Research

Ecography 39: 001–012, 2016
doi: 10.1111/ecog.02414

Subject Editor: Christine Meynard. Editor-in-Chief: Miguel Araújo. Accepted 10 August 2016

Performance tradeoffs in target-group bias correction for species distribution models

Nathan Ranc, Luca Santini, Carlo Rondinini, Luigi Boitani, Françoise Poitevin, Anders Angerbjörn and Luigi Maiorano

N. Ranc (nathan.ranc@gmail.com), Organismic and Evolutionary Biology Dept, Harvard Univ., MA, Cambridge, USA, and Dept of Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy. – L. Santini, C. Rondinini, L. Boitani and L. Maiorano (http://orcid.org/0000-0002-2957-8979), Dept of Environmental Sciences, Radboud Univ., GL Nijmegen, the Netherlands. – F. Poitevin, CEFE UMR 5175, CNRS – Univ. de Montpellier – Univ. Paul-Valéry Montpellier – EPHE – laboratoire Biogéographie et Ecologie des vertébrés, Montpellier, France. – A. Angerbjörn, Dept of Zoology, Stockholm Univ., Stockholm, Sweden.

Species distribution models (SDMs) are often calibrated using presence-only datasets plagued with environmental sampling bias, which leads to a decrease of model accuracy. In order to compensate for this bias, it has been suggested that background data (or pseudoabsences) should represent the area that has been sampled. However, spatially-explicit knowledge of sampling effort is rarely available. In multi-species studies, sampling effort has been inferred following the target-group (TG) approach, where aggregated occurrence of TG species informs the selection of background data. However, little is known about the species- specific response to this type of bias correction.

The present study aims at evaluating the impacts of sampling bias and bias correction on SDM performance. To this end, we designed a realistic system of sampling bias and virtual species based on 92 terrestrial mammal species occurring in the Mediterranean basin. We manipulated presence and background data selection to calibrate four SDM types. Unbiased (unbiased presence data) and biased (biased presence data) SDMs were calibrated using randomly distributed background data. We used real and TG-estimated sampling efforts in background selection to correct for sampling bias in presence data.

Overall, environmental sampling bias had a deleterious effect on SDM performance. In addition, bias correction improved model accuracy, and especially when based on spatially-explicit knowledge of sampling effort. However, our results highlight important species-specific variations in susceptibility to sampling bias, which were largely explained by range size: widely-distributed species were most vulnerable to sampling bias and bias correction was even detrimental for narrow-ranging species. Furthermore, spatial discrepancies in SDM predictions suggest that bias correction effectively replaces an underestimation bias with an overestimation bias, particularly in areas of low sampling intensity. Thus, our results call for a better estimation of sampling effort in multispecies system, and cautions the uninformed and automatic application of TG bias correction.

Species distribution models (SDMs) are widely used in ecology and conservation biology to answer multiple research questions such as forecasting changes in distributions under global change scenarios and spatial priority setting for conservation (Guisan and Thuiller 2005, Guisan et al. 2013). In an ideal setting, a species distribution is modeled using presence and absence data (P-A; see Guisan and Zimmermann 2000 for a review). However, absences are lacking in most datasets, and in any case their reliability for mobile and/or elusive species may be highly questionable. As a result, many advances in SDMs have focused on presence-only modeling (P-O), in which background points (or pseudo-absences) represent available environmental conditions against which presence points are contrasted (see Pearce and Boyce 2006 for a review). It has now been widely acknowledged that the selection procedure of these background points is a central question to the fitting of statistical distribution models to P-O datasets (i.e. model calibration; Chefaoui and Lobo 2008, Vanderwal et al. 2009, Barbet-Massin et al. 2012).

A fundamental assumption in SDM calibration is the unbiased sampling of available conditions in the environmental space (Araújo and Guisan 2006, Pearce and Boyce 2006). However, sampling bias is a widespread issue since the vast majority of SDMs are calibrated with opportunistic data, which are not collected following strict random or systematic sampling methods (Yackulic et al. 2012). This issue is even more problematic in global datasets (Loiselle et al. 2003, 2008, Newbold 2010, Boitani et al. 2011). If the existing geographic bias correlates with gradients in environmental conditions, P-O SDM predictions and performance are likely affected (Kadmon et al. 2004, Phillips et al. 2009). This might happen when observers have

a higher likelihood to visit a given set of areas (e.g. along a road network or inside protected areas; Varela et al. 2014, Fernández and Nakamura 2015). At larger spatial scales, this may arise from very heterogeneous knowledge of species distributions (Boitani et al. 2011) or when investigators have unequal access to regional datasets (Reddy and Davalos 2003). Furthermore, P-O SDMs are particularly vulnerable to sampling bias, when compared to P-A SDMs, since background points are generally randomly drawn across the area of interest instead of representing the environmental conditions of the sampled area (Phillips et al. 2009, Yackulic et al. 2012). The resulting predictions from these naïve SDMs reflect the joint distributions of species probability of presence and sampling effort (Soberón and Nakamura 2009, Elith et al. 2011, Merow et al. 2013, Guillera-Arroita et al. 2015).

Methods to control for sampling bias have received increasing interest in recent years, and mainly focused on three approaches: model manipulation, occurrence filtering and background manipulation. Statistical models have been modified to account for spatial autocorrelation in occurrence data (De Marco et al. 2008, Václavík et al. 2012) or to explicitly model sampling bias using relevant covariates (e.g. accessibility; Warton et al. 2013, Fithian et al. 2015). Such model manipulation is a promising avenue of research, but poses analytical complexities and increased data requirements, which is not always available to all SDM users. Alternatively, many studies have focused on reducing occurrence data aggregation via data filtering (or thinning), conducted both in the geographic (Veloz 2009, Anderson and Raza 2010, Carroll 2010, Verbruggen et al. 2013, Beck et al. 2014, Aiello-Lammens et al. 2015) and environmental spaces (de Oliveira et al. 2014, Varela et al. 2014). Similarly, occurrence records have been attributed weights to reduce bias intensity (Elith et al. 2010, Ancillotto et al. 2016).

Here we focus on background manipulation methods. These aim at selecting background points in areas effectively sampled, thereby controlling for survey effort (Phillips et al. 2009). If sampling effort is known, which is seldom the case, then background points can simply be generated accordingly. Otherwise, sampling intensity can be surrogated following a target-group (TG) approach (Ponder et al. 2001, Anderson 2003, Phillips et al. 2009), often the only option available when modeling multiple species in the same area. The assumption behind TGs is that several species can be pooled according to a common search pattern or interest from observers (e.g. all birds in an area share the same bias in sampling effort because of well-known bird watching spots). If this assumption holds, TG presence data can be used as background points to approximate the unknown sampling effort (Phillips et al. 2009). Similar TG bias-corrections have been successfully employed to improve SDM performances and predictions in a variety of contexts (Mateo et al. 2010, Milanovich et al. 2010, Syfert et al. 2013, Hertzog et al. 2014). However, species-specific responses to both sampling bias and potential corrections remain largely unknown (Warton et al. 2013, Stolar and Nielsen 2015).

