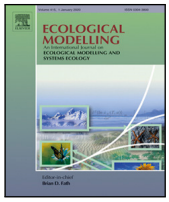
Ecological Modelling 477 (2023) 110248

Contents lists available at ScienceDirect

Ecological Modelling

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



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

Elisa Marchetto [a,*], Daniele Da Re [b], Enrico Tordoni [c], Manuele Bazzichetto [d,e], Piero Zannini [a,g], Simone Celebrin [a], Ludovico Chieffallo [a], Marco Malavasi [e,f], Duccio Rocchini [a,e]

[a] BIOME Lab, Department of Biological, Geological and Environmental Sciences (BiGeA), Alma Mater Studiorum University of Bologna, Bologna, Italy
[b] George Lemaître Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, Belgium
[c] Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
[d] Department of Ecology and Global Change, Desertification Research Centre (CSIC/UV/GV), Valencia, Spain
[e] Department of Spatial Sciences, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha - Suchdol, Czech Republic
[f] Department of Chemistry, Physics, Mathematics and Natural Sciences, University of Sassari, Sassari, Italy
[g] LifeWatch Italy, Lecce, Italy

ARTICLE INFO

Keywords:
Biodiversity
Ecological informatics
Spatial bias
Spatial ecology
Species distribution modelling

ABSTRACT

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

1. Introduction

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

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

* Correspondence to: BIOME Lab, Department of Biological, Geological and Environmental Sciences (BiGeA), Alma Mater Studiorum University of Bologna, Piazza di Porta S. Donato, 1, 40126 Bologna, Italy.
E-mail address: elisa.marchetto5@unibo.it (E. Marchetto).

https://doi.org/10.1016/j.ecolmodel.2022.110248
Received 24 June 2022; Received in revised form 29 November 2022; Accepted 8 December 2022

of species response. Favourability can be then calculated as follows:

$$F = \frac{\frac{P}{(1-P)}}{\frac{n_1}{n_0} + \frac{P}{(1-P)}} \quad (1)$$

being P the probability and n_1 and n_0 the respective number of presences and absences sampled where the ratio is defined as sample prevalence.

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

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

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

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

2. Materials and methods

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

2.1. Generating virtual species

In order to compare favourability-based and probability-based SDMs we used virtual species that were created by the `virtualspecies` R package (Leroy et al., 2016). We derived a virtual species using a subset of the WorldClim Bioclimatic variables at the European extent. We used the `generateRandomSp` function to create the environmental suitability for the virtual species distribution which was generated from a random subsampling (5 replicates) of the 19 bioclimatic variables (https://www.worldclim.org/data/bioclim.html) with 10 arc-minutes of spatial resolution. The environmental suitability was calculated using an additive approach to the response functions of each bioclimatic variable, where the possible types of response function implemented are "gaussian", "linear", "logistic" and "quadratic". The obtained environmental suitability was then rescaled between 0 and 1 (i.e., range of possible probability values of the virtual species distribution). We used the `convertToPA` function to convert the raster layer reporting the environmental suitability into a probability of occurrence; the weighted probability of occurrence was then used to sample the presence or absence in each cell. We transformed the environmental suitability with a logistic conversion setting α and β parameters that determine the shape of the logistic curve (Meynard and Kaplan, 2013). β controls the inflexion point and α drives the 'slope' of the curve, the latter was set equal to −0.05 such that the function detects an appropriate conversion by testing different values of β ; the species prevalence, i.e., the proportion of sites occupied by the species (Meynard and Kaplan, 2012), was fixed at 0.2.

