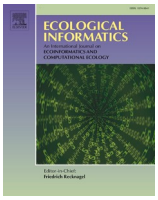
Ecological Informatics 97 (2026) 103937

Contents lists available at ScienceDirect

Ecological Informatics

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



When accessibility distorts niche estimates: A virtual species framework across species prevalence and sample size gradients

Rocìo Beatriz Cortès Lobos [a,1,*], Lorenzo Ricci [a,b,1], Matilde Martini [c], Riccardo Testolin [a], Michele Di Musciano [a,b], Vítězslav Moudrý [d], Quentin Groom [e], Duccio Rocchini [a]

[a] BIOME Lab, BiGeA Department, Alma Mater Studiorum - University of Bologna, Italy
[b] Department of Life, Health & Environmental Science, University of L'Aquila, L'Aquila, Italy
[c] Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Florence, Italy
[d] Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha-Suchdol, Czech Republic
[e] Meise Botanic Garden, Nieuwelaan 38, 1860 Meise, Belgium

ARTICLE INFO

Keywords:
Virtual species
Species distribution models
Species niche
Sampling bias
Species prevalence

ABSTRACT

Species distribution models are widely used in ecology and conservation biology to predict species' habitat suitability, but their performance is often compromised by biases in occurrence data. One of the most pervasive sources of bias arises from geographically uneven sampling, particularly the overrepresentation of areas close to roads. In this work, we outline a reproducible workflow to assess whether such geographic biases affect the representation of species' ecological niches in environmental space. We simulated 250 virtual species within the Abruzzo region (Italy) and evaluated the effects of sampling bias by systematically varying sample size and species prevalence (i.e., geographic range size). For each species, we compared random sampling from within the species' range with sampling constrained to areas near roads.

Rather than revealing a uniform effect of road-biased sampling, our results show that its consequences depend strongly on species prevalence, sample size, and the spatial correspondence between species' ranges and the accessible areas near the roads. The extent to which niche estimates are distorted therefore varies among species, implying that the effects of road-based sampling are species-dependent: sampling near roads can either misrepresent or adequately capture environmental conditions depending on the species considered. These findings highlight the need to be aware of how sampling patterns interact with species' spatial distributions when interpreting model outputs.

1. Introduction

Species distribution models (SDMs) are widely used to address ecological, biogeographical, and conservation questions by estimating species' environmental suitability from relationships between observed occurrences and environmental conditions (Guisan and Zimmermann, 2000; McCune, 2016; Franklin, 2023; Rathore and Sharma, 2023). They play a central role in predicting invasive species spread, assessing biodiversity responses to environmental change, and informing conservation planning (Bazzichetto et al., 2021; Ehrlén and Morris, 2015; Guisan et al., 2013; Haesen et al., 2023).

Despite important methodological advances, substantial limitations remain in modelling species distributions (Araújo et al., 2019; Soley-Guardia et al., 2024; Moudrý et al., 2024). A fundamental assumption underlying most SDMs is that occurrence data provide an unbiased and representative sample of species-environment relationships (Araújo and Guisan, 2006). In practice, however, observed occurrences often reflect only a subset of the environmental conditions that are suitable for a species. This mismatch may arise from ecological processes such as biotic interactions, dispersal limitation, and demographic constraints, which can prevent species from occupying all suitable environments (Pulliam, 2000; Roubicek et al., 2010; Sillero, 2011). In addition, biodiversity datasets are frequently affected by biases associated with how and where occurrence data are collected. Many occurrence records originate from opportunistic or subjective sampling rather than from formal sampling designs or standardized observation protocols,

* Corresponding author.
E-mail address: rocio.corteslobos2@unibo.it (R.B.C. Lobos).
[1] Rocio Beatriz Cortes Lobos and Lorenzo Ricci contributed equally to this study.

https://doi.org/10.1016/j.ecoinf.2026.103937
Received 30 January 2026; Received in revised form 15 July 2026; Accepted 16 July 2026
Available online 16 July 2026

resulting in uneven sampling effort, spatial imbalances, and detection errors (Lobo et al., 2007; Pärtel, 2014; Bazzichetto et al., 2021). This non-random collection of data can contribute to sampling bias and uneven environmental representation. Sampling bias determines where and how often species are recorded, leading to the overrepresentation of certain environments and the underrepresentation of others in occurrence datasets. Consequently, large biodiversity inventories may provide a distorted and incomplete representation of species–environment relationships, with direct implications for ecological inference and conservation decision-making (Hortal et al., 2008).

From a statistical perspective, biodiversity occurrence datasets often originate from non-probability sampling processes, in which inclusion probabilities are unknown and may depend on accessibility, observer behavior, or logistical constraints rather than on a formal sampling design (Boyd et al., 2023). Species distribution models generally assume that sampled occurrences adequately represent the environmental conditions occupied by a species. Under this assumption, occurrence records provide a reasonable representation of species-environment relationships, allowing the estimation of relative habitat suitability. However, when sampling is geographically uneven, sampled occurrences may disproportionately represent only a restricted subset of environmental conditions. In such cases, accessibility-driven sampling may distort the environmental space represented in occurrence datasets, potentially biasing inferred species-environment relationships and habitat suitability predictions if these mechanisms are not explicitly acknowledged (Diggle et al., 2010; Dumelle et al., 2025).

Spatial accessibility commonly contributes to non-random sampling in biodiversity data. Species occurrences are disproportionately recorded near roads, urban areas, and research centers, while remote or inaccessible regions remain systematically undersampled (Kadmon et al., 2004; Albert et al., 2010; Amano et al., 2016; Petersen et al., 2021; Rocchini et al., 2023). Additional structural factors, including disparities in research funding, technological capacity, and data sharing, can further reinforce these geographic biases, particularly in under-resourced regions (Hughes et al., 2021). Road-based sampling therefore represents a clear example of accessibility-driven sampling, in which the probability of observing a species is associated with proximity to infrastructure rather than ecological processes alone (Diggle et al., 2010).

Because environmental conditions are not uniformly distributed across space (Guisan and Thuiller, 2005), geographically uneven sampling effort can lead to environmentally unrepresentative occurrence datasets (Kadmon et al., 2004; Rocchini et al., 2023; Rocchini et al., 2026). Areas near roads may capture only a limited and environmentally unrepresentative subset of the landscape, leading SDMs to misestimate species' environmental tolerances and misrepresent habitat suitability under both current and future scenarios (Arlé et al., 2024; Meyer et al., 2016). In addition to potentially biasing species-environment relationships within the sampled environmental space, incomplete environmental coverage may also require SDMs to extrapolate beyond the range of observed covariates. Importantly, these effects can persist even when sample sizes are large, highlighting that bias cannot be resolved simply by increasing data volume alone (Boyd et al., 2021; Boyd et al., 2023).

Accordingly, a range of theoretical and methodological approaches has been developed to investigate how data limitations and species-related factors influence SDM performance (Araújo and Guisan, 2006). These include model-comparison studies that evaluate algorithmic sensitivity to sample size or prevalence, data manipulation techniques such as resampling or spatial thinning, background or pseudo-absence selection strategies designed to account for uneven sampling effort (e.g., target-group background; Phillips et al., 2009), and simulation-based approaches using virtual species (Austin and Meyers, 1996; Boria et al., 2014b; Fourcade et al., 2014; Meynard and Kaplan, 2013). While model-comparison studies provide valuable insights, they are constrained by real-world datasets and cannot fully disentangle ecological processes from sampling effects. Data filtering and resampling methods offer pragmatic tools to reduce spatial bias but often oversimplify sampling processes and may remove ecologically meaningful information, with limited improvements in model performance (Gábor et al., 2020; Dubos et al., 2022; Moudrý et al., 2024). Moreover, reducing biased occurrences can substantially decrease sample size, potentially leading to underestimation of species' range sizes (Smith et al., 2023). Virtual species simulations provide a useful framework for exploring the consequences of known ecological and sampling processes under controlled conditions, allowing explicit control over species' ecological niches, distributional properties, and sampling mechanisms and enabling attribution of model responses to predefined sources of bias (Meynard et al., 2019; Miller, 2014). However, as with any simulation-based approach, their realism depends on the assumptions imposed and may not fully capture the complexity of empirical systems (Santini et al., 2021). Consequently, recent research has increasingly emphasized approaches that evaluate the environmental representativeness of occurrence datasets and the implications of uneven sampling for ecological inference, rather than attempting to remove potentially unrepresentative observations (Varela et al., 2014; Gueta and Carmel, 2016; Rocchini et al., 2023; Moudrý et al., 2024).

In predictive modelling, sample size and species-related properties are well-established factors influencing SDM performance. Sample size affects model complexity, algorithm performance, and parameterization (Santini et al., 2021; Moudrý et al., 2024) and is shaped by both sampling effort and species detectability (Botella et al., 2020). Species prevalence, defined as the proportion of occupied locations relative to the available area, represents an intrinsic property of species distributions that is closely linked to geographic range size and interacts with effective sample size, particularly at very low or very high values (Jiménez-Valverde et al., 2009; De Marco Junior and Nobrega, 2018; Meynard et al., 2019).

Although the roles of sampling patterns, species prevalence, and environmental coverage are widely recognized, these components are often examined in isolation. Studies such as Baker et al. (2022) have used virtual species to investigate how sampling bias alters the representation of ecological niches and environmental space. However, the combined effects of geographically biased sampling, sample size, and intrinsic properties of species distributions on environmental representativeness remain relatively underexplored (Meynard et al., 2019; Rocchini et al., 2023). Indeed, a recent review by Moudrý et al. (2024) explicitly called for integrative analyses that jointly evaluate these interacting sources of uncertainty together with species' ecological characteristics, emphasizing that such studies are urgently needed to support better-informed modelling choices.

Recent advances in multidimensional niche characterization, especially hypervolume-based approaches (Blonder et al., 2018; Mammola, 2019), provide a direct way to quantify how occurrence data represent environmental space. When combined with virtual species and environmental dissimilarity metrics, these methods offer a practical and interpretable framework to apply selection-bias theory within ecological niche space, complementing existing bias diagnostics that focus primarily on spatial, taxonomic, or environmental coverage (Boyd et al., 2021; Boyd et al., 2022). Rather than proposing new estimators of bias, niche-based diagnostics allow the visualization and quantification of how biased sampling truncates or shifts the environmental conditions represented by occurrence data, independently of specific SDM algorithms (Zurell et al., 2020).