Investigating the impact of sampling bias on SDM performance and the potential benefits of bias correction relies on availability of comprehensive and unbiased evaluation data (Phillips et al. 2009, Syfert et al. 2013). This issue can be circumvented by using virtual species (Hirzel et al. 2001), for which species distributions are simulated (i.e. truth is known). Virtual species are increasingly used to study methodological aspects of SDMs e.g. sampling strategies (Kent and Carmel 2011, Thibaud et al. 2014), pseudoabsence selection (Barbet-Massin et al. 2012), and sampling bias corrections (Kramer-Schadt et al. 2013, Fourcade et al. 2014, Varela et al. 2014, Stolar and Nielsen 2015). In this study, we designed a system of multiple virtual species aiming at 1) evaluating the impacts of sampling bias on SDM performance, 2) investigating species-specific predictors of vulnerability to sampling bias and usefulness of bias correction, and 3) assessing the performance of TG bias correction. To this end, we designed a realistic system of 92 range-based virtual species and sampling bias based on real mammal species occurring in the Mediterranean basin as well as several independent, true virtual species. We generated multiple sets of presence and background points to calibrate unbiased, biased and bias-corrected SDMs. All SDMs were then evaluated against the corresponding virtual distributions. Finally, we investigated the different ecological correlates associated with the severity of sampling bias and the usefulness of bias correction.

Methods

Species data

Our study area covers more than 2 million km^2 and stretches over the Mediterranean biodiversity hotspot (Supplementary material Appendix 1, Fig. A1; Myers et al. 2000). In order to inform our realistic virtual species system, we first gathered all available presence points (n = 250 476) for native terrestrial mammals, from multiple data providers: online resources, museum collections, atlases, scientists, peer-reviewed articles, environmental consultancies and NGOs (see Supplementary material Appendix 2, Table B1 for details). Following Boitani et al. (2011), we filtered the dataset by removing possible duplicated records, occurrences outside known species distribution ranges or with unknown coordinate precision or coarser than 10 km, and data of species whose identification may have been uncertain (only for citizen science database and only considering species potentially hard to distinguish from field-based samples e.g. *Martes martes* vs *Martes foina*). We limited our analyses to species with at least 30 distinct locations, obtaining a final set of 92 species (34% of the native mammals in the area, Supplementary material Appendix 3, Table C1). The minimum number of 30 locations was chosen heuristically considering also the number of environmental predictors, and it is well in the range of what is commonly suggested as the minimum sample size useful for calibrating SDMs (Kadmon et al. 2003, 2004, Boitani et al. 2011, van Proosdij et al. 2016).

Environmental predictor variables

We initially considered for model calibration all 19 WorldClim variables (Hijmans et al. 2005). These have been recognized as important in characterizing the climate of the

Mediterranean basin (Maiorano et al. 2011). All variables were resampled at 5 km resolution. In order to limit multicollinearity, we first identified all possible sets of uncorrelated covariates (|Pearson's r| < 0.7). Among these different combinations, we selected the particular sets containing the maximum number of uncorrelated predictors, and with lowest mean correlation among covariates. The final set of bioclimatic variables contained six predictors: mean temperature, annual range, mean temperature of the wettest quarter, mean temperature of the driest quarter, precipitation of the wettest quarter, precipitation of the warmest quarter. All analyses were conducted in an Albers Equal Area Conic projection, at a resolution of 5×5 km.

Range-based virtual species approach

We generated 92 range-based virtual species distributions using an SDM approach based on real-world species geographic ranges obtained from the IUCN Red List (<www.iucnredlist.org>). Specifically, we calibrated Maxent SDMs using centroids of all raster cells within the geographic ranges as presences, all the cells of the study area as background, and the pre-selected set of environmental variables, to obtain continuous virtual species distributions. In subsequent analyses, we considered that the descriptions generated by Maxent represented the full reality of our virtual species. This approach has the advantage of generating realistic virtual ranges with known environment–occurrence relationships. However, a complementary virtual species approach was also implemented. Full details on the procedure are available in Supplementary material Appendix 4, Fig. D1. We then pooled these 92 range-based virtual species into three TGs: bats TG (27 species), small mammal TG (37 species of rodents and insectivores), and large mammal TG (28 species, mainly carnivores, artiodactyls and lagomorphs; see Supplementary material Appendix 3, Table C1 for details). The assumption is that species composing each TG are likely targeted by similar observers with relatively homogeneous search patterns. In our dataset, bat presence data were largely collected by means of bat detectors, mist nets and roost inventories; small mammals by live-trapping and owl pellet analysis; and large mammals by opportunistic observations and road-kills. Four rodents (*Castor fiber*, *Hystrix cristata*, *Marmota marmota* and *Sciurus vulgaris*) and one insectivore (*Erinaceus europaeus*) were pooled in the large mammal TG as their identification and detection do not require specific training or live trapping.

For each TG, we simulated a virtual sampling effort considering the spatially explicit density of the empirical occurrences available for each species. In particular, we interpolated the species occurrences using a normally distributed kernel of 10 km standard deviation (value chosen as a pragmatic compromise between movement abilities of different mammals) and then rescaled between 0 and 1 (Fig. 1). This method was used to produce a realistic spatially explicit sampling bias input for our virtual species system (hereafter referenced to as true or known sampling effort) instead of relying on arbitrary sampling intensities. Therefore, we did not aim here to estimate empirical TG sampling intensities per se, as this would require additional factors (e.g. TG probability of presence, species richness, and species- and habitat-specific detectability).

True virtual species

Since the species ranges and species occurrence data respectively used to generate the range-based virtual species and the realistic sampling intensities may be prone to similar sampling biases, we generated a set of independent virtual species (hereafter referred to as true virtual species), using the R package 'virtualspecies' (Leroy et al. 2016). Specifically, we simulated two species with different relative occurrence areas (ROA): a wide-ranging (ROA = 0.72, presence sample size = 1000) and a narrow-ranging virtual species (ROA = 0.22, presence sample size = 100) for each TG (see Supplementary material Appendix 4 for details).

SDM calibration sets

We generated four SDM scenarios to test the effects of sampling bias and bias correction on SDM performance (Fig. 2). Unbiased SDMs represented the ideal case in which presence data are unbiased and background points are randomly

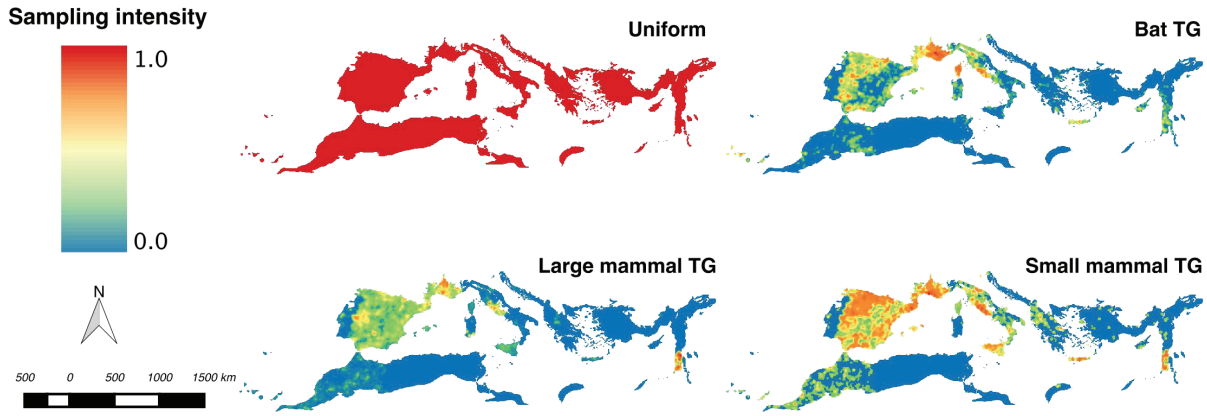

Figure 1. Geographic patterns of sampling strategies: uniform, and biased sampling intensities for the three target-groups (TG) i.e. bats, large mammals and small mammals.

