Received: 14 February 2023 | Accepted: 17 August 2023

DOI: 10.1111/2041-210X.14209

Methods in Ecology and Evolution

RESEARCH ARTICLE

USE it: Uniformly sampling pseudo-absences within the environmental space for applications in habitat suitability models

Daniele Da Re[1] | Enrico Tordoni[2] | Jonathan Lenoir[3] | Jonas J. Lembrechts[4] | Sophie O. Vanwambeke[1] | Duccio Rocchini[5,6] | Manuele Bazzichetto[6]

[1]Georges Lemaître Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, Belgium; [2]Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia; [3]UMR CNRS 7058 «Ecologie et Dynamique des Systèmes Anthropisés» (EDYSAN), Université de Picardie Jules Verne, Amiens, France; [4]Research Group Plants and Ecosystems, University of Antwerp, Antwerp, Belgium; [5]BIOME Lab, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Bologna, Italy and [6]Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Prague, Czech Republic

Correspondence
Daniele Da Re
Email: daniele.dare@uclouvain.be

Funding information
Estonian Research Council, Grant/Award Number: MOBJD1030; FRS-FNRS ASP, Grant/Award Number: 34766961; Horizon Europe Marie Skłodowska-Curie Actions, Grant/Award Number: 101066324; SHOWCASE, Grant/Award Number: 862480; National Recovery and Resilience Plan (NRRP); Horizon Europe projects Earthbridge and B3

Handling Editor: Luis Cayuela

Abstract

1. Habitat suitability models infer the geographical distribution of species using occurrence data and environmental variables. While data on species presence are increasingly accessible, the difficulty of confirming real absences in the field often forces researchers to generate them in silico. To this aim, pseudo-absences are commonly sampled randomly across the study area (i.e. the geographical space). However, this introduces sample location bias (i.e. the sampling is unbalanced towards the most frequent habitats occurring within the geographical space) and favours class overlap (i.e. overlap between environmental conditions associated with species presences and pseudo-absences) in the training dataset.
2. To mitigate this, we propose an alternative methodology (i.e. the uniform approach) that systematically samples pseudo-absences within a portion of the environmental space delimited by a kernel-based filter, which seeks to minimise the number of false absences included in the training dataset.
3. We simulated 50 virtual species and modelled their distribution using training datasets assembled with the presence points of the virtual species and pseudo-absences collected using the uniform approach and other approaches that randomly sample pseudo-absences within the geographical space. We compared the predictive performance of habitat suitability models and evaluated the extent of sample location bias and class overlap associated with the different sampling strategies.
4. Results indicated that the uniform approach: (i) effectively reduces sample location bias and class overlap; (ii) provides comparable predictive performance to

Daniele Da Re, Enrico Tordoni and Manuele Bazzichetto equally contributed to this work.

sampling strategies carried out in the geographical space; and (iii) ensures gathering pseudo-absences adequately representing the environmental conditions available across the study area. We developed a set of R functions in an accompanying R package called `USE` to disseminate the uniform approach.

KEYWORDS
background points, class overlap, ecological niche models, presence-only models, reproducibility, sample location bias, species distribution models

1 | INTRODUCTION

Habitat suitability models (hereafter, HSMs) are a class of statistical models used to describe the relationship between species attributes (e.g. presence–absence and abundance) and a set of spatially explicit variables chiefly representing abiotic, biotic and human-related factors (e.g. climate, soil, demographic parameters and land-use). These models are rooted in the niche theory (i.e. *Hutchinsonian* niche, see Guisan et al., 2017) and rely on both theoretical and practical assumptions: (i) species are assumed to be at (quasi)equilibrium with their environment (Hattab et al., 2017); (ii) the set of predictors used to fit HSMs includes all necessary information to capture the ecological niche of the species; and (iii) species distribution attributes, used as the response variable, need to be appropriate for the intended model purpose (e.g. biodiversity conservation, forecasting biological invasions, assessing the effects of global change; Tessarolo et al., 2021; but see also Guisan et al., 2017 for a thorough review on the theoretical assumptions underpinning HSMs). Some of these assumptions are hardly, if ever, met in nature since species are seldom at equilibrium with their environment (Svenning & Skov, 2004), posing several limitations to the use and interpretation of HSMs' outputs. Acknowledging and, when possible, addressing these limitations still makes HSMs a powerful toolbox for understanding the drivers of the species' realised and potential distributions (sensu Jackson & Overpeck, 2000). For this reason, HSMs are still widely applied in several research fields, including biogeography (Duffy et al., 2017; Wasof et al., 2015), climate change ecology (Jarvie & Svenning, 2018), conservation biology (Newbold, 2018; Santini et al., 2021), invasion ecology (Bazzichetto et al., 2021; Da Re et al., 2020; Hattab et al., 2017) and pathogen risk assessment (Batista et al., 2023).

One of the most critical assumptions underpinning HSMs is the appropriateness of biological data for modelling the ecological niche of the species, which means that species distribution attributes, being either presence–absence or abundance data, should allow an effective description of the true species–environment relationship (Baker et al., 2022; Guisan et al., 2017). However, while information on species occurrence (i.e. presence) is usually readily accessible through field-collected observations or museum/herbaria records, trustworthy absence data are by far more difficult to gather or confirm in the field (Jiménez-Valverde et al., 2008), as their sampling requires labour-intensive and costly field campaigns (Hattab et al., 2017). The usual lack of true absence data has led to the development of HSM approaches that either rely solely on presence data (so-called 'presence-only models', such as the BIOCLIM model; Booth et al., 2014) or combine presence data with pseudo-absences or background points for modelling species distributions (e.g. the MaxEnt algorithm; Phillips et al., 2017).

Pseudo-absences and background points are terms often used interchangeably in the scientific literature (Sillero & Barbosa, 2020), but they may represent different conditions. Pseudo-absences are sampled from locations considered unsuitable for the species (Barbet-Massin et al., 2012). In contrast, background points encompass the full range of environmental conditions, including potential suitable locations for the species (presence locations; Hallgren et al., 2019; Phillips et al., 2009). The choice between pseudo-absences and background points indicates the user's uncertainty about the ecological preferences of the species, with background points used when there is no prior knowledge of unsuitable environmental conditions. Despite recognising the distinction, we will henceforth use the term pseudo-absences to refer to both pseudo-absences and background points for simplicity and alignment with our study.

The most common approaches for sampling pseudo-absences involve (i) randomly surveying a large number of points across the study area (e.g. 10,000; Barbet-Massin et al., 2012; Hysen et al., 2022; Iturbide et al., 2015; Støa et al., 2019) or (ii) sampling them within or (iii) outside buffers created around presence locations (Bedia et al., 2013; VanDerWal et al., 2009). These approaches share the characteristic of deploying pseudo-absences randomly across the geographic space, which often leads to oversampling of the most common habitat conditions that are widespread in the study area (Ronquillo et al., 2020; Tessarolo et al., 2014, 2021). This sample location bias negatively impacts HSMs in multiple ways. First, it can introduce a bias in the sampling of environmental conditions experienced by a species, potentially affecting the accurate estimation of the species response curve, particularly in heterogeneous areas (Albert et al., 2010; Austin, 2007; Bazzichetto et al., 2023; Beck et al., 2014; Hortal et al., 2008). Second, it influences the predictive performance of HSMs, as reflected in the evaluation metrics used (Jiménez-Valverde et al., 2013; Sillero & Barbosa, 2020).

To overcome this issue, previous studies (Hattab et al., 2017; Varela et al., 2014) proposed to sample species presence and (true) absence data throughout a systematic sampling of the environmental conditions available across the study area, thus limiting the artificial constraint imposed by the random sampling towards

the most widespread environments. More specifically, Varela et al. (2014), Hattab et al. (2017) and Perret and Sax (2022) suggested collecting species' presence and/or absence within 2- or 3-dimensional environmental spaces obtained using ordination techniques. Such approaches significantly contributed to the improvement and standardisation of the way species observations, including pseudo-absences, can be collected to calibrate HSMs reducing sample location bias. Yet, they do not explicitly consider class overlap, another relevant methodological issue encountered when collecting pseudo-absences through random sampling across the geographical space. Class overlap refers to the overlap between environmental conditions associated with both species' presence and absence, thus hindering the concept of pseudo-absences itself. It has negative effects on the predictive performance of HSMs and it is particularly critical for machine-learning techniques, while regression techniques such as generalised linear models seem to be less affected (Barbet-Massin et al., 2012; Grimmett et al., 2020; Valavi et al., 2021). So far, class overlap has been addressed using resampling techniques more oriented to adjusting an unbalanced number of classes in the response variable (i.e. the 'up-' or 'down-sampling' approach; Valavi et al., 2021), irrespective of the technique used to obtain pseudo-absences.

