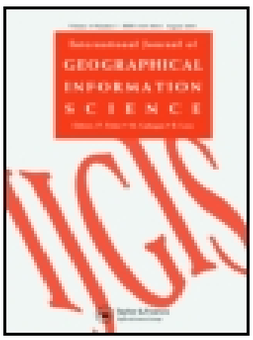

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

How do species and data characteristics affect species distribution models and when to use environmental filtering?

Lukáš Gábor, Vítězslav Moudrý, Vojtěch Barták & Vincent Lecours

To cite this article: Lukáš Gábor, Vítězslav Moudrý, Vojtěch Barták & Vincent Lecours (2019): How do species and data characteristics affect species distribution models and when to use environmental filtering?, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2019.1615070

To link to this article: https://doi.org/10.1080/13658816.2019.1615070

Published online: 14 May 2019.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2019.1615070

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

How do species and data characteristics affect species distribution models and when to use environmental filtering?

Lukáš Gábor[a], Vítězslav Moudrý[a], Vojtěch Barták[a] and Vincent Lecours[b]

[a]Department of Applied Geoinformatics and Spatial Planning, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha – Suchdol, Czech Republic; [b]School of Forest Resources and Conservation, University of Florida, Gainesville, FL, USA

ABSTRACT

Species distribution models (SDMs) are widely used in ecology and conservation. However, their performance is known to be affected by a variety of factors related to species occurrence characteristics. In this study, we used a virtual species approach to overcome the difficulties associated with testing of combined effects of those factors on performance of presence-only SDMs when using real data. We focused on the individual and combined roles of factors related to response variable (i.e. sample size, sampling bias, environmental filtering, species prevalence, and species response to environmental gradients). Results suggest that environmental filtering is not necessarily helpful and should not be performed blindly, without evidence of bias in species occurrences. The more gradual the species response to environmental gradients is, the greater is the model sensitivity to an inappropriate use of environmental filtering, although this sensitivity decreases with higher species prevalence. Results show that SDMs are affected to the greatest degree by the species response to environmental gradients, species prevalence, and sample size. Models' accuracy decreased with sample size below 300 presences. Furthermore, a high level of interactions among individual factors was observed. Ignoring the combined effects of factors may lead to misleading outcomes and conclusions.

ARTICLE HISTORY
Received 12 July 2018
Accepted 1 May 2019

KEYWORDS
MaxEnt; Schoener's D; species rarity; spatial data filtering; virtual species

Introduction

Many of the modeling techniques developed in the last two decades are now recognized to play an important role in monitoring of biodiversity and its conservation (Guisan and Zimmermann 2000, Honrado *et al*. 2016). Species distribution models (SDMs) have become a common tool for the assessment of species-environment relationships. The objective of SDMs is to relate species occurrence data (i.e. response variable) and environmental data (i.e. predictor variables) in order to either describe relationships between them ('explanatory modeling') or predict probabilities of species occurrences at unsampled sites or times ('predictive modeling') (see review by Ferrier *et al*. 2017). SDMs are now routinely used, for example to assess the spread of invasive

CONTACT Vítězslav Moudrý moudry@fzp.czu.cz

species (Gillard *et al*. 2017, Bazzichetto *et al*. 2018), the impact of climate change on biodiversity (Sun *et al*. 2017), or species ranges (Williams and Crouch 2017). High-quality species occurrence records (i.e. unbiased, positionally accurate data without false presences and absences) are essential to generate informative and accurate SDMs (Osborne and Leitão 2009; Duputié *et al*. 2014, Moudrý *et al*. 2017). However, acquisition of such data is often challenging and the underlying challenge in SDMs is to derive response curves from incomplete and biased datasets.

In practice, the most commonly available species records are usually non-systematic observations (see Bino *et al*. 2014), such as collections of individual observations from various sources (e.g. museums, citizen science data) available through global databases (e.g. the Global Biodiversity Information Facility – GBIF; www.gbif.org). This type of species observations are referred to as presence-only records (presence-background records *sensu* Guillera-Arroita *et al*. 2015). Most of the species in global databases are however under-sampled, particularly rare and endangered species (i.e. those of the highest importance from a conservation perspective), resulting in a sample size that is too low to provide reliable models (but see Breiner *et al*. 2015, 2018 for possibilities of overcoming limitations of modeling species with few occurrences). The effects of sample size on model performance have been studied extensively (e.g. Jiménez-Valverde *et al*. 2009, Moudrý and Šímová 2012), although no consensus has been reached; some studies concluded that even very small sample sizes can provide reliable models (Guisan *et al*. 2007, Varela *et al*. 2014, Proosdij *et al*. 2016) while others have shown the opposite (Wisz *et al*. 2008, Tessarolo *et al*. 2014).

Furthermore, species occurrence records are often spatially biased (Isaac and Pocock 2015), which is usually caused by uneven sampling efforts or data sharing. Such bias has been reported for data collected in easily accessible areas (Reddy and Dávalos 2003), protected areas (Boakes *et al*. 2010), or heavily populated areas (Geldmann *et al*. 2016). It is important to account for spatial bias in SDMs because it may affect model calibration and cause an overestimation of SDM performance (Leitão *et al*. 2011, Hijmans 2012, Boria *et al*. 2014). Various methods have been proposed to compensate for sampling bias in species occurrence records, including manipulation of background data (Phillips *et al*. 2009) and spatial filtering (Veloz 2009, Anderson and Raza 2010, Boria *et al*. 2014, Tessarolo *et al*. 2014). Spatial filtering is used to reduce the negative influence of sampling bias in geographic space. Recently, however, Varela *et al*. (2014) suggested that this approach could fail because species occurrences with unique environmental conditions could be removed. Instead, they suggested the use of environmental filtering to down-weight repeated species occurrences in similar environmental conditions, which we also adopted in this study. Increasing attention has also been given to comparison or evaluation of those methods (Kramer-Schadt *et al*. 2013, Varela *et al*. 2014, Ranc *et al*. 2016). Filtering necessarily reduces the sample size, and although Varela *et al*. (2014) suggested that a filtered subsample of occurrences can be better than using all available records to calibrate models, the tradeoff between lower sample size after filtering and higher sample size without filtering has yet to be tested.

In addition to the quality of occurrence data (e.g. sample size, sampling bias) and methods used to filter the data (e.g. environmental filtering, geographic filtering), species characteristics also need to be considered. Studies have shown that commonness and rarity or prevalence may influence the ability to predict species distribution;

models for rare species (i.e. species with low prevalence) tend to have higher prediction accuracy than models generated for more common species (i.e. with high prevalence; Syphard and Franklin 2009, Sor *et al*. 2017).

