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], and Manuele Bazzichetto[7,8†]

[1] Georges Lemaître Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Place Louis Pasteur 3, 1348 Louvain-la-Neuve, Belgium.

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

[3] UMR CNRS 7058 «Ecologie et Dynamique des Systèmes Anthropisés» (EDYSAN), Université de Picardie Jules Verne, 1 rue des Louvels, 80000 Amiens, France

[4] Research Group Plants and Ecosystems, University of Antwerp, Belgium

[5] BIOME Lab., Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Via Irnerio 42, 40126 Bologna, Italy

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

[7] Department of Ecology and Global Change, Centro de Investigaciones sobre Desertificacion, Moncada (Valencia), Spain

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

†DDR, ET and MB equally contributed to this work

Corresponding author: Daniele Da Re, daniele.dare@uclouvain.be

Abstract

1. Correlative 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 to confirm real absences in the field often forces researchers to generate them *in silico*. To this aim, pseudo-absences are commonly randomly sampled 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 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., uniform approach) that systematically samples pseudo-absences within a portion of the environmental space delimited by a kernel-based filter, which minimises the number of false-absences included in the training set.

3. We simulated 50 virtual species and modelled their distribution using training datasets assembled with the occurrences 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 the 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 accuracy than sampling

strategies carried out in the geographic space; (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, pseudo-absence, ecological niche models, environmental space, habitat suitability models, presence-only models, sample location bias, class overlap, species distribution models, reproducibility.

1 Introduction

Correlative habitat suitability models (hereafter, HSMs) are a class of statistical models used to describe the relationship between species attributes (e.g., presence-absence, abundance) and a set of spatially-explicit variables chiefly representing abiotic and human-related factors (e.g., climate, soil, 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 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; 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 and 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' realized and potential distributions (*sensu* Jackson and Overpeck, 2000). For this reason, HSMs are still widely applied in several research fields, including biogeography (Wasof et al., 2015; Duffy et al., 2017), climate change ecology (Jarvie and Svenning, 2018), conservation biology (Newbold, 2018; Santini et al., 2021), and invasion ecology (Hattab et al., 2017; Da Re et al. 2020; Bazzichetto et al. 2021).

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 attributes, being either presence-absence or abundance data, should allow effectively describing the true species-environment relationship (Guisan et al., 2017; Baker et al., 2022). 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 to 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 HSMs 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). The terms pseudo-absences and background points are often used as synonyms in the scientific literature (Sillero and Barbosa, 2020), yet these two concepts reflect rather different conditions. On the one hand, pseudo-absences are sampled from geographical locations that are thought to feature unsuitable environmental conditions for the species to establish (Barbet-Massin et al., 2012). On the other hand, background points are collected from the whole spectrum of environmental conditions present in the study area, thereby possibly including suitable locations for the species (i.e., presence locations; Phillips et al., 2009; Hallgren et al., 2019). Therefore, the use of pseudo-absences rather than background points reflects the user's degree of uncertainty about the species' ecological preferences, with background points being used when there is no *a priori* knowledge about the unsuitable environmental conditions for the species. Although we acknowledge the difference between pseudo-absences and background points, for the sake of simplicity and because we feel the concept of pseudo-absence adheres more to what we propose in this study, hereafter we will always refer to pseudo-absences.

To date, the most common approaches for sampling pseudo-absences involve (i) surveying a large sample of points randomly located across the study area (e.g., 10,000; Barbet-Massin et al. 2012; Iturbide et al., 2015; Støa et al., 2019) or sampling them either (ii) within or (iii) outside the area covered by buffers built around presence locations (VanDer Wal et al., 2009; Bedia et al., 2013). Beyond the pros and cons of each individual approach, a common thread is that they all randomly deploy pseudo-absences across the geographical space, which usually results in the oversampling of the most common habitat conditions, namely those that are more geographically widespread throughout the area under investigation (Tessarolo et al., 2014, 2021; Ronquillo et al., 2020). This phenomenon, which is generally known as sample location bias (Elith et al. 2011), has detrimental effects on HSMs for different reasons. First, it determines the incomplete sampling of the environmental conditions actually experienced by a species (i.e., the realised environment *sensu* Jackson and Overpeck, 2000), possibly leading to the estimation of truncated species response curves (Hortal et al., 2008; Albert et al., 2010; Beck et al., 2014). Second, it affects the predictive performance of HSMs (Acevedo et al., 2012), which is reflected in the behaviour of the metrics used to evaluate them (Jiménez-Valverde et al., 2013; Sillero and Barbosa, 2020).

To overcome this issue, previous studies (Varela et al. 2014; Hattab et al., 2017) proposed to sample species presence and (true) absence data through a systematic sampling of the environmental conditions available throughout 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 species presence and absence, thus hindering the concept of pseudo-absences itself. It has negative effects on the predictive performance of HSMs and is particularly critical for machine learning techniques, while regression techniques such as GLMs seem to be less affected (Barbet-Massin et al., 2012; Grimmett, Whitsed and Horta, 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 to obtain pseudo-absences.

As far as we know, there are no approaches for sampling pseudo-absences that are able to mitigate both sample location bias and class overlap. Thus, here we present an alternative sampling strategy, which we called 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: (i) random (i.e., pseudo-absences randomly sampled within the geographical space); (ii) buffer-in and iii) buffer-out (i.e., pseudo-absences randomly collected within or outside buffers built around presence locations, respectively). 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) as they provide the great advantage of knowing the true generative process underlying the species geographical distribution (Meynard et al., 2019). The realised environmental space (*sensu* Jackson and Overpeck 2000) of the 50 virtual species was created using the bioclimatic variables gathered from the WorldClim database (www.worldclim.org; spatial resolution ~18.6 km at the Equator; Fick and Hijmans, 2017). We restricted the distribution of the simulated VS (and those of the climatic 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 climatic 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 those climatic variables (e.g., linear, quadratic relationships). This way, we obtained a raster layer reporting the habitat suitability index (HSI, Fig. 1a), which we then converted to a binary (i.e., presence-absence) map using the function `convertToPA`. Further details about parameters setting can be found in the R code available at https://github.com/danddr/USE_paper.

2.2 Sampling of 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 an equal number of presences and pseudo-absences were collected. In practice, each of the VS-specific training dataset 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 four 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 (Fig. 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, here intended as the PC-scores projected onto the environmental space, using the function `uniformSampling`. More specifically, each pair of PC-scores represents the position of a given geographical location, as defined by the bioclimatic conditions it features, within the environmental space. Below, we present a step-by-step description of the uniform sampling performed by the function `paSampling`, which internally calls `uniformSampling`, in the `USE` R package:

1. First, kernel density estimation is used to calculate the probability density function of

the presence data within the 2-dimensional environmental space (Fig. 1c). 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 and Cardoso, 2020 and reference therein). All pseudo-absences associated with a probability threshold equal to or greater than 0.75 (i.e., the default threshold value used in the `paSampling` function) are excluded, since these points are likely to bear environmental conditions associated with presence locations and can therefore introduce false-absences in the training dataset. 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. A sampling grid constituted by a pre-selected resolution (e.g., 10 X 10 cells) is overlaid on the 2-dimensional environmental space (Fig. 1d). The optimal resolution of the sampling grid can be found 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;
- The same measure is computed in each cell of the sampling grid and the average of the cell-specific averages is computed (hereafter, grid average);
- The procedure above is separately repeated on sampling grids of increasing resolution (i.e., increasing number of cells);
- The resulting grid averages 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., 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 resolution is set, the sampling grid is sequentially scanned (i.e., cell by cell) by the `uniformSampling` function called via `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). Notice 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 behaviour happens because no points are collected in empty cells, and less points than n.tr are available within the cells at the boundary of the environmental space (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 set sample prevalence is achieved.