As far as we know, there are no approaches for sampling pseudo-absences that seek to mitigate both sample location bias and class overlap. Here, we present an alternative sampling strategy, which we call the 'uniform' approach, that builds upon existing strategies for systematically sampling the environmental space to select pseudo-absences. The novel aspect of the uniform approach is that, beyond reducing sample location bias, it also minimises class overlap by implementing a kernel-based filter that is used to delineate the portion of the environmental space where to collect pseudo-absences. To test our approach, we simulated 50 virtual species and compared the predictive performance of HSMs trained on pseudo-absences sampled using the uniform approach as well as other sampling strategies traditionally carried out within the geographical space: random (i.e. pseudo-absences randomly sampled within the geographical space) and buffer-out (i.e. pseudo-absences randomly collected outside buffers built around presence locations). To foster reproducibility, we provide an accompanying R package called `USE` (Uniform Sampling of the Environmental space), which bundles the R functions needed to implement the uniform approach. The package is available at https://github.com/danddr/USE. Finally, we provide a tutorial to explain how to apply the uniform approach to real case studies, using the European beech *Fagus sylvatica* L. as a target species.

2 | METHODS

2.1 | Simulation of virtual species

We used virtual species (hereafter VS), a simulation tool that provides the great advantage of knowing the true generative process underlying the species geographical distribution (Meynard et al., 2019). We created the realised environmental space (sensu Jackson & Overpeck, 2000) of 50 different virtual species using the bioclimatic variables gathered from the WorldClim database (www.worldclim.org; spatial resolution ~18.6 km at the Equator; Fick & Hijmans, 2017). We restricted the distribution of the simulated VS (and those of the bioclimatic variables) to the geographical extent spanning from -12° W to 25° E and from 36° to 60° N (approximately Western and Southern Europe) to significantly reduce the computational effort to process the entire workflow. Each VS was generated using a random set of five bioclimatic variables (out of the 19) through the function `generateRandomSp` from the R package `virtualspecies` (Leroy et al., 2016), which randomly assigns relationships between the VS and the bioclimatic variables (e.g. linear, quadratic relationships). This way, we obtained a raster layer reporting the habitat suitability index of each VS (HSI, Figure 1a), which we then converted to a binary (i.e. presence–absence) map using the function `convertToPA`. Further details about parameter settings can be found in the R code available at https://github.com/danddr/USE_paper.

2.2 | Sampling the pseudo-absences

Regardless of the sampling approach and modelling technique used to calibrate the HSMs, the ratio between the number of presences and pseudo-absences in the training datasets (i.e. sample prevalence) was kept equal to 1, which means that an equal number of presences and pseudo-absences were collected. In practice, each of the VS-specific training datasets included 300 presences, which were randomly sampled within the geographical extent using the function `sampleOccurrences` from the `virtualspecies` R package. Consequently, we collected an equal number of pseudo-absences according to the three sampling strategies presented below.

2.2.1 | Uniform approach: Pseudo-absences sampled within the environmental space

For each VS (i.e. iteration), we built a 2-dimensional environmental space by keeping the first two axes of a principal component analysis (PCA) performed on the correlation matrix of the five randomly selected bioclimatic variables used to generate the realised environment (Figure 1b). Each time, we checked that the first two principal component axes accounted for at least 70% of the total bioclimatic variability. Then, we uniformly sampled pseudo-absences in the environmental space using the `uniformSampling` function. In short, each pseudo-absence is associated with a geographical location (i.e. a pixel of the environmental layers), which is in turn characterised by the set of environmental conditions encountered at that location. Such a combination of environmental conditions determines the position of the pseudo-absence within the environmental space. A pseudo-absence can thus be defined as the projection of a geographical location onto the environmental space generated through the PCA (i.e. a PC score).

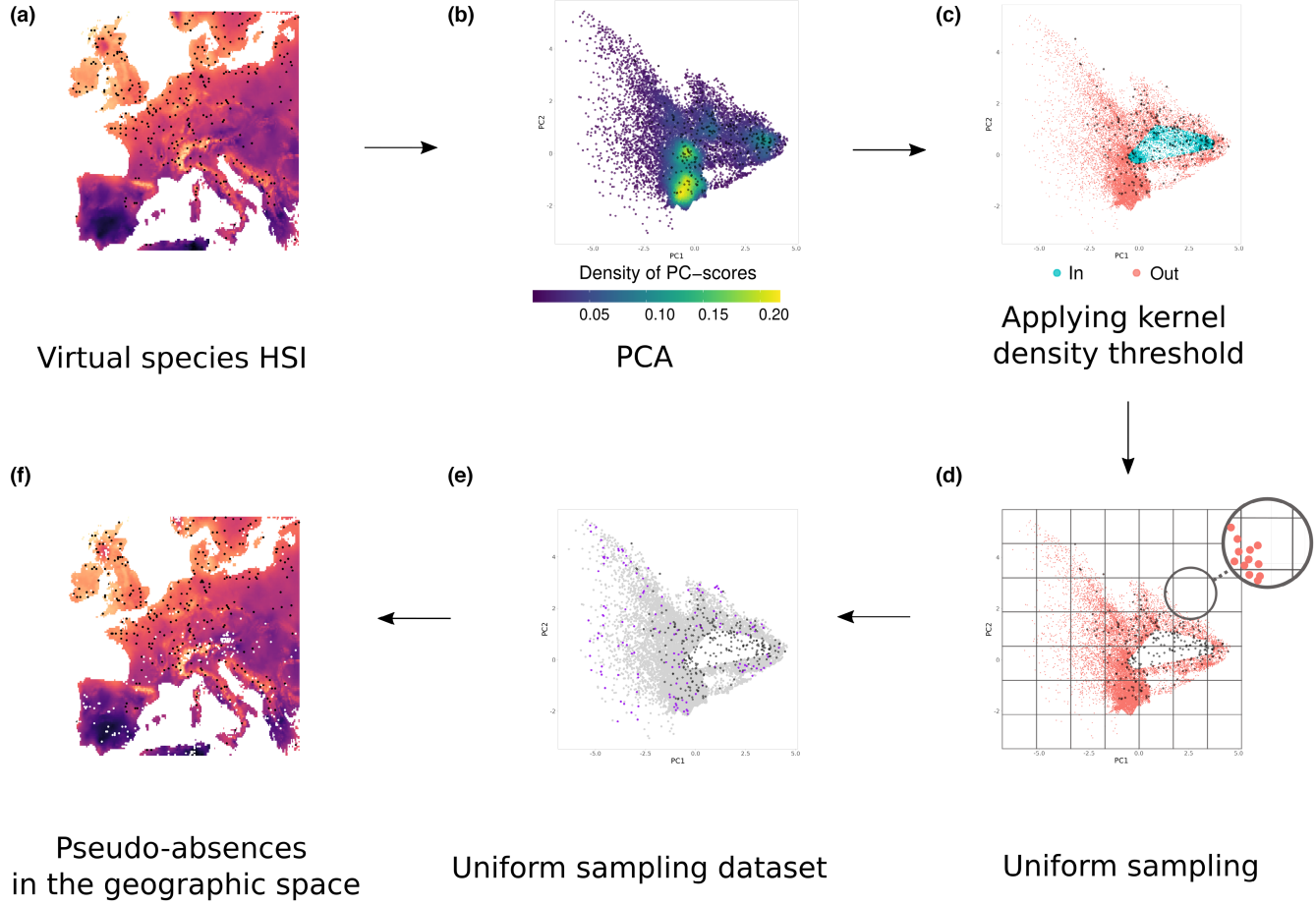

FIGURE 1 Flowchart representing the step-by-step procedure for implementing the uniform approach: (a) habitat suitability index (HSI) of the *i*-th virtual species (VS; lighter colours indicate higher habitat suitability and black dots represent presence points in the geographical space); (b) Principal component analysis (PCA) performed on the environmental variables in the study region (lighter colours indicate high PC scores densities and black dots represent the presence points within the environmental space); (c) application of the kernel-based filter, which splits the environmental space into two subspaces associated with either the environmental conditions more suitable for the species (in blue) or those associated with less/not suitable environmental conditions (in red; with black dots still depicting presence points); (d) pseudo-absences are uniformly sampled across a sampling grid of a chosen resolution overlaid to the 2-dimensional environmental space. Specifically, pseudo-absences are sampled within each cell of the 2-D grid. The inset map shows an example of a grid cell at the boundary of the environmental space (i.e. a grid cell containing a low density of pseudo-absences), black dots represent presence points; (e) the purple dots represent the pool of randomly selected pseudo-absences after running the uniform sampling approach; (f) the white dots represent the selected set of pseudo-absences after running the uniform sampling approach, but displayed in the geographical space this time, black dots still represent presence points from the focal virtual species. The sample prevalence and the number of pseudo-absences sampled within each cell of the sampling grid were defined as `prev` = 1 and `n.tr` = 5, respectively, in the `paSampling` function.