Species characteristics (e.g. prevalence, response to environmental gradients) and data characteristics (e.g. bias, filtering, sample size) are usually studied separately or in combinations of two or three factors (but see e.g. Thibaud *et al*. 2014, Fernandes *et al*. 2018, Liu *et al*. 2019). It is therefore difficult to determine a characteristic affecting SDMs performance the most, as well as to evaluate potential interactions between species and data characteristics. In this study, we used a virtual species approach to assess the effects of prevalence, response to environmental gradients, sampling bias, sample size, and samples filtering, as well as their interactions, on SDMs performance. The use of virtual species approach enables full control over the factors influencing models and the disentanglement of confounding effects (Zurell *et al*. 2010, Miller 2014). Consequently, this approach is increasingly used to evaluate SDMs performance (e.g. Václavík and Meentemeyer 2012, Qiao *et al*. 2015, Moudrý *et al*. 2018). To test how the five species and data characteristics affect SDMs performance, we produced SDMs for virtual species with different responses to environmental gradients (abrupt, nearly abrupt, nearly smooth, smooth), different levels of prevalence (very rare, rare, common), different sample sizes, unbiased and biased samples, and non-filtered and environmentally filtered datasets (Figure 1). Our specific objectives were to (i) determine the role of the sample size and species prevalence in SDMs, (ii) assess whether environmental filtering improves models based on biased samples, and (iii) evaluate the effect of species response to environmental gradients used to generate virtual species on factors under study (i.e. sample size, sampling bias, environmental filtering, species prevalence).

Material and methods

Figure 1 illustrates the workflow used in this study. First, virtual species distributions were modeled for the study area, encompassing to the Iberian Peninsula. The species distributions were modeled using various responses to environmental gradients and various species prevalences. Subsets of species occurrences were subsequently extracted from the species presence/absence distributions, using various sample sizes, sampling patterns and with/without application of filtering. Finally, SDMs were produced using those different subsets and their performance was evaluated and compared.

Simulating ecological patterns with virtual species

Data derived from Worldclim (www.worldclim.org) database are often adopted in SDMs (e.g. Moudrý and Šímová 2013). To build virtual species distributions we used the same variables downloaded from Wordclim that were adopted in the study by Varela *et al*. (2014) who first presented the idea of environmental filtering. However, their study used virtual species generated with threshold approach for evaluation, which was recently criticized (Meynard and Kaplan 2012, Moudrý 2015). In our study, we used a probability approach (see Meynard and Kaplan 2012, 2013) to generate virtual species (see the next paragraph). Using the same variables and study area allowed us to directly compare our results with theirs. The adopted variables included the maximum temperature of the warmest month

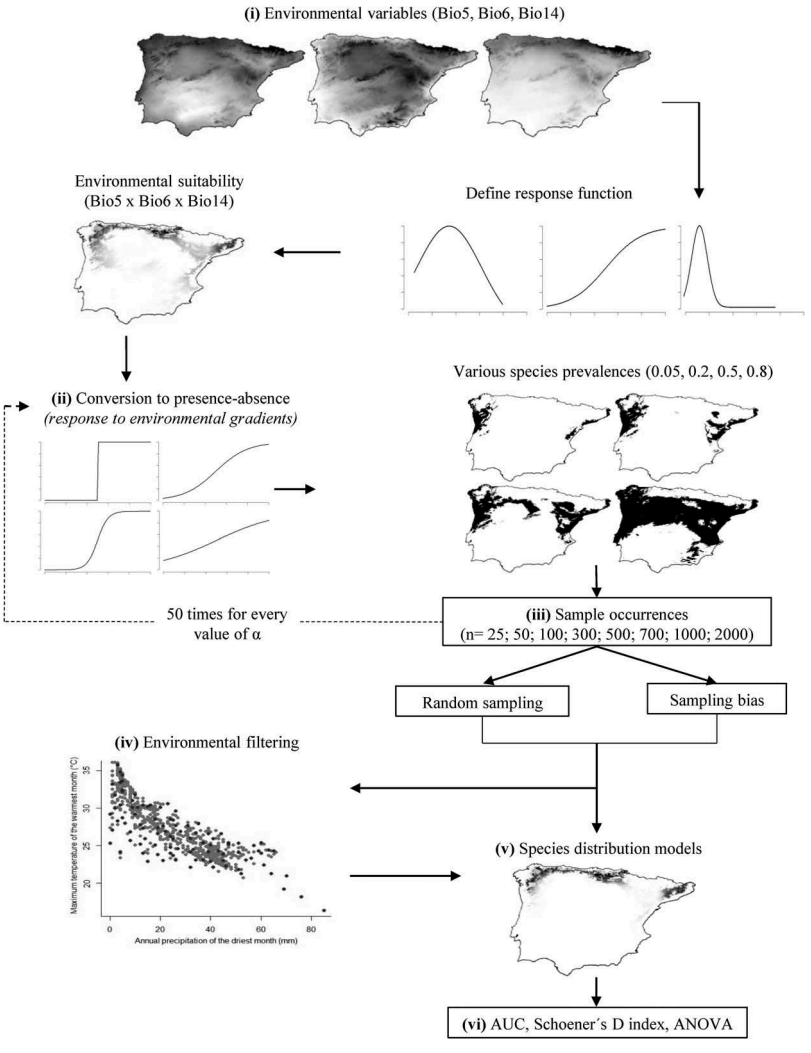

Figure 1. General modeling process. (i) Generating a map of probability occurrence for virtual species (environmental suitability map). (ii) Translating the probability of occurrence into a presence-absence map for various species prevalences. (iii) Sampling species occurrences randomly or with uneven sampling intensity and repeating the sampling 50 times for every α value and species prevalence. (iv) Applying environmental filter. (v) Creating models of species distribution with and without filtered occurrences. (vi) Quantifying SDMs performance using AUC and Schoener´s D index and performing ANOVA to statistically compare SDMs performance.

(Bio5), minimum temperature of the coldest month (Bio6), and annual precipitation of the driest month (Bio14). Those were downloaded at a resolution of 30 arc seconds (approximately 1 km^2) from WorldClim and clipped to the extent of the Iberian Peninsula.

Various software and packages have been developed to facilitate the use of virtual species in SDMs (e.g. Duan *et al.* 2015, Leroy *et al.* 2016, Qiao *et al.* 2016). There are currently two main methods for generating virtual species: a threshold approach and a probability approach (Meynard and Kaplan 2012). The threshold approach generates occurrences

where species always occur above a given threshold and never below it. On the other hand, the probability approach allows to generate species presences and absences along the whole environmental gradients (i.e. the shape of the logit function used to transform the occurrence probability to presences/absences). Moreover, the probability approach allows to take species prevalence into consideration. It is thus closer to ecological theories supporting the idea of dynamic occupancy patterns in space and time (see Hanski 1998, Meynard and Kaplan 2012, 2013). Therefore, the probability approach has been deemed more appropriate for generating virtual species than the threshold approach (Meynard and Kaplan 2012, 2013, Moudrý 2015). Virtual species distributions were created with the package *virtualspecies* (Leroy *et al*. 2016) in the statistical software R (version 3.4.4).

