F) PlantNet weight (PlantW), (G) iNaturalist weight (iNaturW), and (H) weight of other sources (OthersW) (10×10 km^2 grid).

sdNDVI (Fig. 4B), similar to the temporal bias (Fig. 4E). Completeness remained highly variable at low TotalSE values but became more stable as sampling effort increased, thereby reducing taxonomic bias (Fig. 4C). An increase in MR corresponded to a marked reduction in temporal bias (Fig. 4F) but a moderate increase in spatial bias (Fig. 4H). The effect of WikiW on temporal bias followed a non-linear, hump-shaped pattern, with the lowest bias observed at intermediate values (Fig. 4G). Additionally, a strong positive correlation indicated that a higher proportion of records from PlantW contributed to a more spatially uniform distribution, mitigating spatial bias (Fig. 4I).

4. Discussion

In this study, we examined the taxonomic, temporal, and spatial biases present in the GBIF vascular plant records of Sardinia. We also investigated the relationship between these biases and predictors, as well as the contribution of each data source within GBIF. We found that in Sardinia, the biases were moderate and varied throughout the island. Overall, the GBIF dataset of vascular plants from Sardinia is fairly balanced, although it exhibits spatial biases and, to a lesser extent, temporal and taxonomic biases (Fig. 3). Considering the three biases altogether, the data collection was mostly homogeneous across vascular plant species, relatively evenly distributed across the 74 year, but clustered within specific areas of the island, indicating the need for sampling to address these spatial gaps. If such gaps are not addressed, a significant spatial bias can distort the current representation of community composition, environmental conditions, and the potential species distribution (Bazzichetto et al., 2023). Taxonomic bias may have been driven by various factors, in particular the presence of numerous endemic species on which collections and studies have been concentrated mainly in the last decades of the previous century (Arrigoni and Nardi, 1977; Arrigoni, 1983). This may have led to an oversampling of endemic species, which are often located in mountain areas, while other species were underrepresented during the same limited time frame.

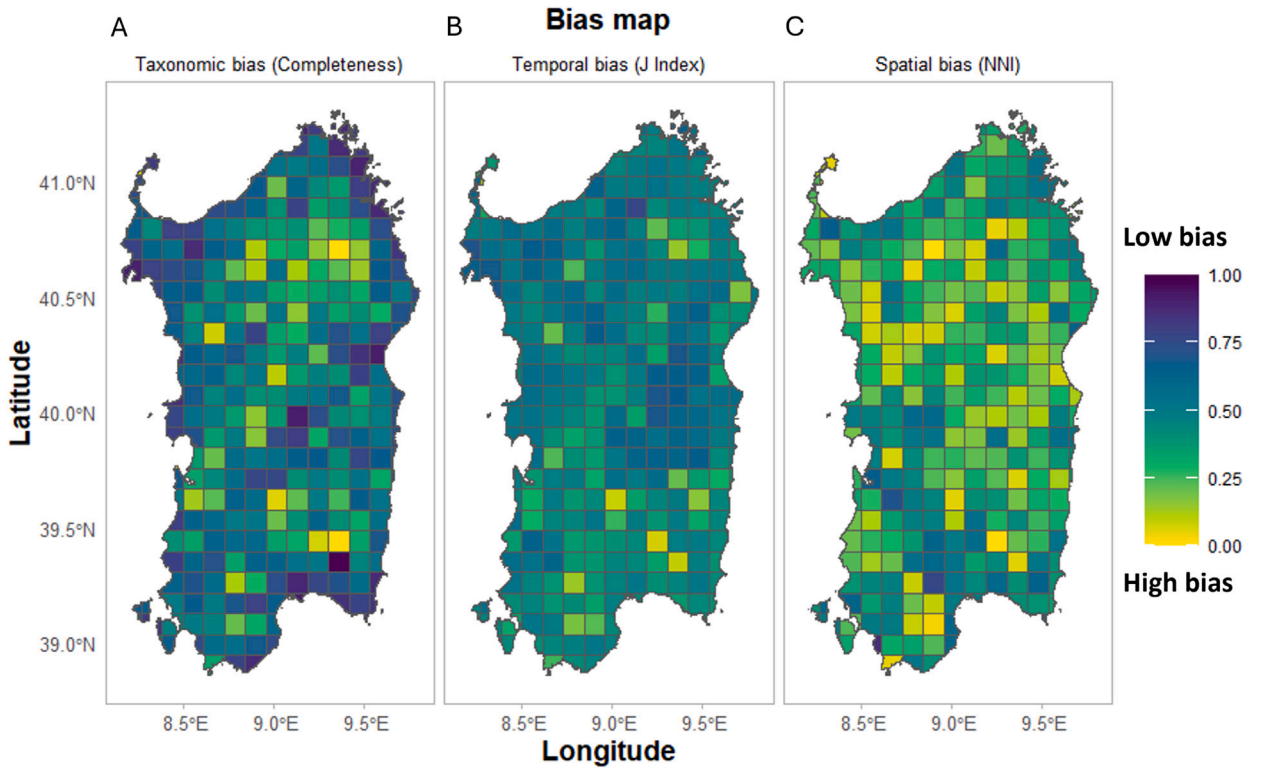


Fig. 2. Vascular plants bias map of Sardinia based on GBIF records from the period 1950–2023. (A) Completeness of the species richness for the taxonomic bias (Completeness), (B) Pielou's evenness for the temporal bias (J Index), and (C) Nearest Neighbor Index for the spatial bias (NNI) (10 × 10 km^2 grid).