In this study, we investigate how geographically biased sampling interacts with sample size and species prevalence to shape the representation of species' ecological niches, with a particular focus on occurrences concentrated near roads. Taking the Abruzzo region in Italy as a study area, we simulated 250 virtual species responding solely to climatic variables, each characterized by a specific prevalence and sample size. For each species, we generated two occurrence datasets: one sampled across the full geographic range and one restricted to sampling

near the secondary road network. We then compared the resulting ecological niches, quantified as hypervolumes, and assessed environmental representativeness using a standardized dissimilarity index. All analyses were implemented within a reproducible computational framework integrating species simulation, niche quantification, and spatial comparison of environmental coverage.

Specifically, we tested the following hypotheses: (i) the proportion of occurrences sampled near roads varies with sample size and species prevalence, with narrow-range species being more sensitive to road-based bias; (ii) road-biased sampling leads to an underestimation of species' ecological niches relative to full-range samples; (iii) environmental conditions at road-biased occurrence points are less representative of those available across the broader landscape; and (iv) species prevalence modulates the severity of road-based sampling bias, with low-prevalence species being particularly susceptible. The contribution of this study lies not in demonstrating that geographically uneven, accessibility-driven sampling can alter the environmental representation of occurrence datasets, an already established result, but in quantifying how its ecological consequences in environmental space depend jointly on species prevalence and sampling intensity.

2. Methods

2.1. Study area

Abruzzo (Fig. 1) was used as the study area. Located in central Italy, the region encompasses a marked environmental gradient, ranging from the Apennine Mountains to the Adriatic coastal lowlands (Cutini et al., 2021, Cutini et al., 2012). This combination of mountainous terrain and densely settled coastal plains makes Abruzzo broadly representative of Mediterranean coastal systems, where sharp climatic and land-use gradients are common. Importantly, the regional road network is unevenly distributed, being dense along the coast and nearly absent in mountainous regions. This spatial configuration results in a road network that does not fully capture the region's environmental heterogeneity (Kadmon et al., 2004), making Abruzzo well suited for simulating geographically biased sampling.

All analyses were conducted in R (v4.1.1; R Core Team, 2021). Spatial data processing was performed using the packages *terra* (Hijmans, 2025) and *sf* (Pebesma, 2018). Bioclimatic predictors were obtained from WorldClim v2.1 (Fick and Hijmans, 2017) at a 30 arc-second (~1 km^2) resolution using the worldclim_country() function from the *geodata* package (Hijmans, 2025). All spatial layers used in the analyses shared the same 1 km^2 resolution.

To reduce multicollinearity among predictors, which can inflate redundancy in environmental space and complicate the interpretation of environmental representativeness, we assessed pairwise Pearson correlations and visualized them using the *corrplot* package. Following a commonly used threshold ($|r| \geq 0.7$), we removed one variable from each highly correlated pair (Dormann et al., 2013). Starting with all 19 bioclimatic variables, the final predictor set included temperature (minimum and maximum), precipitation, water vapor pressure, and wind speed, all summarized as annual means for the period 1980–2010.

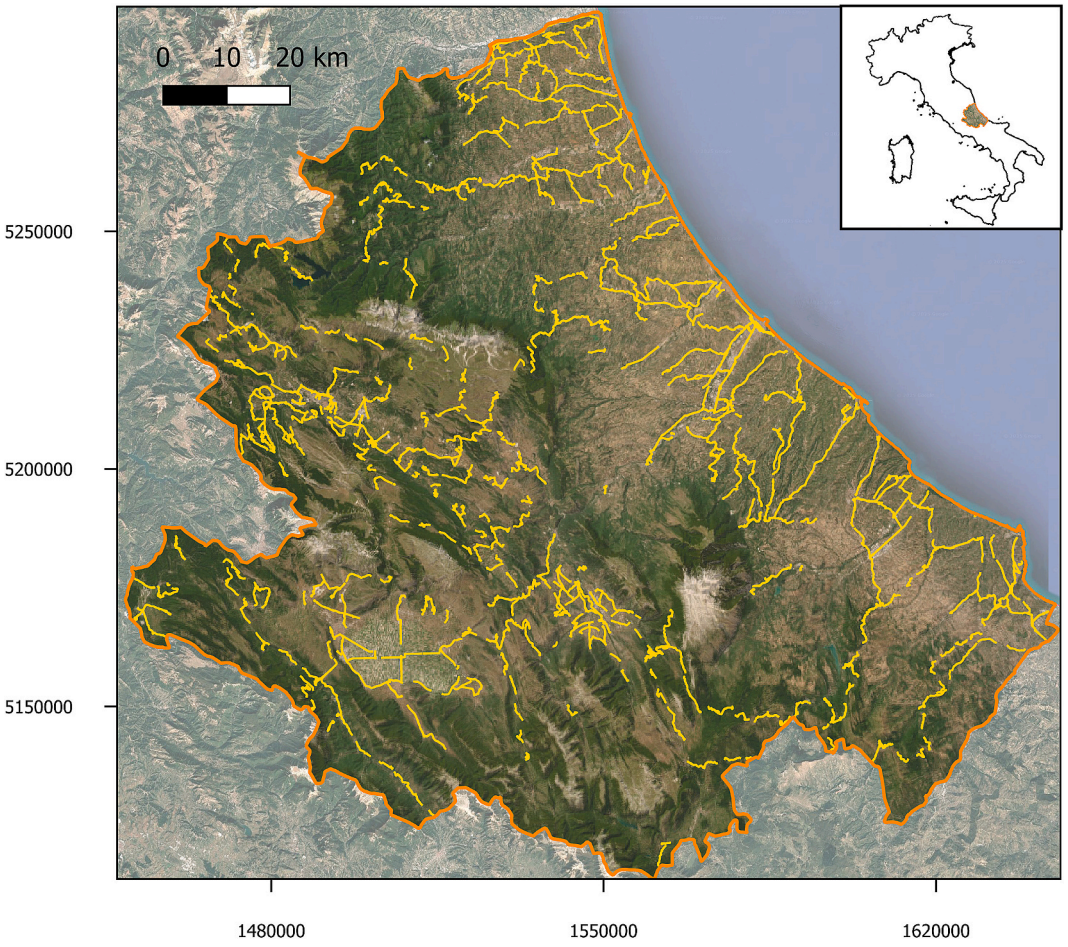


Fig. 1. Study area. The top-right inset shows the location of the study area within Italy. The main map provides a detailed view of the Abruzzo region, with yellow lines representing the secondary road infrastructure (CRS: WGS 84, source: Google Earth). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2. Virtual species generation

Virtual species were generated following the virtual ecologist approach (Zurell et al., 2010), which allows explicit control over species–environment relationships and sampling processes while avoiding confounding ecological and observational effects (Hirzel et al., 2001; Jiménez-Valverde et al., 2008; Zarzo-Arias et al., 2022). Simulations were implemented using the R package *virtualspecies* (Leroy et al., 2016), with the selected bioclimatic variables provided as a raster stack.

Species distributions were generated following a three-step procedure. For each virtual species, we first defined a species–environment relationship by assigning Gaussian response functions to each environmental predictor, using randomly selected means and standard deviations. These responses were combined multiplicatively to produce a continuous environmental suitability map, thereby representing the species' potential niche in environmental space (Leroy et al., 2016; Fig. 2).

Second, the suitability map was converted into a binary presence–absence distribution using a probabilistic transformation that links environmental suitability to occurrence probability. This step explicitly controls species prevalence, defined as the proportion of occupied pixels relative to the total study area (Meynard and Kaplan, 2013), a key intrinsic property of species distributions and is closely related to geographic range size. To simulate species with different range sizes, ten prevalence values were considered: 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, and 0.50.

Finally, occurrence points were randomly sampled within the presence map, representing a complete and unbiased sampling of the species' geographic range. Only presence records were considered, as true absences are difficult to identify in both simulated and real-world contexts (Jiménez-Valverde et al., 2008). Sample sizes ranged from 600 to 1000 occurrences in increments of 100 (i.e., 600, 700, 800, 900, and 1000). Lower sample sizes were not considered because they would yield very few road-biased occurrences, making differences attributable to accessibility-driven sampling difficult to distinguish from stochastic variability associated with small sample sizes (Moudrý et al., 2024). For each combination of prevalence and sample size, five virtual species were generated. Because species–environment relationships were randomly defined in each simulation, this resulted in a total of 250 unique virtual species.

2.3. Non-biased and road-biased sampling

For each virtual species, species prevalence defined the spatial extent of the species' range, within which a fixed number of occurrences (sample size) were selected using simple random sampling, producing a non-biased dataset without constraints imposed by infrastructure or accessibility. In this approach, all candidate cells within the simulated species range had an equal probability of inclusion. To simulate geographic sampling bias, we contrasted this idealized scenario with a road-biased sampling scheme in which occurrences were recorded near roads (Geldmann et al., 2016; Kadmon et al., 2004; Varela et al., 2014).

The secondary road network of the Abruzzo region was extracted from OpenStreetMap using the R package *osmextract* (Gilardi and Lovelace, 2025). Primary roads (highways) were excluded, as they are generally inaccessible to pedestrians and therefore unlikely to influence field sampling. For each 1 km^2 pixel in the study area, we calculated the distance to the nearest secondary road and transformed this distance into a sampling probability surface using Eq. (1):

$$P(d) = 1 - \frac{log(d)}{log(d_{max})}, d > 0 \quad (1)$$

where d is the distance to the nearest road and d_{max} is the maximum observed distance. The logarithmic transformation was used to produce a non-linear decay in sampling probability with increasing distance from roads. The resulting probability raster was generated at the same 1 km^2 resolution as all other spatial layers. Pixels located on roads were assigned a sampling probability of 1, with the probability decreasing as distance from roads increased. Using this probability surface, we derived the road-biased dataset by retaining only occurrences located in road-adjacent pixels (i.e., pixels with a sampling probability of 1), thereby simulating a strongly accessibility-driven sampling scenario. Consequently, the road-biased dataset constituted a strict subset of the original non-biased dataset, and no additional occurrence points were generated beyond those initially sampled. Because the number of road-adjacent occurrences depended on the degree of overlap between simulated species ranges and roads, the size of the biased dataset was not fixed *a priori* and varied among species.