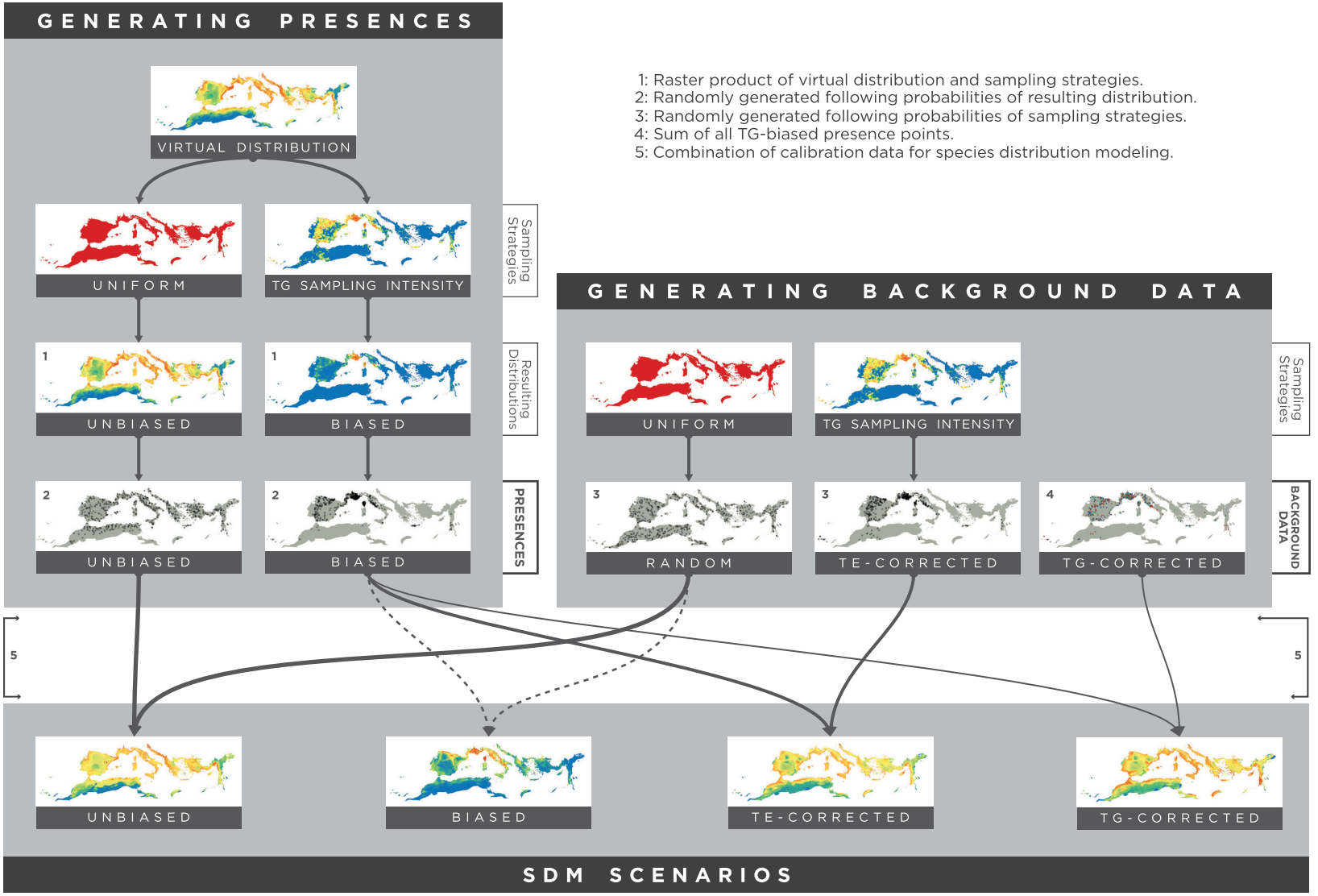

Figure 2. General methods scheme: on the one hand, the intersection of the virtual species distribution with the two sampling strategies provide the species sampling distributions (1) according to which both types of presence records (2) are generated. On the other, two types of background points i.e. random and true effort (TE) corrected (3) are generated based on uniform and target-group (TG) sampling intensities (or biases), respectively. In addition, biased presences of all species across a given TG are combined to obtain the TG-corrected background points (4). These two presence and three background data sets are combined to calibrate four SDM scenarios (5). We illustrated the scheme using data from *Myotis capaccinii* (sample size n = 359), of the bat TG (number of species s = 27). We mapped only a sample of occurrence points to improve figure clarity.

distributed across the study area. Biased SDMs represented the case in which presence data are geographically (and/or environmentally) biased and no attempt is made to address this issue in background points selection (i.e. background points again distributed randomly). Then, we generated two SDM scenarios based on biased presence data for which background points were manipulated to correct for sampling bias. In the first scenario called true effort corrected SDMs (TE-corrected), background points were sampled based on known sampling effort. In the second scenario called target group corrected SDMs (TG-corrected), we used target group presence data as background points to estimate sampling effort.

As for the presence data, species-specific sets were drawn according to the species probability of presence, defined by the range-based virtual species distributions, intersected with two different sampling strategies: 1) uniform for unbiased presences, and 2) TG sampling intensities for biased presences. For each species, we randomly selected 10 sets of n occurrences without replacement, where n corresponded to the number of occurrences available for that species in the real dataset.

As for the background data, we generated a first set of naïve background points (i.e. unbiased) taken at random within the study area (Barbet-Massin et al. 2012). TE-corrected background points were obtained by selecting background data in proportion to TG sampling intensities i.e. the known TG sampling efforts (Merow et al. 2013). Accordingly, the probability of a given raster cell to be selected as background was proportional to its sampling intensity. This is essentially similar to the DebiasAverages approach available in Maxent (Dudík et al. 2005). We defined the TG-corrected background points as simply the aggregation of all biased presences of a given TG used for SDM calibration (i.e. 34 219, 35 317 and 45 740 points for the bat, small mammal and large mammal TGs respectively). Please note that these refer to the presences generated for the virtual system, and not the empirical TG occurrences. As for unbiased and TE-corrected background points we generated 10 sets matching the above quantities of TG bias-corrected background points to control for the effect of background sample size. We combined these two sets of presence and three sets of background data to generate the four SDM scenarios (Fig. 2).

Species distribution modeling

We predicted species distributions using the maximum-entropy algorithm implemented in the MaxEnt software (Phillips et al. 2006). Among the multiple P-O algorithms available, MaxEnt has been recognized to perform well (Elith et al. 2006, Hernandez et al. 2006, Pearson et al. 2007). We used default settings in MaxEnt – i.e. all feature classes, convergence threshold of 10^{-5}, maximum iterations (200) and a regularization multiplier of 1 – which have provided reliable model predictions (Phillips and Dudík 2008). SDMs were calibrated and projected within the entire study area.