The virtual species were created in three steps (Leroy *et al*. 2016). First, we defined a relationship (i.e. the response function) between the artificial species and each variable, using a Gaussian distribution. Response functions were defined as follows (mean ± standard deviation): Bio5 (20 ± 10°C), Bio6 (10 ± 10°C), and Bio14 (20 ± 10 mm). The combination of these three response functions produced environmental suitability rasters for the Iberian Peninsula (function *generateSpFromFun*). Second, a probabilistic approach was used to convert environmental suitability rasters to binary presence-absence rasters (function *convertToPA*); a logistic function was applied to the suitability rasters to model the response to environmental gradients. The logistic function had two parameters, α and β, where 1/α corresponded to the slope of the curve at the inflection point and β to the position of the inflection point. Therefore, using α, one can control the steepness of the species response to the environmental gradients, and with a given value of α, the species prevalence can be controlled by β (see Figure 2). We modeled four species types with respect to their response to environmental gradients: species with an abrupt response (α = −0.000001), species with a nearly abrupt response (α = −0.05), species with a nearly smooth response (α = −0.15), and species with a smooth response (α = −0.3). In addition, to evaluate the effect of prevalence for each type of response, four different species prevalence values were produced (0.05, 0.2, 0.5 and 0.8) by varying the parameter β (Figure 2). We generated species occurrences 50 times for each combination of α and β to produce multiple replications. For each replication, a valid estimation of the true species distribution was provided (Leroy *et al*. 2016). This approach contrasts with the threshold approach, which always generates the same distribution (presences/absences).

The last step consisted of sampling occurrences of virtual species from the modeled distributions using the *sampleOccurrences* function of the package. To test for sampling bias, two different sampling methods (survey designs) were used to generate presence-only data: (i) random sampling across the entire area and (ii) a scheme that extracted samples 40 more times in the 50 largest protected areas of the Iberian Peninsula; this method has been used before to study sampling bias (e.g. Tessarolo *et al*. 2014, Varela *et al*. 2014). The protected areas of the Iberian Peninsula were downloaded from www.protectedplanet.net. Eight different sample sizes were used to test sample size effects on SDMs (n = 25, 50, 100, 300, 500, 700, 1000, 2000).

Environmental filtering of sampled occurrences

We used the *gridSample* function of the *dismo package* to filter the sampled presence-only data (Hijmans *et al*. 2012). Based on a pre-defined grid, the function allows to

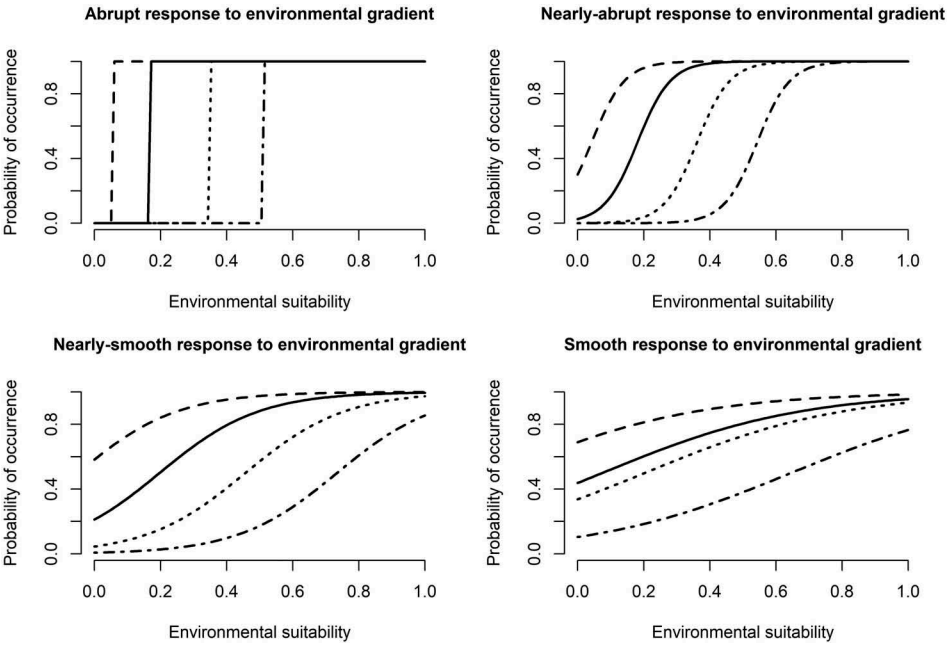

Figure 2. Contrasting examples of conversion curves for all species responses to environmental gradients and species prevalences (dotdash line = 0.05, dotted line = 0.2, solid line = 0.5, dashed line = 0.8).

eliminate repeated occurrences under similar environmental conditions. The environmental filters were defined using only two of the three environmental variables – the maximum temperature of the warmest month and the precipitation of the driest month. This enabled simulating a situation in which some of the environmental characteristics affecting species distribution were unknown, which is often the case in SDMs (Varela *et al*. 2014). The resulting filters were applied to all versions of the generated virtual species (i.e. combination of response to environmental gradients and prevalence; 12 species), to the eight sample sizes, and to both survey designs (random and spatially biased). This totaled 384 unique combinations of model parameters, enabling an in-depth comparison of the effects of the five different characteristics of SDMs under study. For each one of those combinations, 50 models were computed using the previously described probabilistic approach; each of those 50 repetitions can be viewed as a different run of the same stochastic process. A total of 19,200 different virtual species distributions were thus generated to be tested in SDMs.

Species distribution models

While there is currently no consensus on which SDM technique is best, it is widely recognized that every single technique has benefits and drawbacks (Elith *et al*. 2006, Elith and Graham 2009, Fernandes *et al*. 2019). For the purpose of this study, we needed a technique that could be kept consistent across the methodology to allow the comparison of outcomes. We selected the maximum entropy approach (MaxEnt), which is often

adopted in ecological studies as a presence-only modeling technique, due to its good performance when compared to other techniques (Elith *et al*. 2006, Philips *et al*. 2006). SDMs were built in R using the *dismo* package and the same three environmental variables that were used to generate virtual species distributions. To enable comparison of the different SDMs produced, we needed to maintain the parameters of the modeling technique unchanged. Although using MaxEnt with default settings is usually not recommended as it can overfit the models, it is not an issue when using virtually generated species as virtually generated data fit the pre-defined response perfectly and the risk of overfitting therefore is very low (we also employed hinge and linear feature classes and got the same results as with the default settings). Therefore, like many others before (e.g. Phillips *et al*. 2009, Beltrán *et al*. 2013, Ficetola *et al*. 2014, Fourcade *et al*. 2014, Franklin *et al*. 2014, Varela *et al*. 2014, Beaumont *et al*. 2016, Holloway *et al*. 2016, Ranc *et al*. 2016, Tingley *et al*. 2018, Ye *et al*. 2018), we produced the models using the default settings, except for background points.

Since using background points that do not have the same bias as species occurrences (e.g. using random background points when species occurrences are spatially biased) has been shown to negatively affect SDMs performance (Phillips *et al*. 2009, Leroy *et al*. 2018), we did not use randomly generated background points. Instead, based on the artificially generated binary map of the virtual species illustrating true occupied and unoccupied areas, we generated a set of background points (i) across the entire area for randomly drawn species presence data (simulating unbiased dataset) and (ii) with higher sampling intensity in the 50 largest protected areas of the Iberian Peninsula (simulating biased dataset) and use those as background points. We used two times more background points than species occurrences as recommended by Liu *et al*. (2019). For each model replication, a new set of background points was generated. Similarly, as Thibaud *et al*. (2014), we had absence data available. Therefore, we generated background points in locations where species were absent and used Maxent in a nonstandard manner. Hence, the models can be viewed as presence-absence, allowing us to use the area under the receiver operating characteristic curve (AUC) as an appropriate measure for model performance. To evaluate the models, a fivefold cross-validation was used where the data were randomly divided into fifths. Four-fifths of the data were used to train the model and the remaining one fifth was used to quantify the performance.