In general, the biases were primarily explained by variables related to area accessibility, vegetation greenness, and topographic complexity, as well as those derived from bibliographic and citizen science initiatives. More specifically, for taxonomic and temporal biases, the higher presence of roads increases species richness (i.e., lower taxonomic bias) and Pielou's evenness (i.e., lower temporal bias). Although roads are often associated with the overestimation of species or the repeated sampling of the same species (Mair and Ruete, 2016; Maitner et al., 2023), in our study area, a higher road density may have increased accessibility, thereby enhancing the collection of plant records. This could have helped to reduce the gap between observed and true species richness, ultimately improving the completeness of species richness estimates (Fig. 4A). Nevertheless, the presence of roads itself can affect the composition of plant species in a given area, potentially influencing the distribution and abundance of sampled plants (de Beer et al., 2023).

Topographic roughness was a strong predictor of temporal bias, with higher values corresponding to a marked reduction in bias (Fig. 4F). This suggests that areas with higher topographic complexity tended to exhibit a more temporally consistent sampling effort. One possible explanation is that these areas, despite being potentially less accessible, attracted sustained attention from researchers and naturalists due to their ecological richness and perceived conservation value. Topographic roughness drives the formation of diverse habitats and microenvironments that support different species (Stein et al., 2014), and has been shown to significantly influence the distribution, composition, and diversity of vascular plants in various ecosystems, including Mediterranean islands (Raduła et al., 2022; Bátori et al., 2023; Camilleri et al., 2024). Such areas are often considered biodiversity hotspots and may be prioritized for repeated fieldwork and long-term monitoring (Willis et al., 2007), contributing to more uniform temporal coverage in biodiversity data. In contrast, the effect of MR on spatial bias was markedly weaker and even slightly negative. This pattern might reflect the fact that topographically complex areas also present physical constraints, such as steep slopes or fragmented terrain, that limit the spatial coverage of plant surveys. As a result, sampling within these areas tends to be more spatially clustered within accessible zones, leaving other areas under-sampled.

The slightly negative relationship between sdNDVI and both taxonomic and temporal bias suggests that the sampling of areas characterized by higher vegetation variability tends to be more complete, although the effect is not particularly strong. Areas with high variability in vegetation often harbor higher species diversity (Mapfumo et al., 2016; Kumar et al., 2022; Tan et al., 2022), which may attract scientific attention and repeated sampling over time. This interpretation aligns with the pattern observed for MR, where more structurally complex landscapes were also associated with reduced temporal bias.

The relationship between total sampling effort and taxonomic bias revealed a varying trend: at low sampling effort, taxonomic bias was highly variable, but as sampling effort increased, the bias decreased and stabilized. This pattern suggests that areas with low sampling effort can have selective sampling, potentially leading to the underrepresentation of certain taxa. Conversely, higher sampling effort improves the likelihood of capturing a broader range of taxa, resulting in a more taxonomically balanced dataset.

The effect of Wikiplantbase sampling efforts on temporal bias exhibited a non-linear, hump-shaped pattern, with the lowest bias occurring at intermediate levels. This suggests that a moderate integration of bibliographic and herbarium data enhances the temporal consistency of records, likely due to their structured and long-term nature. In contrast, very low or very high contributions from these sources tend to increase temporal bias. Herbarium specimens and literature-based records are often unevenly distributed over time, reflecting periods of intense collecting activity rather than consistent sampling. Additionally, these archival data are not always collected using standardized protocols, and their temporal resolution can be uncertain. These factors may explain the fluctuating levels of temporal bias observed across different proportions of WikiW.

In the case of spatial bias, as PlantNet sampling efforts increased, the spatial distribution of the plant records was less clustered (i.e., less