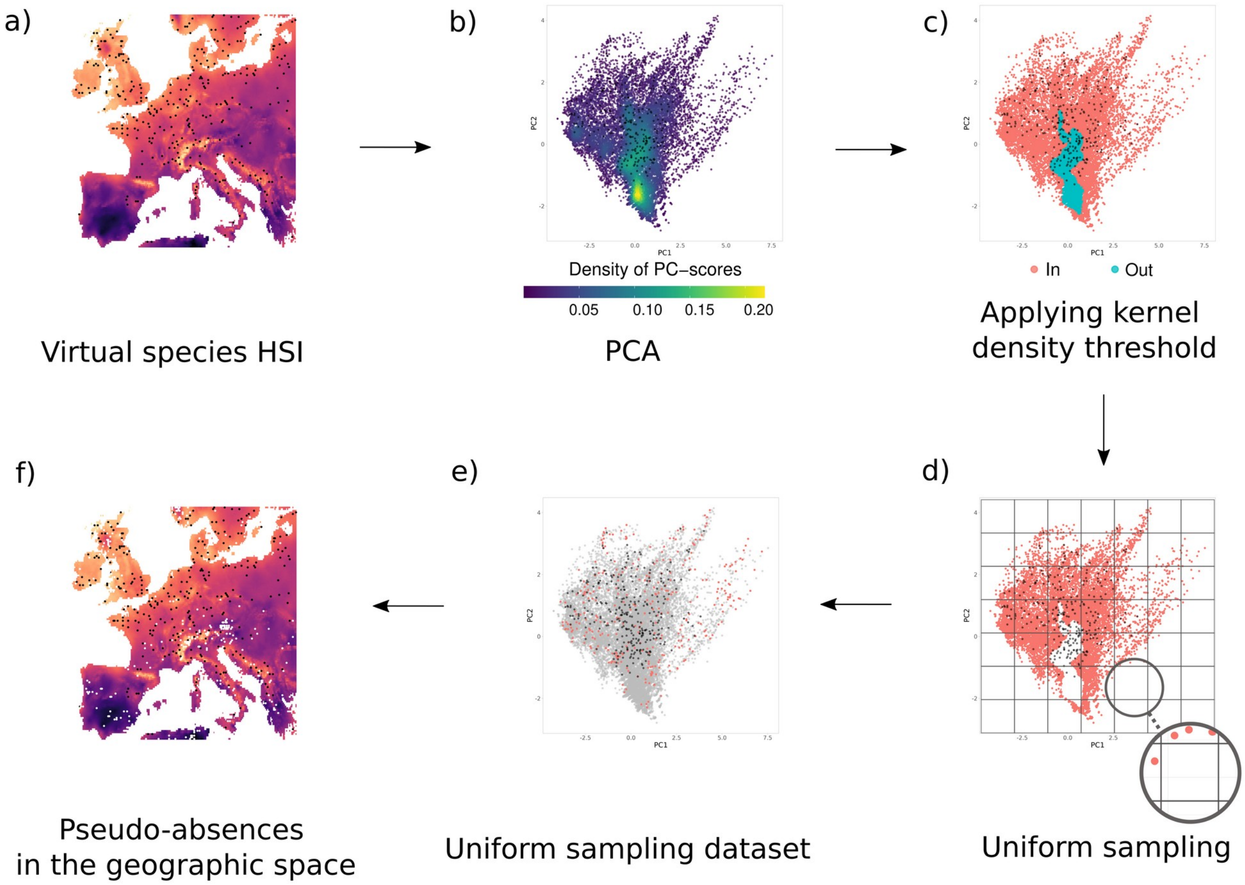


Figure 1: Flowchart representing the step-by-step procedure for implementing the uniform approach: a) habitat suitability index of the *i-th* virtual species (lighter colours indicate higher habitat suitability and black dots represent presence points in the geographical space); b) PCA performed on the environmental variables in the study region, lighter colours indicate high PC-scores densities and black dots represent the presence points in the environmental space; c) application of the kernel-based filter, which splits the environmental space in sub-spaces associated with either the environmental conditions more suitable for the species (blue) or those associated with less/not suitable environmental conditions (red; black dots represent presence points); d) pseudo-absences are uniformly sampled across a sampling

grid 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 an empty cell (i.e., a grid cell containing no pseudo-absences; black dots represent presence points); e) the red dots represent the dataset of pseudo-absences collected within the environmental space using the uniform approach; f) the white dots represent the pseudo-absences collected within the environmental space using the uniform approach displayed in the geographical space, black dots represent VS presence points.

2.2.2 Pseudo-absences sampled within the geographical extent

The sampling of pseudo-absences within the geographical extent defined above was conducted using the random, buffer-in and buffer-out approaches. For the random approach (Barbet-Massin et al. 2012; Iturbide et al., 2015; Støa et al., 2019), we simply located 300 random pseudo-absences across the geographical extent. For the other approaches, we created a buffer of 50 km radius around each presence location, and then we randomly sampled the pseudo-absences within (cf. buffer-in; VanDer Wal et al., 2009) and outside (cf. buffer-out; Bedia et al., 2013) the buffers. Notice that for the buffer-out approach pseudo-absences were collected within the convex hull of the species distribution (i.e., the convex hull that connects the outer occurrences of the species and thus delimits the range actually covered by the species in the geographical space).

2.3 Comparison among sampling strategies

2.3.1 Predictive performance comparison

The overall workflow of the analyses is described in Fig. 2. For each of the 50 VS and for each of the four sampling strategies (i.e., uniform, random, buffer-in, buffer-out), we built a specific dataset combining the presence records with the pseudo-absences sampled within the environmental and the geographical space (Fig. 1e). First, we modelled the presence–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) in 5 training (70% observations) and testing (30%) sets, which we used to calibrate and validate five modelling algorithms: (i) generalised linear models (GLMs); (ii) generalised additive models (GAMs); (iii) random forests (RFs); (iv) boosted regression trees (BRTs); and (v) MaxEnt. In total, we fitted 5,000 HSMs (50 VS species × 4 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 and Araújo, 2016). Although we acknowledge the importance of fine-tuning HSMs (Fourcade, 2021), we leave model settings at their default value since it would have been unfeasible to individually parameterise each algorithm for all 50 VS and sampling strategies.

Then, we compared the predictive performance of each combination of sampling approaches and modelling techniques computing the following metrics: (i) area under the receiver operating characteristic curve (AUC); (ii) continuous Boyce index (CBI); (iii) sensitivity; (iv) specificity; and (v) true skill statistics (TSS). A detailed description of the five modelling techniques and five 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 differences among the predictive performance of the sampling strategies using Kruskall-Wallis tests, followed by Dunn's post hoc rank sum comparisons using the `dunn.test` R package (Dinno, 2017) and correcting p-values for multiple comparisons with the Holm correction.

To test the potential effect of varying sample prevalences on our comparison, we repeated the entire workflow on 10 VS using two different prevalence values: 0.5 and 0.1. Specifically, for each VS, we generated two training datasets with 300 presences, but we combined them with 600 and 3,000 pseudo-absences, respectively.

2.3.2 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 map their spatial aggregation within the environmental space using bivariate density plots. The aim was to identify which, among the four 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 and Ripley, 2002) and displays the results with contours.

To assess the effectiveness of the uniform approach for mitigating class overlap, we simulated 10 further VS, sampled their presences and pseudo-absences using the four 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 hypervolume of the presences and pseudo-absences using the `hypervolumes` R package (Blonder, 2022) and calculated the overlap between them. Significant differences in the degree of overlap were tested using one-way ANOVA and Tukey HSD test.

2.4 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 (hereafter, Western Europe). We chose *F. sylvatica* as a target species because its distribution and biogeographic history is well-known across Europe (Magri et al., 2006; Poli et al., 2022). The whole procedure is described in S4.