2.2. Sampling methods

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

2.3. Models settings

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

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

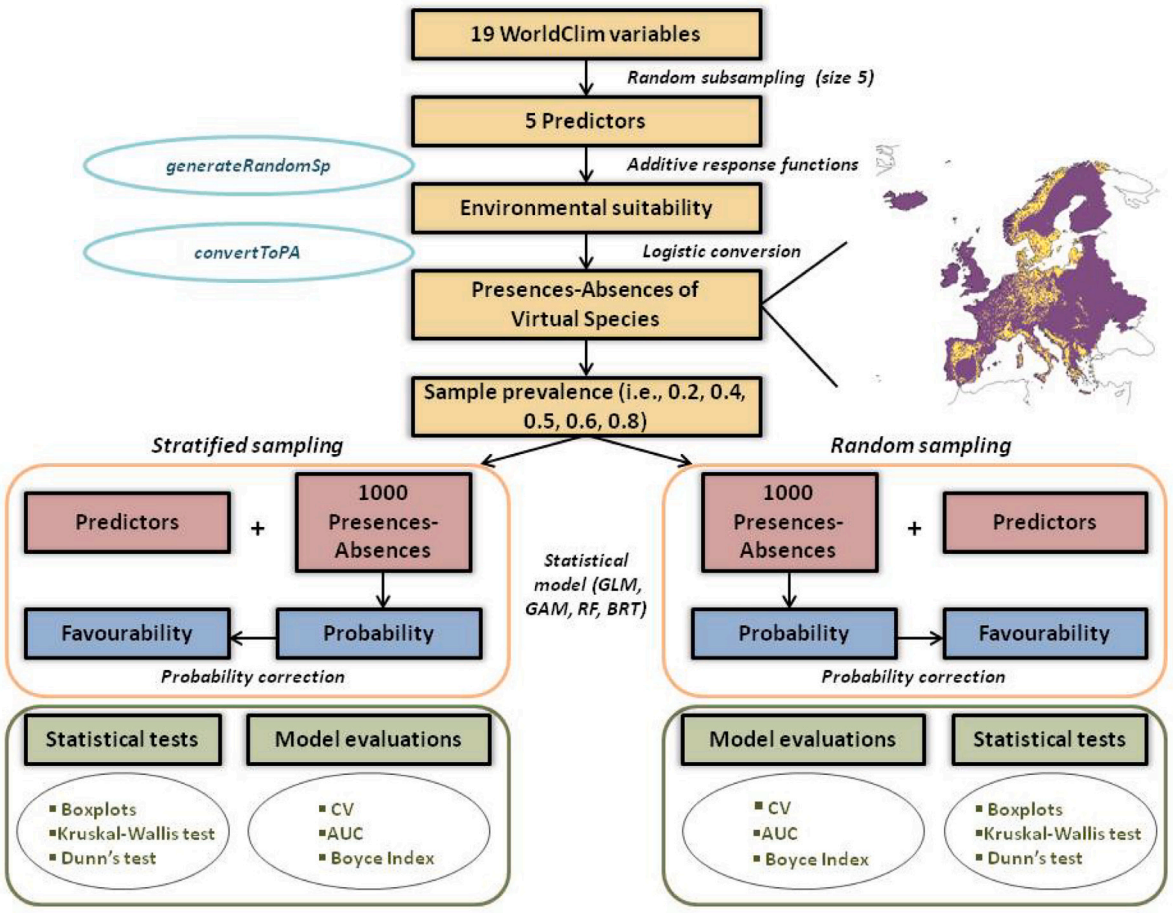


Fig. 1. Workflow methodology for a single virtual species. First, we generated the virtual species (yellow boxes), then we fitted the statistical models in accordance with the sampling method and the sample prevalence being used to derive probability-based and favourability-based SDMs (orange square).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

2.4. Models evaluation

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

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

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

2.5. Statistical tests

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

3. Results

3.1. Models performance evaluation

For more than half of the sample prevalences, the favourability model had slightly higher median Continuous Boyce index values (i.e., better performances) than the probability model for all of the statistical models and for both sampling methods, except for RF trained using a random sampling of presences and absences. Especially, GLM had higher performances using the favourability-based approach for both the sampling methods and for all of the sample prevalences. Overall, the sampling method (i.e., random and stratified) did not have a great impact on the model performances Fig. 2.

Furthermore, for all probability-based SDMs over the set of sample prevalence, calibrated using both the random sampling method and the stratified sampling method, the model performances, estimated with the Area Under the Curve (AUC) of the receiver operator characteristic