Below, we present a step-by-step description of the uniform sampling performed by the function `paSampling`, which internally calls `uniformSampling` (both functions are included in the `USE` R package):

1. First, kernel density estimation (a statistical technique used to estimate the underlying probability distribution of a set of data points by smoothing them with a kernel function; Scott, 1992) is used to calculate the probability density function of the presence data within the 2-dimensional environmental space. Similar uses of kernel density estimation have become popular in recent years, especially due to their increasing use in trait-based ecology to compute probabilistic hypervolumes and trait probability densities (Mammola & Cardoso, 2020 and reference therein). The PC scores associated with a probability threshold equal to or greater than 0.75 (i.e. the default threshold value used in the `paSampling` function) are likely to bear environmental conditions associated with presence locations. Thus, we selected these presence locations and we generated the convex hull delimiting the portion of the environmental space mostly associated with this set of presence points within the environmental space (Figure 1c). The kernel bandwidth (i.e. the width of the kernel density function that defines its shape) can be either defined by the user or automatically estimated by the function `paSampling`. In the latter case, the function uses a bandwidth selector by internally calling the function `Hpi` of the R package `ks` (Duong, 2021).

2. The portion of the environmental space defined by the above-mentioned convex hull is removed from the whole environmental space. Then, a sampling grid was generated from a preselected resolution (e.g. 10×10 cells) and overlaid on the 2-dimensional environmental space (Figure 1d). The optimal resolution of the sampling grid within the environmental space can be determined using the function `optimRes` from the `USE` package. This function operates as follows:

- Within each cell of the sampling grid, the average (squared) Euclidean distance between the pseudo-absences (PC scores) in the cell and the centroid of their convex hull is computed;
- Once this metric is computed across all cells of the sampling grid, the average mean value is computed across all cells (hereafter, grid average);
- The procedure above is separately repeated on different sampling grids of increasing resolution (i.e. increasing number of cells);
- The resulting set of grid averages (one per resolution) are used as a measure of the aggregation among pseudo-absences within the cells of the sampling grids. This value is compared across resolutions, and the best grid is chosen as the one providing the best trade-off between resolution and average distance among points within cells (i.e. the resolution that allows uniformly sampling the environmental space without overfitting it). More specifically, the best grid is the one whose resolution is just below that which would not allow the average distance among pseudo-absences to be reduced by more than 10% (other values can be set by the user).

3. Once the optimal resolution is set, the sampling grid is sequentially scanned (i.e. cell by cell) by the `uniformSampling` function called via the `paSampling` function and, from each grid cell, a given number of pseudo-absences is randomly collected. At this stage, the pseudo-absences associated with environmental conditions too close to those of the presence locations are already excluded (see step 1). Note that the pseudo-absences are randomly selected within the area of each cell of the sampling grid, and not at the centroid nor at the nodes.

The total number of pseudo-absences sampled within each cell of the sampling grid can be set by the user (using the argument `n.tr`, default `n.tr` = 5), who can also indicate a desired sample prevalence. If the sample prevalence is not specified, fewer pseudo-absences are likely to be eventually sampled than expected (i.e. `n.tr` × number of cells). This happens because (i) no pseudo-absence points are collected in empty cells, and (ii) fewer pseudo-absence points than `n.tr` are available within the cells at the boundary of the environmental space (see zooming window in Figure 1d). Similarly, no pseudo-absences are collected within the core area of the presences (excluded in step 1). If a sample prevalence is set by the user, the sampling grid is surveyed until the chosen sample prevalence is reached by the algorithm.

2.2.2 | Pseudo-absences sampled within the geographical extent

The sampling of pseudo-absences within the geographical extent was conducted using the random and buffer-out approaches. For the random approach (Barbet-Massin et al., 2012; Iturbide et al., 2015; Støa et al., 2019), we simply generated 300 random pseudo-absences across the studied geographical extent. For the buffer-out approach (Bedia et al., 2013), we created a buffer of a 50 km radius around each presence location, and we then randomly sampled pseudo-absences outside the presence-specific buffers, but within the convex hull of the species geographical distribution (i.e. the convex hull that connects the outer presences of the species and thus delimits the range actually covered by the species in the geographical space).

2.3 | Habitat suitability models

For each of the 50 VS and for each of the three sampling strategies (i.e. uniform, random and buffer-out), we built a specific dataset combining the presence records with the pseudo-absences sampled within the environmental and the geographical space. First, we modelled the presence and pseudo-absences data as a function of the same five bioclimatic variables used to generate each of the 50 VS. To this aim, we randomly partitioned each dataset (specific for a sampling strategy) into five replicates of both training (70% observations) and testing (30%) sets, which we used to calibrate and validate, respectively, and for each replicate, five modelling algorithms: (i) binomial generalised linear models with 'logit' link (GLMs); (ii) generalised additive models (GAMs); (iii) random forests (RFs); (iv) boosted regression trees (BRTs); and (v) MaxEnt. In total, we fitted 3750 HSMs (50 VS species×3 different sets of pseudo-absences×5 modelling algorithms×5 replicates of 70%–30% partitions). To fit the HSMs, we used the R package `sdm` (Naimi & Araújo, 2016). Although we acknowledge the importance of fine-tuning HSMs (Fourcade, 2021), we kept model settings at their default value since it would have been unfeasible to individually parametrise each algorithm for all 50 VS and sampling strategies. A detailed representation of the workflow of the analyses is shown in Figure 2. Furthermore, we acknowledge that our use of MaxEnt did not conform with the general recommendations for its adequate implementation (e.g. using 10,000 background points; Cobos et al., 2019; Kass et al., 2021). Nonetheless, we included it in the comparison of models' performance due to its wide usage within the HSM community.

2.4 | Comparison among sampling strategies

2.4.1 | Predictive performance comparison

After fitting HSMs for all the 50 VS, we compared the predictive performance associated with each combination of sampling approaches and modelling techniques by computing the following

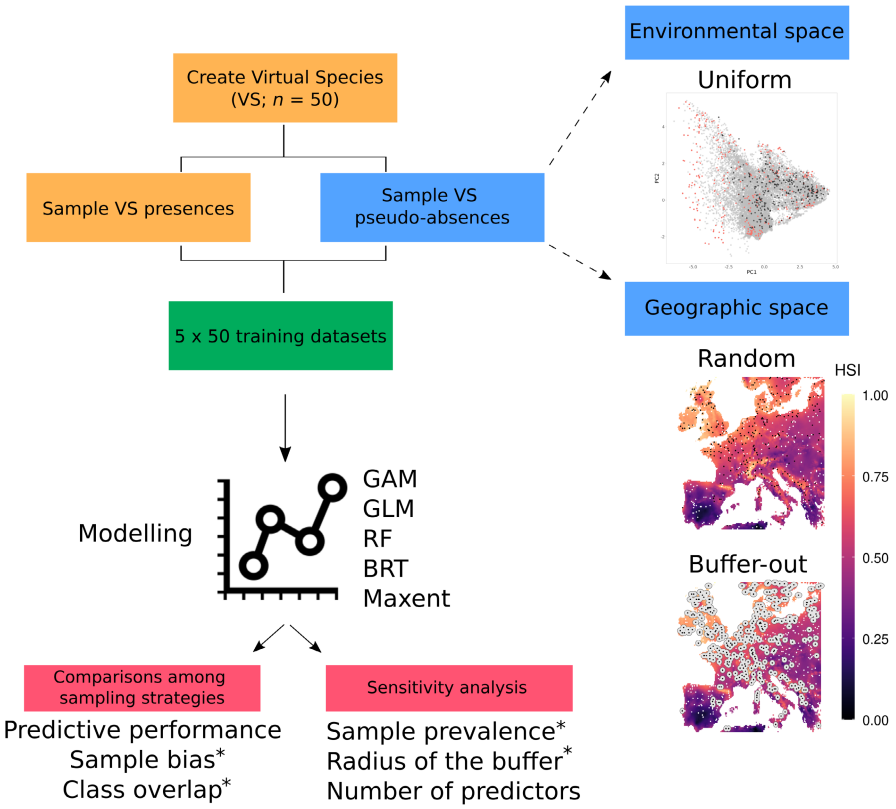

FIGURE 2 The overall workflow of the analysis described in the Methods section. The '*' is associated with analyses (i.e. sample bias, class overlap, sample prevalence and radius of the buffer) performed on $n = 10$ virtual species (VS).

metrics: (i) the area under the receiver operating characteristic curve (AUC); (ii) the continuous Boyce index (CBI); (iii) the sensitivity; (iv) the specificity; (v) the true skill statistics (TSS); and (vi) the root mean squared error (RMSE). The RMSE was computed by comparing the true (i.e. simulated) habitat suitability of the focal VS against the one predicted by each combination of modelling and sampling approach. A detailed description of the above-mentioned modelling techniques and validation metrics can be found in Guisan et al. (2017). To compare the predictive performance of the HSMs fitted under different combinations of sampling strategy and modelling technique, we visually assessed the results of the 50 VS simulations using violin plots reporting the distribution of the values of the predictive performance metrics listed above. Furthermore, we tested for statistical differences between the three sampling strategies for each predictive accuracy metric using the Kruskal–Wallis test, followed by two-tailed Dunn's post hoc rank-sum comparisons using the `dunn.test` R package (Dinno, 2017; *p*-values for multiple comparisons adjusted using Holm correction).