Assessment of model performance

The AUC was calculated to quantify model performance. AUC indicates model performance based on predictions of presences/absences (Fielding and Bell 1997) and varies between 0 and 1 where values 0.9–1 indicate excellent models. In addition, we calculated Schoener's D index (Schoener 1968) to compare modeled probabilities of occurrence. Schoener's D is considered one of the best measures of evaluation of SDMs outputs (Rödder and Engler 2011). This metric measures the absolute spatial conformity between continuous predictions of the species as,

$$D = 1 - \frac{1}{2}\sum\nolimits_{ij} \left| z_{1ij} - z_{2ij} \right|$$

where z_{1ij} is entity 1 occupancy (virtual reality) and z_{2ij} is entity 2 occupancy (model prediction) (Renkonen 1938). It varies between 0 (no overlap/agreement) and 1 (complete overlap/agreement). An analysis of variance (ANOVA) was used to assess the individual and combined effects of species response to environmental gradients, species prevalence, sample size, sampling bias and environmental filtering on SDM's performance. We fitted separate ANOVA models for AUC and Schoener's D index as a response, including all possible interactions among all five factors in both models.

Results

The ANOVA including all possible interactions explained 96% of Schoener's D and 89% of AUC variability (variance of the models explained by used characteristics) (Table 1). All analyses were highly significant given the large number of iterations (n = 19,200). For both Schoener's D and AUC, most of their variability was explained by the species response to environmental gradients (from abrupt to smooth), species prevalence, and sample size; these factors together (and disregarding their interactions) explained 53% of Schoener's D variability and 64% of AUC variability, with species prevalence being more influential for Schoener's D (17%) than for AUC (5%) (see Table 1).

The effect of sample size on SDMs was relatively constant across other factors' (e.g. species prevalence, species response to environmental gradients) levels (see generally low R^2 values for its interaction terms in Table 1). Results show an initial steep increase in performance with increasing sample size, generally stabilizing around 300 samples after which more samples do not necessarily result in better models (see Figures 3 and 4). The initial increase was considerably steeper for AUC metric than for Schoener's D. The three exceptions to this general pattern were (1) almost constant values of Schoener's D across sample sizes for species with smooth or nearly smooth response to environmental gradients and with higher prevalence, (2) decreasing Schoener's D for abrupt and

Table 1. Degrees of freedom (Df), R^2, and F statistics for ANOVA of Schoener's D index and AUC performance metrics.

	Schoener's D index			AUC		
	Df	R^2 (%)	F	Df	R^2 (%)	F
Main effects:						
Sampling method (Samp)	1	0.2	1127.3	1	0.2	334.3
Species (Spec)	3	19.9	32,479.2	3	30.8	17,153.6
Species prevalence (Prev)	2	17.0	41,424.6	2	5.1	4227.1
Sample size (Size)	7	16.3	11,353.3	7	28.1	6697.3
Filter application (Filter)	1	2.4	11,474.9	1	1.6	2665.9
Pair-wise interactions:						
Samp : Spec	3	0.3	452.7	3	0.2	124.3
Samp : Prev	2	0.0	96.6	2	0.1	81.3
Spec : Prev	6	8.9	7206.9	6	2.5	702.6
Samp : Size	7	0.1	58.7	7	0.0	10.0
Spec : Size	21	3.6	835.1	21	2.2	170.8
Prev : Size	14	2.0	699.5	14	1.0	116.3
Samp : Filter	1	1.2	5859.4	1	0.0	31.6
Spec : Filter	3	5.7	9280.1	3	8.0	4472.8
Prev : Filter	2	5.1	12,448.0	2	1.3	1093.8
Size : Filter	7	1.4	993.1	7	1.4	335.1
Higher-order interactions:	-	12.0	-	-	6.3	-
Total:	-	96.2	-	-	88.8	-

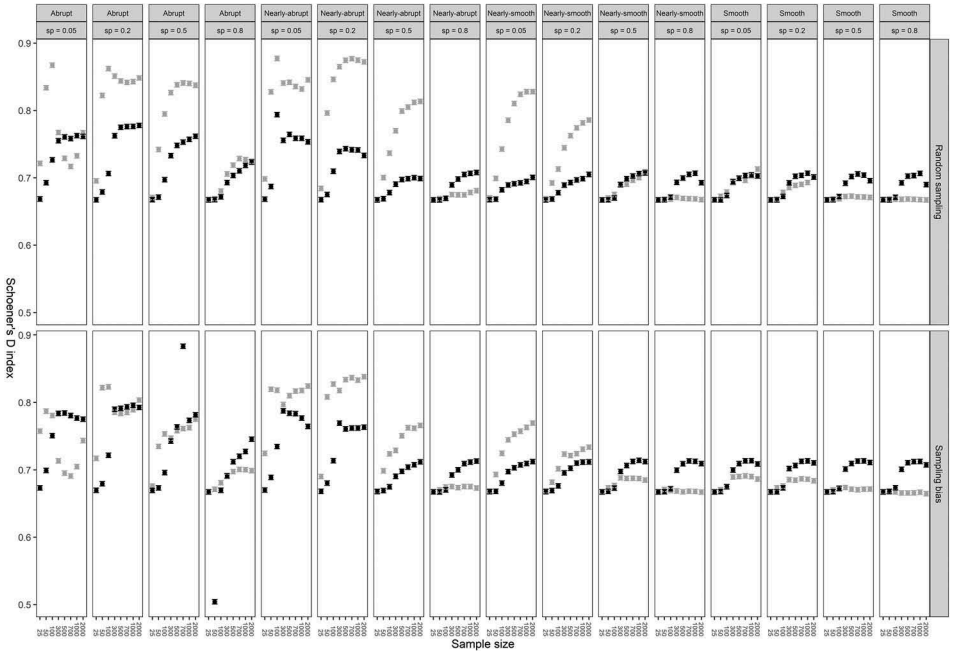

Figure 3. Resulting Schoener´s D index values according to different species responses to environmental gradients (abrupt, nearly abrupt, nearly smooth, smooth), species prevalence (sp = 0.05, 0.2, 0.5, 0.8), various methods of sampling occurrences (random, sampling bias) and different sample size (n = 25, 50, 100, 300, 500, 700, 1000, 2000). Gray color indicates results for non-filtered models, and the black color shows results for models where the environmental filter was applied.

nearly abrupt species with species prevalence 0.05 and non-filtered models (Figure 3) and (3) no stabilization for AUC values for species with nearly smooth or smooth response (Figure 4).