SDMs are usually evaluated using discrimination performance metrics (e.g. area under the curve, AUC, or true skill statistic, TSS) based on evaluation data. In a virtual system, using such metrics would be of little relevance since we can compare the SDM outputs with the true distributions (Fourcade et al. 2014). Agreement between SDM outputs and the range-based virtual species distributions was calculated using the Schoener's D index (D_{GEO}; Schoener 1968). This metric measures the absolute spatial agreement between these two continuous predictions (Fourcade et al. 2014), and ranges from 0 (no overlap) to 1 (perfect overlap).

Analyses

Environmental sampling bias

We evaluated the intensity of environmental bias introduced in the virtual system by measuring niche overlap in the environmental space between sets of points randomly drawn from the four sampling regimes (uniform and three TG sampling intensities i.e. one for each TG; for a total of 6 comparisons). We used the PCA-env method described in Broennimann et al. (2012) to calculate the niche overlap, which ranges from 0 (no overlap) to 1 (equivalent niches). Specifically, we calculated mean niche overlap (and 95% confidence interval, CI) across 10 replicates of 40 000 random points for a total of 100 overlap measures, and considering all six environmental predictor variables mentioned previously.

Testing differences in SDM scenarios

We used a generalized linear mixed model to test the difference in SDM performance (D_{GEO}) between the four SDM scenarios, and their interaction with TGs. The model assumed a Gaussian error distribution and an identity link function. The fixed part of the formula was given by TGs scenarios nested under SDM, whereas the random part was given by SDM scenarios nested under species.

Ecological correlates of sampling bias and bias correction performances

We used Random Forest algorithm to quantify the importance of different potential predictors on the SDM vulnerability to 1) sampling bias, 2) the extent of improvements after TE bias correction, and 3) TG bias correction. Specifically, we used 1000 trees and sampled two random variables as candidates at each split of the trees. The three dependent variables were obtained as follows:

$$\text{Bias ratio} = \text{biased } D_{GEO}/\text{unbiased } D_{GEO} \quad (1)$$

$$\text{Improvement through TE-correction} = \text{TE-corrected } D_{GEO}/\text{biased } D_{GEO} \quad (2)$$

$$\text{Improvement through TG-correction} = \text{TG-corrected } D_{GEO}/\text{biased } D_{GEO} \quad (3)$$

In addition to TG, we used several dependent variables known to affect SDM performance and likely to influence the effect of sampling bias: presence sample size, average probability of presence across the study area (i.e. relative occurrence area, ROA), ratio between the presence sample size and relative occurrence area and niche breadth. Species-specific covariate values are provided in Supplementary material Appendix 3, Table C1. Presence sample size is a known factor affecting SDM performance, and it is usually assumed that SDMs perform better with higher sample size (Kadmon

et al. 2003, Chefaoui et al. 2011, Tessarolo et al. 2014). We calculated the relative occurrence area as average probability of presence of the given virtual species across the study area. SDMs calibrated for wide-ranging species (i.e. high relative occurrence area) have usually lower performance (Segurado and Araújo 2004, Newbold et al. 2009, Chefaoui et al. 2011, Tessarolo et al. 2014), and tend to be more affected by sampling bias (Loiselle et al. 2008, Stolar and Nielsen 2015). In addition, we predict species with higher ratio between the presence sample size and relative occurrence area to be less affected by sampling bias. Finally, we included an estimation of niche breadth in the model, since generalist species are usually more difficult to model (Kadmon et al. 2003) and are likely to be more affected by sampling bias. Niche breadth was calculated as mean tolerance value of unbiased calibration data (average over 10 sets) using the 'adehabitat' package in R (Calenge 2006).

Spatial discrepancies

We calculated the spatial discrepancies in predicted species probability of presence for each combination of model scenario and target group as follows

$$\text{Discrepancy} = \log_{10}\left(\frac{\text{predicted probability of presence}}{\text{virtual species probability of presence}}\right)$$

A positive log ratio indicates a model propensity to overestimate species probability of presence. Conversely, a negative ratio suggests a tendency to underestimate probability of presence. We rescaled both the predicted and the range-based virtual species probability of presence between 0.01 and 1 to obtain log ratio values bounded to [–2, 2].

Softwares and packages

All statistical analyses were performed in R (R Core Team), using the packages 'nlme' (Pinheiro et al. 2015). All spatial analyses were performed using GRASS GIS (GRASS Development Team 2014), QGIS (Quantum GIS Development Team 2014) and PostGIS library in PostgreSQL (PostgreSQL Development Core Team 2014). Database operations were conducted in PostgreSQL.

Results

Environmental sampling bias

The geographic biases introduced into the virtual system translated into strong environmental biases as demonstrated by the low environmental niche overlap between points drawn from a uniform (i.e. no bias) and the three TG biased effort distributions: 0.375 (95% CI = 0.374–0.376) for bats, 0.532 (0.531–0.534) for small mammals, and 0.483 (0.482–0.484) for large mammals. The environmental biases were relatively similar among TGs as revealed by high niche overlap between points drawn from pairs of TGs: mean overlap between large and small mammal TGs was 0.718 (0.717–0.720); mean overlap between the large mammal and bat TGs was 0.689 (0.688–0.691); and mean overlap between the small mammal and bat TGs was 0.573 (0.571–0.575).

SDM performance

The linear mixed model indicated that model performance varied significantly among SDM scenarios ($F = 53.49$; $DF = 265$; $p < 0.0001$, Fig. 3). Unbiased SDMs achieved the highest performance (mean $D_{GEO} = 0.92$, $\sigma^2 = 0.039$) and biased SDMs the lowest (0.82, 0.162). Performance was intermediate for TE-corrected (0.90, 0.047) and TG-corrected SDMs (0.87, 0.064). Tukey's post hoc test revealed that unbiased and TE-corrected SDMs were not significantly different ($\alpha = 0.05$), and so were TE-corrected and TG-corrected. Unbiased, TE-corrected and TG-corrected SDMs achieved higher performance than biased SDMs, and unbiased SDMs perform significantly better than TG-corrected SDMs. All post hoc comparison are reported in Supplementary material Appendix 5, Table E1.

Within SDM scenarios, performance significantly changed across target groups ($F = 19.96$; $DF = 265$; $p < 0.0001$, Fig. 3). Across all four scenarios, SDMs achieved highest performances for small mammals (mean $D_{GEO} = 0.91$, $\sigma^2 = 0.055$) and lowest for bats (0.82, 0.123). Species-specific results are reported in Supplementary material Appendix 3, Table C1. Differences among TGs were especially strong for the biased SDMs. All post hoc comparison are reported in Supplementary material Appendix 5, Table E2. Bias-corrected SDMs greatly outperformed biased SDMs for bats and large mammals. However, biased SDMs did better than TG-corrected and similarly to TE-corrected SDMs for small mammals. Noticeably, bias correction reduced SDM performance for some species, but not in the bat TG. Irrespective of SDM scenario, differences were significant between all pairs of target groups. Similarly, model accuracy for the true virtual species was highly dependent on both the SDM scenarios and the individual virtual species (Fig. 3).

Covariates of model's bias and bias correction

The relative occurrence area (ROA) was the single most important predictor of vulnerability to sampling bias and improvement through bias correction, with wide-ranging species being more affected by bias (i.e. lower bias ratio) and also benefiting from the highest improvement through bias correction (Fig. 4). The other potential predictors were less important, but showed expected relationships. Species with large niche breadth and presence sample size were more vulnerable to sampling bias and more likely to benefit from bias correction, whereas the ratio between sample size and occurrence area (rNROA) showed an opposite trend. Bats were more affected by sampling bias and had higher improvements following correction than small and large mammals (Supplementary material Appendix 6, Fig. F1–F4). All the three models explained a high proportion of variance (97.96–98.63%).