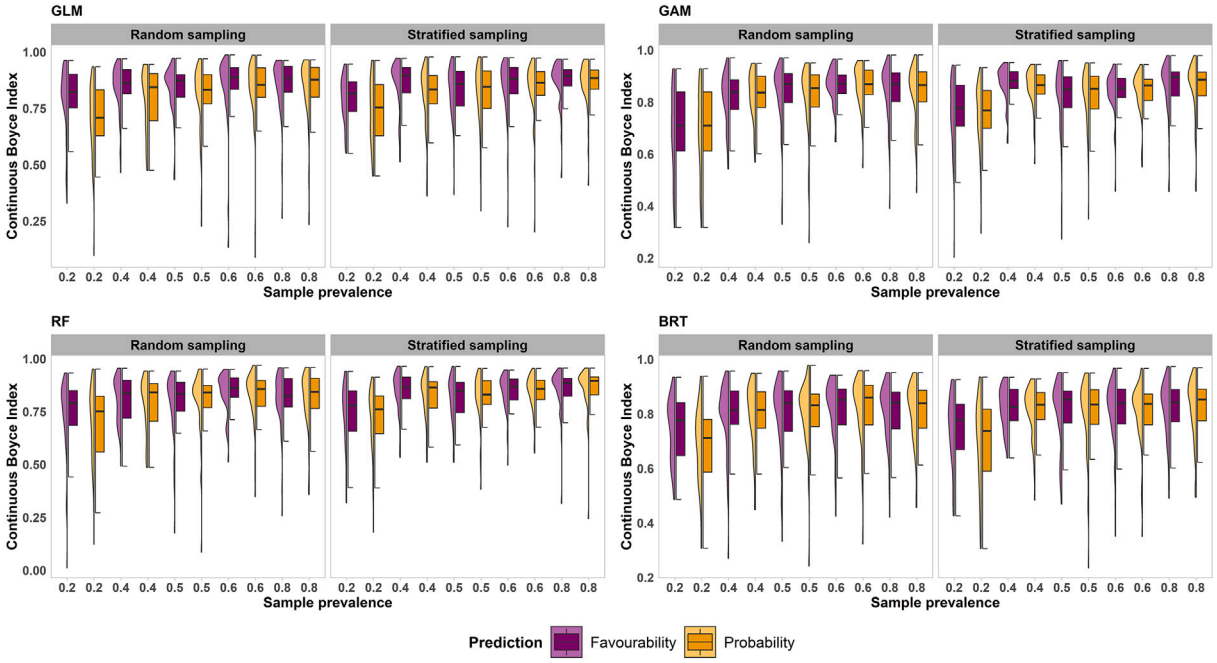


Fig. 2. Distribution between first and third quartiles of continuous Boyce index values of favourability-based and probability-based SDMs of 50 virtual species. The graphs show the distribution values of Continuous Boyce indices estimated by applying Generalized Linear Models, Generalized Additive Models,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

(ROC), had a good accuracy ranging between 0.80 and 0.95 (Appendix, Fig. S6–S9).

3.2. Effect of the sample prevalence on the predictions

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

Besides, the variability of the predictions for a single species – i.e., the pixels variability – calculated as Coefficient of Variation across the sample prevalence values, showed higher stability (i.e., lower CV) for the favourability-based SDM both for the random sampling and for the stratified sampling Figs. 3 and 4. Moreover, the difference between the pixels variability of the probability predictions and the pixels variability of the favourability predictions confirmed that the favourability SDM generates higher pixels stability as the sample prevalence changes. However, the generalized linear model showed a larger decrease in the pixels variability once the sample prevalence was removed from the probability predicted values Fig. 5.

3.3. Effect of the sampling method on the predictions

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

4. Discussion

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

Concerning models' performance, the Continuous Boyce index did not show a great difference in the performance efficiency between favourability and probability models. This behaviour could depend on the fact that the models have been calibrated with the default parameters in order to be extended to 50 different species. Indeed, several authors showed that the model parametrization has an impact on SDMs output (e.g., Fourcade, 2021). However, for more than half of sample prevalences we considered, median Continuous Boyce Index values were slightly higher for favourability-based SDMs than for probability-based SDMs. Besides, although van Proosdij et al. (2016) and Tessarolo et al. (2021) report a linear relationship between the model performance and the sample prevalence, our outcomes of AUCs indicate an independence of the accuracy of predictive models with respect to prevalence values (Guo et al., 2015).

Concerning the spatial variability of predictions of a single virtual species, the pixels variability across the degrees of sample prevalence was lower for the favourability-based SDMs than for the probability-based. By comparing the distribution values of 50 virtual species' predictions this pattern was also retained; favourability-based predictions showed steadier values across different sample prevalences than the probability-based predictions, although they retain a certain degree of variability. Indeed, the Kruskal Wallis rank sum tests highlighted a difference in the values of both probability and favourability predictions across the different degrees of sample prevalence. Hence, favourability-based SDMs do not maintain unchanged the prediction values across different sample prevalence values (Real et al., 2006; Acevedo et al., 2010; Romero et al., 2019), since they are created with a posteriori removal of the sample prevalence after statistical model calibration, so that the correction is made on the probability predictions. Consequently, the favourability model does not lose any information about sample prevalence and, therefore, about species or species-environment interactions, since the statistical model is still dependent on the prevalence value. Furthermore, it allows obtaining more