The relatively low main-effect R^2 values of filter application term (see Table 1) resulted from a reverse effect of this factor in different species types (and also species prevalences in case of Schoener's D) (see Figures 3 and 4). Indeed, taking into account also its pair-wise interactions, filter application explained 16% of the Schoener's D and 12% of the AUC variability. For Schoener's D, these interactions can be summarized as follows (see Figure 3): the performance of the non-filtered SDMs was the highest for species with abrupt response to environmental gradients (about 0.85) and decreased to approximately 0.66 for those with smooth response. On the contrary, the performance of filtered SDMs was much more stable, ranging from approx. 0.78 for species with abrupt response to 0.70 for those with smooth response. This led to a significantly better performance of non-filtered SDMs for abruptly responding species but a slightly better performance of filtered SDMs for those responding smoothly. This relationship was further influenced by a significant decrease of non-filtered SDMs performance with increasing species prevalence, which was more striking for species with abrupt or nearly abrupt response. The exception from this general pattern was models with sample size higher or equal 100 for abrupt and nearly abrupt species and species prevalence 0.05. In this case, filtered models achieved better results than nonfiltered models. Moreover,

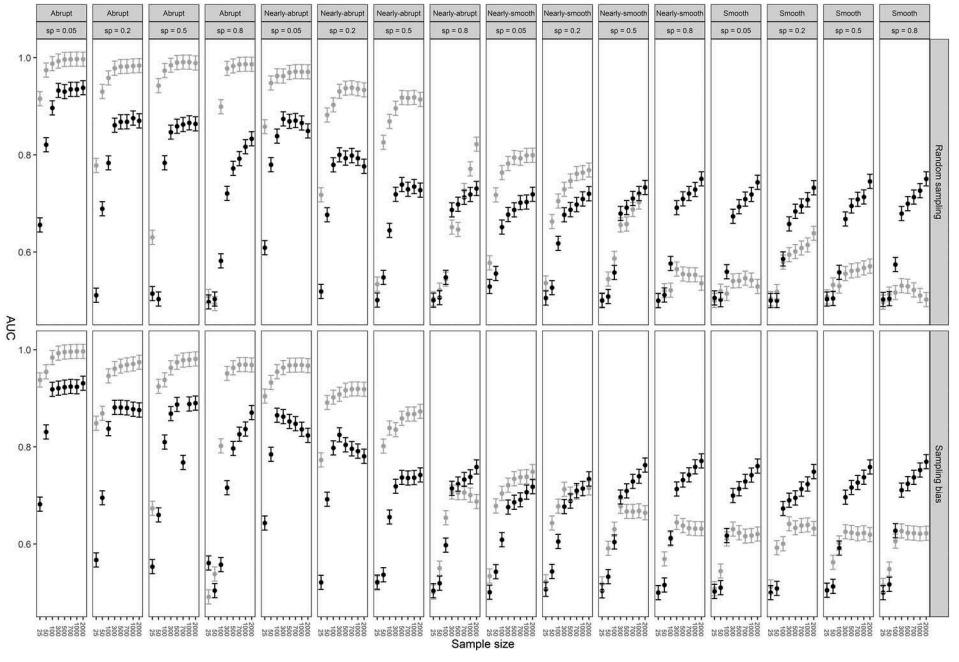

Figure 4. Resulting AUC values according to different species responses to environmental gradients (abrupt, nearly abrupt, nearly smooth, smooth), species prevalence (sp = 0.05, 0.2, 0.5, 0.8), different methods of sampling occurrences (random, sampling bias) and different sample sizes (n = 25, 50, 100, 300, 500, 700, 1000, 2000). Gray color indicates results for non-filtered models, and the black color shows results for models where the environmental filter was applied.

their resulting Schoener's D was even lower in comparison to smoothly or nearly smoothly responding species. For AUC metric (Figure 4), the pattern was similar, with generally larger performance ranges for non-filtered SDMs (from almost 1.0 to approximately 0.5 for random sampling), which led to larger differences between non-filtered and filtered SDMs for species with smooth response. Sampling method (random vs. biased sampling) showed the least importance, both as main effect and in interactions with other effects (maximum R^2 being 1.2% but typically from 0.0% to 0.3%; see Table 1).

Discussion

Our results show that species prevalence and sample size have an equivalent effect on variability in model performance when using the MaxEnt modeling technique. Models performance increased with sample size (often up to a certain level), and where the sample size was constant, the model performance decreased with increasing prevalence. Moreover, our results show that both effects are independent of sampling bias. As opposed to what is often done in other studies, here we were changing the steepness of the response to environmental gradient (i.e. logistic curve) to create virtual species from abrupt (i.e. similar to what would be done with the threshold approach) to very smooth (see Figure 3). Generally, the more abrupt the response of species to the environmental gradient was, the greater the effect of species prevalence, sample size,

and environmental filtering was. Since both measures (AUC and Schoener´s D) showed similar trends, the following discussion is based mostly on the behavior of the Schoener ´s D. We highlight the differences in AUC behavior where necessary.

Sample size

It has been shown many times that the performance of SDMs depends on sample size (see review by Moudrý and Šímová 2012). Prior studies examined sample sizes that varied from a few occurrences up to thousands of occurrences. While Guisan *et al*. (2007) or later Proosdij *et al*. (2016) have shown that a few occurrences may suffice to produce reliable models, other studies argued that it is best to use larger sample sizes (Pearson *et al*. 2007, Wisz *et al*. 2008, Tessarolo *et al*. 2014). Such opposing suggestions can be explained by the differences in data characteristics and model selection in these studies. One reason can be the complexity of species responses to environmental variables. It is clear that the more complex is the species response to environmental variables, the higher is the number of species occurrences required to achieve high model performance (e.g. Barry and Elith 2006). Studies using virtually generated data (e.g. Jiménez-Valverde *et al*. 2009, Varela *et al*. 2014, Proosdij *et al*. 2016) that have occurrences perfectly following the adopted response to environmental variables (e.g. Gaussian) suggested that reliable models can be developed with very small sample size (10 or even 5 samples). In contrast, studies with real species occurrence data (Wisz *et al*. 2008, Tessarolo *et al*. 2014) suggested the opposite. In addition, the effect of sample size could be also affected by species prevalence. Proosdij *et al*. (2016) concluded that increasing species prevalence decreases the influence of sample size. It has also been shown that some modeling techniques are less sensitive to sample size than others (Guisan *et al*. 2007, Tessarolo *et al*. 2014). Besides, Wisz *et al*. (2008) also show that the influence of sample size is changing across different spatial extents and resolutions of environmental variables (sites with resolution of 100 x 100 m performed better in comparison with those with resolution of 1000 × 1000 m). Our results show that a larger sample size has a significant positive effect on SDMs performance, although with a threshold after which more samples do not necessarily improve performance. In our case, that threshold was usually at 300 or 500 samples. This effect, however, was only consistent when measured by AUC, which was expected due to the sensitivity of AUC to the ratio of sample prevalence and species prevalence (see Meynard and Kaplan 2012). Our results are, moreover, in accordance with prior studies by Thibaud *et al*. (2014) and Fernandes *et al*. (2018) who also tested the impact of various factors affecting SDMs using virtual species and concluded that sample size is one of the most important factors. For Schoener's D, the effect of sample size considerably varied with species response to the environmental gradients, species prevalence and the use of environmental filtering.