Similar patterns were obtained for the true virtual species. The wide-ranging species were strongly affected by sampling bias and benefited most from bias-correction. By contrast, the narrow-ranging species tended to be negatively affected by bias-correction. In addition, the bias ratio and improvement through both TE- and TG-corrections as a function

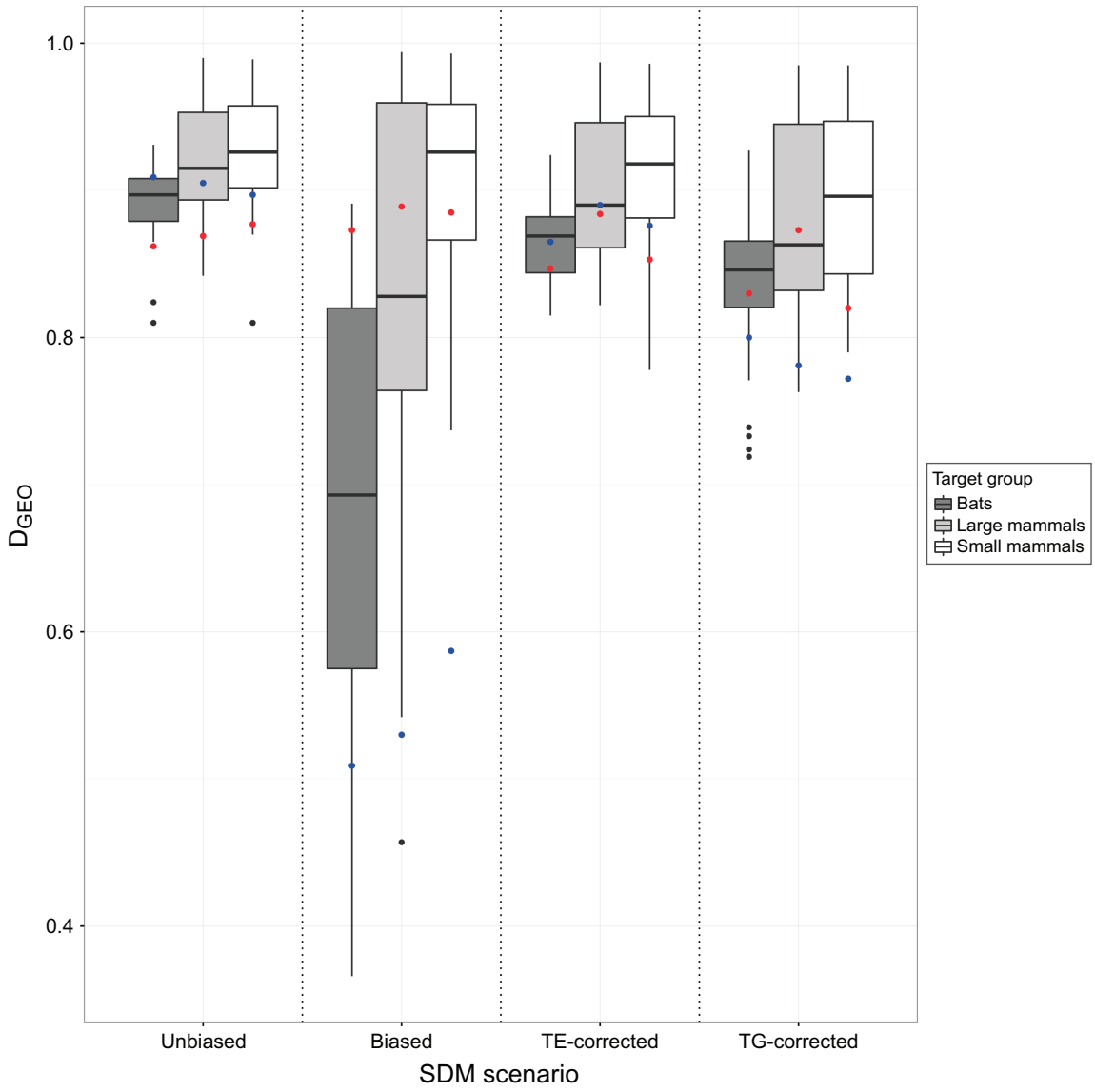

Figure 3. Overall SDM performance (DGEO, y-axis) as a function of SDM scenario (x-axis) and target-group (key). Results for the true virtual species are plotted in blue (wide-ranging species) and red (narrow-ranging species).

of ROA fitted well the results obtained for the range-based virtual species (Fig. 4).

Spatial discrepancies

Spatial discrepancies between predicted and virtual species probability of presence varied considerably between model scenario and target groups (Fig. 5). Overall, unbiased models overestimated species probability of presence (mean log ratio = 0.18 for bats, 0.18 for large mammals and 0.21 for small mammals), and so did TE-corrected (0.22, 0.22 and 0.28, respectively) and TG-corrected SDMs (0.23, 0.23 and 0.34, respectively). By contrast, discrepancies for the biased models depended more heavily on the TG under consideration (Fig. 5): severe underestimation of species probability of presence for bats (–0.45), relatively strong underestimation as well as some overestimation for large mammals (–0.17), and both overestimation and underestimation for small mammals translating into a mean log ratio close to 0 (–0.05). The observed discrepancies between predicted and true species probability of presence (Fig. 5) tended to be located in areas of low sampling intensity (Fig. 1). This was especially pronounced for the biased SDMs, and to a lower extent for TE- and TG-corrected models.

Discussion

Although target-group bias correction has largely been applied to opportunistic datasets (e.g. museum collections, citizen-based data), which likely suffer from strong environmental biases, surprisingly few studies have investigated species-specific responses to bias correction (but see Warton et al. 2013, Stolar and Nielsen 2015). These can be difficult to appreciate with real biodiversity data (Lobo and Tognelli 2011) since it is often challenging to obtain unbiased and sufficient data for reliable evaluation. If SDMs were to be