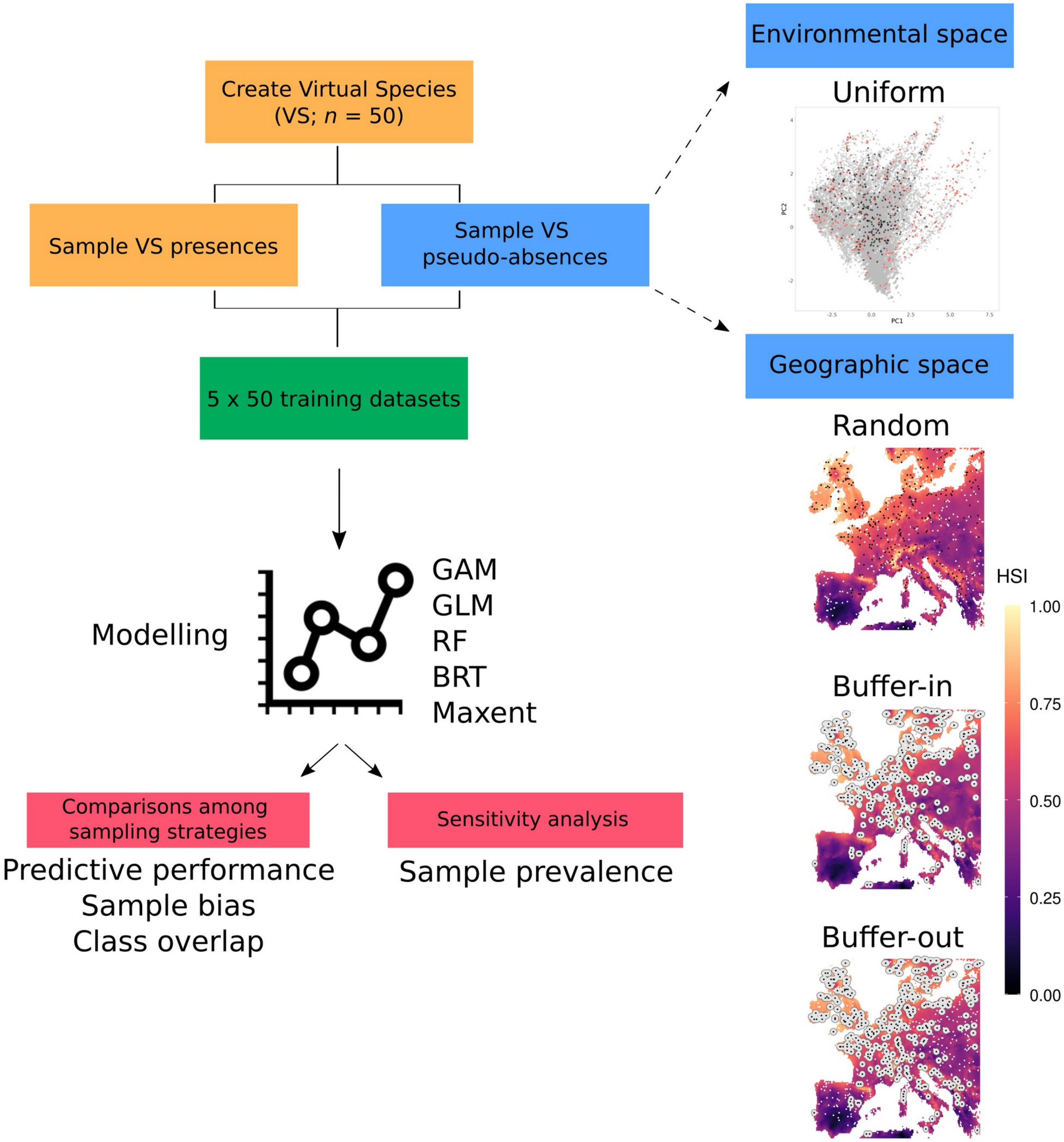
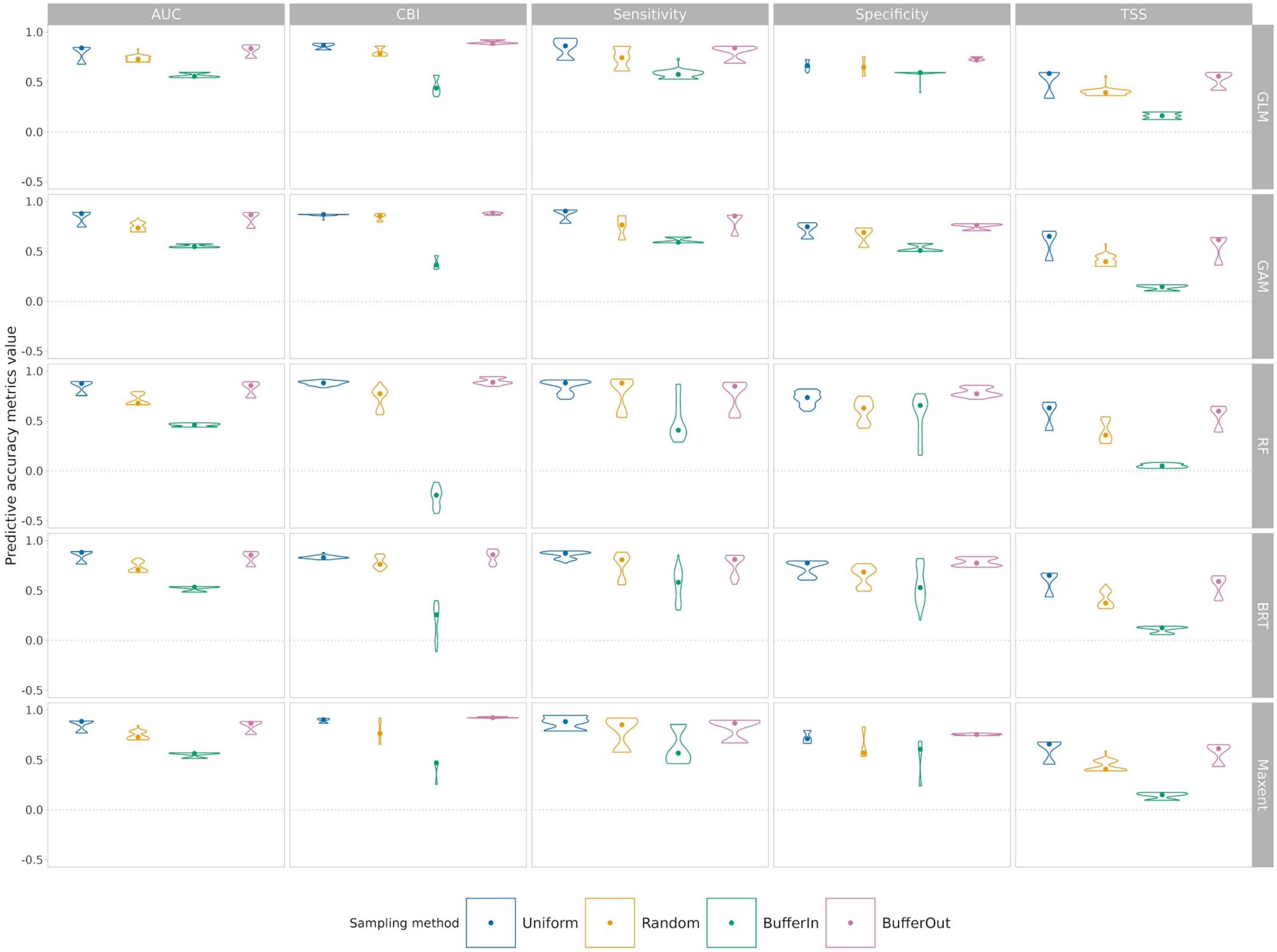


Figure 2 Overall workflow of the analysis described in the Methods section.

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. In particular, the uniform and buffer-out strategies showed, on average, the highest predictive accuracy and their performances were not significantly different (Fig. 3). Pairwise comparisons between the performance of the random and buffer-out approaches against the uniform approach showed statistically significant differences in 92% and 72% of the cases (5,000 models obtained from 50 VS x 4 sampling strategies x 5 algorithms x 5 replicates), respectively. However, these differences were algorithm- and metric- dependent and did not point to a higher predictive performance of the uniform approach (Fig. 3). The buffer-out and uniform approaches exhibited the most similar values for AUC, sensitivity and TSS, while CBI values tended to be higher with respect to random sampling (i.e., Dunn's test: p-value > 0.05; Tab. S1, Fig. S1.1). Finally, the buffer-in consistently showed the lowest performance in all comparisons regardless of the algorithm and predictive performance metric used. The observed pattern of the difference among predictive performances was consistent across sample prevalences (Fig. S3.1-3.2).



356

357 **Figure 3**: Violin plots reporting the distribution of the values of the metrics of predictive
358 performance for the HSMs of the 50 VS modelled as a function of 5, randomly selected bioclimatic
359 predictors, and setting sample prevalence equal to 1 (i.e., same number of presences and pseudo-
360 absences). Dots represent median values of the metrics of predictive accuracy. Columns indicate
361 the different performance metrics, while rows the modelling techniques used to compute HSMs.

362

363

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 (Fig. 4, see Figure S2.1 for a detailed overview of the density of pseudo-absences sampled by the uniform approach). In contrast, the random, buffer-in and buffer-out strategies appeared to be prone to sample location bias, with peaks of high density of pseudo-absences occurring in specific areas of the environmental space, i.e. those associated with the most frequent habitat conditions encountered within the geographical space.

Regarding class overlap, we detected a significant difference in the overlap between ranges occupied by presence and pseudo-absence points within the environmental space (one-way ANOVA F(3,36) = 39, p-value < 0.001). Specifically, the uniform approach exhibited the lowest overlap in comparison to the other sampling strategies (Fig. 5).

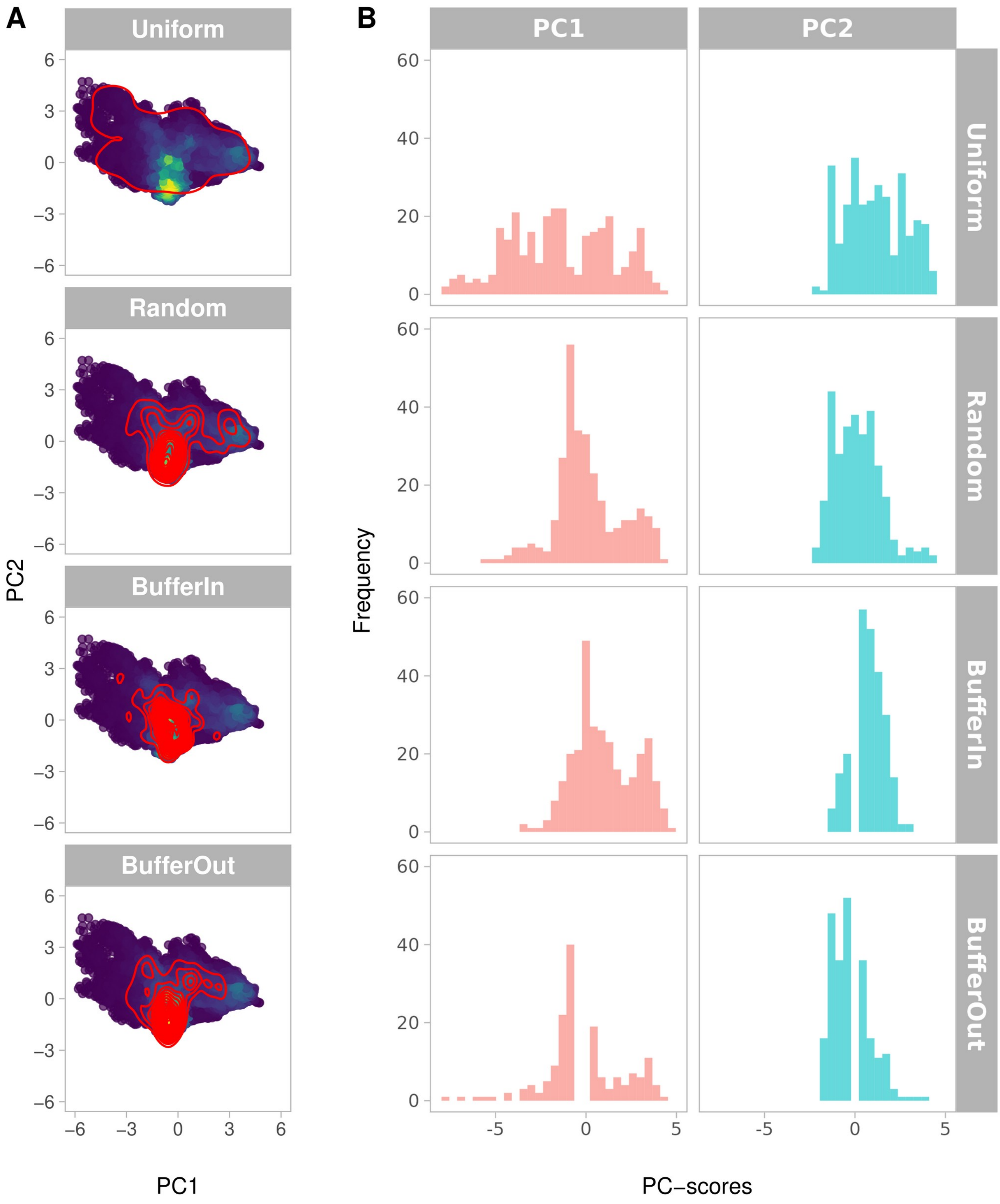


Figure 4: A) Bivariate plots showing the environmental space generated by a PCA carried out on 5 bioclimatic variables. Red lines represent the density of pseudo-absences within the environmental

space for an individual virtual species. A more detailed representation of the density of pseudo-absences sampled by the uniform approach is available in Figure S2.1. B) Histograms showing the frequency distribution of the first two PCs.