Species prevalence

Our results show that species prevalence is one of the most important factors affecting SDMs, having generally a negative effect on both model performance metrics (i.e. model performance was generally decreasing with increasing species prevalence). While this negative effect has been observed by a number of previous studies looking at AUC (e.g.

Manel *et al*. 2001, Allouche *et al*. 2006, Lobo and Tognelli 2011, Meynard and Kaplan 2012, Syfert *et al*. 2013), it is to be noted that Proosdij *et al*. (2016) have found an opposite trend for Schoener's D. However, the authors did not provide any explanation or hypothesis for that trend, making its comparison with our study difficult. A potential explanation for that difference could be that their sample sizes (5 to 50 occurrences) were much smaller than the ones used in the current study (25 to 2000). In addition, our results show that this negative effect only applies to species with abrupt or nearly abrupt response to environmental gradients, a factor that was not specified in Proosdij *et al*. (2016).

Sampling bias

Sampling bias, caused by uneven sampling of species occurrences, is often considered as one of the major factors that have a negative impact on SDMs (e.g. Araujo and Guisan 2006, Leitão *et al*. 2011, Duputié *et al*. 2014, Guillera-Arroita *et al*. 2015). Prior studies have demonstrated that the presence of sampling bias decreases model performance (e.g. Loiselle *et al*. 2008, Leitão *et al*. 2011, Sánchez-Fernández *et al*. 2011, Fourcade *et al*. 2014, Ranc *et al*. 2016), While our results agree with that, they show that the contribution of sampling bias to the overall, combined effects of the different studied factors on SDMs is relatively low, explaining no more than 2% of the variability of the performance metrics. This demonstrates the importance of simultaneously studying multiple factors and their impacts on SDMs, whereas other studies focused solely on the effect of sampling bias and did not provide measures of explained variability, our study compared its effect with the effect of other factors, finding it statistically significant but relatively negligible. Our results are in agreement with the study by Tessarolo *et al*. (2014) who also concluded that sampling bias has rather minor effects on model performance compared to other factors (species characteristics, sampling method, sample size, SDMs technique). Interestingly, they used the same study area as our study (i.e. Iberian Peninsula), the difference lied in the use of 34 real species (amphibians, reptiles, mammals). Nevertheless, the effect of sampling bias may be related to autocorrelation in the predictor variables, which is relatively high in interpolated climate data (such as Worldclim used in both their and our study).

Environmental filtering

Another goal of our study was to test the applicability of environmental filtering on models generated with spatially biased data. According to Varela *et al*. (2014), environmental filtering consistently improves model performance. Our results show that the measured effect of environmental filtering was significant, they however also showed that that effect was relatively unimportant when compared to other factors (see Table 1). Moreover, its positive or negative effect strongly depended on the type of species response to environmental gradients, species prevalence, and sample size. We only confirmed the positive effect for species with smooth or nearly smooth response, whereas for species with abrupt or nearly abrupt response the effect was negative (except models with species prevalence 0.05). This contradicts the results of Varela *et al*. (2014) as their positive effect was observed for species generated using a threshold approach (i.e. the equivalent of our abrupt-responding species).

In addition, the positive effect was much stronger when assessed by AUC (up to more than 20% increase, see Figure 4) than by Schoener's D (only approx. 5% increase, see Figure 3). This is in accordance with previous concerns about using AUC as the only model performance measure (Jiménez-Valverde 2012, Moudrý 2015, Fernandes *et al*. 2019).

We recognize that SDMs may be affected by many other factors (see Thibaud *et al*. 2014, Fernandes *et al*. 2019). Thus, we recommend that further studies focus on interactions of environmental filtering with other factors, such as the effects of spatial scale (extent and resolution) (e.g. Connor *et al*. 2018; Šímová *et al*. 2019), spatial autocorrelation (Thibaud *et al*. 2014) or modeling technique (Fernandes *et al*. 2018).

Conclusions

We focused on several factors related to species occurrences (response variable) in SDMs (i.e. environmental filtering, sampling bias, sample size, species prevalence and species response to environmental gradient). We found that both sample size and species prevalence equivalently affect performance (measured by AUC and Schoener´s D) of SDMs (in general, increasing sample size positively, increasing species prevalence negatively). Our results also highlighted the importance of using a probability approach to the generation of virtual species distribution, which allowed us to model species with different response to environmental gradient from abrupt to smooth, as opposed to a threshold approach, which is still commonly used. Indeed, our results showed that the response of a species to environmental gradients has a strong effect not only on the model performance itself but also on the effects of other factors. The unprecedented complexity of our study enabled us to recognize the importance not only of each of the factors themselves but also of their interactions. Ignoring such interactions, which is almost inevitable in studies focusing on one or two factors only, may lead to substantially misleading conclusions.

Our results suggest that environmental filtering is not always a good idea and should not be performed blindly without evidence of bias in species occurrences. Environmental filtering down-weights repeated observations of the same environmental conditions and reduces sample size. Therefore, sampling must be dense enough to characterize the curve and the algorithms must be able to uncover the true form of the relationship. Our results show that at least 300 presences are necessary for accurate predictions when using presence-only models fitted by MaxEnt. We suggest that models using original, unfiltered data should be always fitted. We highlight that the more gradual is the species response to environmental gradients (except species with prevalence 0.05), the greater is the model sensitivity to inappropriate use of environmental filtering, although the sensitivity decreases with higher species prevalence. Finally, we advocate that additional data and species characteristics (e.g. resolution, extent, positional error) should be evaluated using more complex virtual species (e.g. with more complex response curves) to improve SDM use in biodiversity monitoring and conservation.

Author contributions

VM is author of the main idea of the research and supervised whole research. LG and VL further improved the idea and performed all GIS and statistical analyses. VB supervised statistical analyses. All authors discussed the results and equally contributed to the final text.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was funded by the Internal Grant Agency of Faculty of Environmental Sciences, Czech University of Life Sciences Prague [grant no. 20174241].

Notes on contributors

Lukáš Gábor is a PhD student at the Department of Applied Geoinformatics and Spatial planning, Czech University of Life Sciences Prague. His research focuses on how quality of spatial data affects species distribution models. He is enthusiastic about using fine-scale environmental data which are increasingly available for modelling species distribution. Lukáš Gábor is a member of the Spatial Science in Ecology and Environment research group (https://www.fzp.czu.cz/ssee).

Vítězslav Moudrý is an assistant professor at the Department of Applied Geoinformatics and Spatial planning, Czech University of Life Sciences Prague. His research interests include various applications of Earth observation data to monitor the environment and to study species-environment relationships. He is also interested in methodological issues related to species distribution modelling, particularly to spatial scale and spatial data quality. Vítězslav Moudrý is a member of the Spatial Science in Ecology and Environment research group (https://www.fzp.czu.cz/ssee).