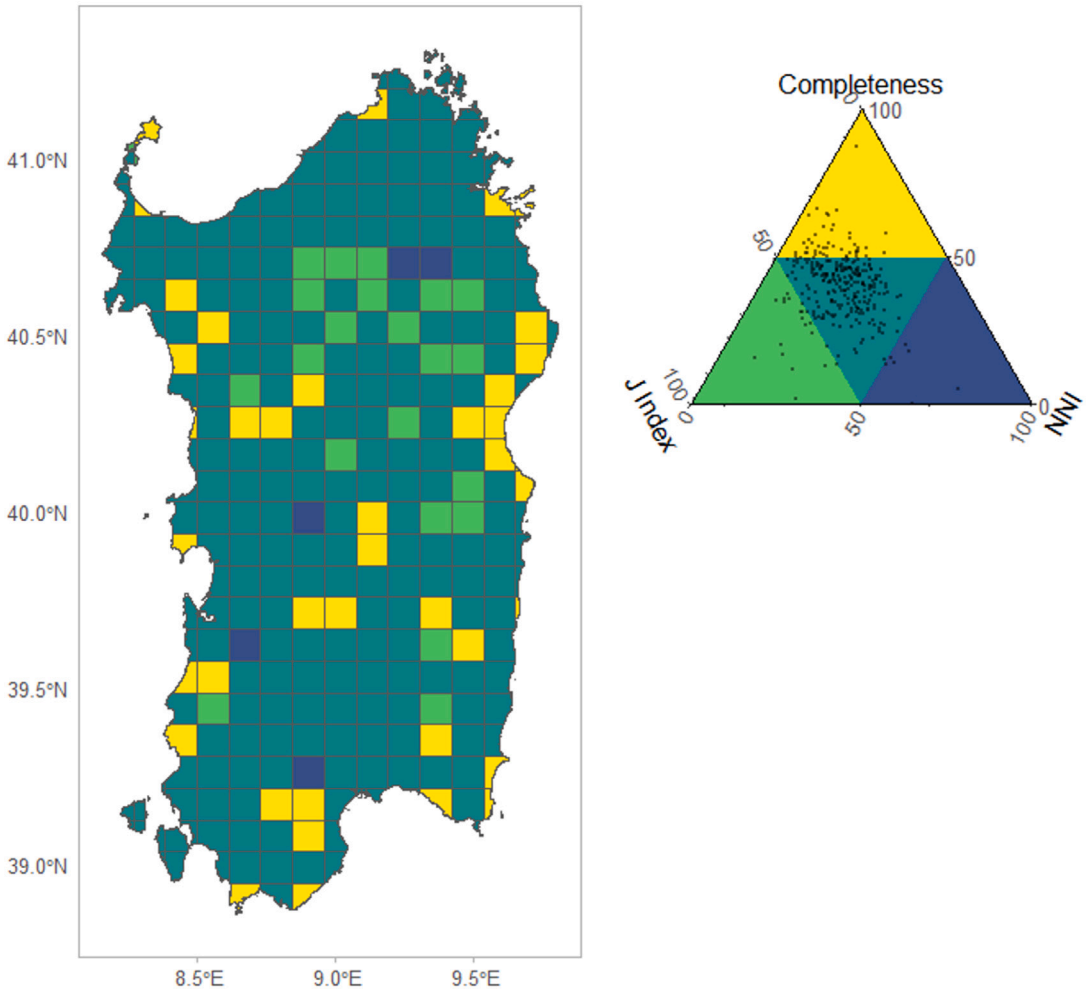


Fig. 3. Trivariate map of bias metrics. The map on the left shows the grid cells colored according to combined taxonomic, temporal, and spatial bias. The legend on the right indicates the contribution of each bias metric: Completeness (taxonomic bias), Pielou's evenness index (J index, temporal bias), and Nearest Neighbor Index (NNI, spatial bias) (10 × 10 km^2 grid).

spatial bias). PlantNet is a direct (i.e., using mobile applications) plant species identification and recognition platform, making data collection more accessible for users and thus increasing the probability of recording more species present in different areas within grid cells. In other words, the sampling effort would be more evenly dispersed. On the contrary, Wikiplantbase also includes older observations for which original detailed geographic coordinates are not available. In such cases, the distribution records would appear clustered since the toponym is the only geographic reference available to locate such old observations, thus resulting in several records with the same coordinates. PlantNet is a recent educational platform for participatory and collaborative research to produce, aggregate, and disseminate botanical observations at large scale using mobile applications for the identification and recognition of plant species (Li et al., 2022a). This means that PlantNet has the capacity to collect data faster than Wikiplantbase. This is not so surprising if we look at the details of our dataset (Table 1); while we found a similar proportion of data from PlantNet and Wikiplantbase, the duration of data collection differs significantly, with 13 years of observations from PlantNet and 70 years from Wikiplantbase, respectively. On the contrary, Wikiplantbase includes data collected by plant scientists, and the registration process of georeferenced records is more formal, i.e., it requires access credentials, the data entered are verified by the project coordinators and then are publicly visible on the website (Peruzzi et al., 2017).

The use of citizen science in biodiversity monitoring has been highlighted as a valuable approach, allowing greater participation in data collection and increasing awareness of plant biodiversity (Chozas et al., 2023; Forti and Szabo, 2024). To ensure the reliability of citizen science data, initiatives have been taken to align citizen science practices with best practices in biodiversity conservation, underscoring the importance of sharing data with national and international biodiversity repositories (Steven et al., 2019). However, different platforms – such as Wikiplantbase, PlantNet, and iNaturalist – features different degree in the reliability of species identification (Hart et al., 2023). Moreover, the growing volume of data poses a challenge due to the limited number of experts available for validation (Saoud et al., 2020). By implementing protocols to enhance data quality, considering historical biases in surveys, and addressing sampling bias in ecological niche modeling, researchers can mitigate the impact of bias on biodiversity data analysis and interpretation. Efforts to promote inclusivity and diversity in biodiversity data collection, such as citizen science initiatives, can help to improve the overall quality of biodiversity data (Paleco et al., 2021). However, before implementing biodiversity estimation measures, one must first understand the potential biases in their data, and our methodological approach is a useful tool for this purpose.

In conclusion, our analysis identifies essential focal points for forthcoming sampling initiatives. To reduce taxonomic bias, prioritizing central Sardinia over coastal regions is recommended. Addressing spatial

Table 3
Model comparison for taxonomic, temporal and spatial bias. The statistical significance values of the predictors were shown for the best candidate models.

Taxonomic bias			
Candidate models	No. var	AIC	Adj. R^2
Completeness = s(RD) + s(sdNDVI) + s(MR) + s(Lon,Lat)	4	−192.1362	0.43
Completeness = s(RD) + s(sdNDVI) + s(MR) + s(TotalSE) + s(Lon,Lat)	**5**	**−251.0304**	**0.54**
Completeness = s(RD) + s(sdNDVI) + s(MR) + s(WikiW) + s(Lon,Lat)	5	−194.3102	0.43
Completeness = s(RD) + s(sdNDVI) + s(MR) + s(PlantW) + s(Lon,Lat)	5	−186.6619	0.43
Completeness = s(RD) + s(sdNDVI) + s(MR) + s(OthersW) + s(Lon,Lat)	5	−188.0749	0.43
Completeness = s(RD) + s(sdNDVI) + s(MR) + s(iNaturW) + s(Lon,Lat)	5	−195.8028	0.45
Best predictors	F	p-value	
s(RD)	2.28	<**0.001*****	
s(sdNDVI)	0.29	**0.043***	
s(MR)	0.19	0.082	
s(TotalSE)	10.75	<**0.001*****	
s(Lon,Lat)	1.96	<**0.001*****	
Temporal bias			
Candidate models	No. var	AIC	Adj. R^2
J Index = s(RD) + s(sdNDVI) + s(MR) + s(Lon,Lat)	4	−498.2321	0.32
J Index = s(RD) + s(sdNDVI) + s(MR) + s(TotalSE) + s(Lon,Lat)	5	−520.9024	0.38
J Index = s(RD) + s(sdNDVI) + s(MR) + s(WikiW) + s(Lon,Lat)	**5**	**−646.6940**	**0.59**
J Index = s(RD) + s(sdNDVI) + s(MR) + s(PlantW) + s(Lon,Lat)	5	−576.1263	0.49
J Index = s(RD) + s(sdNDVI) + s(MR) + s(OthersW) + s(Lon,Lat)	5	−571.8692	0.49
J Index = s(RD) + s(sdNDVI) + s(MR) + s(iNaturW) + s(Lon,Lat)	5	−538.4304	0.42
Best predictors	F	p-value	
s(RD)	1.35	<**0.001*****	
s(sdNDVI)	0.71	**0.003****	
s(MR)	1.86	<**0.001*****	
s(WikiW)	27.59	<**0.001*****	
s(Lon,Lat)	1.69	<**0.001*****	
Spatial bias			
Candidate models	No. var	AIC	Adj. R^2
NNI = s(RD) + s(sdNDVI) + s(MR) + s(Lon,Lat)	4	−312.1457	0.32
NNI = s(RD) + s(sdNDVI) + s(MR) + s(TotalSE) + s(Lon,Lat)	5	−423.3246	0.55
NNI = s(RD) + s(sdNDVI) + s(MR) + s(WikiW) + s(Lon,Lat)	5	−455.9686	0.58
NNI = s(RD) + s(sdNDVI) + s(MR) + s(PlantW) + s(Lon,Lat)	**5**	**−510.2732**	**0.65**
NNI = s(RD) + s(sdNDVI) + s(MR) + s(OthersW) + s(Lon,Lat)	5	−312.1461	0.32
NNI = s(RD) + s(sdNDVI) + s(MR) + s(iNaturW) + s(Lon,Lat)	5	−323.2458	0.36
Best predictors	F	p-value	
s(RD)	0.21	0.079	
s(sdNDVI)	0.02	0.271	
s(MR)	0.42	**0.030***	
s(PlantW)	39.63	<**0.001*****	
s(Lon,Lat)	0.38	**0.006****	