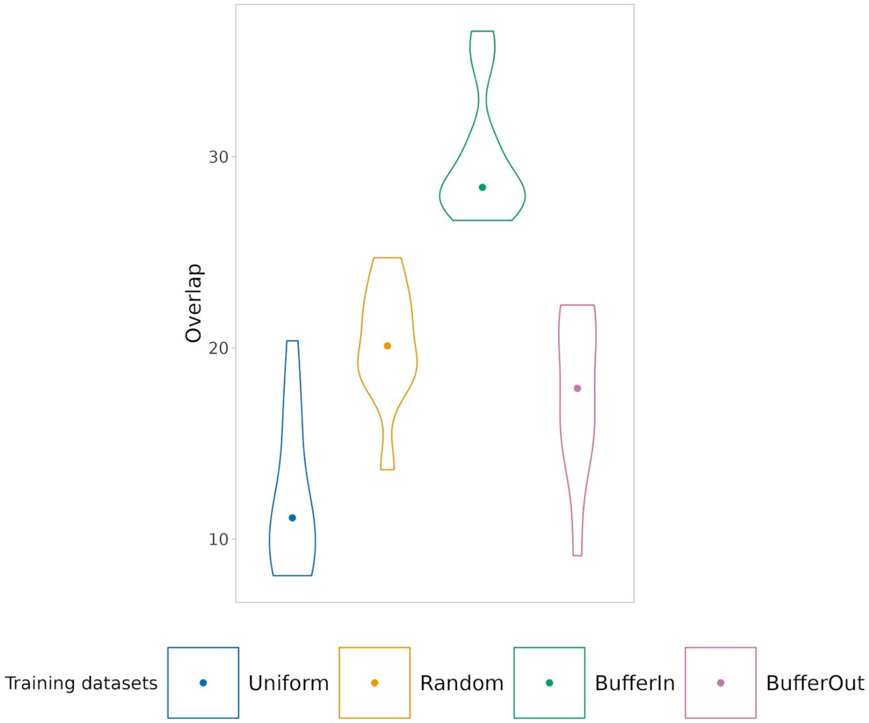


Figure 5: Violin plots showing the overlap in the environmental space between species presences and pseudo-absences. Colours represent samples of pseudo-absences generated using four different strategies, dots represent median values of overlap across 50 VS.

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 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 showed that the uniform approach represents a valid strategy for gathering pseudo-absences, as it performs equally or better than sampling strategies implemented within the geographical space in terms of model out-of-sample predictive accuracy. Also, 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 lets the user free to set parameters (e.g., kernel bandwidth, sample prevalence, sampling grid resolution) that control how pseudo-absences are sampled within the environmental space. This is particularly valuable, as it makes this approach adaptable for modelling species with different ecological properties (e.g., generalist vs specialist species). By generating informative pseudo-absences, the uniform approach allows satisfying one of the most critical assumptions underpinning habitat suitability modelling: the need of adequate species attributes to model the species-environment relationship (Guisan et al., 2017).

4.1 Effect of the sampling approaches on models' predictive performances

Results of the VS simulations showed that the uniform approach performed well in terms of out-of-sample prediction regardless of the modelling technique, metric of predictive performance, and sample prevalence. HSMs calibrated on pseudo-absences sampled with the uniform approach consistently showed high predictive performance, especially for the accuracy metrics related to the capacity of the model to correctly predict presences (i.e., sensitivity and CBI). Concerning the metrics associated with the models' ability to predict absences (e.g., specificity), the uniform sampling showed values comparable to the other strategies, except for the buffer-in approach, which always scored the lowest values. This clearly 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 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, were particularly encouraging: the uniform approach scored, together with the buffer-out approach, the highest values across all modelling techniques. The high predictive performance associated with the uniform approach can be attributed to how it operates: the systematic sampling of the environmental space and the kernel-based filter. In particular, the specular trends of the predictive accuracy metrics (Fig. 3) and the environmental overlap among pseudo-absences collected through the different sampling approaches (Fig. 5) highlight the importance of the kernel-based filter to favour the discrimination between the environmental features associated with presences and pseudo-absences.

Notwithstanding the positive results obtained in terms of predictive performance, we argue that comparing model predictive accuracy may not be the best choice for evaluating the adequacy of sampling 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), sample bias (Dubos et al., 2022) 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 HSMs 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 along the two principal component axes of the bivariate density plot (Fig. 3ab, Fig S1.2 in Supplementary Materials). On the contrary, the sampling approaches based on geographical space, which all perform a random sampling of the pseudo-absences, exhibited prominent peaks of density of pseudo-absences in correspondence of the most frequently environmental conditions available within the geographical space. As a consequence, the random, buffer-in and buffer-out approaches are likely to provide sub-optimal pseudo-absences for modelling the species-environment relationship, potentially resulting in the estimation of truncated species response curves (Thuiller et al. 2004; Austin 2007). This aspect is increasingly relevant as environmental conditions are more heterogeneously distributed across the geographic space. Therefore, HSMs calibrated on training datasets adequately representing environmental variability rather than wide geographical coverage represent a crucial step to better capture and discriminate the niche variability of a species (Tessarolo et al., 2014, 2021; Varela et al., 2014; Perret and Sax 2022).

The uniform approach proved to also significantly reduce class overlap. The `thres` argument in the `paSampling` function controls the portion of the environmental space associated with the species presence, thus inherently limiting the class overlap issue by excluding environmental conditions more favourable to the study species (see Fig. 1c, 5 and Fig. S2.2). 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 pseudo-absences sampling under particular scenarios. As an example, in case of source-sink dynamics, setting more conservative thresholds for the kernel functions may allow excluding pseudo-absences from environmentally suitable areas, while not excluding areas where a sink population is present due to accidental or mass dispersal close to a source population.

4.4 Limitations and usage notes

4.4.1 Limitations

A first limitation of the uniform approach is that its effectiveness depends on the amount (sample size) and quality (e.g., geographically unbiased data *sensu* Fourcade 2014) of presence data. Indeed, if few presence data are available and/or are geographically biased, the kernel-based filter might not accurately delimit the area associated with suitable conditions for the species. As a consequence, the discrimination between suitable and not suitable conditions within the environmental space might be sub-optimal.

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) (Fig. S3.1-3.2). Since the uniform approach samples a user-defined number of pseudo-absences within a grid overlaid to the 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 of the geographic space. Finally, from a more practical perspective, the uniform approach can currently only operate across 2-dimensional environmental spaces, but 3-dimensional spaces might be supported in the future.

4.4.2 Usage notes

We here used the uniform approach to sample climatic spaces, although we stress the importance of not only using bioclimatic variables, but also information on soil, land-use as well as other relevant variables when modelling species distributions. Also, we invite potential users of the uniform approach to always check that the first two axes of the principal component analysis used to generate the environmental space explains a large portion of the variance 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 where it is not possible to find the target species due to ecophysiological limitations (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 a responsibility of the end user to carefully verify if such information is ecologically meaningful.

5 Conclusion

In this study, we evaluated the predictive performance of four sampling strategies, of which one implemented within the environmental space, to collect pseudo-absences for HSMs applications. Also, we compared the sampling approaches in terms of their vulnerability to sample location bias and class overlap. The sampling strategy which we proposed, the uniform approach, proved to (i) have good predictive performances, and (ii) to reduce sample location bias and class overlap. The uniform approach is openly available to users at https://github.com/danddr/USE.

6 Declaration

- Ethics approval and consent to participate: Not applicable.
- Competing interests: No conflict of interest has been declared by the authors.
- Funding: DDR is supported by a FRS-FNRS Belgian grant, ET is supported by an Estonian Research Council grant (MOBJD1030), MB acknowledges funding from the European Union’s Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101066324.
- Authors’ contribution: MB conceived the idea of the Uniform approach and wrote the related R functions, while ET and DDR integrated the kernel density-based estimation of presences and the prevalence-related settings. DDR, ET and MB performed the simulations, analysed the data and assembled the `USE` R package. JL, JJL, SOV, and DR critically commented on the results of the analyses and their interpretation; DDR, ET and MB 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 Joaquin Hortal, who provided constructive feedback and commented on a previous version of this manuscript. Simulations were carried out using the facilities of the High-Performance Computing Center of the University of Tartu.