Vojtěch Barták is an assistant professor at the Department of Applied Geoinformatics and Spatial planning, Czech University of Life Sciences Prague. His research focuses on spatial ecology of animals (including home range analysis and modelling), species distribution modelling, and digital terrain analysis. Vojtěch Barták is a member of the Spatial Science in Ecology and Environment research group (https://www.fzp.czu.cz/ssee).

Vincent Lecours is an assistant professor of remote sensing and geospatial analysis in the Geomatics program and the Fisheries & Aquatic Sciences program at the University of Florida. His research focuses on bridging the spatial sciences and ecology, particularly by studying the roles of spatial concepts such as spatial scale and spatial data quality in habitat mapping and species distribution modelling.

ORCID

Lukáš Gábor http://orcid.org/0000-0001-6137-0994
Vítězslav Moudrý http://orcid.org/0000-0002-3194-451X
Vojtěch Barták http://orcid.org/0000-0001-9887-1290

References

Allouche, O., Tsoar, A., and Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43 (6), 1223–1232. doi:10.1111/jpe.2006.43.issue-6

Anderson, R.P. and Raza, A., 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, 37 (7), 1378–1393. doi:10.1111/jbi.2010.37.issue-7

Araujo, M.B. and Guisan, A., 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33 (10), 1677–1688. doi:10.1111/j.1365-2699.2006.01584.x

Barry, S. and Elith, J., 2006. Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43 (3), 413–423. doi:10.1111/jpe.2006.43.issue-3

Bazzichetto, M., *et al*., 2018. Modeling plant invasion on Mediterranean coastal landscapes: an integrative approach using remotely sensed data. *Landscape and Urban Planning*, 171, 98–106. doi:10.1016/j.landurbplan.2017.11.006

Beaumont, L.J., *et al*., 2016. Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? *Ecological Modelling*, 342, 135–146. doi:10.1016/j.ecolmodel.2016.10.004

Beltrán, B.J., *et al*., 2013. Effects of climate change and urban development on the distribution and conservation of vegetation in a Mediterranean type ecosystem. *International Journal of Geographical Information Science*, 28 (8), 1561–1589. doi:10.1080/13658816.2013.846472

Bino, G., Ramp, D., and Kingsford, R.T., 2014. Identifying minimal sets of survey techniques for multi-species monitoring across landscapes: an approach utilising species distribution models. *International Journal of Geographical Information Science*, 28 (8), 1674–1708. doi:10.1080/13658816.2013.871016

Boakes, E.H., *et al*., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, 8 (6), e1000385. doi:10.1371/journal.pbio.1000418

Boria, R.A., *et al*., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275, 73–77. doi:10.1016/j.ecolmodel.2013.12.012

Breiner, F.T., *et al*., 2015. Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, 6, 1210–1218. doi:10.1111/2041-210X.12403

Breiner, F.T., *et al*., 2018. Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution*, 9, 802–808. doi:10.1111/2041-210X.12957

Connor, T., *et al*., 2018. Effects of grain size and niche breadth on species distribution modeling. *Ecography*, 41 (8), 1270–1282. doi:10.1111/ecog.03416

Duan, R.Y., *et al*., 2015. SDMvspecies: a software for creating virtual species for species distribution modelling. *Ecography*, 38 (1), 108–110. doi:10.1111/ecog.01080

Duputié, A., Zimmermann, N.E., and Chuine, I., 2014. Where are the wild things? Why we need better data on species distribution. *Global Ecology and Biogeography*, 23 (4), 457–467. doi:10.1111/geb.2014.23.issue-4

Elith, J., *et al*., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129–151. doi:10.1111/j.2006.0906-7590.04596.x

Elith, J. and 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, 66–77. doi:10.1111/eco.2009.32.issue-1

Fernandes, R.F., Scherrer, D., and Guisan, A., 2018. How much should one sample to accurately predict the distribution of species assemblages? A virtual community approach. *Ecological Informatics*, 48, 125–134. doi:10.1016/j.ecoinf.2018.09.002

Fernandes, R.F., Scherrer, D., and Guisan, A., 2019. Effects of simulated observation errors on the performance of species distribution models. *Diversity and Distributions*, 25 (3), 400–413. doi:10.1111/ddi.12868

Ferrier, S., Jetz, W., and Scharlemann, J., 2017. Biodiversity modelling as part of an observation system. *In*: Walters, M. and Scholes, R., J., eds. *The GEO handbook on biodiversity observation networks*. Cham: Springer International Publishing, 239–257.

Ficetola, G.F., *et al*., 2014. How many predictors in species distribution models at the landscape scale? Land use versus LiDAR-derived canopy height. *International Journal of Geographical Information Science*, 28 (8), 1723–1739. doi:10.1080/13658816.2014.891222

Fielding, A.H. and Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24 (1), 38–49. doi:10.1017/S0376892997000088

Fourcade, Y., *et al*., 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PLoS One*, 9 (5), e97122. doi:10.1371/journal.pone.0097122

Franklin, J., Regan, H.M., and Syphard, A.D., 2014. Linking spatially explicit species distribution and population models to plan for the persistence of plant species under global change. *Environmental Conservation*, 41 (2), 97–109. doi:10.1017/S0376892913000453

Geldmann, J., *et al*., 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22 (11), 139–1149. doi:10.1111/ddi.12477

Gillard, M., *et al*., 2017. Present and future distribution of three aquatic plants taxa across the world: decrease in native and increase in invasive ranges. *Biological Invasions*, 19 (7), 2159–2170. doi:10.1007/s10530-017-1428-y

Guillera-Arroita, G., *et al*., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, 24 (3), 276–292. doi:10.1111/geb.2015.24.issue-3

Guisan, A., *et al*., 2007. What matters for predicting the occurrences of trees: techniques, data or species characteristics? *Ecological Monographs*, 77 (4), 615–630. doi:10.1890/06-1060.1

Guisan, A. and Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135 (2–3), 147–186. doi:10.1016/S0304-3800(00)00354-9

Hanski, I., 1998. Metapopulation dynamics. *Nature*, 396 (6706), 41–49. doi:10.1038/23876

Hijmans, R. J., et al. 2012. Package 'dismo'. *Species Distribution Modeling. R Package Version*, 0.8–11.

Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, 93 (3), 679–688.

Holloway, P., Miller, J.A., and Gillings, S., 2016. Incorporating movement in species distribution models: how do simulations of dispersal affect the accuracy and uncertainty of projections? *International Journal of Geographical Information Science*, 30 (10), 2050–2074.

Honrado, J.P., Pereira, H.M., and Guisan, A., 2016. Fostering integration between biodiversity monitoring and modelling. *Journal of Applied Ecology*, 53 (5), 1299–1304. doi:10.1111/1365-2664.12777

Isaac, N.J. and Pocock, M.J., 2015. Bias and information in biological records. *Biological Journal of the Linnean Society*, 115 (3), 522–531. doi:10.1111/bij.12532

Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21 (4), 498–507. doi:10.1111/geb.2012.21.issue-4

Jiménez-Valverde, A., Lobo, J., and Hortal, J., 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, 10 (2), 196–205. doi:10.1556/ComEc.10.2009.2.9

Kramer-Schadt, S., *et al*., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19 (11), 1366–1379. doi:10.1111/ddi.12096

Leitão, P.J., Moreira, F., and 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. *International Journal of Geographical Information Science*, 25 (3), 439–454. doi:10.1080/13658816.2010.531020

Leroy, B., *et al*., 2016. Virtualspecies, an R package to generate virtual species distributions. *Ecography*, 39 (6), 599–607. doi:10.1111/ecog.01388

Leroy, B., Delsol, R., Hugueny, B., Meynard, Ch.N., Barhoumi, Ch., 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.