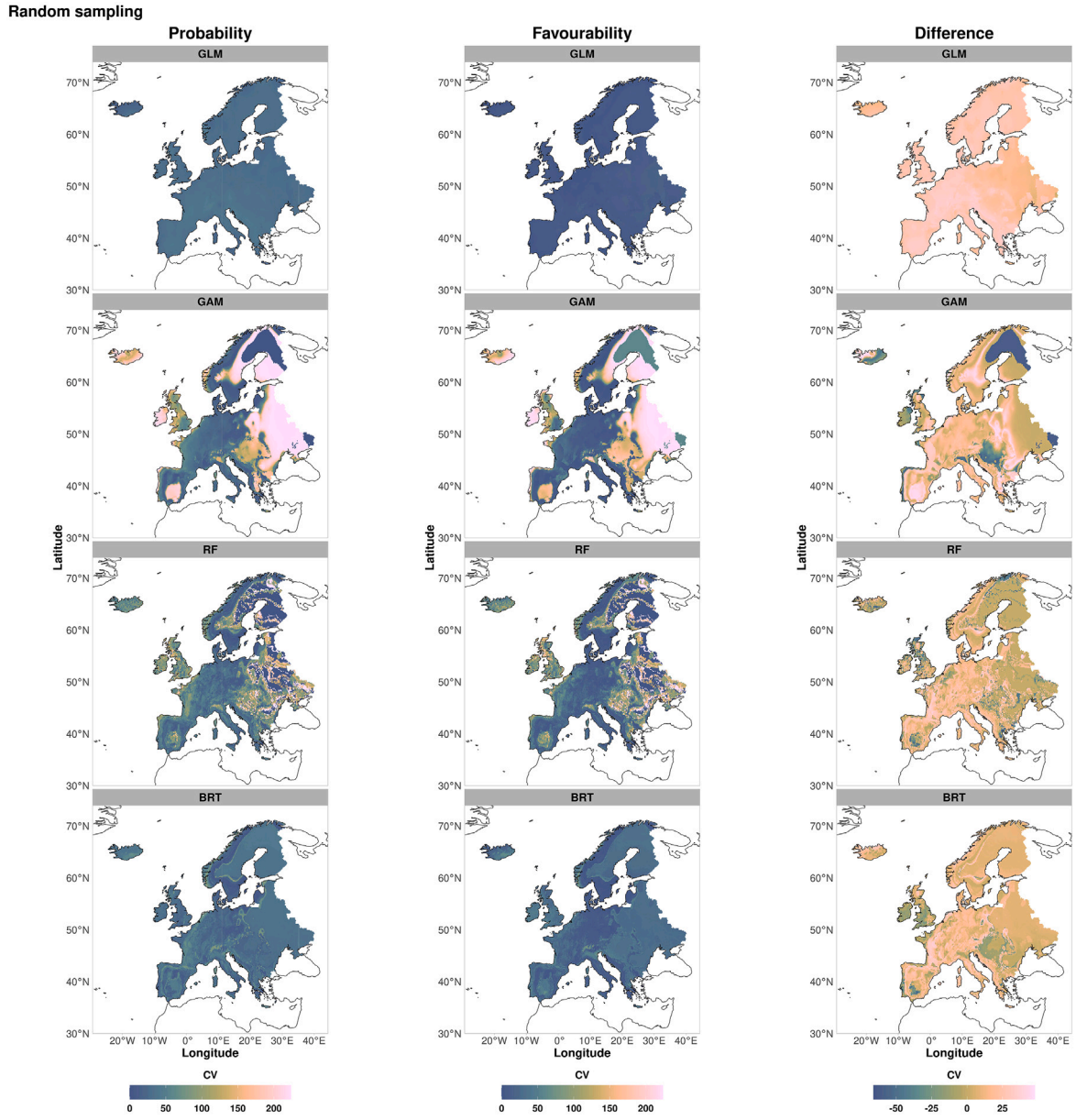


Fig. 3. Predictions variability, i.e., pixels variability, of a single virtual species calculated as Coefficient of Variation (CV) of favourability and probability SDMs related to a random sampling of presence–absence points. The left column shows the probability-based coefficients of variation for each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

effective comparisons among SDMs of different species or populations and time scales as a consequence of a lower pixels variability. This makes the favourability an extremely powerful tool to broaden our understanding of ecological trends such as the ecological niches pattern between species (Pulido-Pastor et al., 2021), the environmental factors that favour the spread of an invasive species (Romero et al., 2014; Baquero et al., 2021) or an epidemiological vector (Aliaga-Samanez et al., 2021), the areal shift range under land and climate changes (Muñoz et al., 2005; Chamorro et al., 2020).

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

According to our results, the sampling method of presences and absences does not have a decisive impact on the predictions variability