Abbreviations: F, explained variance; adj. R^2, adjusted R^2; Completeness, completeness of the species richness for the taxonomic bias; J Index, Pielou's evenness for the temporal bias; NNI, Nearest Neighbor Index for the spatial bias; RD, road density; sdNDVI, standard deviation of the Normalized Difference Vegetation Index; MR, mean topographic roughness; TotalSE, total sampling effort; WikiW, Wikiplantbase weight; PlantW, PlantNet weight; iNaturW, iNaturalist weight; OthersW, weight of other sources; Lon, longitude; Lat, latitude (10 × 10 km^2 grid).

bias requires meticulous pre-planning of vegetation surveys, coupled with randomized sampling techniques. Conversely, temporal bias is largely inevitable, emphasizing that only select areas may reliably indicate species turnover. Additionally, careful evaluation of data sources within the GBIF dataset is essential, as the dominance of specific datasets can lead to varying levels of bias.

CRediT authorship contribution statement

Raimondo Melis: Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Marco Malavasi:** Writing – review & editing, Writing – original draft, Supervision, Resources, Conceptualization. **Manuele Bazzichetto:** Writing – review & editing, Validation, Methodology, Data curation. **Maria Carmela Caria:** Writing – review & editing. **Giovanni Rivieccio:** Writing – review & editing. **Agnese Denaro:** Writing – review & editing. **Elisa Marchetto:** Writing – review & editing. **Michela Perrone:** Writing – review & editing. **Martina Livornese:** Writing – review & editing, Data curation. **Duccio Rocchini:** Writing – review & editing, Visualization, Methodology. **Simonetta Bagella:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Code availability

The code is available at https://github.com/Redkenn/BiasEstimationGBIF.

Funding

This research was funded by Fondazione di Sardegna annualità 2022–2023, project "Verso una ricerca floristica di nuova generazione: dalla gap analysis alla valutazione della biodiversità con il supporto della citizen science", D.R. n. 751 prot. n 29253 del 03/03/2021.

Declaration of competing 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.