2.4.2 | Sample location bias and class overlap

To assess the intensity of sample location bias associated with the different sampling strategies, we extracted the pseudo-absences of a single VS and mapped their aggregation within the environmental space using bivariate density plots. The aim was to identify which, among the three sampling strategies, was more subject to oversampling particular environmental conditions within the geographical space. In principle, the sampling strategies more affected by sample location bias would exhibit a clear aggregation of pseudo-absences within the environmental space. We visually assessed the areas of the environmental space sampled by the different sampling strategies using the function `geom_density_2d` of the `ggplot2` R package (Wickham, 2016). This function performs a 2D kernel density estimation using the `kde2d` function of the `MASS` R package (Venables & Ripley, 2002) and displays the results with contours. In addition, for 10 new VS, we calculated the total range (i.e. max PC score – min PC score) of the two principal component axes associated with the pseudo-absences collected through the different sampling strategies. We then derived the 95% confidence interval of the total range through a nonparametric bootstrap ($n = 2000$) using the function `smean.cl.boot` from the `Hmisc` R package for each principal component axis and sampling strategies (Harrell, 2021). We tested for statistical differences for each principal component axis among sampling strategies using the Kruskal–Wallis test followed by two-tailed Dunn's post hoc rank-sum comparisons with Holm's correction. To assess the effectiveness of the uniform approach for mitigating class overlap, we simulated 10 new VS, sampled their presences and pseudo-absences using the three sampling strategies and mapped the position of the presence and pseudo-absence points within the environmental space following the procedure explained in Section 2.2.1 and Figure 1a,b. Then, we computed the Gaussian hypervolume of the presences and pseudo-absences using the `hypervolumes` R package (Blonder et al., 2014, 2022)

and calculated the overlap between them. Statistically significant differences in the degree of overlap were tested using one-way ANOVA and Tukey HSD test.

2.5 | Sensitivity analyses

In our analytical framework, we kept the value of the following parameters fixed: sample prevalence, the size of the buffer for the buffer-out approach and the number of bioclimatic variables used as predictors to fit the HSMs for the VS. To test the potential effect on our results of varying these parameters, we conducted the following sensitivity analyses:

- To test the effect of changing sample prevalence on the predictive performance of the different sampling strategies, we repeated the entire workflow on 10 VS using two additional prevalence values, namely 0.5 and 0.1. Specifically, for each VS, we generated two additional training datasets with 300 presences, but we combined them with 600 and 3000 pseudo-absences to achieve a sample prevalence of 0.5 and 0.1, respectively.
- To test the effect of the size of the buffer on the predictive performance of the buffer-out approach, we repeated the entire workflow on 10 VS considering the 100 and 200 km buffer radius lengths, in addition to the 50 km buffer radius length.
- To test how using a different number of bioclimatic variables would affect the predictive performance of the sampling strategies, we repeated the entire workflow on 50 VS using all 19 bioclimatic variables to both define the environmental space to generate the VS and as predictors to fit the related HSMs.

2.6 | Real case study

To illustrate how to apply the uniform approach with the USE R package, we modelled the realised distribution of *Fagus sylvatica* in Italy, France and Spain. We chose *F. sylvatica* as a target species because its distribution and biogeographic history are well-known across Europe (Magri et al., 2006; Poli et al., 2022). The whole analysis of *F. sylvatica* is described in Appendix S5, and the R code to replicate it can be found at: https://github.com/danddr/USE_paper.

3 | RESULTS

3.1 | Comparison of the predictive performance associated with geographical vs environmental sampling

Overall, the uniform approach performed equal to or better than the geographical approaches in terms of out-of-sample prediction (Figure 3). Pairwise comparisons between the predictive accuracy performance of the uniform approach against the random and buffer-out approaches showed statistically significant differences in 73% and 47% of the combinations, respectively. However, these differences were algorithm- and metric-dependent and did not point to an overall higher predictive performance of the uniform approach (Figure 3, Table S1, Figure S1.1). The pattern of the differences among predictive performance metrics was consistent among the prevalence values (Figures S2.1–2.) and the number of bioclimatic variables used in the models (Figure S3). Increasing the buffer radius length (Figure S4) resulted in higher predictive performance of the buffer-out approach for some metrics (AUC, TSS and specificity), while for CBI, sensitivity and RMSE results remained comparable with those presented in Figure 3.

3.2 | Effect of sample location bias and class overlap

The bivariate density plots of the pseudo-absences sampled within the environmental and geographical space highlighted that the uniform approach had the widest and most homogeneous coverage of environmental conditions throughout the environmental space (Figure 4, see Figure S1.2 for a more detailed representation of the density of pseudo-absences sampled within the environmental space when running the uniform approach; Figure S1.3). In contrast, the random and buffer-out approaches appeared to be prone to sample location bias, with peaks of high density of pseudo-absences occurring in specific areas of the environmental space, that is, those associated with the most frequent habitat conditions encountered within the geographical space, and a narrow mean range of PC scores sampled along both principal component axes compared with the uniform approach (Figure 4, Figure S1.3; Kruskal–Wallis test for PC1: $\chi^2 = 21.54$, df = 2, p-value < 0.001; Kruskal–Wallis test for PC2: $\chi^2 = 14.91$, df = 2, p-value < 0.001).

Regarding class overlap, we detected a statistically significant difference in the overlap between the portions of the environmental space occupied by presences and pseudo-absences sampled through different approaches (one-way ANOVA $F(2, 27) = 5.83$, p-value = 0.008). Specifically, the uniform approach exhibited the lowest overlap in comparison with the other sampling strategies (Figure 5). The post hoc Tukey HSD test showed that the uniform approach exhibited a significantly lower overlap than the random sampling ($p < 0.001$), whereas the uniform-buffer-out and buffer-out-random comparisons did not show significant differences ($p = 0.09$, $p = 0.47$).

4 | DISCUSSION

In this study, we proposed the uniform approach as an alternative strategy to sample pseudo-absences within the environmental space. In contrast to existing techniques, our approach systematically samples pseudo-absences from portions of the environmental space excluding the conditions that are likely to