7 Code and Data availability

The scripts for replicating the analyses presented in this paper are available at https://github.com/danddr/USE_paper, as well as all the raw outputs of the simulations and statistical analysis, which are available as an .RDS file.

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 in S4. The R script of the tutorial is available at https://github.com/danddr/USE_paper.

References

Acevedo, P., Jiménez-Valverde, A., Lobo, J. M., and Real, R. (2012). Delimiting the geographical background in species distribution modelling. *Journal of biogeography*, 39(8):1383–1390.

Albert, C. H., Yoccoz, N. G., Edwards Jr, T. C., Graham, C. H., Zimmermann, N. E., and Thuiller, W. (2010). Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography*, 33(6):1028–1037.

Austin, M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological modelling*, *200*(1-2), 1-19.

Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo absences for species distribution models: how, where and how many? *Methods in ecology and evolution*, 3(2):327–338.

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.

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.

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

Bedia, J., Herrera, S., and 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.

Blonder B, Morrow wcfCB, Harris DJ, 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., and 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.

Da Re, D., Tordoni, E., De Pascalis, F., Negrín-Pérez, Z., Fernández-Palacios, J. M., Arévalo, J. R., ... & 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.

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., ... & Luque, S. (2022). Assessing the effect of sample bias correction in species distribution models. Ecological Indicators, 145, 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.

Duong, T. (2021). *ks: Kernel Smoothing*. R package version 1.13.3.

Fei, S. and 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.

Fick, S. E. and 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.

Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian Obama nungara. *Ecological Modelling*, *457*, 109686.

Fourcade, Y., Besnard, A. G., and 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.

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.

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.

Guisan, A., Thuiller, W., and 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., and Mackey, B. (2019). Species distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, 408:108719.

Hattab, T., Garzón-López, C. X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., Brasseur, B., Gallet-Moron, E., Spicher, F., Decocq, G., et al. (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.

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M., and Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117(6):847–858.

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., and Gutiérrez, J. M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312:166–174.

Jackson, S. T. and Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late quaternary. *Paleobiology*, 26(S4):194–

220.

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.

Jiménez-Valverde, A. (2021). Prevalence affects the evaluation of discrimination capacity in presence-absence species distribution models. *Biodiversity and Conservation*, *30*(5), 1331-1340.

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.

Jiménez-Valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M., and 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.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., and 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.

Leroy, B., Meynard, C. N., Bellard, C., and Courchamp, F. (2016). virtualspecies, an r package to generate virtual species distributions. *Ecography*, 39(6):599–607.

Lobo, J. M., Jiménez-Valverde, A., and Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114.

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145– 151.

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., ... & 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.

Mammola, S. and Cardoso, P. (2020). Functional diversity metrics using kernel density n-dimensional hypervolumes. *Methods in Ecology and Evolution*, 11(8):986–995.

Meynard, C. N., Leroy, B., and 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.

Naimi, B. and Araújo, M. B. (2016). sdm: a reproducible and extensible r platform for species distribution modelling. *Ecography*, 39(4):368–375.

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*, 285(1881):20180792.

Perret, D. L. and Sax, D. F. (2022). Evaluating alternative study designs for optimal sampling of species' climatic niches. *Ecography*.

Poli et al. (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

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., and Blair, M. E. (2017). Opening the black box: An open-source release of maxent. *Ecography*, 40(7):887–893.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19(1):181–197.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ronquillo, C., Alves-Martins, F., Mazimpaka, V., Sobral-Souza, T., Vilela-Silva, B., Medina, N. G., and Hortal, J. (2020). Assessing spatial and temporal biases and gaps in the publicly available distributional information of iberian mosses. *Biodiversity Data Journal*, 8.

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

Sillero, N. and Barbosa, A. M. (2020). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, pages 1–14.

Støa, B., Halvorsen, R., Stokland, J. N., and 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.

Svenning, J.-C. and Skov, F. (2004). Limited filling of the potential range in European tree species. *Ecology Letters*, 7(7):565–573.

Tessarolo, G., Lobo, J. M., Rangel, T. F., and Hortal, J. (2021). High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, 121:107147.

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

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., and Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37(11):1084–1091.

Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Wasof et al. (2015) Disjunct populations of European vascular plant species keep the same climatic niches. *Global Ecology and Biogeography*, 24: 1401-1412

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Supplementary Material 1

Tab. S1: Post-hoc multiple comparisons with Dunn's rank sum test (α = 0.05; omnibus test was always significant with P < 0.05, data not shown). All the comparisons were performed comparing the Uniform dataset with the other different sampling strategies. P-values were adjusted using Holm correction.

Model	Metric	Comparisons	χ2	P.val
BRT	AUC	Buffer IN - Uniform	146.28	p<0.001
BRT	AUC	Buffer OUT - Uniform	146.28	0.176
BRT	AUC	Random - Uniform	146.28	p<0.001
BRT	BoyceI	Buffer IN - Uniform	131.66	p<0.001
BRT	BoyceI	Buffer OUT - Uniform	131.66	0.0356
BRT	BoyceI	Random - Uniform	131.66	0.009
BRT	Sensitivity	Buffer IN - Uniform	104.47	p<0.001
BRT	Sensitivity	Buffer OUT - Uniform	104.47	p<0.001
BRT	Sensitivity	Random - Uniform	104.47	p<0.001
BRT	Specificity	Buffer IN - Uniform	76.62	p<0.001
BRT	Specificity	Buffer OUT - Uniform	76.62	0.016
BRT	Specificity	Random - Uniform	76.62	p<0.001
BRT	TSS	Buffer IN - Uniform	150.78	p<0.001
BRT	TSS	Buffer OUT - Uniform	150.78	0.0087
BRT	TSS	Random - Uniform	150.78	p<0.001
GAM	AUC	Buffer IN - Uniform	141.11	p<0.001
GAM	AUC	Buffer OUT - Uniform	141.11	0.0336
GAM	AUC	Random - Uniform	141.11	p<0.001
GAM	BoyceI	Buffer IN - Uniform	144.02	p<0.001
GAM	BoyceI	Buffer OUT - Uniform	144.02	0.0044
GAM	BoyceI	Random - Uniform	144.02	0.0033
GAM	Sensitivity	Buffer IN - Uniform	131.32	p<0.001
GAM	Sensitivity	Buffer OUT - Uniform	131.32	p<0.001
GAM	Sensitivity	Random - Uniform	131.32	p<0.001

Model	Metric	Comparisons	χ2	P.val
GAM	Specificity	Buffer IN - Uniform	128.72	p<0.001
GAM	Specificity	Buffer OUT - Uniform	128.72	0.1586
GAM	Specificity	Random - Uniform	128.72	p<0.001
GAM	TSS	Buffer IN - Uniform	145.45	p<0.001
GAM	TSS	Buffer OUT - Uniform	145.45	0.0028
GAM	TSS	Random - Uniform	145.45	p<0.001
GLM	AUC	Buffer IN - Uniform	132.53	p<0.001
GLM	AUC	Buffer OUT - Uniform	132.53	0.0822
GLM	AUC	Random - Uniform	132.53	0.003
GLM	BoyceI	Buffer IN - Uniform	175.57	p<0.001
GLM	BoyceI	Buffer OUT - Uniform	175.57	p<0.001
GLM	BoyceI	Random - Uniform	175.57	p<0.001
GLM	Sensitivity	Buffer IN - Uniform	128.02	p<0.001
GLM	Sensitivity	Buffer OUT - Uniform	128.02	p<0.001
GLM	Sensitivity	Random - Uniform	128.02	p<0.001
GLM	Specificity	Buffer IN - Uniform	98.02	p<0.001
GLM	Specificity	Buffer OUT - Uniform	98.02	p<0.001
GLM	Specificity	Random - Uniform	98.02	0.1366
GLM	TSS	Buffer IN - Uniform	141.06	p<0.001
GLM	TSS	Buffer OUT - Uniform	141.06	0.0333
GLM	TSS	Random - Uniform	141.06	p<0.001
Maxent	AUC	Buffer IN - Uniform	151.46	p<0.001
Maxent	AUC	Buffer OUT - Uniform	151.46	0.0099
Maxent	AUC	Random - Uniform	151.46	p<0.001
Maxent	BoyceI	Buffer IN - Uniform	178.36	p<0.001
Maxent	BoyceI	Buffer OUT - Uniform	178.36	p<0.001
Maxent	BoyceI	Random - Uniform	178.36	p<0.001
Maxent	Sensitivity	Buffer IN - Uniform	64.45	p<0.001
Maxent	Sensitivity	Buffer OUT - Uniform	64.45	0.0677

Model	Metric	Comparisons	χ2	P.val
Maxent	Sensitivity	Random - Uniform	64.45	0.0099
Maxent	Specificity	Buffer IN - Uniform	66.81	p<0.001
Maxent	Specificity	Buffer OUT - Uniform	66.81	0.046
Maxent	Specificity	Random - Uniform	66.81	0.0035
Maxent	TSS	Buffer IN - Uniform	151.49	p<0.001
Maxent	TSS	Buffer OUT - Uniform	151.49	0.0098
Maxent	TSS	Random - Uniform	151.49	p<0.001
RF	AUC	Buffer IN - Uniform	147.3	p<0.001
RF	AUC	Buffer OUT - Uniform	147.3	0.0747
RF	AUC	Random - Uniform	147.3	p<0.001
RF	BoyceI	Buffer IN - Uniform	166.26	p<0.001
RF	BoyceI	Buffer OUT - Uniform	166.26	0.1462
RF	BoyceI	Random - Uniform	166.26	p<0.001
RF	Sensitivity	Buffer IN - Uniform	89.75	p<0.001
RF	Sensitivity	Buffer OUT - Uniform	89.75	p<0.001
RF	Sensitivity	Random - Uniform	89.75	0.1444
RF	Specificity	Buffer IN - Uniform	108.22	p<0.001
RF	Specificity	Buffer OUT - Uniform	108.22	p<0.001
RF	Specificity	Random - Uniform	108.22	p<0.001
RF	TSS	Buffer IN - Uniform	147.11	p<0.001
RF	TSS	Buffer OUT - Uniform	147.11	0.079
RF	TSS	Random - Uniform	147.11	p<0.001