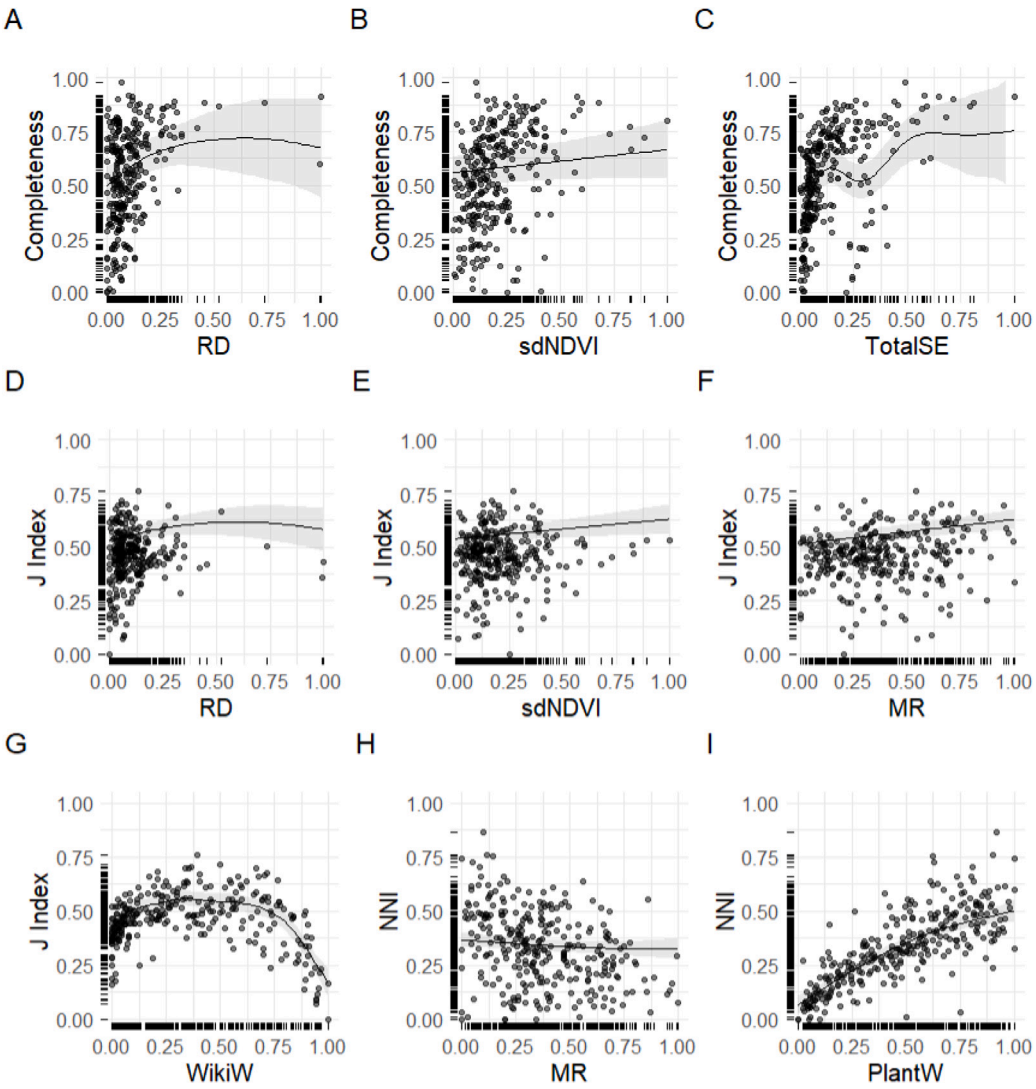


Fig. 4. Partial effects plot showing the relationship between response variables and significant predictors of GAMs models. The tick marks on the *x*-axis are observed data points. The *y*-axis represents the partial effect of each variable. The shaded areas indicate the 95% confidence intervals. (A) completeness of the species richness versus road density (RD); (B) completeness of the species richness versus standard deviation of the Normalized Difference Vegetation Index (sdNDVI); (C) completeness of the species richness versus total sampling effort (TotalSE); (D) Pielou's evennes (J Index) versus road density (RD); (E) Pielou's evennes (J Index) versus standard deviation of the Normalized Difference Vegetation Index (sdNDVI); (F) Pielou's evennes (J Index) versus mean topographic roughness (MR); (G) Pielou's evennes (J Index) versus Wikiplantbase weight (WikiW); (H) Nearest Neighbor Index (NNI) versus mean topographic roughness (MR); (I) Nearest Neighbor Index (NNI) versus PlantNet weight (PlantW).

Data availability

The data is available at https://github.com/Redkenn/BiasEstimationGBIF.

References

Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60 (2), 255–265, Available from: https://doi.org/10.1093/biomet/60.2.255.

Arceo-Gómez, G., Alonso, C., Ashman, T.-L., Parra-Tabla, V., 2018. Variation in sampling effort affects the observed richness of plant–plant interactions via heterospecific pollen transfer: implications for interpretation of pollen transfer networks. Am. J. Bot. 105 (9), 1601–1608, Available from: https://doi.org/10.1002/ajb2.1144.

Arrigoni, P.V., 1983. Aspetti corologici della flora sarda. Biogeographia– J. Integr. Biogeogr. 8 (1), Available from: https://doi.org/10.21426/b68110118.

Arrigoni, P., Nardi, E., 1977. Le piante endemiche della sardegna: 1. Aquilegia barbaricina spoecies nova. Boll. Della SocietÀ Sarda di Sci. Nat. 16, 265–268, Available from: https://doi.org/10.21426/b68110118.

Backstrom, L.J., Callaghan, C.T., Worthington, H., Fuller, R.A., Johnston, A., 2025. Estimating sampling biases in citizen science datasets. Ibis 167 (1), 73–87, Available from: https://doi.org/10.1111/ibi.13343.

Banks-Leite, C., Ewers, R.M., Pimentel, R.G., Metzger, J.P., 2012. Decisions on temporal sampling protocol influence the detection of ecological patterns. Biotropica 44 (3), 378–385, Available from: https://doi.org/10.1111/j.1744-7429.2011.00801.x.

Bátori, Z., Tölgyesi, C., Li, G., Erdős, L., Gajdács, M., Kelemen, A., 2023. Forest age and topographic position jointly shape the species richness and composition of vascular plants in karstic habitats. Ann. For. Sci. 80 (1), 1–20, Available from: https://doi.org/10.1186/s13595-023-01183-x.

Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., Barták, V., Sperandii, M.G., 2023. Sampling strategy matters to accurately estimate response curves' parameters in species distribution models. Glob. Ecol. Biogeogr. 32 (10), 1717–1729, Available from: https://doi.org/10.1111/geb.13725.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Informatics 19, 10–15, Available from: http://dx.doi.org/10.1016/j.ecoinf.2013.11.002.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., Courchamp, F., 2012. Impacts of climate change on the future of biodiversity. Ecol. Lett. 15 (4), 365–377, Available from: https://doi.org/10.1111/j.1461-0248.2011.01736.x.

Berendsohn, W.G., Chavan, V., Macklin, J., 2010. Summary of recommendations of the GBIF task group on the global strategy and action plan for the digitisation of natural history collections. Biodiversity Informatics 7 (2), Available from: http://dx.doi.org/10.17161/bi.v7i2.3989.

Bolker, B.M., 2008. Ecological Models and Data in R. Princeton University Press, URL: http://www.jstor.org/stable/j.ctvcm4g37.

Braat, L., Ten Brink, P., Bakkes, J., Bolt, K., Braeuer, I., Ten Brink, B., Chiabai, A., Ding, H., Gerdes, H., Jeuken, M., Kettunen, M., Kirchholtes, U., Klok, C., Markandya, A., Nunes, P., Van Oorschot, M., Peralta-Bezerra, N., Rayment, M., Travisi, C., Walpole, M., 2008. The cost of policy inaction: The case of not meeting the 2010 biodiversity target. In: Alterra-rapport 1718, Alterra, Wageningen, 314 pages, 85 figures, 45 tables, 140 references.