To quantify the degree to which biased sampling approximates the original sampling within environmental space, we computed $NB_{no rm}$, defined as the ratio between the number of biased occurrences and the total number of occurrences for each species. Higher values of $NB_{no rm}$ indicate that road-biased sampling more closely approximates the original (random) sampling within the species' niche. To assess whether species prevalence (SP) influences the magnitude and variability of sampling bias, we first applied Levene's test (Glass, 1966) to evaluate the homogeneity of variances in $NB_{no rm}$ across prevalence levels. We then used linear regression to relate species prevalence to the residual variance of $NB_{no rm}$, treating residual variance as an indicator of variability in sampling bias.

2.4. Species' niche: hypervolume

Following Hutchinson (1957), a species' ecological niche can be represented as a hypervolume in an n-dimensional environmental space, where each dimension corresponds to an environmental variable. Operationally, this hypervolume can be estimated from the environmental conditions associated with species occurrences. We quantified ecological niches using the R package *hypervolume* (Blonder et al., 2018), applying Gaussian kernel density estimation (hypervolume_gaussian), which assumes that species' environmental niches are finite. Because the road-biased dataset contained fewer occurrences than the non-biased dataset, and its size varied depending on overlap with road-accessible areas, we first resampled the non-biased dataset to improve comparability between sampling schemes (Benkendorf et al., 2023). Specifically, for each species, the resampled non-biased dataset was constructed by retaining all N_β biased occurrences and adding an additional $0.2 \cdot N_\beta$ occurrences randomly drawn from the remaining non-biased observations. This additional 20% was intended to represent a scenario in which predominantly road-based sampling is supplemented by a limited number of observations collected outside highly accessible areas (Hughes et al., 2021). This resampling procedure was applied consistently across all species.

For both datasets (resampled non-biased and biased), hypervolume accumulation curves were generated to assess how estimated niche breadth changeed as occurrences accumulated. Prior to hypervolume estimation, all bioclimatic variables were standardized (Cardillo et al., 2019) according to Eq. (2) to transform predictors measured on different scales into a common, unitless space and ensure comparable contributions to environmental distance calculations:

$$S_{val} I_{i,j = \frac{val_{i,j} - \mu(val)}{\sigma(val)}} \quad (2)$$

where $val_{i,j}$represents the original values of the variables, $\mu(val)$ and $\sigma(val)$ are the average and the standard deviation of each of them.

Starting from a single randomly selected occurrence, we progressively increased the number of occurrences and recalculated the hypervolume at each step, without replacement. This procedure was repeated 10 times per species to account for stochastic variability, and the results were averaged across repetitions (Kennedy, 2019).

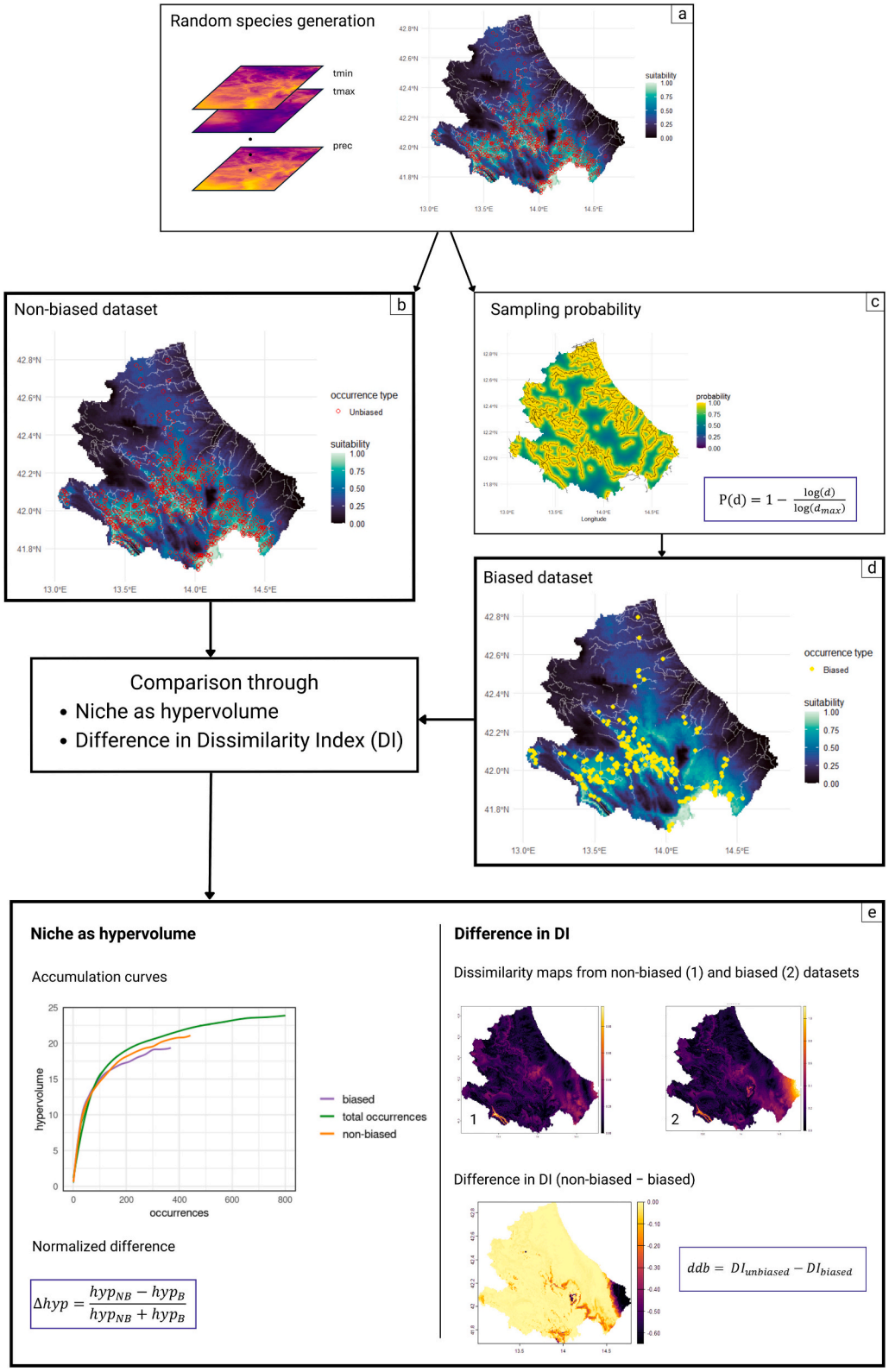


Fig. 2. Overview of the main steps in the workflow, illustrated for a single species. A virtual species is generated based on a defined species prevalence and sample size (a). The resulting set of occurrences constitutes the unbiased dataset (b). A sampling probability surface is then simulated as a function of distance from roads (c). By filtering the unbiased dataset to retain only occurrences with 100% sampling probability (i.e., located on or within 1 km of the road network), a biased subset is obtained (d). The unbiased and biased datasets are compared using two approaches (e). First, a niche-based comparison via hypervolume analysis, which includes standardizing environmental variables. Accumulation curves are computed as occurrences are added, and the normalized difference between the resulting hypervolumes (Δhyp) is calculated. Second, a spatial comparison based on the Dissimilarity Index (DI), computed for both datasets across the study area. The spatial difference between the two DI maps (ddb) is calculated, and the number of pixels where ddb = 0 (i.e., no change) is used to identify areas unaffected by sampling bias.

Accumulation curves were smoothed using LOESS (Jacoby, 2000). Curve saturation was interpreted as evidence that the sampled occurrences adequately represented the environmental space captured by the sampling scheme (Arlé et al., 2024).

For each species, we extracted hypervolume values at the point of curve convergence and calculated for all species a relative hypervolume index (3):

$$\Delta hyp = \frac{hyp_{NB} - hyp_{B}}{hyp_{NB} + hyp_{B}} \quad (3)$$

where hyp_{NB} and hyp_{B} are the hypervolume values for the resampled non-biased and biased datasets, respectively. The index ranges from −1 to 1, with positive values indicating niche underestimation under road-biased sampling. When the index is equal to 0, it indicates that both datasets yield similar hypervolume estimates, meaning road-based sampling does not strongly bias niche representation.

As an exploratory step, hypervolume estimates were summarized across groups defined by species prevalence and quartiles of NB_{norm}. To model the relationship between the hypervolume index (Δhyp), species prevalence (SP), and NB_{norm}, we first evaluated the appropriate model through a sensitivity analysis. We fitted both generalized linear models (GLMs) with a quasibinomial family (Guisan et al., 2002; Dunn et al., 2018) and generalized additive models (GAMs; Wood, 2018) to allow for potential nonlinear responses of Δhyp to species prevalence, sampling bias, and their interaction. For both modelling approaches, *the ggeffects* R package (Lüdecke, 2018) was employed. Model performance and complexity were compared using the root mean square error (RMSE), explained deviance, and Akaike's Information Criterion (AIC). Based on this comparison, the final model was selected to balance model fit, parsimony, and interpretability within the controlled simulation context. Because the hypervolume index (Δhyp) ranges from −1 to 1, whereas proportion-type response models require values bounded between 0 and 1, we applied the linear transformation (4) to rescale the index prior to modelling:

$$\Delta hyp_{rescaled} = \frac{\Delta hyp + 1}{2} \quad (4)$$

This transformation preserves the relative differences among observations while ensuring compatibility with the selected modelling frameworks. SP and NB_{norm} were used in their raw, untransformed form.

To interpret model outputs and visualize interactions between predictors, we computed model-based expected values of $\Delta hyp_{rescaled}$ using the *ggeffects* R package (Lüdecke, 2018). Predicted responses were explored under two complementary scenarios:

(i) varying species prevalence (SP) while holding NB_{norm} constant at three representative values corresponding to its empirical quartiles (0.24, 0.34, 0.43), and.

(ii) varying NB_{norm} while fixing SP at three representative levels (0.15, 0.30, 0.45). Model diagnostics and predictor relevance were assessed using standard metrics applicable to both modelling approaches. These included Variance Inflation Factors (VIFs, Fox and Monette, 1992) to evaluate multicollinearity among predictors, estimated coefficients and associated standard errors to assess effect direction and magnitude, and deviance-based tests (F-tests) to evaluate the contribution of each explanatory variable to variation in the response.