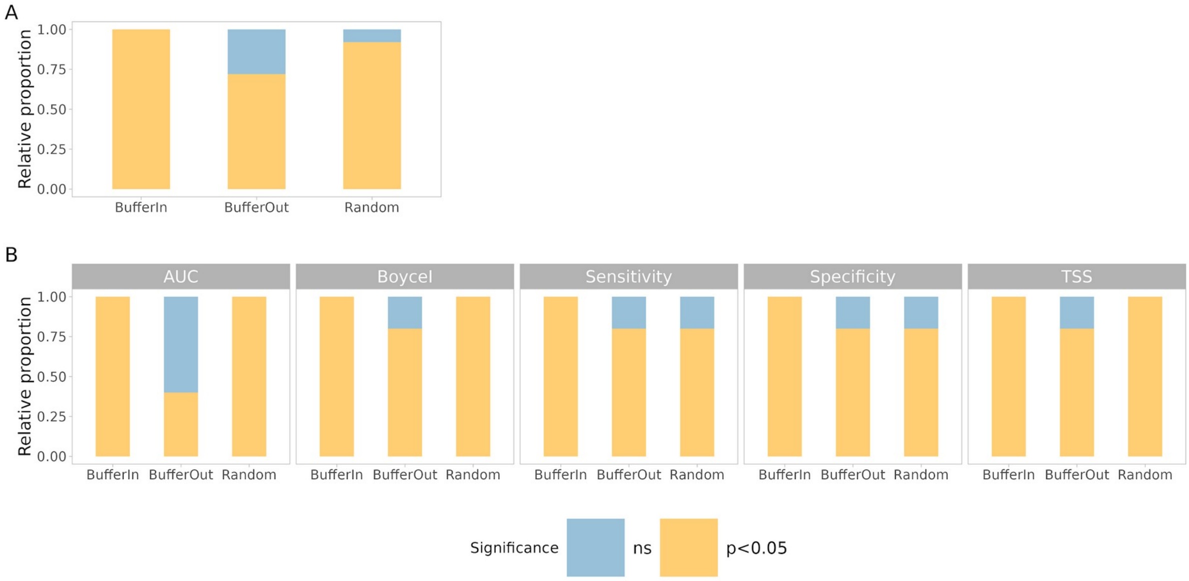


Figure S1.1: Post-hoc multiple comparisons with Dunn’s rank sum test (α = 0.05; omnibus test was always significant with P < 0.05, data not shown). All the comparisons were performed comparing the Uniform dataset with the other different sampling strategies: A) relative proportion of the significant comparisons aggregated by sampling strategy; B) relative proportion of the significant comparisons aggregated by sampling strategy and metric.

Supplementary Material 2

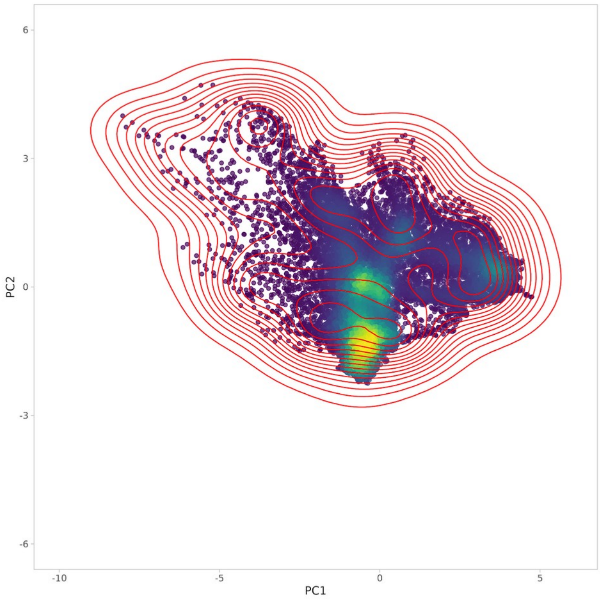


Figure S2.1: Bivariate plots showing the environmental space generated by a PCA carried out on 5 bioclimatic variables. Red lines represent the density of pseudo-absences within the environmental space for an individual virtual species and for the uniform approach only.

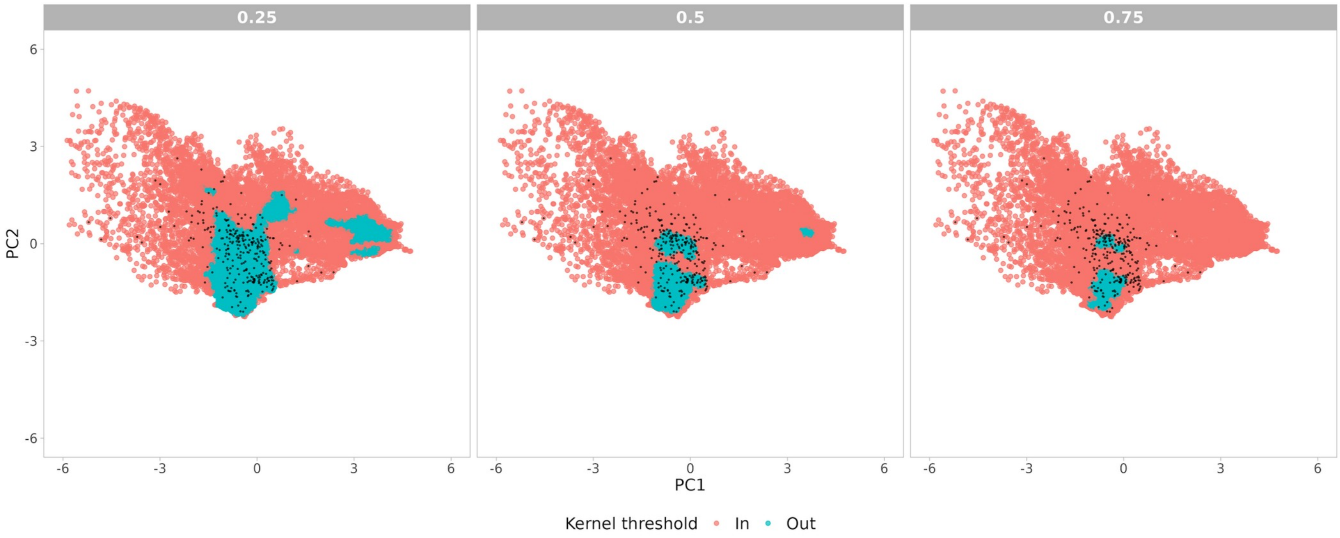


Figure S2.2: Effect of the kernel threshold in the inclusion/exclusion of the environmental space to sample. Black dots are the real VS occurrences plotted in the environmental space.

Supplementary Material 3

Sensitivity analyses on the sample prevalence

To test the potential effect of different sample prevalence, we also repeated the entire workflow on 10 VS with two different prevalence values. Specifically, in both cases we kept a training dataset consisting of 300 presences but we used alternatively 600 and 3,000 pseudo-absences (sample prevalence = 0.5 and sample prevalence = 0.1, respectively).