Camilleri, L., Debono, K., Grech, F., Bellia, A.F., Pace, G., Lanfranco, S., 2024. Topographic complexity is a principal driver of plant endemism in mediterranean islands. Plants 13 (4), 546, Available from: https://doi.org/10.3390/plants13040546.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A., et al., 2012. Biodiversity loss and its impact on humanity. Nature 486 (7401), 59–67, Available from: https://doi.org/10.1038/nature11148.

Chao, A., 1984. Nonparametric estimation of the number of classes in a population. Scand. J. Stat. 265–270.

Chao, A., Gotelli, N.J., Hsieh, T., Sander, E.L., Ma, K., Colwell, R.K., Ellison, A.M., 2014. Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monograph. 84 (1), 45–67, Available from: https://doi.org/10.1890/13-0133.1.

Chao, A., Kubota, Y., Zelenỳ, D., Chiu, C.-H., Li, C.-F., Kusumoto, B., Yasuhara, M., Thorn, S., Wei, C.-L., Costello, M.J., et al., 2020. Quantifying sample completeness and comparing diversities among assemblages. Ecol. Res. 35 (2), 292–314, Available from: https://doi.org/10.1111/1440-1703.12102.

Chozas, S., Nunes, A., Serrano, H.C., Ascensão, F., Tapia, S., Máguas, C., Branquinho, C., 2023. Rescuing botany: using citizen-science and mobile apps in the classroom and beyond. Npj Biodivers. 2 (1), 6, Available from: https://doi.org/10.1038/s44185-023-00011-9.

Clark, P.J., Evans, F.C., 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35 (4), 445–453, Available from: https://doi.org/10.2307/1931034.

Colli-Silva, M., Vasconcelos, T.N., Pirani, J.R., 2019. Outstanding plant endemism levels strongly support the recognition of campo rupestre provinces in mountaintops of eastern south america. J. Biogeogr. 46 (8), 1723–1733, Available from: https://doi.org/10.1111/jbi.13585.

Cosentino, F., Maiorano, L., 2021. Is geographic sampling bias representative of environmental space? Ecol. Informatics 64, 101369, Available from: https://doi.org/10.1016/j.ecoinf.2021.101369.

de Beer, I.W., Hui, C., Botella, C., Richardson, D.M., 2023. Drivers of compositional turnover of narrow-ranged versus widespread naturalised woody plants in South Africa. Front. Ecol. Evol. 11, 1106197, Available from: http://dx.doi.org/10.3389/fevo.2023.1106197.

De Groot, R.S., Wilson, M.A., Boumans, R.M., 2002. A typology for the classification, description and valuation of ecosystem functions, goods and services. Ecol. Econom. 41 (3), 393–408, Available from: https://doi.org/10.1016/S0921-8009(02)00089-7.

Díaz, S., Malhi, Y., 2022. Biodiversity: Concepts, patterns, trends, and perspectives. Annu. Rev. Environ. Resour. 47 (1), 31–63, Available from: https://doi.org/10.1146/annurev-environ-120120-054300.

El-Gabbas, A., Van Opzeeland, I., Burkhardt, E., Boebel, O., 2021. Dynamic species distribution models in the marine realm: Predicting year-round habitat suitability of baleen whales in the southern ocean. Front. Mar. Sci. 8, 802276, Available from: https://doi.org/10.3389/fmars.2021.802276.

Fiske, I.J., Bruna, E.M., Bolker, B.M., 2008. Effects of sample size on estimates of population growth rates calculated with matrix models. PloS One 3 (8), e3080, Available from: http://dx.doi.org/10.1371/journal.pone.0003080.

Forti, L.R., Szabo, J.K., 2024. Raising awareness of plant biodiversity and combating zoocentrism with citizen science: A case study of undergraduate students pursuing animal-related degrees in northeast Brazil. Hum. Ecol. 1–8, Available from: https://doi.org/10.1007/s10745-024-00539-9.

Fourcade, Y., 2016. Comparing species distributions modelled from occurrence data and from expert-based range maps. Implication for predicting range shifts with climate change. Ecol. Informatics 36, 8–14, Available from: https://doi.org/10.1016/j.ecoinf.2016.09.002.

Gamfeldt, L., Hillebrand, H., Jonsson, P.R., 2008. Multiple functions increase the importance of biodiversity for overall ecosystem functioning. Ecology 89 (5), 1223–1231, Available from: https://doi.org/10.1890/06-2091.1.

Gaul, W., Sadykova, D., White, H.J., Leon-Sanchez, L., Caplat, P., Emmerson, M.C., Yearsley, J.M., 2020. Data quantity is more important than its spatial bias for predictive species distribution modelling. PeerJ 8, e10411, Available from: https://doi.org/10.7717/peerj.10411.

Gazis, I.-Z., Greinert, J., 2021. Importance of spatial autocorrelation in machine learning modeling of polymetallic nodules, model uncertainty and transferability at local scale. Minerals 11 (11), 1172, Available from: https://doi.org/10.3390/min11111172.

Glad, A., Monnet, J.-M., Pagel, J., Reineking, B., 2019. Importance and effectiveness of correction methods for spatial sampling bias in species with sex-specific habitat preference. Ecol. Evol. 9 (23), 13188–13201, Available from: https://doi.org/10.1002/ece3.5765.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27, Available from: https://doi.org/10.1016/j.rse.2017.06.031.

Hart, A.G., Bosley, H., Hooper, C., Perry, J., Sellors-Moore, J., Moore, O., Goodenough, A.E., 2023. Assessing the accuracy of free automated plant identification applications. People Nat. 5 (3), 929–937, Available from: https://doi.org/10.1002/pan3.10460.

Hijmans, R.J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J.A., Lamigueiro, O.P., Bevan, A., Racine, E.B., Shortridge, A., et al., 2015. Package 'raster'. R Packag. 734, 473, Available from: 10.32614/CRAN.package.raster.