2.5. Geographic space: dissimilarity index

While hypervolume analyses quantify differences in ecological niche representation in environmental space, they do not provide spatially explicit information on where model predictions may become unreliable. To address this, we used the Dissimilarity Index (DI), a unitless metric that quantifies how dissimilar the environmental conditions at prediction locations are relative to those represented in the training data (Meyer and Pebesma, 2021). The DI accounts for both the variability and relative importance of predictor variables, providing a standardized measure of environmental distance. In the original approach, the DI was combined with a cross-validation-derived threshold to define an Area of Applicability (AOA). Here, we computed the DI but did not apply such a threshold, instead using it as a continuous measure of environmental dissimilarity.

For each species, we computed DI values by comparing the environmental conditions at occurrence locations with those across the entire study area. This procedure was applied separately to the non-biased and road-biased datasets, resulting in two DI raster maps at 1 km^2 spatial resolution. To quantify this difference, we defined the DI difference (*DDB)* as

$$DDB = DI_{non-biased} - DI_{biased} \quad (5)$$

This value was calculated for each pixel in the study area. The expectation is that *DDB* will generally be negative ($DI_{biased} > DI_{non-biased}$), indicating that the *DI* for the biased dataset is higher than that for the non-biased one. In particular, the most interesting case occurs when $DI_{biased} = DI_{non-biased}$ (i.e., $DDB = 0$), highlighting that road-biased sampling does not introduce a difference in environmental predictions. To provide an initial overview of how the Dissimilarity Index (DI) varied across different values of species prevalence and sample size, we computed summary statistics (mean $DI_{non-biased}$, mean DI_{biased} and mean number of pixels with $DDB = 0$) grouped by SP and NB_{norm}. Next, for each species, we counted the number of pixels where $DDB = 0$ and normalized it by the total number of pixels in the study area; this proportion, hereafter referred to as $(DDB = 0)_{norm}$, indicates the spatial extent to which biased and non-biased samples produce identical environmental distances.

To assess how $(DDB = 0)_{norm}$ varies with species prevalence and sampling bias, we modelled its relationship with SP, NB_{norm}, and their interaction using generalized linear models (GLMs) with a quasibinomial family. As for the hypervolume analysis, generalized additive models (GAMs) were also fitted as a sensitivity analysis to evaluate potential nonlinear effects, and model comparison was conducted using the same criteria described above.

Model-based expected values of $(DDB = 0)_{norm}$ were extracted using the *ggeffects* package. Predictions were explored under two scenarios: (i) varying species prevalence while holding NB_{norm} constant at representative quartile values (0.24, 0.34, 0.43), and (ii) varying NB_{norm} while fixing species prevalence at representative levels (0.15, 0.30, 0.45).

Model performance and predictor relevance were evaluated using coefficient estimates and their associated standard errors, deviance-based tests, and overall goodness-of-fit metrics. Predictive accuracy was assessed using pseudo-R^2 (Faraway, 2016), calculated as:

$$\text{pseudo} - R^2 = 1 - \text{model deviance/null deviance}$$

where model deviance represents the residual deviance of the fitted model and null deviance corresponds to the deviance of the intercept-only model. In addition, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were calculated using 10-fold cross-validation. Discrimination-based metrics, such as AUC or ROC, were not used because they are designed for binary classification rather than continuous response variables (Zuur et al., 2009).

3. Results

3.1. Relationship between species prevalence and the number of biased occurrences

Levene's test revealed a significant difference in the dispersion of NB_{norm} across levels of species prevalence (SP), indicating that the variability of NB_{norm} was not constant across groups (Supporting Information, Appendix 1). Specifically, NB_{norm} showed greater variability at

low *SP* values, while its dispersion stabilized as SP increased (Fig. 3; see also Supporting Information, Appendix 1). This pattern was confirmed by a linear regression model, where the negative coefficient for SP (estimate = −0.0567, p = 0.0004) indicated that variance in NB_{norm} decreased significantly with increasing species prevalence. The model exhibited a strong fit, with an adjusted R^2 of 0.79, suggesting that nearly 80% of the variation in NB_{norm} variability was explained by SP (Fig. 3).

3.2. Species' niche: hypervolume

Summary statistics showed that, in 39 out of 40 combinations of species prevalence and biased-occurrence ranges (quartiles), mean hypervolume estimates derived from biased occurrences were lower than those obtained from non-biased data. Accordingly, the normalized hypervolume difference index was positive across all groups, ranging from 0.05 to 0.34, indicating a consistent reduction in estimated niche breadth under road-biased sampling (Supporting Information, Appendix 2).

Fig. 4a shows hypervolume accumulation curves for a representative species, averaged across 10 simulations for each dataset. As occurrences were incrementally added, estimated hypervolume values increased and progressively stabilized. The three curves correspond to the total occurrence dataset (green), the resampled non-biased dataset (orange), and the biased dataset (violet). Throughout the accumulation process, the violet curve consistently lay below the green and orange curves, indicating that road-biased sampling yielded systematically lower estimates of ecological niche breadth.

To model the relationship between niche representation, species prevalence, and sampling bias, we first compared generalized linear models (GLMs) and generalized additive models (GAMs). GAMs yielded only marginal improvements in model fit relative to GLMs, with a slight reduction in RMSE (from 0.00152 to 0.00143) and a small increase in explained deviance (from 46.2% to 48.1%). Model comparison based on AIC provided no support for the more complex GAM formulation (ΔAIC ≈ 0). Smooth terms indicated moderate nonlinearity for sampling bias, whereas effects of species prevalence and their interaction were approximately linear or non-significant. Based on these results, the GLM was retained for inference.

GLM results showed a significant negative effect of species prevalence (SP) on Δ*hyp* (coefficient = −0.35). NB_{norm} also had a negative effect on Δ*hyp*, althought is effect was weaker (Supporting Information, Appendix 3). Deviance analysis confirmed that SP explained a substantial portion of the variation in Δ*hyp* (F = 12.12), while NB_{norm} had a smaller but significant effect (F = 3.98). Variance Inflation Factors indicated no collinearity between predictors (VIF = 1.03).

Model predictions showed that Δ*hyp* decreased with increasing NB_{norm} and that species with higher prevalence started from lower hypervolume differences overall (Fig. 4b). Similarly, Δ*hyp* declined with increasing species prevalence, particularly at higher levels of sampling bias (Fig. 4c).

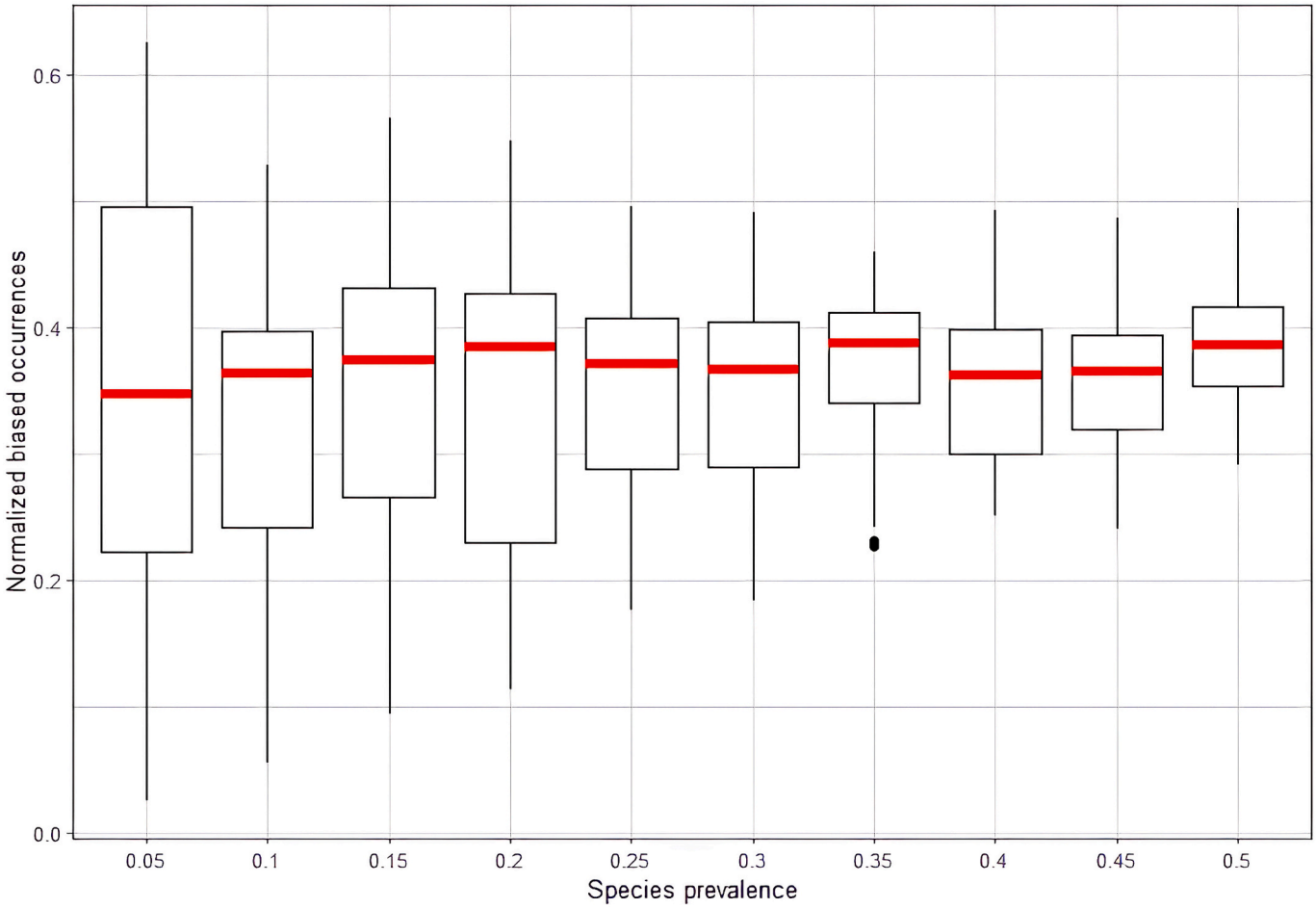
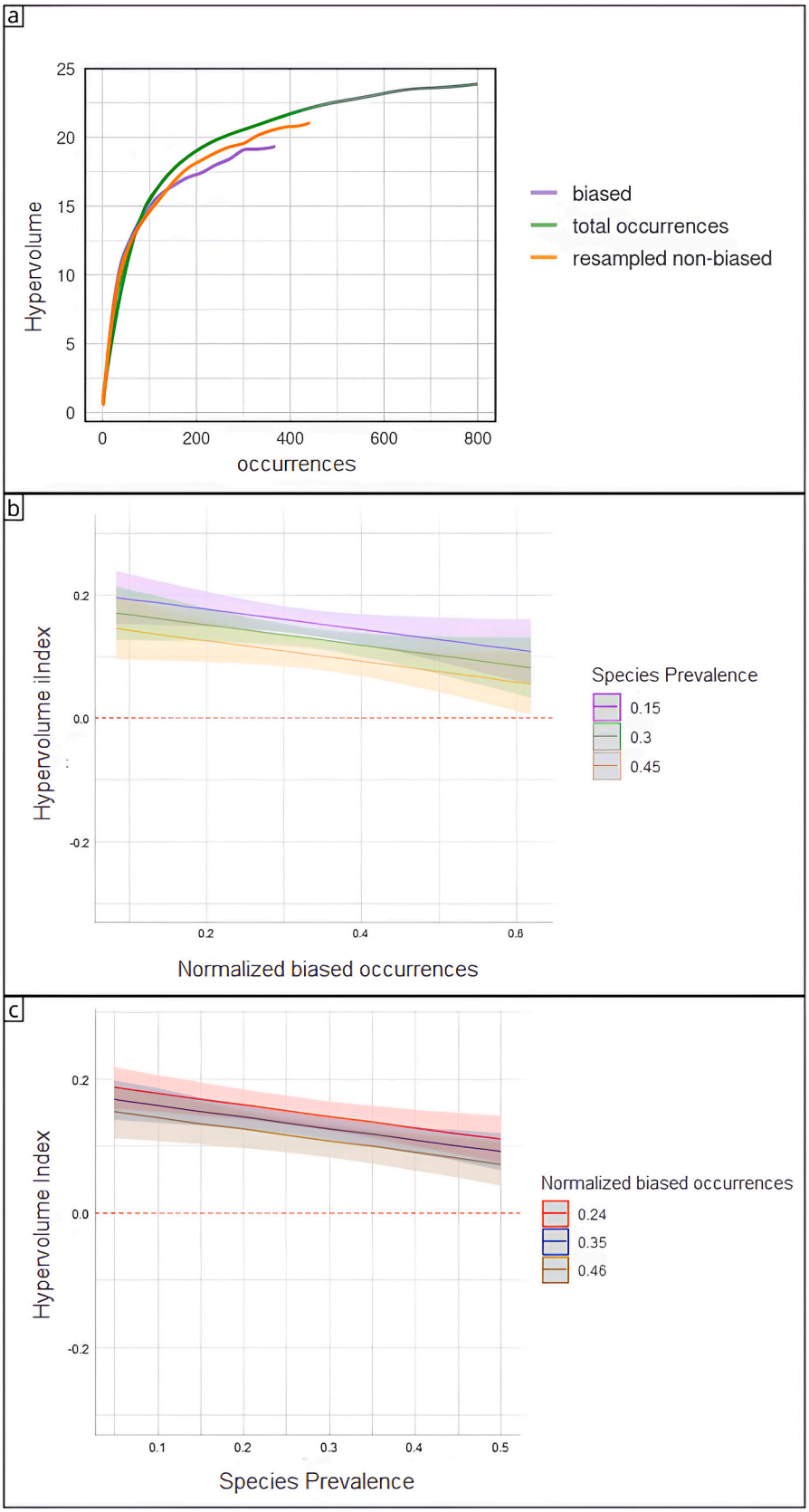