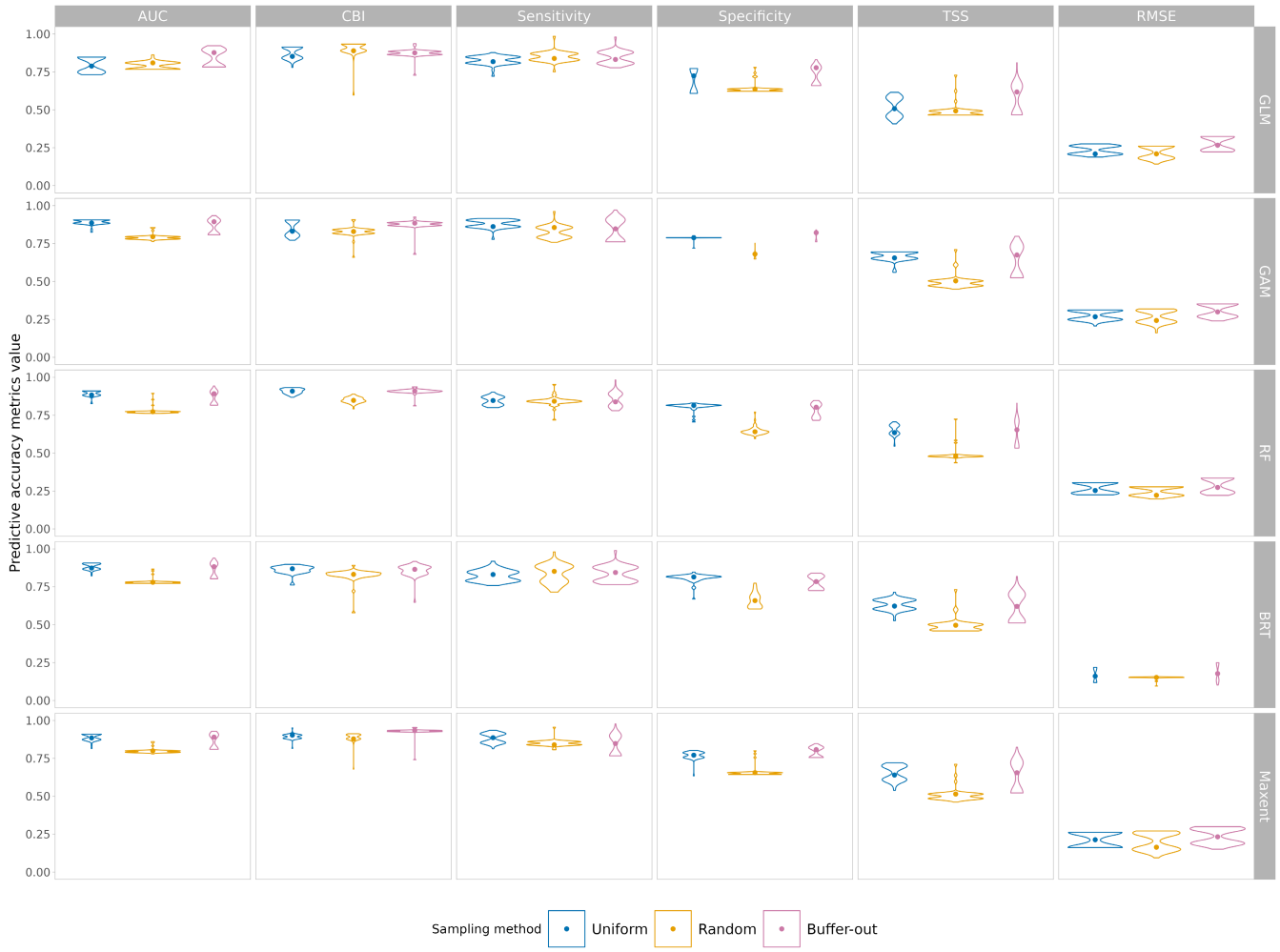

FIGURE 3 Violin plots reporting the distribution of the values of the metrics of predictive performance for the habitat suitability models (HSMs) of the 50 virtual species (VS), as modelled using five randomly selected bioclimatic predictors and setting sample prevalence equal to 1 (i.e. same number of presences and pseudo-absences). Dots represent median values of the metrics of predictive accuracy. Columns indicate the different performance metrics, while rows are associated with the modelling techniques used to fit the HSMs. Higher values in all metrics but RMSE reflect higher predictive performance. AUC = area under the curve; CBI = continuous Boyce index, TSS = true skill statistic; RMSE = root mean squared error; GLM = generalised linear model; GAM = generalised additive model; RF = random forest; BRT = boosted regression trees.

be suitable for the species to establish. As a result, the uniform approach reduces the chance of including false absences in the training dataset. From a more theoretical perspective, data collected after the application of the kernel-based filter are much closer to the concept of pseudo-absences than those obtained through traditional, geographical sampling approaches. Our findings show that the uniform approach represents a valid strategy for gathering pseudo-absences, resulting in out-of-sample predictive accuracy comparable to the sampling strategies implemented within the geographical space. In addition, the uniform sampling significantly reduces sample location bias and class overlap, which is critical to obtain ecologically meaningful pseudo-absences. Importantly, the uniform approach is flexible, as it allows the user to set parameters (e.g. kernel bandwidth, sample prevalence and sampling grid resolution) that control how pseudo-absences are sampled within the environmental space. Such flexibility is particularly valuable to mimic different ecological processes that are easier to capture within the environmental space than within the geographical space (e.g. source–sink dynamics). In all cases, by generating informative pseudo-absences, the uniform approach allows satisfying one of the most critical assumptions underpinning habitat suitability modelling: the need for adequate species distribution attributes (i.e. pseudo-absence data here) to model the species–environment relationship (Guisan et al., 2017).

4.1 | Effect of the sampling approaches on models' predictive performance

Results of the VS' simulations showed that the uniform approach performed well in terms of out-of-sample prediction regardless of

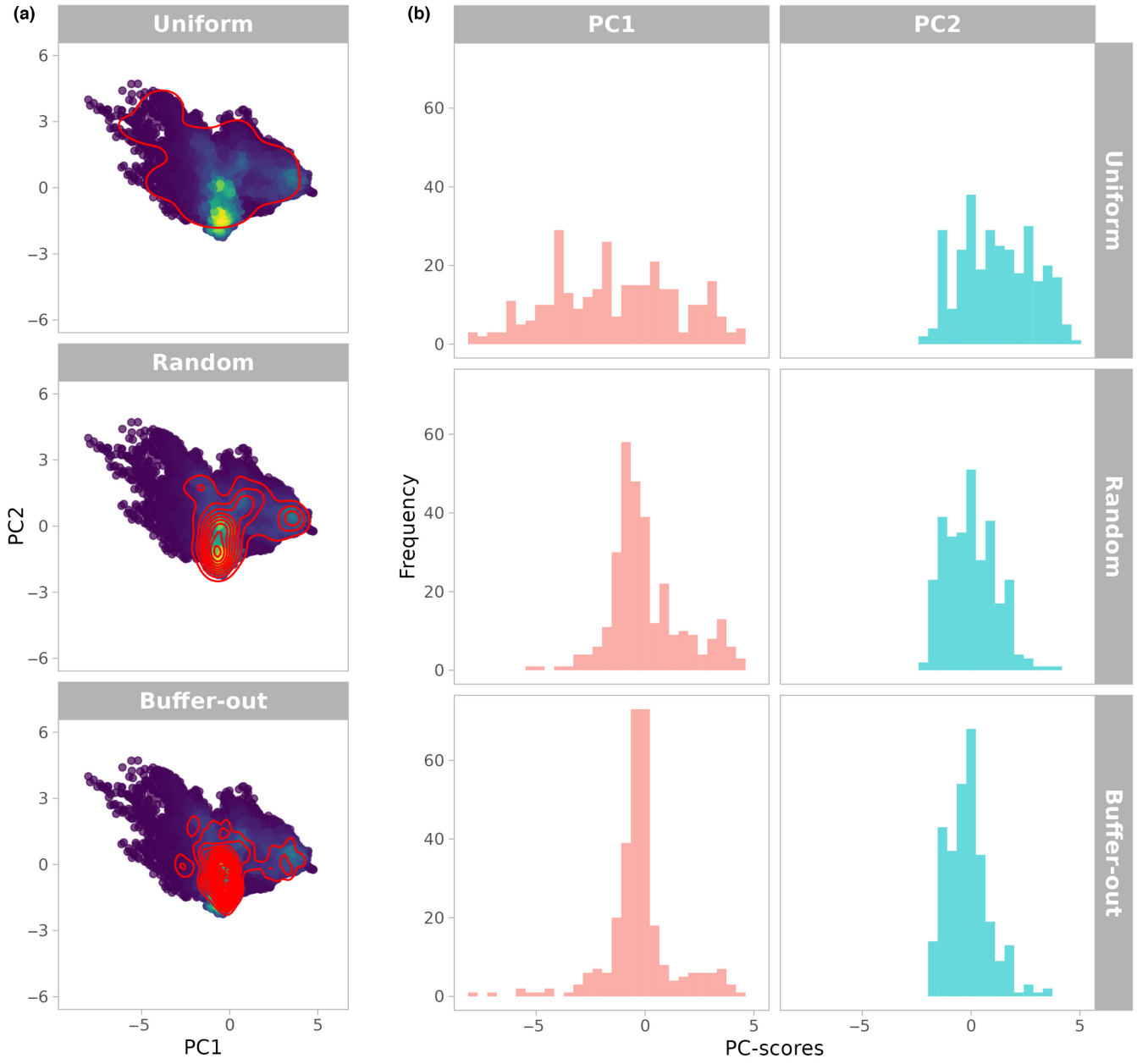

FIGURE 4 (a) Bivariate plots showing the environmental space generated by a principal component analysis carried out on five bioclimatic variables. Red lines represent the density of pseudo-absences for an individual virtual species, as sampled by the random and buffer-out approaches within the geographical space, and by the uniform approach within the environmental space. A more detailed representation of the density of pseudo-absences sampled by the uniform approach is reported in Figure S1.2. (b) Histograms showing the frequency distribution of the first two principal components (columns) associated with the different sampling strategies (rows).

the modelling technique, the metric of predictive performance and the sample prevalence used. All HSMs calibrated on pseudo-absences sampled with the uniform approach consistently showed high predictive performance, especially for the metrics related to the capacity of a model to correctly predict presences (i.e. sensitivity and CBI). Concerning the metrics associated with the model's ability to predict absences (e.g. specificity), the uniform sampling showed values comparable to the other strategies. This suggests that the uniform approach reduces omission error without necessarily increasing commission error. This is coherent with Fei and Yu (2016), who reported an increase in overall model predictive performance when pseudo-absences were systematically collected within the environmental space.

In this sense, results for the CBI, which is currently the go-to accuracy metric for validating HSMs fitted on pseudo-absences (or background points), and for the RMSE were particularly encouraging since the uniform approach scored, together with the buffer-out approach, the highest CBI values and lowest RMSE values across all modelling techniques. The high predictive performance associated with the uniform approach can be attributed to its two main underlying properties: the systematic sampling of the environmental space and the kernel-based filter on the presence observations.