Liu, C., Newell, G., and White, M., 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. *Ecography*, 42 (3), 535–548. doi:10.1111/ecog.2019.v42.i3

Lobo, J.M. and Tognelli, M.F., 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, 19 (1), 1–7. doi:10.1016/j.jnc.2010.03.002

Loiselle, B.A., *et al*., 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, 35 (1), 105–116.

Manel, S., Williams, H.C., and Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38 (5), 921–931. doi:10.1046/j.1365-2664.2001.00647.x

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

Meynard, C.H.N. and Kaplan, D.M., 2013. Using virtual species to study species distributions and model performance. *Journal of Biogeography*, 40 (1), 1–8. doi:10.1111/jbi.12006

Miller, J.A., 2014. Virtual species distribution models: using simulated data to evaluate aspects of model performance. *Progress in Physical Geography*, 38 (1), 117–128. doi:10.1177/0309133314521448

Moudrý, V., 2015. Modelling species distributions with simulated virtual species. *Journal of Biogeography*, 42 (8), 1365–1366. doi:10.1111/jbi.2015.42.issue-8

Moudrý, V., *et al*., 2018. On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs. *Ecological Modelling*, 383, 3–9. doi:10.1016/j.ecolmodel.2018.05.006

Moudrý, V., Komárek, J., and Šímová, P., 2017. Which breeding bird categories should we use in models of species distribution? *Ecological Indicators*, 74, 526–529. doi:10.1016/j.ecolind.2016.11.006

Moudrý, V. and Šímová, P., 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *International Journal of Geographical Information Science*, 26 (11), 2083–2095. doi:10.1080/13658816.2012.721553

Moudrý, V. and Šímová, P., 2013. Relative importance of climate, topography, and habitats for breeding wetland birds with different latitudinal distributions in the Czech Republic. *Applied Geography*, 44, 165–171. doi:10.1016/j.apgeog.2013.08.001

Osborne, P. E., and Leitão, P. J., 2009. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*, 15 (4), 671–681.

Pearson, R.G., *et al*., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34 (1), 102–177. doi:10.1111/j.1365-2699.2006.01594.x

Phillips, S.J., *et al*., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19 (1), 181–197.

Phillips, S.J., Anderson, R.P., and Schapire, R.E., 2006. Maximum entropy modelling of species geographic distribution. *Ecological Modelling*, 190 (3–4), 231–259. doi:10.1016/j.ecolmodel.2005.03.026

Proosdij, A.S.J., *et al*., 2016. Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39 (6), 542–552. doi:10.1111/ecog.01509

Qiao, H., *et al*., 2015. Marble algorithm: a solution to estimating ecological niches from presence-only records. *Scientific Reports*, 5, 14232. doi:10.1038/srep14232

Qiao, H., *et al*., 2016. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, 39 (8), 805–813. doi:10.1111/ecog.2016.v39.i8

Ranc, N., *et al*., 2016. Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, 40 (9), 1076–1087. doi:10.1111/ecog.02414

Reddy, S. and Dávalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 30 (11), 1719–1727. doi:10.1046/j.1365-2699.2003.00946.x

Renkonen, O., 1938. Statistisch-ökologische Untersuchungen über die terrestrische Käfer- welt der finnischen Bruchmoore. *Annales Zoologici*, 6, 1–231.

Rödder, D. and Engler, J.O., 2011. Quantitative metrics of overlaps in Grinnellian niches: advances and possible drawbacks. *Global Ecology and Biogeography*, 20 (6), 915–927. doi:10.1111/j.1466-8238.2011.00659.x

Sánchez-Fernández, D., Lobo, J.M., and Hernández-Manrique, O.L., 2011. Species distribution models that do not incorporate global data misrepresent potential distributions: a case study

using Iberian diving beetles. *Diversity and Distributions*, 17 (1), 163–171. doi:10.1111/ddi.2010.17.issue-1

Schoener, T.W., 1968. The anolis lizards of Bimini: resource partitioning in a complex fauna. *Ecology*, 49 (4), 704–726. doi:10.2307/1935534

Šímová, P., *et al*., 2019. Fine scale waterbody data improve prediction of waterbird occurrence despite coarse species data. *Ecography*, 42, 511–520. doi:10.1111/ecog.03724

Sor, R., *et al*., 2017. Effects of species prevalence on the performance of predictive models. *Ecological Modelling*, 354, 11–19. doi:10.1016/j.ecolmodel.2017.03.006

Sun, Y., *et al*., 2017. Climatic suitability ranking of biological control candidates: a biogeographic approach for ragweed management in Europe. *Ecosphere*, 8 (4), e01731. doi:10.1002/ecs2.1731

Syfert, M.M., Smith, M.J., and 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. doi:10.1371/journal.pone.0055158

Syphard, A.D. and Franklin, J., 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, 32 (6), 907–918. doi:10.1111/eco.2009.32.issue-6

Tessarolo, G., *et al*., 2014. Uncertainty associated with survey design in species distribution models. *Diversity and Distributions*, 20 (11), 1258–1269. doi:10.1111/ddi.12236

Thibaud, E., *et al*., 2014. Measuring the relative effect of factors affecting species distribution model predictions. *Methods in Ecology and Evolution*, 5 (9), 947–955. doi:10.1111/2041-210X.12203

Tingley, R., *et al*., 2018. Integrating transport pressure data and species distribution models to estimate invasion risk for alien stowaways. *Ecography*, 41 (4), 635–646. doi:10.1111/ecog.02841

Václavík, T. and Meentemeyer, R.K., 2012. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity and Distributions*, 18 (1), 73–83. doi:10.1111/j.1472-4642.2011.00854.x

Varela, S., *et al*., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37 (11), 1084–1091.

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, 36 (12), 2290–2299. doi:10.1111/jbi.2009.36.issue-12

Williams, V.L. and Crouch, N.R., 2017. Locating sufficient plant distribution data for accurate estimation of geographic range: the relative value of herbaria and other sources. *South African Journal of Botany*, 109, 116–127. doi:10.1016/j.sajb.2016.12.015

Wisz, M.S., *et al*., 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distribution*, 14 (5), 763–773. doi:10.1111/j.1472-4642.2008.00482.x

Ye, X., *et al*., 2018. Impacts of future climate and land cover changes on threatened mammals in the semi-arid Chinese Altai Mountains. *Science of the Total Environment*, 612, 775–787. doi:10.1016/j.scitotenv.2017.08.191

Zurell, D., *et al*., 2010. The virtual ecologist approach: simulating data and observers. *Oikos*, 119 (4), 622–635. doi:10.1111/j.1600-0706.2009.18284.x