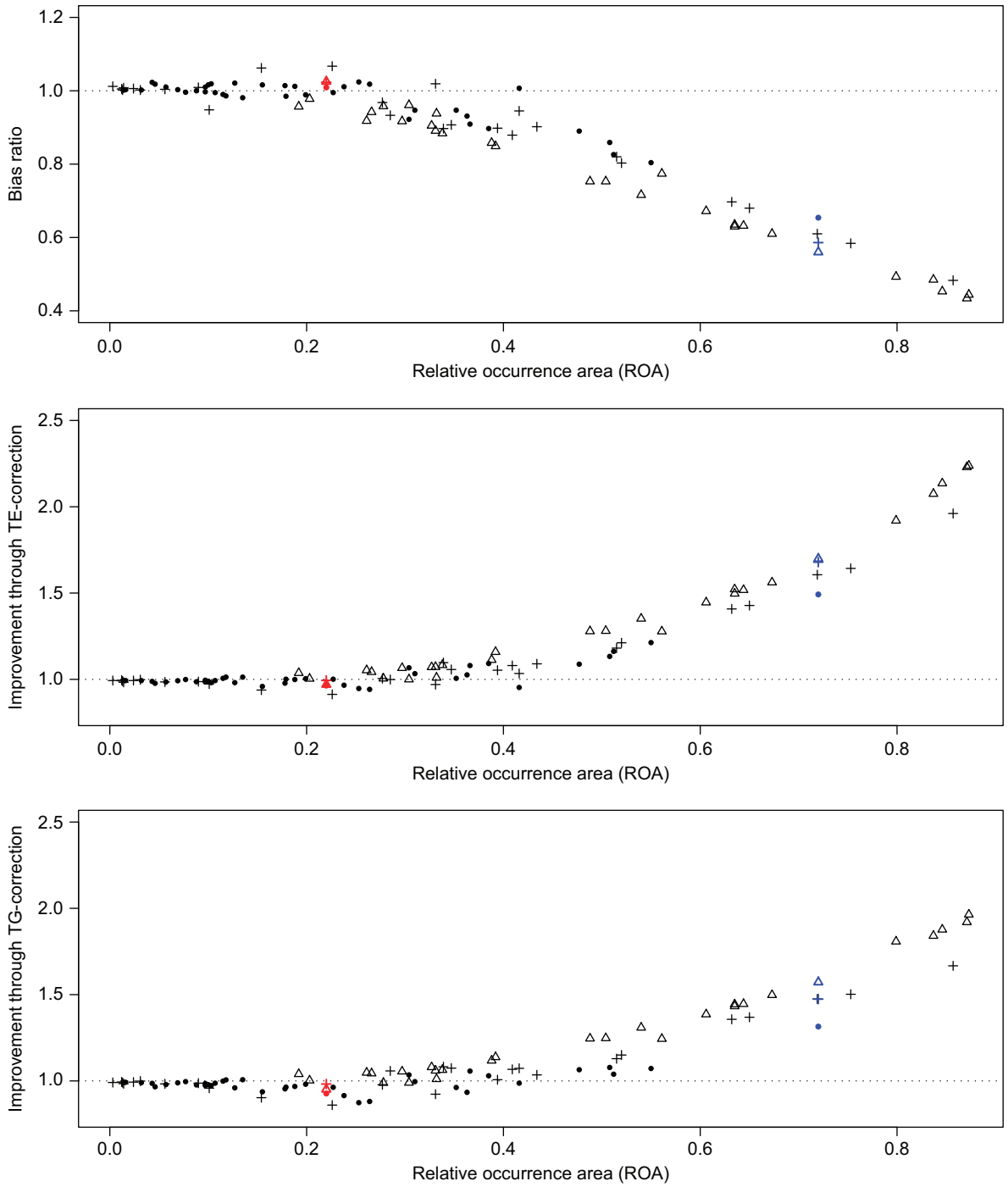

Figure 4. Effect of relative occurrence area on bias ratio (upper panel), improvement through true effort (TE) correction (central panel) and through target group (TG) correction (lower panel). Results are plotted for the three target-groups: bats (triangle), large mammals (cross) and small mammals (point), as well as for the true virtual species (wide-ranging species in blue and narrow-ranging species in red).

evaluated with test data mimicking the biases of the calibration data, biased models would likely identify larger areas as unsuitable (i.e. have a higher specificity), and thus, artificially outperform the corrected SDMs. This limitation of using autocorrelated real-world data to evaluate SDMs has been shown to impact the Area Under the Curve (AUC, Veloz 2009) and to lead to wrong conclusions regarding performance of bias-corrected SDMs (Bystriakova et al. 2012, Syfert et al. 2013, Hertzog et al. 2014). In this context, virtual species offer a very appealing alternative since truth is known.

We designed a system of multiple range-based virtual species based on real-world data in order to represent a realistic situation. Modelling virtual distributions using an SDM approach based on extant species data, as opposed to user-defined characterization of species niches (Hirzel et al. 2001, Duan et al. 2015, Leroy et al. 2016), enabled us to generate a high diversity of realistic species traits (sample

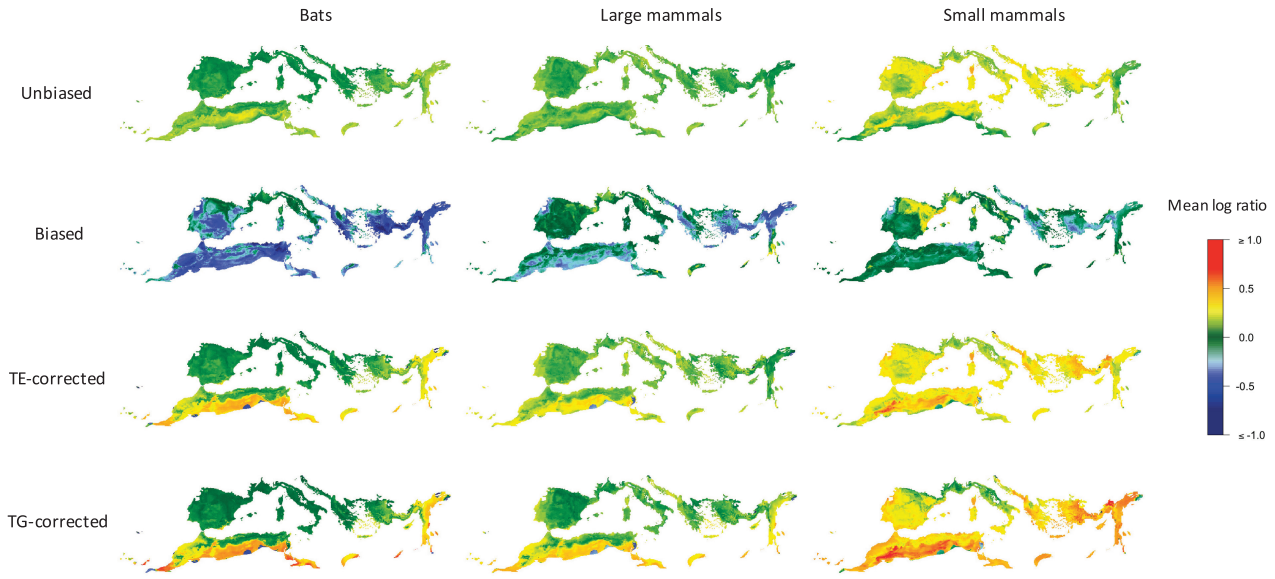

Figure 5. Spatially explicit discrepancies in predicted species probability of presence for different combinations of target groups (n = 3) and SDM scenarios (n = 4), expressed as mean log ratio of predicted over virtual species probability of presence. Positive values indicate overestimation of species probability of presence. Respectively, negative values indicate underestimation of species probability of presence.

size, relative occurrence area, niche breadth) and a plausible species richness pattern. Since we could not assert that species ranges used to model range-based virtual species were not affected by similar underlying biases than those present in the occurrence dataset, which was used to generate the target group sampling biases, we incorporated independent, true virtual species in our analyses. The remarkable congruence in the results obtained with both approaches reinforces our findings discussed hereafter.

The realistic geographic sampling bias introduced into the virtual species system translated into environmental bias, which had two main consequences. First, sampling bias had a deleterious effect on SDM performance, despite presence and background points used for SDM calibration being sampled within the same geographical ranges. In real-world analyses an inadequate choice of study area extent – a known factor influencing SDM predictions and performance (Barve et al. 2011) – would likely exacerbate the sampling bias issue (e.g. by considering portions of the environmental space devoid of actual sampling as background data). Second, bias-corrected SDMs outperformed biased SDMs, which supports previous studies highlighting the benefits of sampling bias correction, and especially those regarding target-group bias correction (Phillips et al. 2009, Mateo et al. 2010, Syfert et al. 2013, Hertzog et al. 2014). However, these aggregated performance results hide complex species-specific response patterns.