Hill, M.O., 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54 (2), 427–432, Available from: https://doi.org/10.2307/1934352.

Hsieh, T., Ma, K., Chao, A., 2016. iNEXT: an r package for rarefaction and extrapolation of species diversity (h ill numbers). Methods Ecol. Evol. 7 (12), 1451–1456, Available from: https://doi.org/10.1111/2041-210X.12613.

Hugo, S., Altwegg, R., 2017. The second southern african bird atlas project: causes and consequences of geographical sampling bias. Ecol. Evol. 7 (17), 6839–6849, Available from: https://doi.org/10.1002/ece3.3228.

Ivanova, N., Shashkov, M., 2021. The possibilities of GBIF data use in ecological research. Russ. J. Ecol. 52 (1), 1–8, Available from: https://doi.org/10.1134/S1067413621010069.

Kumar, P., Dobriyal, M., Kale, A., Pandey, A., Tomar, R., Thounaojam, E., 2022. Calculating forest species diversity with information-theory based indices using sentinel-2A sensor's of mahavir swami wildlife sanctuary. PLoS One 17 (5), e0268018, Available from: https://doi.org/10.1371/journal.pone.0268018.

Li, G., Fang, C., Li, Y., Wang, Z., Sun, S., He, S., Qi, W., Bao, C., Ma, H., Fan, Y., et al., 2022b. Global impacts of future urban expansion on terrestrial vertebrate diversity. Nat. Commun. 13 (1), 1628, Available from: https://doi.org/10.1038/s41467-022-29324-2.

Li, D., Shi, G., Li, J., Chen, Y., Zhang, S., Xiang, S., Jin, S., 2022a. PlantNet: A dual-function point cloud segmentation network for multiple plant species. ISPRS J. Photogramm. Remote Sens. 184, 243–263, Available from: https://doi.org/10.1016/j.isprsjprs.2022.01.007.

Mair, L., Ruete, A., 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. PloS One 11 (1), e0147796, Available from: https://doi.org/10.1371/journal.pone.0147796.

Maitner, B., Gallagher, R., Svenning, J.-C., Tietje, M., Wenk, E.H., Eiserhardt, W.L., 2023. A global assessment of the raunkiæran shortfall in plants: geographic biases in our knowledge of plant traits. New Phytol. 240 (4), 1345–1354, Available from: http://dx.doi.org/10.1111/nph.18999.

Mapfumo, R.B., Murwira, A., Masocha, M., Andriani, R., 2016. The relationship between satellite-derived indices and species diversity across african savanna ecosystems. Int. J. Appl. Earth Obs. Geoinf. 52, 306–317, Available from: https://doi.org/10.1016/j.jag.2016.06.025.

Marchetto, E., Livornese, M., Sabatini, F.M., Tordoni, E., Da Re, D., Lenoir, J., Testolin, R., Bacaro, G., Gatti, R.C., Chiarucci, A., et al., 2024. Addressing multiple facets of bias and uncertainty in continental scale biodiversity databases. Biodiversity Informatics 18, Available from: https://doi.org/10.17161/bi.v18i.21810.

McLeod, A., Leroux, S.J., Gravel, D., Chu, C., Cirtwill, A.R., Fortin, M.-J., Galiana, N., Poisot, T., Wood, S.A., 2021. Sampling and asymptotic network properties of spatial multi-trophic networks. Oikos 130 (12), 2250–2259, Available from: https://doi.org/10.1111/oik.08650.

Meyer, C., Weigelt, P., Kreft, H., 2016b. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19 (8), 992–1006, Available from: http://dx.doi.org/10.1111/ele.12624.

Miller, J., 2022. Evolution of gbif's taxonomic backbone. Biodivers. Inf. Sci. Stand. 6, e91092, Available from: http://dx.doi.org/10.3897/biss.6.91092.

Monsarrat, S., Kerley, G.I., 2018. Charismatic species of the past: biases in reporting of large mammals in historical written sources. Biol. Cons. 223, 68–75, Available from: https://doi.org/10.1016/j.biocon.2018.04.036.

Moua, Y., Roux, E., Seyler, F., Briolant, S., 2020. Correcting the effect of sampling bias in species distribution modeling–a new method in the case of a low number of presence data. Ecol. Informatics 57, 101086, Available from: https://doi.org/10.1016/j.ecoinf.2020.101086.

Nugent, J., 2018. Inaturalist: citizen science for 21st-century naturalists. Sci. Scope 41 (7), 12–15.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., Suggests, M., 2007. The vegan package. Community Ecol. Packag. 10 (631–637), 719.

Oliveira, U., Paglia, A.P., Brescovit, A.D., de Carvalho, C.J., Silva, D.P., Rezende, D.T., Leite, F.S.F., Batista, J.A.N., Barbosa, J.P.P.P., Stehmann, J.R., et al., 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of b razilian terrestrial biodiversity. Diversity and Distributions 22 (12), 1232–1244, Available from: https://doi.org/10.1111/ddi.12489.

Paleco, C., García Peter, S., Salas Seoane, N., Kaufmann, J., Argyri, P., Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., et al., 2021. Inclusiveness and diversity in citizen science. Sci. Citiz. Sci. 261, 261–281, Available from: https://doi.org/10.1007/978-3-030-58278-4.

Pereira, H.M., Navarro, L.M., Martins, I.S., 2012. Global biodiversity change: the bad, the good, and the unknown. Annu. Rev. Environ. Resour. 37, 25–50, Available from: https://doi.org/10.1146/annurev-environ-042911-093511.

Perrone, M., Di Febbraro, M., Conti, L., Divíšek, J., Chytrỳ, M., Keil, P., Carranza, M.L., Rocchini, D., Torresani, M., Moudrỳ, V., et al., 2023. The relationship between spectral and plant diversity: Disentangling the influence of metrics and habitat types at the landscape scale. Remote Sens. Environ. 293, 113591, Available from: https://doi.org/10.1016/j.rse.2023.113591.