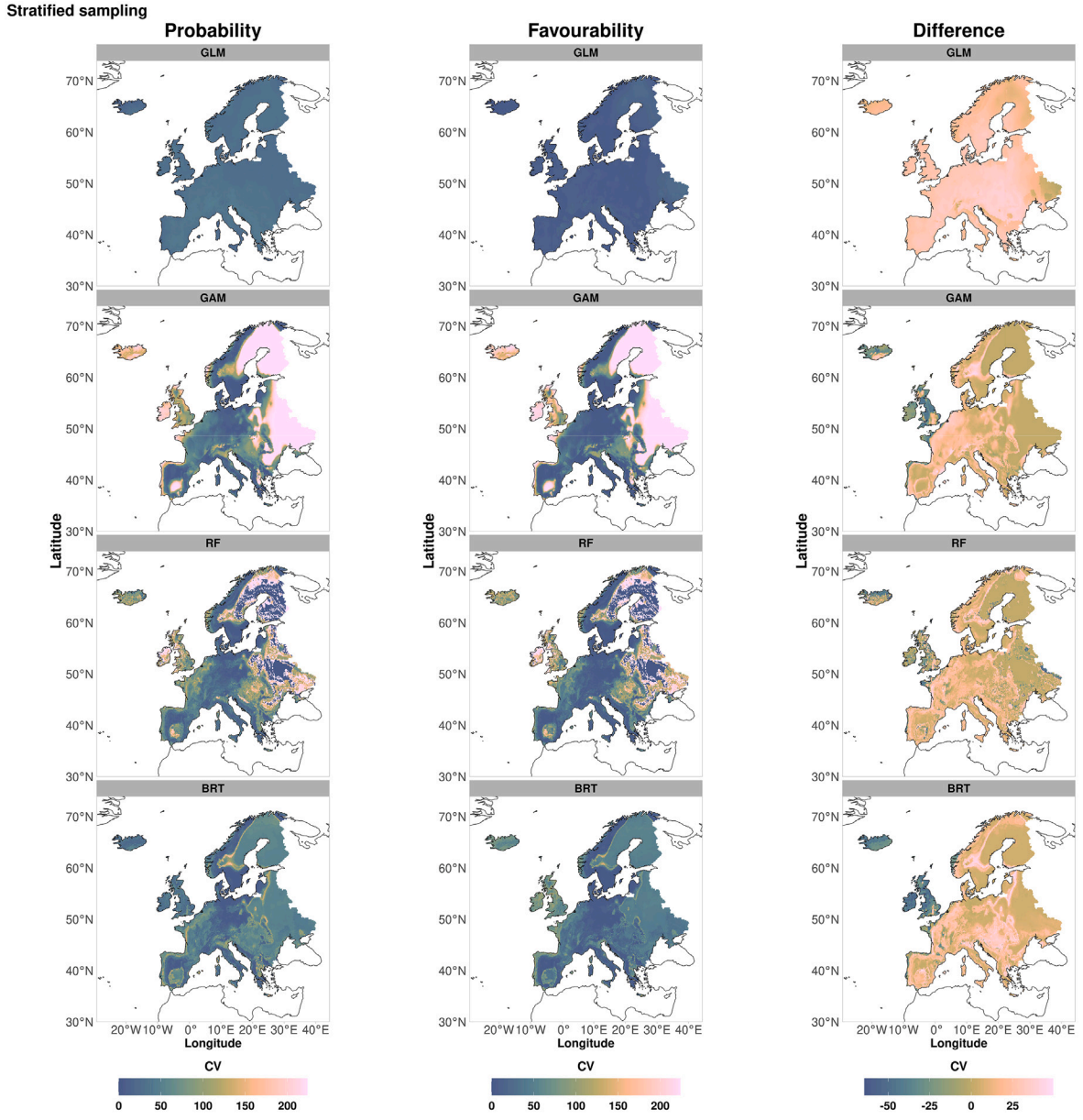


Fig. 4. Predictions variability, i.e. pixels variability, of a single virtual species calculated as Coefficient of Variation (CV) of favourability and probability SDMs related to a stratified sampling of presence–absence points. The left column shows the probability-based coefficients of variation for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

of favourabiliy-based and probability-based SDMs across the degrees of sample prevalences. However, the random sampling determined a greater uniformity around a narrower range of values (Appendix, Fig. S2–S5). Besides, Dunn's test confirmed that the sampling methods generate different prediction values at the spatial scale.

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

5. Conclusion

Favourability might provide an important contribution to map species distribution, especially for the fact that training datasets are often biased, as in the case of rare species of important conservation value. Indeed, favourability-based SDM, although it does not maintain unchanged the predictions values across different sample prevalences, proved to be effective in reducing their variability across different prevalence degrees compared to probability-based SDM. Besides, favourability model showed high model performances for all