Fig. 3. Boxplot showing the relationship between species prevalence levels and the number of normalized biased occurrences to assess how dispersion in occurrence numbers varies with increasing species prevalence. For each prevalence level, the lower and upper hinges correspond to the first and third quartiles, the thick red line indicates the median, and the whiskers extend to the most extreme data points. Values beyond this range are plotted as individual outlier points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(caption on next page)

Fig. 4. a) Stepwise hypervolume accumulation curves for a single species. Each curve represents the mean hypervolume value across 10 simulations, with occurrences added incrementally. The orange curve corresponds to the resampled non-biased dataset, the violet to the biased dataset, and the green to the total occurrence dataset. **b)** Predicted values of the difference in the hypervolume between non-biased and biased dataset (*Δhyp*) as a function of the normalized proportion of road-biased occurrences (NB_{norm}) for three fixed values of species prevalence (SP), represented by different colors. The red dashed line at zero indicates the point where the non-biased and biased hypervolumes are equal. Shaded areas are confidence intervals. **c)** Predicted values of Δhyp as a function of SP, for three fixed values of NB_{norm}, represented by different colors. The red dashed line at zero indicates the point where the non-biased and biased hypervolumes are equal. Shaded areas represent confidence intervals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Dissimilarity index

Taking one species as an example, the *DDB* values were zero in areas covered by both sampling methods (or none of them) but tended to be negative where only the non-biased sampling was present (Fig. 5a).

To assess general patterns across species, we aggregated DI and DDB values by species prevalence and normalized biased occurrences. Mean DI was consistently higher under biased sampling, resulting in negative mean DDB values across all groups. The proportion of pixels with DDB $<$ 0, indicating higher dissimilarity in the biased data, was dominant, typically ranging from 74% to 97%. In contrast, DDB $>$ 0 rarely exceeded 26%, and DDB = 0 remained consistently below 1% (see Supporting Information, Appendix 4).

The GLM results reported that, as NB_{norm} increased, the number of pixels where DDB = 0 tended to rise (coefficient: 3.10), whereas an increase in SP led to a decrease in these pixels (coefficient: −1.96). The combined effect of the two independent variables was strong and positive: a simultaneous increase of NB_{norm} and *SP* visibly raised the number of pixels where DDB = 0 (coefficient: 8.30) (Supporting Information, Appendix 5). The GLM effectively explained a substantial portion of the data variability (pseudo R^2 = 0.461) and provided predictions that closely aligned with observed values (RMSE = 0.00153 and MAE = 0.00117 from 10-fold cross-validation).

Predicted values from the GLM, visualized using the ggeffects package, revealed the following patterns. Fig. 5b shows that the increase in the number of pixels where *DDB* = 0 became more pronounced as NB_{norm} grew, with SP fixed at three values. The highest established SP value reached 2% of the total pixels in the area. In Fig. 5c, where NB_{norm} is held constant at fixed values of NB_{norm}, the number of neutral pixels increased sharply at the highest NB_{norm} threshold (0.46). At the intermediate threshold (0.35), the increase was more gradual, while at the lowest threshold (0.24), the number of these pixels decreased despite the rise in SP. Additionally, when NB_{norm} was low, an increase in SP led to a further reduction in the predicted pixels where *DDB* = 0 (Fig. 5c).

4. Discussion

4.1. Species with low prevalence show greater variability in the number of road-biased occurrences

The relationship between species prevalence and the variability in the number of biased occurrences may reflect an inherent property of species' spatial distributions, specifically how limited range increases their sensitivity to sampling biases. When species prevalence is low, the species occupies a relatively small area, and therefore, the probability that sampled occurrences fall on roads (or other sampling sites) is highly variable (Guisan et al., 2006).

From a sampling-theory perspective, this represents a consistent non-random sampling mechanism in which inclusion probability depends on spatial accessibility and is correlated with the underlying environmental suitability of the species (Diggle et al., 2010; Dumelle et al., 2025). When suitable environments for a species are located near roads, biased sampling yields relatively dense occurrence data; when they are distant from roads, sampling may be extremely sparse. In such cases, models built from few occurrences are inherently unstable and highly sensitive to the sampling configuration (Inman et al., 2021; Jiménez-Valverde et al., 2008).

As species prevalence increases, distributions extend across larger areas, increasing the likelihood of overlap with the road network, regardless of where suitable environments are located. This leads to more consistent numbers of biased occurrences across species, reducing variability without necessarily reducing the bias itself. Importantly, this stabilization reflects increased consistency in sampling outcomes rather than a correction of the underlying accessibility-driven sampling process. Widespread species are therefore more likely to be at least partially sampled under geographically biased sampling strategies simply because they occupy more space, increasing the likelihood of intersecting accessible areas.

4.2. Species with narrow ranges show greater sensitivity to sampling bias in hypervolume-based niche estimation

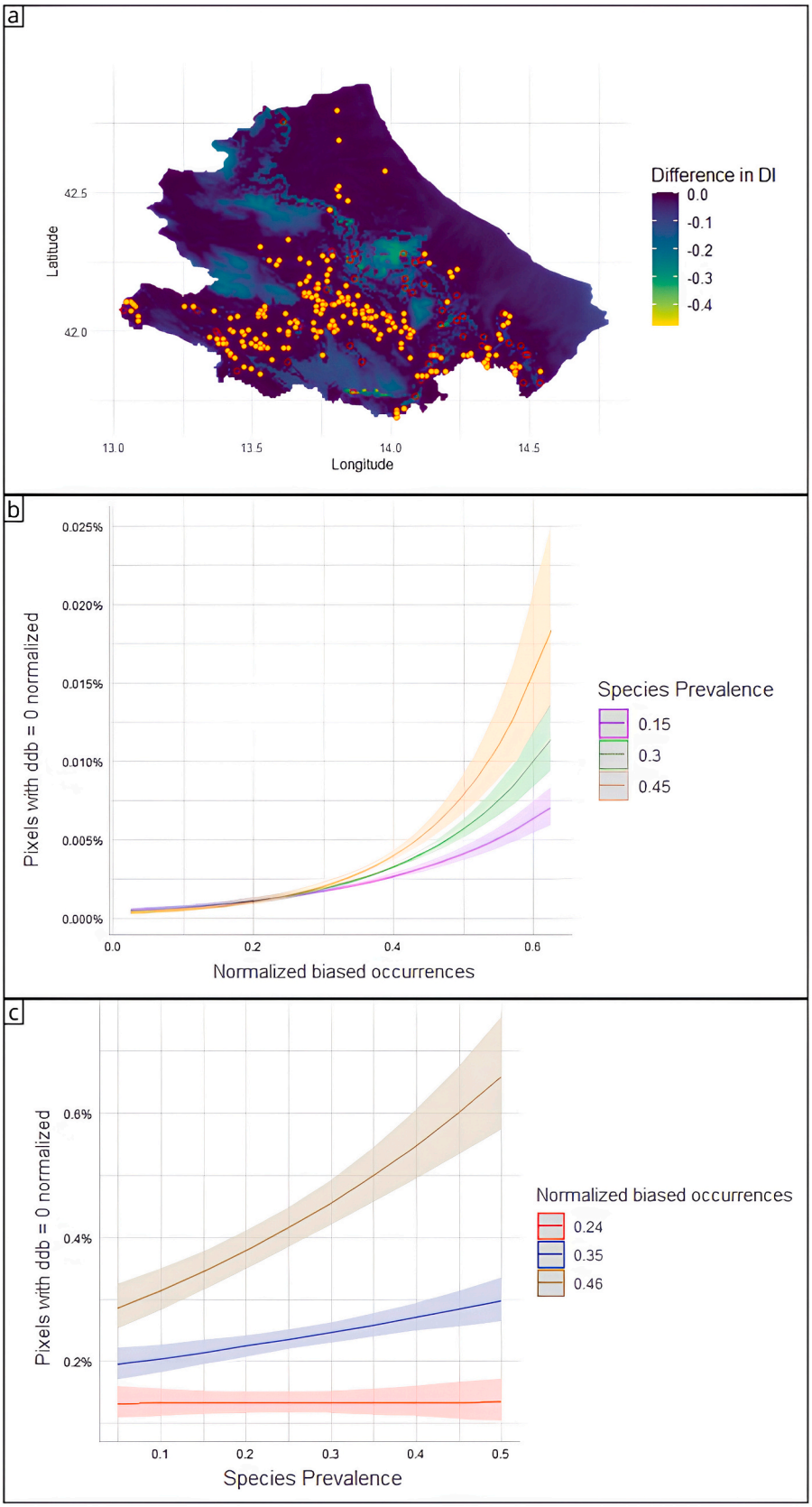
Comparing hypervolumes estimated from biased and unbiased samples provides insights into how sampling strategies influence niche characterization. The difference between these hypervolumes (*Δhyp*) acts as an indicator of the extent to which environmental representation differs between sampling schemes. In this sense, Δhyp can be interpreted as a niche-based diagnostic of environmental truncation associated with accessibility-driven sampling, highlighting how geographically uneven occurrence data may constrain the environmental conditions represented in sampled niches.