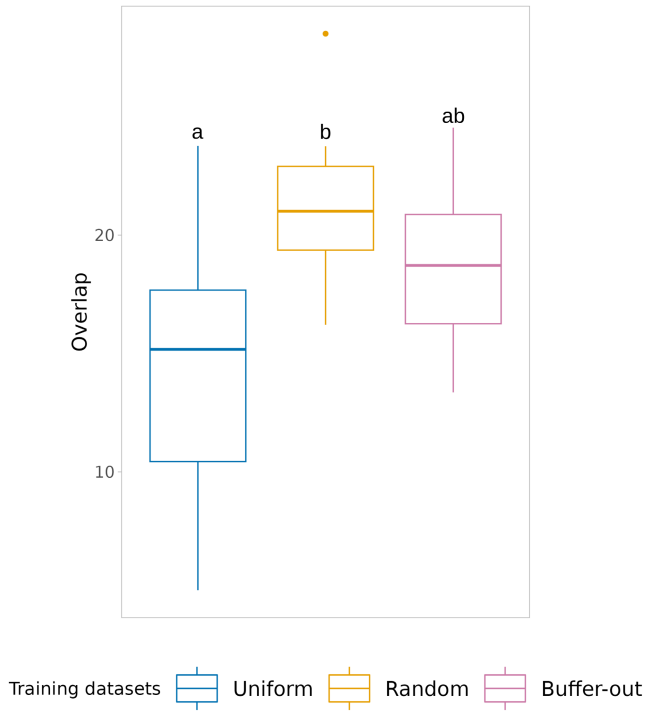

FIGURE 5 Box plots showing the overlap between environmental spaces generated by presences and pseudo-absences of the virtual species. Letters denote significant differences using the Tukey HSD test. Colours are associated with the three sampling strategies used to generate the pseudo-absences (uniform in blue, random in yellow and buffer-out in pink).

Notwithstanding the positive results obtained in terms of predictive performance, we argue that a comparison of metrics of model predictive accuracy may not be the best means for evaluating the adequacy of different sampling strategies carried out within the environmental rather than the geographical space. Indeed, previous studies showed that these metrics are affected by several factors, including sample prevalence (Guisan et al., 2017; Leroy et al., 2018; Marchetto et al., 2023), sample bias (Dubos et al., 2022; Rocchini et al., 2023) or the spatial extent of the study area (Lobo et al., 2008). Moreover, AUC and TSS tend to score high even in case of poor models calibrated on data exhibiting strong sample location bias (Fourcade et al., 2018; Jiménez-Valverde, 2021). Assessing HSM predictive performance using a set of different predictive accuracy metrics might help the user to critically evaluate the outputs of the models.

4.2 | Effect of the uniform sampling on sample location bias and class overlap

The uniform approach proved to significantly reduce sample location bias, since pseudo-absences were homogeneously scattered across the bivariate density plot of the two principal component axes (Figure 4a,b, Figure S1.2 in Supplementary Materials) and collected a wider range of PC scores compared with the random and buffer-out approaches (Figure S1.3). On the contrary, the two sampling approaches carried out within the geographical space exhibited prominent peaks of density of pseudo-absences in correspondence with the most frequently encountered environmental conditions within the geographical space, resulting in a narrower mean of PC scores. As a consequence, the random and buffer-out approaches may provide suboptimal pseudo-absences for modelling the species–environment relationship (Austin, 2007; Thuiller et al., 2004). This aspect gets increasingly relevant as environmental conditions are more heterogeneously distributed across the geographical space (e.g. in mountain regions with high topographic heterogeneity). Therefore, HSMs calibrated on training datasets adequately representing environmental variability rather than wide geographical coverage represent a crucial step to better capture and discriminate species niche breadth (Bazzichetto et al., 2023; Perret & Sax, 2022; Tessarolo et al., 2014, 2021; Varela et al., 2014).

The uniform approach proved to also significantly reduce class overlap. The `thres` argument passed to the `paSampling` function controls the portion of the environmental space associated with the species presence, thus inherently limiting class overlap by the exclusion of environmental conditions suitable to the species (see Figures 1c and 5; Figure S1.4). This results in a set of pseudo-absences theoretically much closer to the species' true absences. Given that presence points are unevenly distributed within the environmental space, different kernel thresholds might also be used to handle the sampling of pseudo-absences under particular scenarios. As an example, setting a low kernel threshold would allow excluding accidental presences from unsuitable locations (e.g. 'sink populations') from the training dataset, while potentially including observations from these areas as pseudo-absences. Unfortunately, there is no a priori choice about the value of the threshold without having preliminary information on the species' ecology, the study area and the goal of the research. For this reason, we provided the `thresh.inspect` function, which produces plots depicting the entire environmental space alongside the portion that would be excluded based on a specific kernel density threshold.

4.3 | Limitations and usage notes

4.3.1 | Limitations

The first limitation of the uniform approach, which is anyway a general limitation in HSMs (e.g. Cayuela et al., 2009), is that its effectiveness depends on the amount (sample size) and quality (e.g. geographically unbiased data sensu Fourcade et al., 2014) of presence data. Indeed, if few presence data are available and/or presence data are geographically biased, the kernel-based filter might not accurately delimit the area associated with suitable conditions for the species. As a consequence, the capacity to discriminate between suitable and unsuitable conditions of the uniform approach might be negatively affected.

A second limitation is that, although the uniform approach proved to be robust to varying sample prevalence, its effectiveness might diminish if a very large number of pseudo-absences is sampled (e.g. in case of low sample prevalence) (Figures S2.1–2.2). Since the uniform approach samples a user-defined number of pseudo-absences within a grid overlaid to a bi-dimensional environmental space, if the number of pseudo-absences grows indefinitely, the advantage of the systematic sampling decreases. Indeed, oversampling the environmental space would generate datasets suffering from sample location bias as much as those based on the random sampling carried out within the geographical space.

From a more practical perspective, the uniform approach can currently operate only across 2-dimensional environmental spaces, but 3-dimensional spaces might be supported in the future.

Finally, although the idea behind `USE` and the uniform sampling approach is to provide users with an easy-to-use tool to generate more ecologically meaningful pseudo-absences, we acknowledge the existence of other techniques designed to avoid generating pseudo-absences altogether. Notable examples are point-process analyses (e.g. Isaac et al., 2020), which model the density of presence-only points per unit area, rather than the probability of presences and (pseudo-)absences. More recently, machine-learning methods based on isolation forests were also proposed, with the R package `ITSDM` specifically dedicated to HSMs (Song & Estes, 2023). We believe, however, that our approach provides a simpler and more intuitive way to deal with the issue of presence-only data and thus has a lower threshold for end-users to implement in their workflow.

4.3.2 | Usage notes

We here used the uniform approach to sample bioclimatic spaces, although we stress the importance of not only using bioclimatic variables but also information on soil, land use and other relevant variables when modelling species distributions. Also, we invite potential users of the uniform sampling approach to always check that the first two axes of the principal component analysis used to generate the environmental space explain a large portion of the variance observed in the data (e.g. ≥70%). Equally important is the choice of the boundaries of the geographical extent for which the 2-dimensional space has to be generated. Indeed, to avoid the 'there are no elephants in the Antarctic' paradox (Lobo et al., 2010), the spatial extent of the study area should be delineated so that it excludes geographical locations, and in turn environmental conditions, less suitable for the species (e.g. collecting pseudo-absences from Mediterranean coastal dunes when modelling the distribution of an alpine plant species). In short, the uniform approach can provide exhaustive information on where the species is likely to not occur, but it remains the responsibility of the end user to carefully verify if such information is ecologically meaningful.

5 | CONCLUSIONS

In this study, we compared the predictive performance of two strategies for sampling pseudo-absences carried out within the geographical space with that of the uniform approach, which operated within the environmental space. Also, we compared geographical and environmental sampling approaches in terms of their vulnerability to sample location bias and class overlap. The uniform approach proved to have good predictive performances and to reduce sample location bias and class overlap, thereby representing a valid alternative to generate pseudo-absences for HSMs. We made the uniform approach openly available to the modellers community at https://github.com/danddr/USE.

AUTHOR CONTRIBUTIONS

Manuele Bazzichetto conceived the idea of the uniform approach and wrote the related R functions, while Enrico Tordoni and Daniele Da Re integrated the kernel density-based estimation of presences and the prevalence-related settings. Daniele Da Re, Enrico Tordoni and Manuele Bazzichetto performed the simulations, analysed the data and assembled the USE R package. Jonathan Lenoir, Jonas J. Lembrechts, Sophie O. Vanwambeke and Duccio Rocchini critically commented on the results of the analyses and their interpretation; Daniele Da Re, Enrico Tordoni and Manuele Bazzichetto led the writing of the manuscript and produced a first draft, which was further improved by all other authors.

ACKNOWLEDGMENTS

The authors are grateful to Prof. Joaquin Hortal, who provided constructive feedback and commented on a previous version of this manuscript. We are also grateful to the MEE's Associate Editor Prof. Luis Cayuela and the two anonymous reviewers for the very constructive comments and suggestions received during the revision process. Simulations were carried out using the facilities of the High-Performance Computing Center of the University of Tartu.