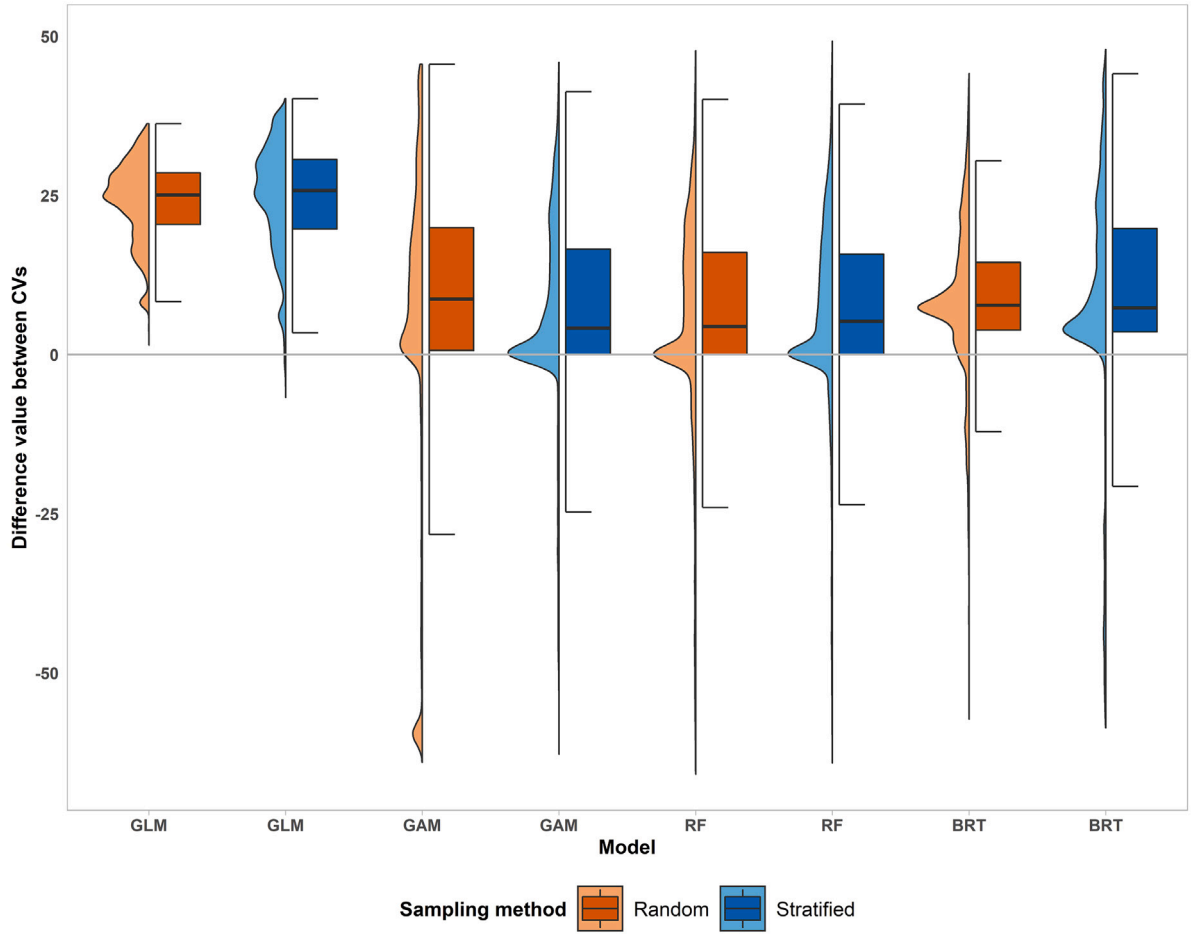


Fig. 5. Distribution values between first and third quartiles of the difference between probability-based and favourability-based coefficients of variation for each statistical model (i.e., GLM, GAM, RF, BRT) and sampling method (i.e., random and stratified sampling) for a single virtual species.

of the modelling techniques being applied. Therefore, according to our results, favourability-based SDM can definitely improve knowledge in community and population dynamics and provide useful tools for biogeography conservation allowing to achieve more effective comparisons among species distributions in space and their possible shifts over time. Nevertheless, being aware that the favourability is not independent of the uncertainty related to the sampling effort, the extent and the resolution of analysis, in a future study, these elements should also be considered. In our study, the sampling methods, i.e., random and stratified, revealed that they have a great impact neither in the variability of the predictions across the set of sample prevalence values nor in the performance of the models if no source of bias is present in the sampling. However, having proved the advantages of favourability-based SDM with virtual species, future studies on real species distribution models can be definitively promising in testing the real empirical power of favourability-based approach.

CRediT authorship contribution statement

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

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecolmodel.2022.110248.

References

Acevedo, P., Real, R., 2012. Favourability: concept distinctive characteristics and potential usefulness. Naturwissenschaften 99 (7), 515–522. http://dx.doi.org/10.1007/s00114-012-0926-0.

Acevedo, P., Ward, A.I., Real, R., Smith, G.C., 2010. Assessing biogeographical relationships of ecologically related species using favourability functions: a case study on British deer. Divers. Distrib. 16 (4), 515–528. http://dx.doi.org/10.1111/j.1472-4642.2010.00662.x.

Aliaga-Samanez, A., Cobos-Mayo, M., Real, R., Segura, M., Romero, D., Fa, J.E., Olivero, J., 2021. Worldwide dynamic biogeography of zoonotic and anthroponotic dengue. PLoS Negl. Trop. Dis. 15 (6), e0009496. http://dx.doi.org/10.1371/journal.pntd.0009496.

Baquero, R.A., Barbosa, A.M., Ayllón, D., Guerra, C., Sánchez, E., Araújo, M.B., Nicola, G.G., 2021. Potential distributions of invasive vertebrates in the iberian peninsula under projected changes in climate extreme events. Divers. Distrib. 27 (11), 2262–2276. http://dx.doi.org/10.1111/ddi.13401.

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

Barbosa, A.M., 2015. FuzzySim: applying fuzzy logic to binary similarity indices in ecology. Methods Ecol. Evolut. 6 (7), 853–858. http://dx.doi.org/10.1111/2041-210X.12372.