At low species prevalence, *Δhyp* values were consistently higher, indicating greater sensitivity of niche estimates to sampling bias (Vilas et al., 2022). This suggests greater uncertainty in niche estimation for species with limited distributions, whose occurrence patterns are more sensitive to the spatial configuration of the sampling network (Moudrý et al., 2024). Such patterns are consistent with accessibility-driven, non-random sampling processes, in which inclusion probabilities depend on accessibility rather than ecological suitability, potentially leading to systematic underrepresentation of portions of the environmental space (Boyd et al., 2023; Diggle et al., 2010). Under such conditions, whether suitable environments happen to overlap with roads strongly determines which environmental conditions are sampled, resulting in greater variability in estimated hypervolumes.

As species prevalence increases, *Δhyp* values tend to decline and converge toward zero, indicating a growing similarity between biased and unbiased niche estimates (Moudrý et al., 2024). Importantly, this convergence should not be interpreted as a reduction in sampling bias. Instead, it reflects increased stability in biased sampling outcomes: widespread species occupy larger areas, increasing the likelihood that geographically biased sampling consistently captures a similar subset of environmental conditions across replicates. This greater consistency reduces fluctuations in hypervolume estimates, even though the underlying sampling process remains biased.

Consequently, the magnitude of niche distortion depends not only on the sampling strategy itself but also on the spatial extent and position of a species' range relative to that strategy (Roleček et al., 2007). For narrowly distributed species, road-based sampling may either yield a reasonably representative niche if suitable environments are located near roads or fail entirely if they are far from accessible areas. In contrast, broader-ranging species are more likely to intersect the road network, resulting in more repeatable, although still biased, environmental representations.

By explicitly measuring how biased sampling limits the range of environmental conditions that are sampled, our niche-based framework helps distinguish between reduced variability in model outcomes and



(caption on next page)

Fig. 5. a) Map showing the distribution of occurrences across the study area and the difference in the Dissimilarity Index between the non-biased and biased sampling. Yellow points are the occurrences near roads, which are included within the non-biased sampling (red circles). b) Predicted values of the difference in the Dissimilarity Index between non-biased and biased datasets (DDB) = 0 obtained from a Generalized Linear Model (GLM) when increasing the species prevalence (SP) and 3 values of the normalized proportion of road-biased occurrences (NB_{norm}) fixed. Each line has its confidence interval (95%). **c)** Predicted values of DDB = 0 as we increase NB_{norm} and keep 3 fixed values of SP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

genuine improvements in environmental representativeness. This perspective complements previous studies showing that approaches based on environmental space are more effective than purely geographic strategies for capturing species' climatic niches (Perret and Sax, 2022). It also emphasizes the need to interpret sampling bias by jointly considering species distributions and sampling design.

4.3. High-prevalence species show greater environmental dissimilarity under low-intensity biased sampling

Because species prevalence influences the spatial extent of simulated ranges and their potential overlap with roads, species prevalence and NB_{norm} are not fully independent within the simulation framework. Consequently, these relationships should be interpreted primarily as associations among interacting simulation components rather than isolated causal effects. Observing the trend for $(DDB = 0)_{norm}$ as NB_{norm} increased while keeping SP fixed at three thresholds, we found that $(DDB = 0)_{norm}$ grows gradually at low SP levels but increases more rapidly at high *SP* values. This supports the hypothesis that increasing the sample size leads to a growing number of locations where biased and unbiased sampling produce no difference in Dissimilarity Index values. Our findings align with previous studies emphasizing sample size as the most critical factor in ensuring reliable predictions (Moudrý et al., 2024). The interaction term between NB_{norm} and species prevalence is significant, indicating that the effect of NB_{norm} on $(DDB = 0)_{norm}$ is not constant, but instead depends on species prevalence. When species prevalence is higher, the relationship between NB_{norm} and $(DDB = 0)_{norm}$ strengthens, reflecting a compounded effect of broader species distribution and increased sampling coverage along roads. However, when we examine the effect of species prevalence alone while keeping NB_{norm} fixed at different values, a different pattern emerges. The negative coefficient associated with species prevalence indicates that, for a given number of biased occurrences, an increase in species prevalence tends to reduce $(DDB = 0)_{norm}$. In particular, when only a small number of occurrences are collected along roads, more widespread species, which occupy a larger area, will exhibit a decrease in the number of environmentally "indifferent" pixels. This is because higher prevalence implies that the species occurs across a broader environmental and spatial gradient, including many regions not covered by biased sampling. In such cases, modelling a generalist species with sparse data leads to considerable uncertainty, as emphasized by Jiménez-Valverde et al. (2008). Similarly, Botella et al. (2020) and Baker et al. (2022) showed that even broadly distributed species can be mischaracterized if occurrence data are geographically biased, often resulting in overestimated specialization or distorted predictions of niche breadth. Overall, our results support previous evidence that geographically biased sampling can distort the representation of environmental space, particularly for widespread species. In this context, the DI represents a useful complementary metric for identifying where sampling bias is most likely to compromise environmental representativeness in geographic space.

4.4. Study limitation and perspective

The present study focuses on environmental representativeness under varying degrees of road-sampling bias, sample size, and species prevalence; however, these effects are also expected to propagate to SDM predictions. In particular, road-based sampling introduces a systematic over-representation of accessible environmental conditions, which can bias the estimation of species–environment relationships and shift inferred ecological niches. This, in turn, may influence predicted probabilities of presence by favoring environments associated with sampling accessibility rather than the full range of suitable conditions. Consequently, SDM outputs derived from such data may exhibit spatial artifacts, such as inflated suitability in well-sampled (road-proximate) areas and reduced predictive performance in environmentally suitable but under-sampled regions. Recent studies have demonstrated that such sampling-induced distortions can substantially affect SDM reliability, transferability, and the estimation of species–environment relationships, and that appropriate bias correction can mitigate these effects (Pili et al., 2025). In addition, spatially explicit evaluation metrics have been shown to be more sensitive to environmentally biased predictions than conventional non-spatial metrics, providing improved diagnostic capability for detecting sampling artifacts (Bracho-Estévanez et al., 2024). Correcting sampling bias directly in environmental space has been shown to improve model performance across multiple objectives, including prediction accuracy and transferability (Pili et al., 2025). However, it is important to note that environmental filtering is not always beneficial and should not be applied indiscriminately without evidence of bias in species occurrence data (Gábor et al., 2020, Gábor et al., 2023). This is particularly relevant because geographic and environmental sampling domains are tightly coupled, and correcting bias in one domain may propagate unintended distortions into the other.

In this context, our framework is intended to complement these existing bias-correction and evaluation approaches. Instead of focusing on model calibration or bias-correction procedures (Balej et al., 2025; Dubos et al., 2022), we explicitly examine how geographically biased occurrence data modify the environmental space represented by observations under different combinations of species prevalence and biased sampling intensity. Whereas Meyer and Pebesma (2021) thresholded the DI to define an Area of Applicability, we retain the DI as a continuous measure of environmental dissimilarity, allowing it to serve as a diagnostic indicator of environmental truncation before model fitting and interpretation. By combining hypervolume-based niche estimation with spatially explicit dissimilarity metrics (Bracho-Estévanez et al., 2024), we provide diagnostic tools that can help identify environmentally truncated occurrence datasets before model fitting and interpretation.

Our results show that species prevalence strongly mediates how geographically biased sampling propagates to environmental niche estimates. Rare or spatially restricted species were particularly vulnerable to environmental mischaracterization, whereas widespread species often exhibited more stable, but not necessarily more representative, environmental coverage. These findings suggest that evaluating environmental representativeness before modelling may help identify situations in which geographically biased occurrence data are likely to distort niche characterization.