Peruzzi, L., Bagella, S., Filigheddu, R., Pierini, B., Sini, M., Roma-Marzio, F., Caparelli, K., Bonari, G., Gestri, G., Dolci, D., et al., 2017. The wikiplantbase project: the role of amateur botanists in building up large online floristic databases. Flora Mediterr. 27, 117–129, Available from: https://doi:10.7320/FlMedit27.117.

Phaka, F.M., Vanhove, M.P., Du Preez, L.H., Hugé, J., 2022. Reviewing taxonomic bias in a megadiverse country: primary biodiversity data, cultural salience, and scientific interest of South African animals. Environ. Rev. 30 (1), 39–49, Available from: https://doi.org/10.1139/er-2020-0092.

Phillips, J., Asdal, Å., Magos Brehm, J., Rasmussen, M., Maxted, N., 2016. In situ and ex situ diversity analysis of priority crop wild relatives in Norway. Diversity and Distributions 22 (11), 1112–1126, Available from: https://doi.org/10.1111/ddi.12470.

Piccolo, R.L., Warnken, J., Chauvenet, A.L.M., Castley, J.G., 2020. Location biases in ecological research on Australian terrestrial reptiles. Sci. Rep. 10 (1), 9691, Available from: https://doi.org/10.1038/s41598-020-66719-x.

Pielou, E.C., 1966. The measurement of diversity in different types of biological collections. J. Theoret. Biol. 13, 131–144, Available from: https://doi.org/10.1016/0022-5193(66)90013-0.

Raduła, M.W., Szymura, T.H., Szymura, M., Swacha, G., 2022. Macroecological drivers of vascular plant species composition in semi-natural grasslands: A regional study from lower silesia (Poland). Sci. Total Environ. 833, 155151, Available from: https://doi.org/10.1016/j.scitotenv.2022.155151.

Robertson, T., Wieczorek, J., Raymond, M., 2022. Diversifying the GBIF data model. Biodivers. Inf. Sci. Stand. Available from: http://dx.doi.org/10.3897/biss.6.94420.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Prog. Phys. Geogr. 35 (2), 211–226, Available from: https://doi.org/10.1177/0309133311399491.

Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Barbosa, A.M., Bazzichetto, M., Beierkuhnlein, C., Castelnuovo, E., Gatti, R.C., Chiarucci, A., et al., 2023. A quixotic view of spatial bias in modelling the distribution of species and their diversity. Npj Biodivers. 2 (1), 10, Available from: https://doi.org/10.1038/s44185-023-00014-6.

Rosário, I.T., Tiago, P., Chozas, S., Capinha, C., 2025. When do citizen scientists record biodiversity? Non-random temporal patterns of recording effort and associated factors. People Nat. 7 (4), 860–870, Available from: https://doi.org/10.1002/pan3.70017.

Saoud, Z., Fontaine, C., Loïs, G., Julliard, R., Rakotoniaina, I., 2020. Miss-identification detection in citizen science platform for biodiversity monitoring using machine learning. Ecol. Informatics 60, 101176, Available from: https://doi.org/10.1016/j.ecoinf.2020.101176.

Schöley, J., 2018. Tricolore. a flexible color scale for ternary compositions. Available from: 10.32614/CRAN.package.tricolore.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Stein, A., Gerstner, K., Kreft, H., 2014. Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. Ecol. Lett. 17 (7), 866–880, Available from: https://doi.org/10.1111/ele.12277.

Steven, R., Barnes, M., Garnett, S.T., Garrard, G., O'connor, J., Oliver, J.L., Robinson, C., Tulloch, A., Fuller, R.A., 2019. Aligning citizen science with best practice: Threatened species conservation in Australia. Conserv. Sci. Pr. 1 (10), e100, Available from: https://doi.org/10.1111/csp2.100.

Tan, X., Shan, Y., Wang, X., Liu, R., Yao, Y., 2022. Comparison of the predictive ability of spectral indices for commonly used species diversity indices and hill numbers in wetlands. Ecol. Indic. 142, 109233, Available from: https://doi.org/10.1016/j.ecolind.2022.109233.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7 (1), 9132, Available from: https://doi.org/10.1038/s41598-017-09084-6.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8 (2), 127–150, Available from: https://doi.org/10.1016/0034-4257(79)90013-0.

Walther, B.A., Moore, J.L., 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), 815–829, Available from: http://dx.doi.org/10.1111/j.2005.0906-7590.04112.x.

Willis, K., 2017. State of the World's Plants 2017. Royal Botanics Gardens Kew.

Willis, K.J., Gillson, L., Knapp, S., 2007. Biodiversity hotspots through time: an introduction. Phil. Trans. R. Soc. B 362 (1478), 169–174. http://dx.doi.org/10.1098/rstb.2006.1976, Available from: https://doi.org/10.1098/rstb.2006.1976.

Wilson, J.P., 2012. Digital terrain modeling. Geomorphology 137 (1), 107–121, Available from: https://doi.org/10.1016/j.geomorph.2011.03.012.

Wood, S.N., 2017. Generalized Additive Models: an Introduction with R. chapman and hall/CRC.

Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol. Model. 157 (2–3), 157–177, Available from: https://doi.org/10.1016/S0304-3800(02)00193-X.

Wood, S., Wood, M.S., 2015. Package 'mgcv'. Available from: https://cran.r-project.org/web/packages/mgcv/index.html.

Yang, W., Ma, K., Kreft, H., 2013. Geographical sampling bias in a large distributional database and its effects on species richness–environment models. J. Biogeogr. 40 (8), 1415–1426, Available from: https://doi.org/10.1111/jbi.12108.

Zhao, X., Tian, Y., Huang, K., Zheng, B., Zhou, X., 2023. Towards efficient index construction and approximate nearest neighbor search in high-dimensional spaces. Proc. the VLDB Endow. 16 (8), 1979–1991, Available from: https://doi.org/10.14778/3594512.3594527.