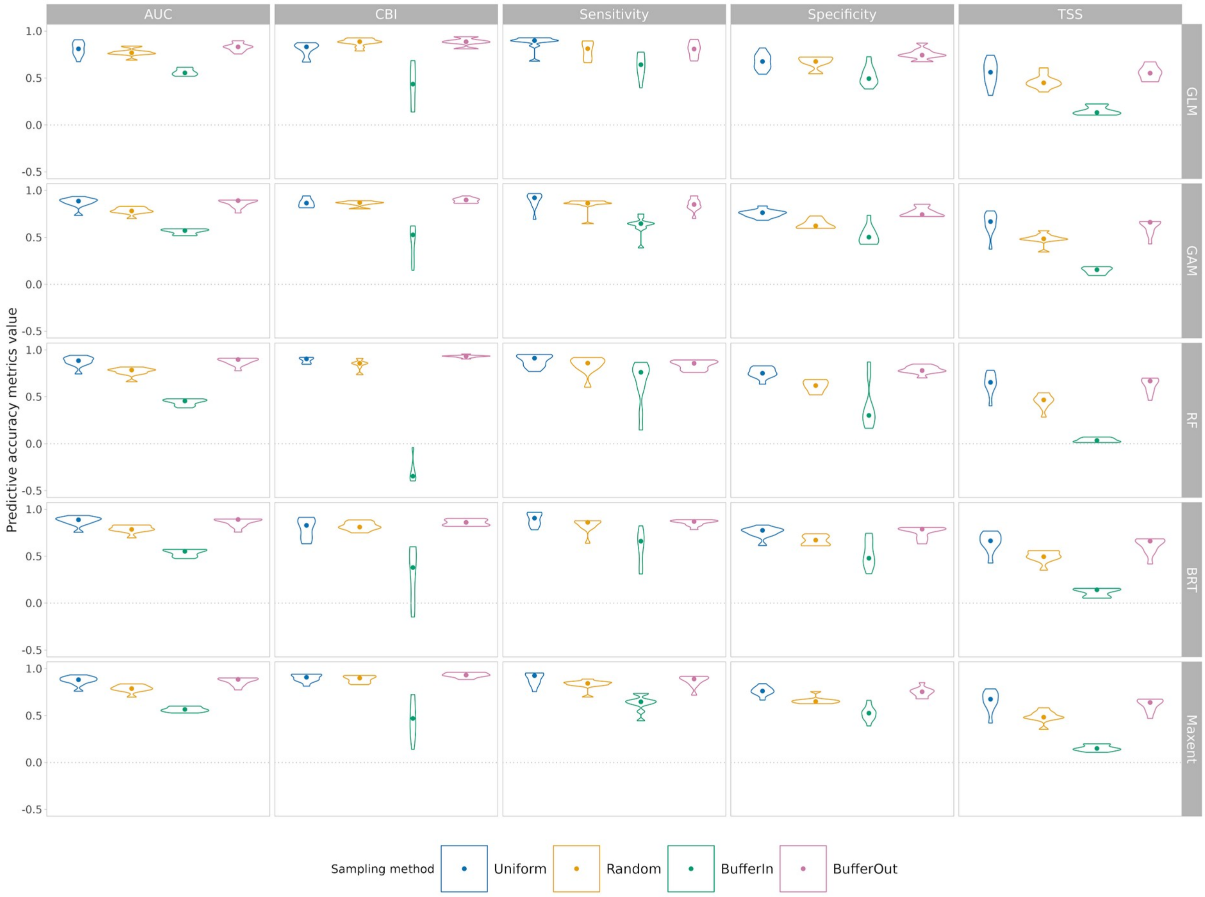


Figure S3.1: Violin plots reporting the distribution of the values of the metrics of predictive performance for the HSMs of the 50 VS modelled as a function of 5, randomly selected bioclimatic predictors, and setting sample prevalence equal to 0.5. Dots represent median values of the metrics of predictive accuracy. Columns indicate the different performance metrics, while rows the modelling techniques used to compute HSMs.

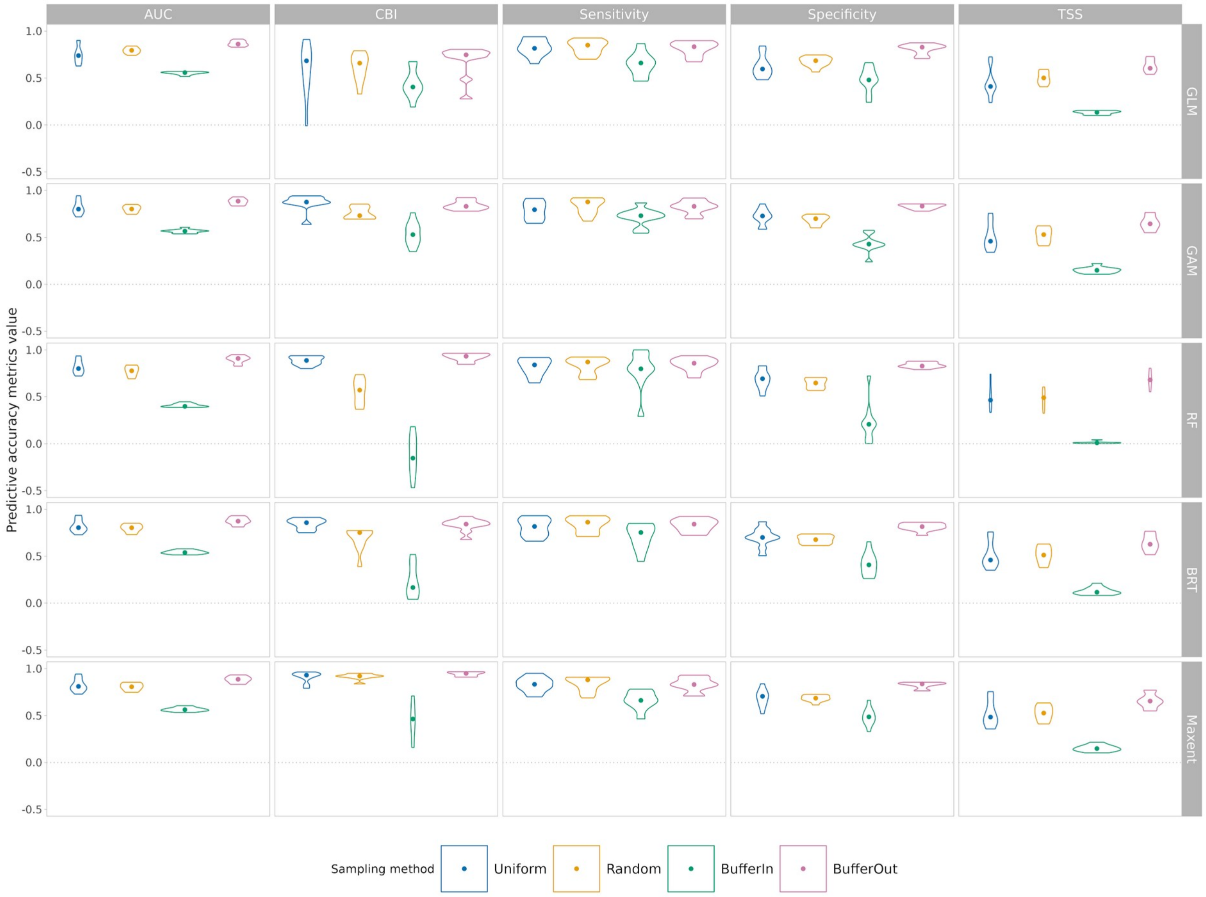


Figure S3.2: Violin plots reporting the distribution of the values of the metrics of predictive performance for the HSMs of the 50 VS modelled as a function of 5, randomly selected bioclimatic predictors, and setting sample prevalence equal to 0.1. Dots represent median values of the metrics of predictive accuracy. Columns indicate the different performance metrics, while rows the modelling techniques used to compute HSMs.

Supplementary Material 4: case study on the realized distribution of *Fagus sylvatica* in Western Europe

Methods

To illustrate how to apply the uniform approach using the `USE` R package, we modelled the realised distribution of *Fagus sylvatica* in Italy, France and Spain (hereafter, western Europe). We chose *F. sylvatica* as an example species because its distribution and biogeographic history is well-known across Europe (Magri et al., 2006; Poli et al., 2022). For the sake of simplicity, we restricted the area of investigation to western Europe and used only two modelling algorithms (i.e., GLM and RF). Indeed, the case study of *F. sylvatica* is only used as a practical example on how to use the `USE` package, while not providing a further comparison of the predictive performance of HSMs fitted on data collected through different sampling strategies (as already done with the VS approach). We gathered data on the presence-absence of *F. sylvatica* from the open EU-Forest dataset (Mauri et al., 2017), which compiles presence data on European tree species from national inventories and other similar sources (see Mauri et al., 2017 for further information about EU-Forest). EU-Forest data consist of presence-absence records of tree species exhaustively collected across Europe, and then aggregated to a 1 × 1 km resolution grid. This let us assume with a certain degree of confidence that the EU-Forest dataset provided a geographically unbiased sample of occurrence records for *F. sylvatica*, and absence data represented ‘true’ absences.

Across our study area, the EU-Forest dataset provided a total of 12,444 presence records for *F. sylvatica*, which we sub-sampled within the environmental space to retrieve both a training and a testing presence (for internal validation) dataset. To this aim, we generated a 2-dimensional environmental space using all 19 bioclimatic variables available from WorldClim. Then, we used the function `uniformSampling` from the `USE` package to uniformly sample occurrence records within the environmental space. Note that this approach is conceptually similar to the spatial-thinning approach proposed by Aiello-Lammens et al. (2015), which aims at reducing the clustering of presences within the geographical space (Sillero and Barbosa, 2020), except that we here applied it uniformly and within the environmental space (see Varela et al., 2014;

Hattab et al. 2017). Once presence records (n = 2,747) were uniformly sub-sampled within the environmental space, we randomly divided them into two sets of training (70%) and testing (30%) occurrences to then derive the two respective sets of training and testing pseudo-absences. To this aim, we first used all 12,444 available presence records to recover the core area of *F. sylvatica*'s bioclimatic niche using the function `paSampling` from the `USE` package and then filtered out the background points likely associated with suitable locations for the species (see step 1 in section 2.2.1 of the main text). Once we removed background points likely associated with the core bioclimatic niche of *F. sylvatica*, the obtained sample sizes were: 1,856 and 906 background points for the training and testing (internal validation) dataset, respectively. Finally, we derived a completely independent testing (external validation) dataset using presence and true absence data from sPlotOpen (Sabatini et al., 2021). The sPlotOpen database is an open-access subset of sPlot, one of the most comprehensive global databases of vegetation records (Sabatini et al., 2021). Here, we used sPlotOpen to gather *F. sylvatica* presences (*n* = 366), and also to derive true absence data from those vegetation plots where *F. sylvatica* was not recorded (*n* = 4039). As done for the EU-Forest dataset, we selected only those vegetation plots data from sPlotOpen included in our study area (Italy, France and Spain) in western Europe.