Spatially explicit dissimilarity metrics such as the DI used here can provide a useful first assessment of the environmental distance between occurrence records and the broader predictor space, helping identify regions where sampling bias may compromise representativeness. When occurrence datasets originate from different sampling schemes, data sources, or temporal snapshots, complementary niche-overlap approaches based on environmental hypervolumes may further help quantify differences in environmental coverage (e.g., Zurell et al., 2024). When available, independent range information, such as International Union for Conservation of Nature (IUCN) distribution maps

(Betts et al., 2020), can also support survey design and help assess whether occurrence records adequately capture the environmental conditions occupied by a species. Although our simulations focused specifically on road-associated sampling bias, similar mechanisms are likely to operate across broader spatial scales (Bystriakova et al., 2012), where environmental representativeness is shaped by multiple interacting factors, including uneven sampling effort, accessibility, monitoring policies, and differences among biodiversity databases. Recent large-scale assessments have shown that biodiversity data sources and conservation networks may differ substantially in their taxonomic, geographic, and environmental representativeness. For example, Di Musciano et al. (2026) found strong variation in how effectively the Natura 2000 network captures regional plant diversity across European countries and biogeographic regions, whereas Santi et al. (2026) showed that major biodiversity repositories such as GBIF, EVA, and GIFT, can systematically differ in species coverage and ecological representation across Mediterranean islands. Here, combining spatial dissimilarity metrics with hypervolume-based niche comparisons may provide a useful framework for evaluating how different data sources represent environmental space for the same taxa and for identifying where incomplete or geographically biased sampling may propagate to large-scale ecological inference.

For poorly known species, the proposed framework may provide an initial exploratory assessment of environmental representativeness before model calibration. Given a set of occurrence records, spatial dissimilarity metrics such as the DI can help identify environmentally similar areas that remain unsampled or underrepresented. In such cases, ecologists may then investigate whether these gaps reflect true ecological constraints, local disturbances, or simply uneven sampling effort and accessibility.

Finally, an additional aspect not explicitly considered in this study is spatial autocorrelation, whereby geographically proximate observations and environmental conditions tend to be more similar than distant ones (Cressie, 2015). Because our simulations were designed to isolate the effects of species prevalence, sample size, and geographic sampling bias, spatial dependence was not incorporated into the modelling workflow. However, in real-world applications, spatial autocorrelation in predictors and occurrence data may further influence environmental representativeness (Naimi et al., 2011) and could be incorporated into future extensions of the framework to refine hypervolume estimation and DI-based assessments.

5. Conclusions

This study demonstrates that geographically biased sampling and roadside-biased data collection in particular can substantially distort inferred ecological niches. When occurrence records are concentrated near roads, estimated Hutchinsonian hypervolumes are systematically constrained, with the strongest effects observed for low-prevalence species. In contrast, unbiased sampling captures a broader and more representative spectrum of environmental conditions. We further show that increasing sample size can partially stabilize model outputs but cannot fully compensate for bias-induced distortions. These effects may become particularly consequential when models are transferred across space and time, as biased sampling markedly limits their extrapolation to different geographic regions and past or future environmental scenarios. Overall, our results reinforce that accessibility-driven bias does not affect all species equally. Even where road networks are dense and encompass most environmental conditions, the spatial distribution and intensity of sampling can be concentrated in particular areas (e.g., near cities or popular sites). Therefore, road-network extent alone is not a reliable proxy for sampling bias. No universally robust sampling strategy exists, and reliable niche estimation requires a species-specific evaluation of how accessibility, prevalence, and sample size jointly shape the representation of ecological reality, even in an era of abundant biodiversity data.

CRediT authorship contribution statement

Rocìo Beatriz Cortès Lobos: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lorenzo Ricci:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration. **Matilde Martini:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Riccardo Testolin:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Michele Di Musciano:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Formal analysis, Conceptualization. **Vítězslav Moudrý:** Writing – review & editing, Visualization, Validation, Methodology. **Quentin Groom:** Writing – review & editing, Visualization, Methodology. **Duccio Rocchini:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition.

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.

Acknowledgements

RBCL and DR were supported by the B3 project (Biodiversity Building Blocks for policy), funded by the European Union's Horizon Europe Research and Innovation Programme (Grant Agreement No. 101059592).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.103937.

Data availability

All data, metadata, code, and figure-generation scripts used in this study are openly available on GitHub: https://github.com/rociobeatrizc/When-accessibility-distorts-niche-estimates

References

Albert, C.H., Yoccoz, N.G., Edwards Jr., T.C., Graham, C.H., Zimmermann, N.E., Thuiller, W., 2010. Sampling in ecology and evolution: bridging the gap between theory and practice. Ecography 33, 1028–1037.
Amano, T., González-Varo, J.P., Sutherland, W.J., 2016. Languages are still a major barrier to global science. PLoS Biol. 14 (12), e2000933.
Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.
Araújo, M.B., Anderson, R.P., Márcia Barbosa, A., Beale, C.M., Dormann, C.F., Early, R., Rahbek, C., 2019. Standards for distribution models in biodiversity assessments. Sci. Adv. 5 (1), eaat4858.
Arlé, E., Knight, T.M., Jiménez-Muñoz, M., Biancolini, D., Belmaker, J., Meyer, C., 2024. The cumulative niche approach: a framework to assess the performance of ecological niche model projections. Ecol. Evol. 14, e11060.
Austin, Mike P., Meyers, Jacqui A., 1996. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. For. Ecol. Manag. 85 (1–3), 95–106.
Baker, D.J., Maclean, I.M.D., Goodall, M., Gaston, K.J., 2022. Correlations between spatial sampling biases and environmental niches affect species distribution models. Glob. Ecol. Biogeogr. 31, 1038–1050.
Balej, P., Moudrý, V., Prajzlerová, D., Gábor, L., Sillero, N., Rocchini, D., Šímová, P., 2025. Species-observer link and kernel density estimation of background points allow for sampling bias correction in bird species distribution models. Ecography 2025 (12), e08202.
Bazzichetto, M., Massol, F., Carboni, M., Lenoir, J., Lembrechts, J.J., Joly, R., et al., 2021. Once upon a time in the far south: influence of local drivers and functional traits on plant invasion in the harsh sub-Antarctic islands. J. Veg. Sci. 32, e13057.
Donald J. Benkendorf, Samuel D. Schwartz, D. Richard Cutler, Charles P. Hawkins, Correcting for the effects of class imbalance improves the performance of machine-