Different species are unlikely to be similarly affected by sampling bias or equally amenable to bias correction (Warton et al. 2013, Stolar and Nielsen 2015) – depending on range size, coverage by sampling, and habitat specificity. We found that relative occurrence area was the single most important factor driving species vulnerability to sampling bias thereby widespread species were more affected and more likely to benefit from bias correction than narrow-ranging species. This finding indicates that relative occurrence area is a critical factor in determining species vulnerability to sampling bias (Loiselle et al. 2008, Stolar and Nielsen 2015), and model performance overall (Jiménez-Valverde et al. 2008, Lobo et al. 2008, Chefaoui et al. 2011). Although species characterised by wide niche breadth and large presence sample size were more affected by sampling bias, and more likely to benefit from bias correction the correlations of these two predictors with relative occurrence area prevented us to isolate their specific contributions. Beyond the species correlates (i.e. proximate factors) analyzed here, the vulnerability of a given species distribution model to sampling bias must ultimately depend on the extent to which geographic sampling bias translates into a bias in the environmental predictors influencing species distributions.

The lower performance of target-group bias correction (TG-corrected SDMs) as compared to true effort bias correction (TE-corrected) can be explained by two factors. First, where both presence and background data spatially overlap, background data may be uninformative in discriminating presences from available conditions, thereby weakening the model discriminatory abilities (Lütolf et al. 2006). TG-corrected SDMs are especially vulnerable to this contamination since background points are directly drawn from the set of occurrences, whereas they are generated from interpolated occurrence densities in TE-corrected SDMs. Although Phillips et al. (2009) found weak impact of contamination, we believe it may severely affect species with large presence sample size. Second, the differences in performance observed between both SDM scenarios suggests that the density of target-group occurrences (TG-corrected points) did not always provide a robust estimate of the sampling effort introduced in the virtual species system. This can be explained by the confounding effect of spatially-heterogeneous species richness patterns. For instance, given a constant sampling effort, it can reasonably be expected to obtain more occurrences in areas of higher species richness. If species richness is highly variable in the study area, TG bias correction is effectively substituting a sampling bias with

a species richness bias (Warton et al. 2013, Fernández and Nakamura 2015). In our study, species richness was highly heterogeneous, which probably lead to an overestimation of sampling effort in areas of high species richness by the TG-corrected SDMs, and ultimately, spatial stratification in SDM errors.

Investigating spatially-explicit predictions indicated that discrepancies between the different SDM scenarios and the virtual species distributions tend to occur in areas of low sampling intensity. There, biased SDMs underestimated species probability of presence whereas bias-corrected SDMs tended to overestimate (especially TG-corrected SDMs). The relative performances of bias-corrected and biased SDMs (as estimated by the D_{GEO} metric) is likely due to relative intensity of overestimation by bias-corrected SDMs compared to the intensity of underestimation by biased SDMs in areas of low sampling intensity. It is plausible that the relative performance of biased and bias-corrected SDMs is largely dependent on the study area and more precisely on the extent of high versus low sampling areas. Where sampling is concentrated in restricted zones of the study area, biased SDMs will inevitably underestimate species probability over large areas, thereby drastically reducing their performance. Such effect of study area extent has been found to cause inflation of the AUC (Lobo et al. 2008). Modelers are thus more likely to see improvements through bias-control where large tracts of the study area correspond to environmentally under-sampled areas. Furthermore, the discrepancies between predicted and virtual species probability of presence show that bias-control tends to replace a strong underestimation by a milder yet noticeable overestimation of species probability of presence – especially in areas of low sampling intensity. This explains the overall better performance of these models as compared to the biased SDMs. Generally speaking, biased SDMs seem to be more prone to underestimation of species probability of presence (omission errors) whereas bias-controlled SDMs seem to be more sensitive to overestimation (commission errors). These two types of errors may not have the same implications depending on the question of interest (Rondinini et al. 2006).

Ultimately, observed presence data result from two concurrent processes: sampling and detection (Fernández and Nakamura 2015). This study has focused on sampling and demonstrated that 1) sampling bias severely affects model performance and 2) that spatially-explicit knowledge of sampling effort can be used to improve model performances for presence-only datasets in conditions of sampling bias. However, detectability is an equally important issue confounding occurrence probability in SDMs (Yackulic et al. 2012). Provided that repeated sampling exists, occupancy models (MacKenzie et al. 2006) can be used to model species distributions accounting for imperfect detections and have been shown to improve model predictions (Rota et al. 2011). In the present study, we assumed perfect and uniform detection probabilities when sampling presences for model calibration. Future work should assess the concomitant effect of sampling and detection biases on model performance.

It has been suggested that target-group bias correction should be applied by default where sampling bias may occur. However, our results highlight important species-specific responses to this background manipulation, which, in some cases, can even prove to decrease model performance. These results corroborate previous findings by Warton et al. (2013) and Stolar and Nielsen (2015). It seems that the usefulness of target-group bias correction is highly dependent on the system investigated. The benefits are likely high where environmental bias is strong, species richness patterns relatively homogeneous and species wide-ranging. In this study, these characteristics were met by the bat target-group, but largely violated by the small mammal target-group, thereby explaining important target-group differences in SDM performance.

Based on the significant negative effects of bias on SDM performance measures should be undertaken to correct for sampling bias in presence-background settings. However, we caution the uninformed and ubiquitous application of TG bias correction as well as the belief that it provides an alternative to rigorous sampling schemes. Instead, we advocate for a more mechanistic understanding of this method. We suggest that the attention should be focused on a more essential question: how to estimate environmental sampling bias from an opportunistic dataset containing multiple species, while considering species richness patterns?

Acknowledgements – We are very grateful to all data contributors and especially H. Ambarli, S. Aulagnier, P. Bayle, F. Bego, P. Benda, P. Bergier, P. Georgiakakis, B. Gharaibeh, J. Hall, S. Heude, Y. Iliopoulos, Y. Kayser, M. Krofel, B. Krystufek, P. Lymberakis, T. Marijnissen, S. Meiri, D. Mertzanidou, Y. Mertzanis, J. Palomo, E. Papadatou, R. Pavisse, E. Rogozi, P. Rigaux, I. Selanec, N. Yigit, the ECO-MED ecology-consulting organization, the French Bat Inventory Initiative (MNHN, ONF, SFEPM), GBIF, the 'Groupe Chiroptères de Corse', the 'Groupe Chiroptères de Provence', the Israel Nature and Parks Authority, Meridionalis, the Natural History Museum of Crete, Observado, Silene PACA and the Tel Aviv Univ. Natural History Museum. In addition, we would like to acknowledge the LPO, whose partnership enabled us to access three major databases: Faune Drôme, Faune Ardèche and Faune PACA. We thank Stéphane Aulagnier, Daniele Baisero, Samy Gaiji and Piero Visconti for early comments on the manuscript, and Andrew R. Gipe (Diakron Inst.) for improving the graphics. Nathan Ranc also wish to thank GBIF for their support through the Young Researcher Award 2013 and Swarovski Optik for support.

References

Aiello-Lammens, M. E. et al. 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. – Ecography 38: 541–545.