The realised distribution of *F. sylvatica* was modelled as a function of WorldClim bioclimatic variables (resolution of 2.5 minutes at the Equator). For simplicity, we solely focused on the climatic niche of *Fagus sylvatica*, although we acknowledge that other factors different from climatic drivers may equally contribute in shaping the distribution of this species, especially so at local scales (Mellert et al., 2018). As modelling techniques, we used binary generalised linear models and random forests (ranger function available in ranger R package; Wright and Ziegler, 2017). To avoid multicollinearity issues, we selected a subset of the 19 bioclimatic variables using the findCorrelation function from the caret R package (Kuhn, 2021) (pairwise-correlation threshold: 0.6). The bioclimatic variables finally selected for *F. sylvatica* were: BIO6 (minimum temperature of the coldest month); BIO7 (temperature annual range); and BIO8 (mean temperature of the wettest quarter). We also used the latitudinal position of the presence and pseudo-absence records as an additional predictor to account for the effect of factors affecting *F. sylvatica* that correlates with its latitudinal gradient of occurrence and were not included in the model, such as its biogeographic history of post-glacial recolonization towards northern Europe (Magri et al., 2006). To account for non-linearity in the profile of Pearson’s residuals and improve the fit of the binary GLM, we introduced second order polynomial terms for BIO6, BIO7 and latitude. Statistically non-significant predictors were dropped from the original full model to reach a most parsimonious model. The predictive performance of the fitted models was assessed using TSS and CBI on three different types of data: (i) the testing dataset derived from the EU-Forest dataset; (ii) 5 partitions of the training dataset (i.e., a 5-fold cross-validation); and (iii) the independent testing dataset derived from sPlotOpen. We often assume TSS>0.5 to indicate good predictions, while CBI positive values indicate a model which present predictions are consistent with the distribution of presences in the evaluation dataset, values close to zero mean that the model is not different from a random model, negative values indicate counter predictions, i.e., predicting poor quality areas where presences are more frequent (Hirzel et al. 2006).

Beyond model predictive metrics, we computed the following measures of goodness-of-fit: Tjur’s R^2 for the binary GLM and the R^2 for the RF.

A full description of the modelling procedure (from the sub-sampling of the presence and background points to the assessment of the model predictive performance) is reported at https://github.com/danddr/USE_paper.

Results

Both the GLM and the RF for *F. sylvatica* showed high predictive performances, regardless of the dataset used for testing (Table 1). Concerning the GLM, the TSS was always equal to or above 0.40, with the lowest value obtained for the sPlotOpen testing dataset (0.40) and the highest for the EU-Forest dataset (0.6). The lowest CBI was scored for the sPlotOpen dataset (0.88), while the highest for the EU-Forest dataset (0.98).

Similar results were obtained for the RF, for which the lowest TSS was obtained for the sPlotOpen testing dataset (0.48), while comparable values resulted from the EU-Forest dataset and the 5-fold cross validation (0.79 and 0.76, respectively). With respect to the CBI, the highest value was observed for the EU-Forest dataset (average = 0.98), while the lowest was obtained for the sPlotOpen dataset (0.96).

Goodness-of-fit measures seemed to be affected by the modelling technique, with the R^2 of the RF being 0.67, and the Tjur's R^2 for the GLM being 0.35 (Tab. S4.1).

The pseudo-absences of *F. sylvatica* collected using the uniform approach are uniformly distributed in the environmental space, reducing the sample location bias (Fig. S4.1a) and the class overlap, since in the geographical space they are distributed in areas far from the occurrences (Fig. S4.1b).

Table S4.1: Results of the two HSMs for *Fagus sylvatica* (GLM and RF). Models' predictive performance was assessed through internal (5-fold CV and EU Forest) and external (sPlotOpen) validation. TSS: True Skill Statistics; Boyce I: Boyce's Index; R-sq: Tjur's R^2 for the GLM, and R^2 for RF.

Validation dataset	**GLM**			**RF**		
	TSS	*CBI*	*Tjur's R^2*	*TSS*	*CBI*	*R^2*
5-fold CV	0.52	0.92		0.76	0.97	
EU-Forest	0.60	0.98	0.35	0.79	0.98	0.67
sPlotOpen	0.40	0.88		0.48	0.96	

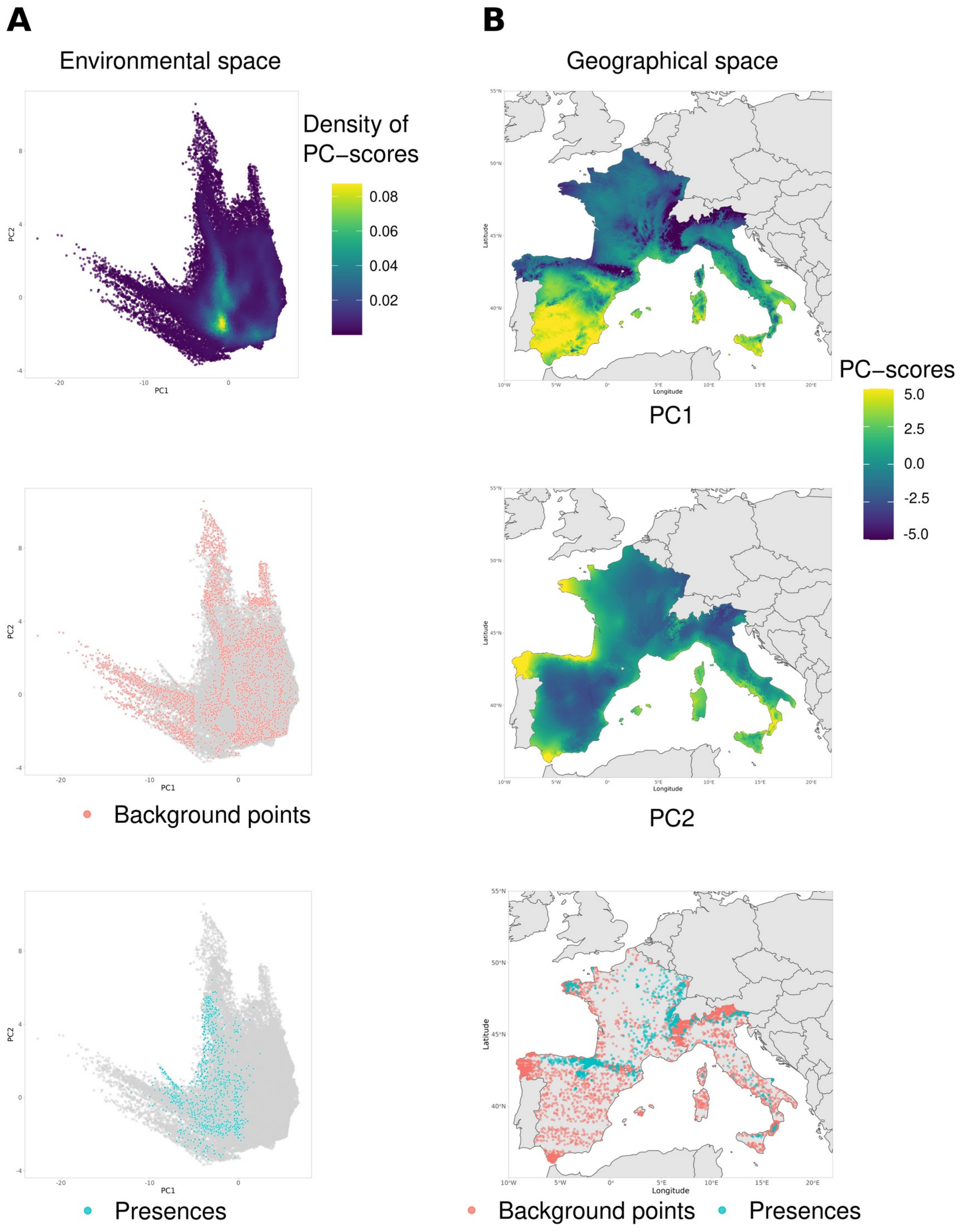


Figure S4.1: (A) the environmental space available for *Fagus sylvatica* in Italy, Spain and France, and the position of presences and pseudo-absences sampled within the environmental space using the Uniform approach; (B) the distribution of PC-scores in the geographical space and the geographical location of presences (blue) and pseudo-absences (red) sampled within the environmental space using the Uniform approach.

References

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). spthin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5):541–545.

Hattab, T., Garzón-López, C. X., Ewald, M., Skowronek, S., Aerts, R., Horen, H., Brasseur, B., Gallet-Moron, E., Spicher, F., Decocq, G., et al. (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.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological modelling*, 199(2), 142-152.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-88.

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., ... & 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.

Mauri, A., Strona, G., and San-Miguel-Ayanz, J. (2017). Eu-forest, a high-resolution tree occurrence dataset for europe. *Scientific data*, 4(1):1–8.

Mellert et al. (2018) Soil water storage appears to compensate for climatic aridity at the xeric margin of European tree species distribution. *European Journal of Forest Research*, 137: 79-92.

Poli et al. (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

Sabatini, F. M., Lenoir, J., Hattab, T., Arnst, E. A., Chytr`y, M., Dengler, J., De Ruffray, P., Hennekens, S. M., Jandt, U., Jansen, F., et al. (2021). splotopen–an environmentally balanced, open-access, global dataset of vegetation plots. *Global Ecology and Biogeography*.

Sillero, N. and Barbosa, A. M. (2020). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, pages 1–14.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37(11):1084–1091.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.