learning based species distribution models, Ecol. Model., 483, 2023, 110414, ISSN 0304-3800, doi:https://doi.org/10.1016/j.ecolmodel.2023.110414.
Betts, J., Young, R.P., Hilton-Taylor, C., Hoffmann, M., Rodríguez, J.P., Stuart, S.N., Milner-Gulland, E.J., 2020. A framework for evaluating the impact of the IUCN red list of threatened species. Conserv. Biol. 34 (3), 632–643.
Blonder, B., Morrow, C.B., Maitner, B., et al., 2018. New approaches for delineating n-dimensional hypervolumes. Methods Ecol. Evol. 9, 305–319.
Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77.
Botella, C., Joly, A., Monestiez, P., Bonnet, P., Munoz, F., 2020. Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection. PLoS ONE 15, e0232078.
Boyd, R.J., Powney, G.D., Carvell, C., Pescott, O.L., 2021. occAssess: an R package for assessing potential biases in species occurrence data. Ecol. Evol. 11 (22), 16177–16187.
Boyd, R.J., Powney, G.D., Burns, F., Danet, A., Duchenne, F., Grainger, M.J., Jarvis, S.G., Martin, G., Nilsen, E.B., Porcher, E., Stewart, G.B., 2022. ROBITT: a tool for assessing the risk-of-bias in studies of temporal trends in ecology. Methods Ecol. Evol. 13 (7), 1497–1507.
Boyd, R.J., Powney, G.D., Pescott, O.L., 2023. We need to talk about nonprobability samples. Trends Ecol. Evol. 38 (6), 521–531.
Bracho-Estévanez, C.A., Arenas-Castro, S., González-Varo, J.P., González-Moreno, P., 2024. Spatially explicit metrics improve the evaluation of species distribution models facing sampling biases. Eco. Inform. 84, 102916.
Bystriakova, N., Peregrym, M., Erkens, R.H., Bezsmertna, O., Schneider, H., 2012. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. Syst. Biodivers. 10 (3), 305–315.
Cardillo, M., Dinnage, R., McAlister, W., 2019. The relationship between environmental niche breadth and geographic range size across plant species. J. Biogeogr. 46, 97–109.
Cressie, N., 2015. Statistics for Spatial Data. John Wiley & Sons.
Cutini, M., Iocchi, M., Theurillat, J.P., Petriccione, B., 2012. 1.2. Appennino Centrale: Velino-Duchessa.
Cutini, M., Flavio, M., Giuliana, B., Guido, R., Jean-Paul, T., 2021. Bioclimatic pattern in a Mediterranean mountain area: assessment from a classification approach on a regional scale. Int. J. Biometeorol. 65, 1085–1097.
De Marco Junior, P., Nobrega, C.C., 2018. Evaluating collinearity effects on species distribution models: an approach based on virtual species simulation. PLoS ONE 13 (9), e0202403.
Di Musciano, M., Zannini, P., Testolin, R., Sabatini, F.M., Santovito, D., Jiménez-Alfaro, B., et al., 2026. Representativeness of the Natura 2000 network for preserving plant biodiversity in the European Union. Conserv. Biol. 40 (2), e70158.
Diggle, P.J., Menezes, R., Su, T.L., 2010. Geostatistical inference under preferential sampling. J. R. Stat. Soc.: Ser. C: Appl. Stat. 59 (2), 191–232.
Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.
Dubos, N., Préau, C., Lenormand, M., Papuga, G., Monsarrat, S., Denelle, P., Luque, S., 2022. Assessing the effect of sample bias correction in species distribution models. Ecol. Indic. 145, 109487.
Dumelle, M., Trangucci, R., Nahlik, A.M., Olsen, A.R., Irvine, K.M., Blocksom, K., Ver Hoef, J.M., Fuentes, C., 2025. Missing data in ecology: syntheses, clarifications, and considerations. Ecol. Monogr. 95 (4), e70037.
Dunn, P.K., Smyth, G.K., Dunn, P.K., Smyth, G.K., 2018. Chapter 9: models for proportions: binomial GLMs. In: Generalized Linear Models With Examples in R, pp. 333–369.
Ehrlén, J., Morris, W.F., 2015. Predicting changes in the distribution and abundance of species under environmental change. Ecol. Lett. 18 (3), 303–314.
Faraway, J.J., 2016. Extending the Linear Model With R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman and Hall/CRC.
Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315.
Fourcade, Yoan, et al., 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. PLoS ONE 9 (5), e97122.
Fox, J., Monette, G., 1992. Generalized collinearity diagnostics. J. Am. Stat. Assoc. 87 (417), 178–183.
Franklin, Janet, 2023. Species distribution modelling supports the study of past, present and future biogeographies. J. Biogeogr. 50 (9), 1533–1545.
Gábor, Lukáš, et al., 2020. How do species and data characteristics affect species distribution models and when to use environmental filtering? Int. J. Geogr. Inf. Sci. 34 (8), 1567–1584.
Gábor, L., Jetz, W., Zarzo-Arias, A., Winner, K., Yanco, S., Pinkert, S., Marsh, C.J., Rogan, M.S., Mäkinen, J., Rocchini, D., Barták, V., Malavasi, M., Balej, P., Moudrý, V., 2023. Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable. Ecography 2023, e06358.
Geldmann, J., Heilmann-Clausen, J., Holm, T.E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., Tøttrup, A.P., 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Divers. Distrib. 22, 1139–1149.
Gilardi, A., Lovelace, R., 2025. Osmextract: Download and Import Open Street Map Data Extracts R Package Version 0.5.3.9000. https://docs.ropensci.org/osmextract/.
Glass, G.V., 1966. Testing homogeneity of variances. Am. Educ. Res. J. 3 (3), 187–190. https://doi.org/10.3102/00028312003003187.
Gueta, Tomer, Carmel, Yohay, 2016. Quantifying the value of user-level data cleaning for big data: a case study using mammal distribution models. Eco. Inform. 34, 139–145. ISSN 1574–9541.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009.
Guisan, A., Zimmermann, N. E., Predictive habitat distribution models in ecology, Ecol. Model., Volume 135, Issues 2–3, 2000, Pages 147–186, ISSN 0304–3800.
Guisan, A., Edwards, T. C., Hastie, T., Generalized linear and generalized additive models in studies of species distributions: setting the scene, Ecol. Model., Volume 157, Issues 2–3, 2002, Pages 89–100, ISSN 0304–3800.
Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A., Zimmermann, N.E., 2006. Using niche-based models to improve the sampling of rare species. Conserv. Biol. 20, 501–511.
Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16 (12), 1424–1435.
Haesen, S., Lenoir, J., Gril, E., De Frenne, P., Lembrechts, J.J., Kopecký, M., Van Meerbeek, K., 2023. Microclimate reveals the true thermal niche of forest plant species. Ecol. Lett. 26 (12), 2043–2055.
Hijmans, R., 2025. Raster: Geographic Data Analysis and Modeling. R Package Version 3.6–32. https://rspatial.org/raster.
A.H. Hirzel, V. Helfer, F. Metral, Assessing habitat-suitability models with a virtual species, Ecol. Model., Volume 145, Issues 2–3, 2001, Pages 111–121, ISSN 0304–3800.
Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117, 847–858.
Hughes, A.C., Orr, M.C., Ma, K., Costello, M.J., Waller, J., Provoost, P., Qiao, H., 2021. Sampling biases shape our view of the natural world. Ecography 44 (9), 1259–1269.
Hutchinson, G.E., 1957. Concluding remarks. Cold Spring Harb. Symp. Quant. Biol. 22, 415–427.
Inman, Richard D., Franklin, Janet, Esque, Todd C., Nussear, Kenneth E., 2021. Comparing sample bias correction methods for species distribution modeling using virtual species. Ecosphere 12, e03422.
Jacoby, William G., 2000. Loess:: a nonparametric, graphical tool for depicting relationships between variables. Elect. Stud. 19 (4), 577–613. ISSN 0261–3794.
Jiménez-Valverde, A., Lobo, J.M., Hortal, J., 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Divers. Distrib. 14, 885–890.
Jiménez-Valverde, A., Lobo, J., Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Commun. Ecol. 10 (2), 196–205.
Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside BIAS on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14, 401–413.
Kennedy, M.C., 2019. Experimental design principles to choose the number of Monte Carlo replicates for stochastic ecological models. Ecol. Model. 394, 11–17.
Leroy, B., Meynard, C.N., Bellard, C., Courchamp, F., 2016. Virtualspecies, an R package to generate virtual species distributions. Ecography 39, 599–607.
Lobo, J.M., Baselga, A., Hortal, J., Jiménez-Valverde, A., Gómez, J.F., 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? Divers. Distrib. 13, 772–780.
Lüdecke, D., 2018. Ggeffects: tidy data frames of marginal effects from regression models. J. Open Source Softw. 3 (26), 772.
Mammola, S., 2019. Assessing similarity of n-dimensional hypervolumes: which metric to use? J. Biogeogr. 46, 2012–2023.
McCune, J.L., 2016. Species distribution models predict rare species occurrences despite significant effects of landscape context. J. Appl. Ecol. 53 (6), 1871–1879.
Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633.
Meyer, C., Weigelt, P., Kreft, H., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19 (8), 992–1006.
Meynard, C.N., Kaplan, D.M., 2013. Using virtual species to study species distributions and model performance. J. Biogeogr. 40, 1–8.
Meynard, C.N., Leroy, B., Kaplan, D.M., 2019. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? Ecography 42 (12), 2021–2036.
Miller, Jennifer A., 2014. Virtual species distribution models: using simulated data to evaluate aspects of model performance. Prog. Phys. Geogr. 38 (1), 117–128.
Moudrý, V., Bazzichetto, M., Remelgado, R., Devillers, R., Lenoir, J., Mateo, R.G., Lembrechts, J.J., Sillero, N., Lecours, V., Cord, A.F., Barták, V., Balej, P., Rocchini, D., Torresani, M., Arenas-Castro, S., Man, M., Prajzlerová, D., Gdulová, K., Prošek, J., Marchetto, E., Zarzo-Arias, A., Gábor, L., Leroy, F., Martini, M., Malavasi, M., Cazzolla Gatti, R., Wild, J., Šímová, P., 2024. Optimising occurrence data in species distribution models: sample size, positional uncertainty, and sampling bias matter. Ecography 12, e07294.
Naimi, B., Skidmore, A.K., Groen, T.A., Hamm, N.A., 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. J. Biogeogr. 38 (8), 1497–1509.
Pärtel, M., 2014. Community ecology of absent species: hidden and dark diversity. J. Veg. Sci. 25 (5), 1154–1159.
Pebesma, E., 2018. Simple features for R: standardized support for spatial vector data. R J. 10 (1), 439–446.
Perret, D.L., Sax, D.F., 2022. Evaluating alternative study designs for optimal sampling of species' climatic niches. Ecography 2022, e06014.

Petersen, T.K., Speed, J.D.M., Grøtan, V., Austrheim, G., 2021. Species data for understanding biodiversity dynamics: the what, where and when of species occurrence data collection. Ecol Solut Evid. 2 (e12048), 8.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197.
Pili, A., Leroy, B., Zurell, D., 2025. Correcting environmental sampling bias improves transferability of species distribution models. Ecography 2025 (10), e08002.
Pulliam, H.R., 2000. On the relationship between niche and distribution. Ecol. Lett. 3, 349–361.
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. Austria. https://www.R-project.org/.
Rathore, Mahima Kanwar, Sharma, Laxmi Kant, 2023. Efficacy of species distribution models (SDMs) for ecological realms to ascertain biological conservation and practices. Biodivers. Conserv. 32 (10), 3053–3087.
Rocchini, D., Tordoni, E., Marchetto, E., et al., 2023. A quixotic view of spatial bias in modelling the distribution of species and their diversity. NPJ Biodivers. 2, 10.
Rocchini, Duccio, Garzon-Lopez, Carol X., Marcantonio, Matteo, Amici, Valerio, Bacaro, Giovanni, Bastin, Lucy, Brummitt, Neil, Chiarucci, Alessandro, Foody, Giles M., Hauffe, Heidi C., He, Kate S., Ricotta, Carlo, Rizzoli, Annapaola, Rosà, Roberto, 2026. Anticipating Species Distributions: Handling Sampling Effort Bias Under a Bayesian Framework.
Roleček, J., Chytrý, M., Hájek, M., et al., 2007. Sampling design in large-scale vegetation studies: do not sacrifice ecological thinking to statistical purism! Folia Geobot. 42, 199–208. https://doi.org/10.1007/BF02893886.
Roubicek, A.J., VanDerWal, J., Beaumont, L.J., 2010. Does the choice of climate baseline matterin ecological niche modelling? Ecol. Model. 221 (19), 2280–2286.
Santi, F., Di Musciano, M., Testolin, R., Zannini, P., Sabatini, F.M., Jiménez-Alfaro, B., et al., 2026. Plant diversity estimates of Mediterranean islands differ among biodiversity databases. Ecography 2026 (4), e08128.
Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., Huijbregts, M.A., 2021. Assessing the reliability of species distribution projections in climate change research. Divers. Distrib. 27, 1035–1050.
Sillero, Neftalí, 2011. What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecol. Model. 222 (8), 1343–1346. ISSN 0304–3800.
Smith, A.B., Armstrong, M.P., Meehan, T.D., et al., 2023. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. Glob. Ecol. Biogeogr. 32, 342–355.
Soley-Guardia, Mariano, Alvarado-Serrano, Diego F., Anderson, Robert P., 2024. Top ten hazards to avoid when modeling species distributions: a didactic guide of assumptions, problems, and recommendations. Ecography 2024, e06852.
Varela, Sara, et al., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37 (11), 1084–1091.
Vilas, D., Fletcher Jr., R.J., Siders, Z.A., Chagaris, D., 2022. Understanding the temporal dynamics of estimated environmental niche hypervolumes for marine fishes. Ecol. Evol. 12 (12), e9604. https://doi.org/10.1002/ece3.9604.
Wood, S.N., 2018. Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation and GAMMs by REML/PQL. R package version 1.8-23.
Zarzo-Arias, A., Penteriani, V., Gábor, L., Šímová, P., Grattarola, F., Moudrý, V., 2022. Importance of data selection and filtering in species distribution models: a case study on the Cantabrian brown bear. Ecosphere 13 (12), e4284.
Zurell, Damaris, et al., 2020. A standard protocol for reporting species distribution models. Ecography 43 (9), 1261–1277.
Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., Grimm, V., 2010. The virtual ecologist approach: simulating data and observers. Oikos 119, 622–635.
Zurell, D., Zimmermann, N.E., Brun, P., 2024. The niche through time: considering phenology and demographic stages in plant distribution models. J. Ecol. 112, 1926–1939.
Zuur, Alain F., et al., 2009. Mixed Effects Models and Extensions in Ecology with R, Vol. 574. Springer, New York.