FUNDING INFORMATION

Daniele Da Re was supported by a FRS-FNRS ASP Belgian grant No 34766961, Enrico Tordoni is supported by the Estonian Research Council grant (MOBJD1030), Manuele Bazzichetto acknowledges funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101066324. This study has received funding from the project SHOWCASE (SHOWCASing synergies between agriculture, biodiversity and ecosystems services to help farmers capitalising on native biodiversity) within the European Union's Horizon 2020 Researcher and Innovation Programme under grant agreement No. 862480. Duccio Rocchini was partially funded by a research project implemented under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and

Research funded by the European Union – NextGenerationEU. Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP J33C22001190001, Project title "National Biodiversity Future Center - NBFC". Duccio Rocchini was also partially funded by the Horizon Europe projects Earthbridge and B3.

CONFLICT OF INTEREST STATEMENT

No conflict of interest has been declared by the authors.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14209.

DATA AVAILABILITY STATEMENT

The scripts for replicating the analyses presented in this paper are available at https://github.com/danddr/USE_paper (https://zenodo.org/badge/latestdoi/565763084), as well as all the raw outputs of the simulations and statistical analyses (which are available as an RDS file). The USE package is available on GitHub at https://github.com/danddr/USE (https://zenodo.org/badge/latestdoi/381982533). We provide a general tutorial to explain how to apply the USE package at https://danddr.github.io/USE/articles/USE_vignette.html. In addition, we provide a tutorial on how to apply the uniform approach based on a real species (the European beech, *Fagus sylvatica* L.) in S5. The R script related to the tutorial is available at https://github.com/danddr/USE_paper.

ORCID

Daniele Da Re https://orcid.org/0000-0002-3398-9295
Enrico Tordoni https://orcid.org/0000-0002-9722-6692
Jonathan Lenoir https://orcid.org/0000-0003-0638-9582
Jonas J. Lembrechts https://orcid.org/0000-0002-1933-0750
Sophie O. Vanwambeke https://orcid.org/0000-0001-6620-6173
Duccio Rocchini https://orcid.org/0000-0003-0087-0594
Manuele Bazzichetto https://orcid.org/0000-0002-9874-5064

REFERENCES

Albert, C. H., Yoccoz, N. G., Edwards, T. C., Jr., Graham, C. H., Zimmermann, N. E., & Thuiller, W. (2010). Sampling in ecology and evolution—Bridging the gap between theory and practice. *Ecography*, *33*(6), 1028–1037. https://doi.org/10.1111/j.1600-0587.2010.06421.x

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*(1–2), 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005

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. *Global Ecology and Biogeography*, *00*, 1–13. https://doi.org/10.1111/geb.13491

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Batista, E., Lopes, A., Miranda, P., & Alves, A. (2023). Can species distribution models be used for risk assessment analyses of fungal plant pathogens? A case study with three Botryosphaeriaceae species. *European Journal of Plant Pathology*, *165*(1), 41–56. https://doi.org/10.1007/s10658-022-02587-7

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. *Global Ecology and Biogeography*, *32*, 1717–1729. https://doi.org/10.1111/geb.13725

Bazzichetto, M., Massol, F., Carboni, M., Lenoir, J., Lembrechts, J. J., Joly, R., & Renault, D. (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. *Journal of Vegetation Science*, *32*(4), e13057. https://doi.org/10.1111/jvs.13057

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bedia, J., Herrera, S., & Gutiérrez, J. M. (2013). Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections. *Global and Planetary Change*, *107*, 1–12. https://doi.org/10.1016/j.gloplacha.2013.04.005

Blonder, B., Lamanna, C., Violle, C., & Enquist, B. J. (2014). The n-dimensional hypervolume. *Global Ecology and Biogeography*, *23*(5), 595–609. https://doi.org/10.1111/geb.12146

Blonder, B., Morrow, C. B., Harris, D. J., Brown, S., Butruille, G., Laini, A., & Chen, D. (2022). *Hypervolume: High dimensional geometry, set operations, projection, and inference using kernel density estimation, support vector machines, and Convex Hulls*. R package version 3.0.4. https://CRAN.R-project.org/package=hypervolume

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). BIOCLIM: The first species distribution modelling package, its early applications and relevance to most current maxent studies. *Diversity and Distributions*, *20*(1), 1–9. https://doi.org/10.1111/ddi.12144

Cayuela, L., Golicher, D. J., Newton, A. C., Kolb, M., De Alburquerque, F. S., Arets, E. J. M. M., Alkemade, J. R. M., & Pérez, A. M. (2009). Species distribution modeling in the tropics: Problems, potentialities, and the role of biological data for effective species conservation. *Tropical Conservation Science*, *2*(3), 319–352. https://doi.org/10.1177/194008290900200304

Cobos, M. E., Peterson, A. T., Barve, N., & Osorio-Olvera, L. (2019). kuenm: An R package for detailed development of ecological niche models using Maxent. *PeerJ*, *7*, e6281. https://doi.org/10.7717/peerj.6281

Da Re, D., Tordoni, E., De Pascalis, F., Negrín-Pérez, Z., Fernández-Palacios, J. M., Arévalo, J. R., Rocchini, D., Medina Félix, M., Otto, R., Arlé, E., & Bacaro, G. (2020). Invasive fountain grass (*Pennisetum setaceum* (Forssk.) Chiov.) increases its potential area of distribution in Tenerife Island under future climatic scenarios. *Plant Ecology*, *221*(10), 867–882. https://doi.org/10.1007/s11258-020-01046-9

Dinno, A. (2017). *dunn.test*: *Dunn's test of multiple comparisons using rank sums*. R package version 1.3.5. https://CRAN.R-project.org/package=dunn.test

Dubos, N., Préau, C., Lenormand, M., Papuga, G., Monsarrat, S., Denelle, P., Le Louarn, M., Heremans, S., Roel, M., Roche, P., & Luque, S. (2022). Assessing the effect of sample bias correction in species distribution models. *Ecological Indicators*, *145*, 109487. https://doi.org/10.1016/j.ecolind.2022.109487

Duffy, G. A., Coetzee, B. W., Latombe, G., Akerman, A. H., McGeoch, M. A., & Chown, S. L. (2017). Barriers to globally invasive species are weakening across the Antarctic. *Diversity and Distributions*, *23*(9), 982–996. https://doi.org/10.1111/ddi.12593

Duong, T. (2021). *ks*: *Kernel smoothing*. R package version 1.13.3. https://cran.r-project.org/web/packages/ks/index.html

Fei, S., & Yu, F. (2016). Quality of presence data determines species distribution model performance: A novel index to evaluate data quality. *Landscape Ecology*, *31*(1), 31–42. https://doi.org/10.1007/s10980-015-0272-7

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian *Obama nungara*. *Ecological Modelling*, *457*, 109686. https://doi.org/10.1016/j.ecolmodel.2021.109686

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. https://doi.org/10.1111/geb.12684

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (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. https://doi.org/10.1371/journal.pone.0097122

Grimmett, L., Whitsed, R., & Horta, A. (2020). Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecological Modelling*, *431*, 109194. https://doi.org/10.1016/j.ecolmodel.2020.109194

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge University Press.

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., & Mackey, B. (2019). Species distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, *408*, 108719. https://doi.org/10.1016/j.ecolmodel.2019.108719

Harrell, F., Jr. (2021). *Hmisc: Harrell miscellaneous*. R package version 4.6-0. https://CRAN.R-project.org/package=Hmisc

Hattab, T., Garzón-López, C. X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., Brasseur, B., Gallet-Moron, E., Spicher, F., Decocq, G., Feilhauer, H., Honnay, O., Kempeneers, P., Schmidtlein, S., Somers, B., van de Kerchove, R., Rocchini, D., & Lenoir, J. (2017). A unified framework to model the potential and realized distributions of invasive species within the invaded range. *Diversity and Distributions*, *23*(7), 806–819. https://doi.org/10.1111/ddi.12566

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*(6), 847–858. https://doi.org/10.1111/ddi.12566

Hysen, L., Nayeri, D., Cushman, S., & Wan, H. Y. (2022). Background sampling for multi-scale ensemble habitat selection modeling: Does the number of points matter? *Ecological Informatics*, *72*, 101914. https://doi.org/10.1016/j.ecoinf.2022.101914

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., & Gutiérrez, J. M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, *312*, 166–174. https://doi.org/10.1016/j.ecolmodel.2015.05.018

Jackson, S. T., & Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late quaternary. *Paleobiology*, *26*(S4), 194–220. https://doi.org/10.1017/S0094837300026932

Jarvie, S., & Svenning, J. C. (2018). Using species distribution modelling to determine opportunities for trophic rewilding under future scenarios of climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1761), 20170446. https://doi.org/10.1098/rstb.2017.0446

Jiménez-Valverde, A. (2021). Prevalence affects the evaluation of discrimination capacity in presence-absence species distribution models. *Biodiversity and Conservation*, *30*(5), 1331–1340. https://doi.org/10.1007/s10531-021-02144-4

Jiménez-Valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M., & Real, R. (2013). Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. *Global Ecology and Biogeography*, *22*(4), 508–516. https://doi.org/10.1111/geb.12007