Ancillotto, L. et al. 2016. Extraordinary range expansion in a common bat: the potential roles of climate change and urbanisation. – Sci. Nat. 103: 1–8.

Anderson, R. P. 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. – J. Biogeogr. 30: 591–605.

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. – J. Biogeogr. 37: 1378–1393.

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.

Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.

Barve, N. et al. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. – Ecol. Model. 222: 1810–1819.
Beck, J. et al. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. – Ecol. Inform. 19: 10–15.
Boitani, L. et al. 2011. What spatial data do we need to develop global mammal conservation strategies? – Phil. Trans. R. Soc. B 366: 2623–2632.
Broennimann, O. et al. 2012. Measuring ecological niche overlap from occurrence and spatial environmental data. – Global Ecol. Biogeogr. 21: 481–497.
Bystriakova, N. et al. 2012. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. – Syst. Biodivers. 10: 305–315.
Calenge, C. 2006. The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. – Ecol. Model. 197: 516–519.
Carroll, C. 2010. Role of climatic niche models in focal-species-based conservation planning: assessing potential effects of climate change on northern spotted owl in the Pacific Northwest, USA. – Biol. Conserv. 143: 1432–1437.
Chefaoui, R. M. and Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – Ecol. Model. 210: 478–486.
Chefaoui, R. M. et al. 2011. Effects of species' traits and data characteristics on distribution models of threatened invertebrates. – Anim. Biodivers. Conserv. 34: 229–247.
De Marco, P. et al. 2008. Spatial analysis improves species distribution modelling during range expansion. – Biol. Lett. 4: 577–580.
de Oliveira, G. et al. 2014. Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: a new approach based on environmentally equidistant records. – Ecography 37: 637–647.
Duan, R.-Y. et al. 2015. SDMvspecies: a software for creating virtual species for species distribution modelling. – Ecography 38: 108–110.
Dudík, M. et al. 2005. Correcting sample selection bias in maximum entropy density estimation. – Adv. Neural Inform. Process. Syst. 17: 1–8.
Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.
Elith, J. et al. 2010. The art of modelling range-shifting species. – Methods Ecol. Evol. 1: 330–342.
Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. – Divers. Distrib. 17: 43–57.
Fernández, D. and Nakamura, M. 2015. Estimation of spatial sampling effort based on presence-only data and accessibility. – Ecol. Model. 299: 147–155.
Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods Ecol. Evol. 6: 424–438.
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: e97122.
GRASS Development Team 2014. Geographic resources analysis support system. – Open Source Geospatial Foundation, USA.
Guillera-Arroita, G. et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. – Global Ecol. Biogeogr. 24: 276–292.
Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.
Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. – Ecol. Lett. 16: 1424–1435.
Hernandez, P. A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.
Hertzog, L. R. et al. 2014. Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. – Divers. Distrib. 20: 1403–1413.
Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.
Hirzel, A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – Ecol. Model. 145: 111–121.
Jiménez-Valverde, A. et al. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. – Divers. Distrib. 14: 885–890.
Kadmon, R. et al. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. – Ecol. Appl. 13: 853–867.
Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.
Kent, R. and Carmel, Y. 2011. Presence-only versus presence–absence data in species composition determinant analyses. – Divers. Distrib. 17: 474–479.
Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. – Divers. Distrib. 19: 1366–1379.
Leroy, B. et al. 2016. virtualspecies, an R package to generate virtual species distributions. – Ecography 39: 599–607.
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. – J. Nat. Conserv. 19: 1–7.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Loiselle, B. A. et al. 2003. Avoiding pitfalls of using species-distribution models in conservation planning. – Conserv. Biol. 17: 1591–1600.
Loiselle, B. A. et al. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? – J. Biogeogr. 35: 105–116.
Lütolf, M. et al. 2006. The ghost of past species occurrence: improving species distribution models for presence-only data. – J. Appl. Ecol. 43: 802–815.
MacKenzie, D. I. et al. 2006. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. – Elsevier.
Maiorano, L. et al. 2011. The future of terrestrial mammals in the Mediterranean basin under climate change. – Phil. Trans. R. Soc. B 366: 2681–2692.
Mateo, R. G. et al. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. – Divers. Distrib. 16: 84–94.
Merow, C. et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.
Milanovich, J. R. et al. 2010. Projected loss of a salamander diversity hotspot as a consequence of projected global climate change. – PLoS One 5: e12189.
Myers, N. et al. 2000. Biodiversity hotspots for conservation priorities. – Nature 403: 853–858.

Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. – Prog. Phys. Geogr. 34: 3–22.
Newbold, T. et al. 2009. Effect of characteristics of butterfly species on the accuracy of distribution models in an arid environment. – Biodivers. Conserv. 18: 3629–3641.
Pearce, J. L. and Boyce, M. S. 2006. Modelling distribution and abundance with presence-only data. – J. Appl. Ecol. 43: 405–412.
Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.
Phillips, S. J. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. – Ecography 31: 161–175.
Phillips, S. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Pinheiro J. et al. 2015. nlme: linear and nonlinear mixed effects models. – R package ver. 3.1-120, <http://CRAN.R-project.org/package=nlme>.
Ponder, W. F. et al. 2001. Evaluation of museum collection data for use in biodiversity assessment. – Conserv. Biol. 15: 648–657.
PostgreSQL Development Core Team 2014. – <www.postgresql.org/developer/core/>.
Quantum GIS Development Team 2014. A free and open source geographic information system. – <www.qgis.org/en/site/>.
Reddy, S. and Davalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Rota, C. T. et al. 2011. Does accounting for imperfect detection improve species distribution models? – Ecography 24: 659–670.
Rondinini, C. et al. 2006. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. – Ecol. Lett. 9: 1136–1145.
Schoener, T. W. 1968. The anolis lizards of Bimini: resource partitioning in a complex fauna. – Ecology 49: 704–726.
Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.
Soberón, J. and Nakamura, M. 2009. Niches and distributional areas: concepts, methods, and assumptions. – Proc. Natl Acad. Sci. USA 106: 19644–19650.
Stolar, J. and Nielsen, S. E. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. – Divers. Distrib. 21: 595–608.
Syfert, M. M. et al. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. – PLoS One 8: e55158.
Tessarolo, G. et al. 2014. Uncertainty associated with survey design in species distribution models. – Divers. Distrib. 20: 1258–1269.
Thibaud, E. et al. 2014. Measuring the relative effect of factors affecting species distribution model predictions. – Methods Ecol. Evol. 5: 947–955.
Václavík, T. et al. 2012. Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). – J. Biogeogr. 39: 42–55.
van Proosdij, A. S. J. et al. 2016. Minimum required number of specimen records to develop accurate species distribution models. – Ecography 39: 542–552.
Vanderwal, J. et al. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? – Ecol. Model. 220: 589–594.
Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. – Ecography 37: 1084–1091.
Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. – J. Biogeogr. 36: 2290–2299.
Verbruggen, H. et al. 2013. Improving transferability of introduced species' distribution models: new tools to forecast the spread of a highly invasive seaweed. – PLoS One 8: e68337.
Warton, D. I. et al. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. – PLoS One 8: e79168.
Yackulic, C. B. et al. 2012. Presence-only modelling using MAXENT: when can we trust the inferences? – Methods Ecol. Evol. 4: 236–243.

Supplementary material (Appendix ECOG-02414 at <www.ecography.org/appendix/ecog-02414>). Appendix 1–6.