Barbosa, A.M., Real, R., Vargas, J.M., 2009. Transferability of environmental favourability models in geographic space: the case of the iberian desman (*Galemys pyrenaicus*) in Portugal and Spain. Ecol. Model. 220 (5), 747–754. http://dx.doi.org/10.1016/j.ecolmodel.2008.12.004.

Bazzichetto, M., Lenoir, J., Re, D.Da., Tordoni, E., Rocchini, D., Malavasi, M., Barták, V., Sperandii, M.G., 2022. Effect of sampling strategies on the response curves estimated by plant species distribution models. EcoEvoRxiv http://dx.doi.org/10.32942/osf.io/rhys3, August 27.

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

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K., 2002. Evaluating resource selection functions. Ecol. Model. 157 (2–3), 281–300. http://dx.doi.org/10.1016/S0304-3800(02)00200-4.

Chamorro, D., Real, R., Muñoz, A.R., 2020. Fuzzy sets allow gaging the extent and rate of species range shift due to climate change. Sci. Rep. 10 (1), 1–14. http://dx.doi.org/10.1038/s41598-020-73509.

d'Amen, M., Dubuis, A., Fernandes, R.F., Pottier, J., Pellissier, L., Guisan, A., 2015. Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. J. Biogeogr. 42 (7), 1255–1266. http://dx.doi.org/10.1111/jbi.12485.

Dunn, O.J., 1964. Multiple comparisons using rank sums. Technometrics 6 (3), 241–252. http://dx.doi.org/10.1080/00401706.1964.10490181.

Elith, J., Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32 (1), 66–77. http://dx.doi.org/10.1111/j.1600-0587.2008.05505.x.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159.

Fourcade, Y., 2021. Fine-tuning niche models matters in invasion ecology a lesson from the land Planarian Obama Nungara. Ecol. Model. 457, 109686. http://dx.doi.org/10.1016/j.ecolmodel.2021.109686.

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. http://dx.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. Ecol. Model. 431, 109194. http://dx.doi.org/10.1016/j.ecolmodel.2020.109194.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., et al., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecol. Biogeogr. 24 (3), 276–292. http://dx.doi.org/10.1111/geb.12268.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009. http://dx.doi.org/10.1111/j.1461-0248.2005.00792.x.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge University Press.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2–3), 147–186. http://dx.doi.org/10.1016/S0304-3800(00)00354-9.

Guo, C., Lek, S., Ye, S., Li, W., Liu, J., Li, Z., 2015. Uncertainty in ensemble modelling of large-scale species distribution: effects from species characteristics and model techniques. Ecol. Model. 306, 67–75. http://dx.doi.org/10.1016/j.ecolmodel.2014.08.002.

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

Hirzel, A., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecol. Model. 157 (2–3), 331–341. http://dx.doi.org/10.1016/S0304-3800(02)00203-X.

Hirzel, A.H., Lay, G.Le., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. 199 (2), 142–152. http://dx.doi.org/10.1016/j.ecolmodel.2006.05.017.

Hortal, J., Bello, F.de., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M., Ladle, R.J., 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu. Rev. Ecol. Evol. Syst. 46, 523–549. http://dx.doi.org/10.1146/annurev-ecolsys-112414-054400.

Inman, R., Franklin, J., Esque, T., Nussear, K., 2021. Comparing sample bias correction methods for species distribution modeling using virtual species. Ecosphere 12, http://dx.doi.org/10.1002/ecs2.3422, e03422.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47 (260), 583–621. http://dx.doi.org/10.1080/01621459.1952.10483441.

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

Leitão, P.J., Moreira, F., Osborne, P.E., 2011. Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. Int. J. Geogr. Inf. Sci. 25 (3), 439–454. http://dx.doi.org/10.1080/13658816.2010.531020.

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

Lomolino, M.V., 2004. Conservation biogeography. Front. Biogeogr. New Direct. Geogr. Nat. 293.

Meynard, C.N., Kaplan, D.M., 2012. The effect of a gradual response to the environment on species distribution modeling performance. Ecography 35 (6), 499–509. http://dx.doi.org/10.1111/j.1600-0587.2011.07157.x.

Meynard, C.N., Kaplan, D.M., 2013. Using virtual species to study species distributions and model performance. J. Biogeogr. 40 (1), 1–8. http://dx.doi.org/10.1111/jbi.12006.

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. http://dx.doi.org/10.1111/ecog.04385.

Muñoz, A.R., Real, R., Barbosa, A.M., Vargas, J.M., 2005. Modelling the distribution of Bonelli's eagle in Spain: implications for conservation planning. Divers. Distrib. 11 (6), 477–486. http://dx.doi.org/10.1111/j.1366-9516.2005.00188.x.