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: The importance of concepts in species distribution modelling. *Diversity and Distributions*, *14*(6), 885–890. https://doi.org/10.1111/j.1472-4642.2008.00496.x

Kass, J. M., Muscarella, R., Galante, P. J., Bohl, C. L., Pinilla-Buitrago, G. E., Boria, R. A., Soley-Guardia, M., & Anderson, R. P. (2021). ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods in Ecology and Evolution*, *12*(9), 1602–1608. https://doi.org/10.1111/2041-210X.13628

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2018). Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, *45*(9), 1994–2002. https://doi.org/10.1111/jbi.13402

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, *33*(1), 103–114. https://doi.org/10.1111/j.1600-0587.2009.06039.x

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., Latałowa, M., Litt, T., Paule, L., Roure, J. M., Tantau, I., van der Knaap, W. O., Petit, R. J., & De Beaulieu, J. L. (2006). A new scenario for the quaternary history of European beech populations: Palaeobotanical evidence and genetic consequences. *New Phytologist*, *171*(1), 199–221. https://doi.org/10.1111/j.1469-8137.2006.01740.x

Mammola, S., & Cardoso, P. (2020). Functional diversity metrics using kernel density n-dimensional hypervolumes. *Methods in Ecology and Evolution*, *11*(8), 986–995. https://doi.org/10.1111/2041-210X.13424

Marchetto, E., Da Re, D., Tordoni, E., Bazzichetto, M., Zannini, P., Celebrin, S., Chieffallo, L., Malavasi, M., & Rocchini, D. (2023). Testing the effect of sample prevalence and sampling methods on probability- and favourability-based SDMs. *Ecological Modelling*, *477*, 110248. https://doi.org/10.1016/j.ecolmodel.2022.110248

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. https://doi.org/10.1111/ecog.04385

Naimi, B., & Araújo, M. B. (2016). Sdm: A reproducible and extensible R platform for species distribution modelling. *Ecography*, *39*(4), 368–375. https://doi.org/10.1111/ecog.01881

Newbold, T. (2018). Future effects of climate and land-use change on terrestrial vertebrate community diversity under different scenarios. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1881), 20180792. https://doi.org/10.1098/rspb.2018.0792

Perret, D. L., & Sax, D. F. (2022). Evaluating alternative study designs for optimal sampling of species' climatic niches. *Ecography*, *2022*(1). https://doi.org/10.1111/ecog.06014

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, *40*(7), 887–893. https://doi.org/10.1111/ecog.03049

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. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Poli, P., Guiller, A., & Lenoir, J. (2022). Coupling fossil records and traditional discrimination metrics to test how genetic information improves species distribution models of the European beech *Fagus sylvatica*. *European Journal of Forest Research*, *141*, 253–265. https://doi.org/10.1007/s10342-021-01437-1

Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Barbosa, A. M., Bazzichetto, M., Beierkuhnlein, C., Castelnuovo, E., Gatti, R. C., Chiarucci, A., Chieffallo, L., da Re, D., di Musciano, M., Foody, G. M., Gabor, L., Garzon-Lopez, C. X., Guisan, A., Hattab, T., Hortal, J., … Malavasi, M. (2023). A quixotic view of spatial bias in modelling the distribution of species and their diversity. *Npj Biodiversity*, *2*, 10. https://doi.org/10.1038/s44185-023-00014-6

Ronquillo, C., Alves-Martins, F., Mazimpaka, V., Sobral-Souza, T., Vilela-Silva, B., Medina, N. G., & Hortal, J. (2020). Assessing spatial and temporal biases and gaps in the publicly available distributional information of Iberian mosses. *Biodiversity Data Journal*, *8*. https://doi.org/10.3897/BDJ.8.e53474

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., & Huijbregts, M. A. (2021). Assessing the reliability of species distribution projections in climate change research. *Diversity and Distributions*, *27*(6), 1035–1050. https://doi.org/10.1111/ddi.13252

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. John Wiley & Sons.

Sillero, N., & Barbosa, A. M. (2020). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, *35*, 1–14. https://doi.org/10.1080/13658816.2020.1798968

Song, L., & Estes, L. (2023). ITSDM: Isolation forest-based presence-only species distribution modelling and explanation in R. *Methods in Ecology and Evolution*, *14*(3), 831–840. https://doi.org/10.1111/2041-210X.14067

Støa, B., Halvorsen, R., Stokland, J. N., & Gusarov, V. I. (2019). How much is enough? Influence of number of presence observations on the performance of species distribution models. *Sommerfeltia*, *39*(1), 1–28. https://doi.org/10.2478/som-2019-0001

Svenning, J.-C., & Skov, F. (2004). Limited filling of the potential range in European tree species. *Ecology Letters*, *7*(7), 565–573. https://doi.org/10.1111/j.1461-0248.2004.00614.x

Tessarolo, G., Lobo, J. M., Rangel, T. F., & Hortal, J. (2021). High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, *121*, 107147. https://doi.org/10.1016/j.ecolind.2020.107147

Tessarolo, G., Rangel, T. F., Araújo, M. B., & Hortal, J. (2014). Uncertainty associated with survey design in species distribution models. *Diversity and Distributions*, *20*(11), 1258–1269. https://doi.org/10.1111/ddi.12236

Thuiller, W., Brotons, L., Araújo, M. B., & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, *27*(2), 165–172. https://doi.org/10.1111/j.0906-7590.2004.03673.x

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731–1742. https://doi.org/10.1111/ecog.05615

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, *220*(4), 589–594. https://doi.org/10.1016/j.ecolmodel.2008.11.010

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11), 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/

Wasof, S., Lenoir, J., Aarrestad, P. A., Alsos, I. G., Armbruster, W. S., Austrheim, G., Bakkestuen, V., Birks, H. J. B., Bråthen, K. A., Broennimann, O., Brunet, J., Bruun, H. H., Dahlberg, C. J., Diekmann, M., Dullinger, S., Dynesius, M., Ejrnæs, R., Gégout, J.-C., Graae, B. J., & Decocq, G. (2015). Disjunct populations of European vascular plant species keep the same climatic niches. *Global Ecology and Biogeography*, *24*(12), 1401–1412. https://doi.org/10.1111/geb.12375

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Table S1. Post-hoc multiple comparisons with Dunn's rank sum test ($\alpha = 0.05$; the omnibus test was always significant with $p < 0.05$, data not shown).

Figure S1.1. Post-hoc multiple comparisons with two-tailed Dunn's rank sum test ($\alpha = 0.05$; the omnibus test was always significant with $p < 0.05$, data not shown).

Figure S1.2. Bivariate density plot of principal component scores associated with the pseudo-asbences sampled for a virtual species using the uniform approach.

Figure S1.3. Mean (points) and 95% confidence interval (error bars) of the principal components' total range (max PC-score – min PC-score) captured by the three sampling strategies.

Figure S1.3. Mean (points) and 95% confidence interval (error bars) of the principal components' total range (max PC-score – min PC-score) captured by the three sampling strategies.

Figure S2.1. Violin plots reporting the distribution of the values of the metrics of predictive performance for the habitat suitability models of 10 virtual species (dots represent median values of the metrics of predictive performance), considering 5 predictors, and using a sample prevalence equal to 0.5. Columns indicate the different performance metrics, while rows are associated with the modelling algorithms used to fit the habitat suitability models.

Figure S2.2. Violin plots reporting the distribution of the values of the metrics of predictive performance for the habitat suitability models of 10 virtual species (the dots represent median values of the metrics of predictive performance), considering 5 predictors, and using a sample prevalence equal to 0.1.

Figure S3. Violin plots reporting the distribution of the values of the metrics of predictive performance for the habitat suitability models of 50 virtual species modelled as a function of 19 bioclimatic predictors, and setting sample prevalence equal to 1 (i.e. same number of presences and pseudo-absences).

Figure S4.1. Violin plots reporting the distribution of the values of the metrics of predictive performance for the habitat suitability models of 10 virtual species modelled as a function of 5 bioclimatic predictors, and setting sample prevalence equal to 1 (i.e. same number of presences and pseudo-absences).

Table S5.1. Results of the habitat suitability models for *Fagus sylvatica* (generalised linear model, GLM, and random forest, RF).

Figure S5.2. (A) environmental space available for *Fagus sylvatica* in Italy, Spain and France, and the position of presences (light blue) and pseudo-absences (red) sampled within the environmental space using the uniform approach; (B) distribution of principal component scores across the geographical space, and location (across western Europe) of presences (light blue) and pseudo-absences (red) sampled using the uniform approach.

How to cite this article: Da Re, D., Tordoni, E., Lenoir, J., Lembrechts, J. J., Vanwambeke, S. O., Rocchini, D., & Bazzichetto, M. (2023). USE it: Uniformly sampling pseudo-absences within the environmental space for applications in habitat suitability models. *Methods in Ecology and Evolution*, *14*, 2873–2887. https://doi.org/10.1111/2041-210X.14209