Oliveira, U., Paglia, A.P., Brescovit, A.D., de Carvalho, C.J., Silva, D.P., Rezende, D.T., et al., 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of b razilian terrestrial biodiversity. Divers. Distrib. 22 (12), 1232–1244. http://dx.doi.org/10.1111/ddi.12489.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19 (1), 181–197. http://dx.doi.org/10.1890/07-2153.1.

van Proosdij, A.S., Sosef, M.S., Wieringa, J.J., Raes, N., 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography 39 (6), 542–552. http://dx.doi.org/10.5061/dryad.8sb8.

Pulido-Pastor, A., Márquez, A.L., Guerrero, J.C., García-Barros, E., Real, R., 2021. Metapopulation patterns of Iberian butterflies revealed by fuzzy logic. Insects 12 (5), 392. http://dx.doi.org/10.3390/insects12050392.

Real, R., Barbosa, A.M., Vargas, J.M., 2006. Obtaining environmental favourability functions from logistic regression. Environ. Ecol. Stat. 13 (2), 237–245. http://dx.doi.org/10.1007/s10651-005-0003-3.

Rocchini, D., Garzon-Lopez, C.X., Marcantonio, M., Amici, V., Bacaro, G., Bastin, L., et al., 2017. Anticipating species distributions: Handling sampling effort bias under a Bayesian framework. Sci. Total Environ. 584, 282–290. http://dx.doi.org/10.1016/j.scitotenv.2016.12.038.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C., et al., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Progr. Phys. Geogr. 35 (2), 211–226. http://dx.doi.org/10.1177/0309133311399491.

Rocchini, D., Marcantonio, M., Arhonditsis, G., Cacciato, A.L., Hauffe, H.C., He, K.S., 2019. Cartogramming uncertainty in species distribution models: a Bayesian approach. Ecol. Complex. 38, 146–155. http://dx.doi.org/10.1016/j.ecocom.2019.04.002.

Romero, D., Báez, J.C., Ferri-Yáñez, F., Bellido, J.J., Real, R., 2014. Modelling favourability for invasive species encroachment to identify areas of native species vulnerability. Sci. World J. http://dx.doi.org/10.1155/2014/519710.

Romero, D., Olivero, J., Real, R., Guerrero, J.C., 2019. Applying fuzzy logic to assess the biogeographical risk of dengue in South America. Parasites Vectors 12 (1), 1–13. http://dx.doi.org/10.1186/s13071-019-3691-5.

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. Biodiver. Data J. 8, http://dx.doi.org/10.3897/BDJ.8.e53474.

S., Wood, 2017. Generalized Additive Models: An Introduction with R, 2 edition Chapman and Hall/CRC.

Schmitt, S., Pouteau, R., Justeau, D., Boissieu, F.De., Birnbaum, P., 2017. Ssdm: An r package to predict distribution of species richness and composition based on stacked species distribution models. Methods Ecol. Evolut. 8 (12), 1795–1803. http://dx.doi.org/10.1111/2041-210X.12841.

Schweiger, A.H., Irl, S.D.H., Steinbauer, M.J., Dengler, J., Beierkuhnlein, C., 2016. Optimizing sampling approaches along ecological gradients. Methods Ecol. Evol. 7, 463–471. http://dx.doi.org/10.1111/2041-210X.12495.

Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C.G., Sousa-Guedes, D., Martínez-Freiría, F., et al., 2021. Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. Ecol. Model. 456, 109671. http://dx.doi.org/10.1016/j.ecolmodel.2021.109671.

Sillero, N., Barbosa, A.M., 2021. Common mistakes in ecological niche models. Int. J. Geogr. Inf. Sci. 35 (2), 213–226. http://dx.doi.org/10.1080/13658816.2020.1798968.

Syfert, M.M., Smith, M.J., Coomes, D.A., 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PLoS One 8 (2), e55158. http://dx.doi.org/10.1371/journal.pone.0055158.

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. Ecol. Indic. 121, 107147. http://dx.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. Divers. Distrib. 20 (11), 1258–1269. http://dx.doi.org/10.1111/ddi.12236.

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A.C., Guisan, A., 2014. Measuring the relative effect of factors affecting species distribution model predictions. Methods Ecol. Evolut. 5 (9), 947–955. http://dx.doi.org/10.1111/2041-210X.12203.

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? Ecol. Model. 220 (4), 589–594. http://dx.doi.org/10.1016/j.ecolmodel.2008.11.010.

Warren, D.L., Glor, R.E., Turelli, M., 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolut. Int. J. Org. Evolut. 62 (11), 2868–2883. http://dx.doi.org/10.1111/j.1558-5646.2008.00482.x.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., NCEAS Predicting Species Distributions Working Group, 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14 (5), 763–773. http://dx.doi.org/10.1111/j.1472-4642.2008.00482.x.

Wright, M.N., Ziegler, A., 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1), 1–17. http://dx.doi.org/10.18637/jss.v077.i01.