ECOGRAPHY

Research

Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent

Ahmed El-Gabbas and Carsten F. Dormann

A. El-Gabbas (http://orcid.org/0000-0003-2225-088X) (elgabbas@outlook.com) and C. F. Dormann, Dept of Biometry and Environmental System Analysis, Univ. of Freiburg, Freiburg, Germany.

Ecography
41: 1161–1172, 2018
doi: 10.1111/ecog.03149

Subject Editor: Robert P. Anderson
Editor-in-Chief: Miguel Araújo
Accepted 4 October 2017

Species distribution modelling (SDM) has become an essential method in ecology and conservation. In the absence of survey data, the majority of SDMs are calibrated with opportunistic presence-only data, incurring substantial sampling bias. We address the challenge of correcting for sampling bias in the data-sparse situations. We modelled the relative intensity of bat records in their entire range using three modelling algorithms under the point-process modelling framework (GLMs with subset selection, GLMs fitted with an elastic-net penalty, and Maxent). To correct for sampling bias, we applied model-based bias correction by incorporating spatial information on site accessibility or sampling efforts. We evaluated the effect of bias correction on the models' predictive performance (AUC and TSS), calculated on spatial-block cross-validation and a holdout data set. When evaluated with independent, but also sampling-biased test data, correction for sampling bias led to improved predictions. The predictive performance of the three modelling algorithms was very similar. Elastic-net models have intermediate performance, with slight advantage for GLMs on cross-validation and Maxent on hold-out evaluation. Model-based bias correction is very useful in data-sparse situations, where detailed data are not available to apply other bias correction methods. However, bias correction success depends on how well the selected bias variables describe the sources of bias. In this study, accessibility covariates described bias in our data better than the effort covariate, and their use led to larger changes in predictive performance. Objectively evaluating bias correction requires bias-free presence–absence test data, and without them the real improvement for describing a species' environmental niche cannot be assessed.

Introduction

Species distribution data often come in the form of presence-only, with information on where species have been recorded, but no reliable information on where they have not, or where people have looked (Pearce and Boyce 2006). Museums, herbaria, personal collections, published literature, and citizen records are valuable sources for presence-only data (Pearce and Boyce 2006, Newbold 2010), especially in developing

© 2017 The Authors. Ecography © 2017 Nordic Society Oikos

NORDIC SOCIETY OIKOS

www.ecography.org

countries where there is a lack of systematic nation-wide surveys. Some initiatives make such species sightings freely available: GBIF (the Global Biodiversity Information Facility – < www.gbif.org >) collates global biodiversity data from different sources. One fundamental problem is that data are often incidental with no information on the sampling efforts and survey method used (Pearce and Boyce 2006). They are biased taxonomically (towards larger, easy to detect, or more charismatic species groups; Newbold 2010), environmentally (less collection effort in areas with harsh environments), temporally (more in summer than winter), and spatially (near populated places, roads, research institutes and protected areas; Phillips et al. 2009, Newbold 2010, Stolar and Nielsen 2015). For example, GBIF-data show huge differences in data contribution among countries (Supplementary material Appendix 1 Fig. A1; on average, more from well-financed than from species-rich countries). Spatial bias is a particular concern for statistical analysis when it leads to environmental bias (Phillips et al. 2009), e.g. when large parts of the environmental space remain unsampled (Merow et al. 2014).

Species distribution models (SDMs) relate species occurrences to the environment to estimate habitat preference, and predict potential distribution and responses to climate change (Phillips and Dudík 2008, Elith et al. 2011). Statistical analyses of presence-only data describe the environment at record locations relative to the background environment, making them more susceptible to sampling bias than presence–absence data from dedicated surveys (Phillips et al. 2009, Fithian et al. 2015). However, such targeted survey-data are rare (Pearce and Boyce 2006), especially in developing countries, explaining why the majority of SDM applications uses presence-only data. Presence-only data may produce sound models if they are efficiently corrected for sampling bias (Elith et al. 2011). Point-process models (PPM) have recently emerged as the most appropriate technique for presence-only data (Renner et al. 2015). PPMs do not use background points as pseudo-absences (as in the naïve logistic regression; Fithian and Hastie 2013), but rather as quadrature points for estimating the spatial integral of the likelihood function, and hence require careful tuning of its number (for details see, Warton and Aarts 2013). The response variable of PPMs is the density of species records per unit area (also called 'intensity'), which should be proportional to the probability of occurrence (which can not be estimated empirically using presence-only data without additional information; Fithian and Hastie 2013, Renner et al. 2015, Phillips et al. 2017). It is mathematically equivalent to methods already commonly used in ecology, e.g. Maxent and some implementations of the generalised linear modelling framework, but differently efficient (Renner and Warton 2013, for details, Renner et al. 2015).

Sampling bias has been addressed by 'spatial filtering' of presence locations (keeping only a limited number of records within a certain distance) to dilute the effect of uneven sampling effort across the study area (Anderson and Raza 2010, Boria et al. 2014). Alternatively, others effectively use location of records from related species as background points to have background points with the same bias as the species records (Elith and Leathwick 2007, 'target-group background': Phillips et al. 2009, see also, Ponder et al. 2001, and 'weighted target group' presented in Anderson 2003). Neither approach is applicable when there are only few data (a typical case in developing countries). For example, to apply the target-group background approach to model the distribution of a bat species in North Africa, all GBIF bat species records (Supplementary material Appendix 1 Fig. A2) are clearly not enough to serve as representative background points in this large study area. Similar to the target-group background, if the pattern of sampling bias is known a priori, it can be used as prior weight for sampling the background proportionally to the sampling effort (e.g. 'bias file' in Maxent; Phillips and Dudík 2008, Warren et al. 2014), so that both presences and background samples have the same bias (see also, Stolar and Nielsen 2015).

As a third strategy, sampling bias can be addressed also by modelling the distribution of the focal species as a function of two additive covariate sets: the environmental covariates and other covariate(s) describing potential sources of sampling bias, hereafter 'bias covariates' (model-based bias correction; Warton et al. 2013). For unbiased predictions, the bias covariates are set to a common level of bias, say 0, at all locations; however, sometimes it is difficult to settle on a meaningful adjustment level (Warton et al. 2013).

The aim of this paper is to address the problem of sampling bias in data-sparse situations and how to correct for it in presence-only SDMs. We apply model-based bias correction, comparing two different sets of bias covariates to model the distribution of Egyptian bat species in areas of their known global distributions. Bias-models for each of three modelling techniques within the PPM framework are calibrated with information on either accessibility or sampling effort and compared to an 'environment-only model'. To maintain a reasonable degree of independence between training and testing datasets, we evaluated the models using a) entirely independent presence-only data (not used to fit any of the models); and b) spatial-block cross-validation.

Material and methods

Species and study area

We are interested in understanding the environmental preferences of Egyptian bats as an example representing the sampling-bias issue in data-sparse situations. We collected records for the entire range of each species (Supplementary material Appendix 1 Fig. A2, latitude: –35° to + 56° and longitude: –20° to + 80°) from the literature and GBIF (see Supplementary material Appendix 2 for list of literature sources). Although GBIF provides a valuable source of data, such opportunistically compiled data bases inevitably contain misidentified or incorrectly georeferenced records (Gaiji et al. 2013), and thus require careful revision before use. Relevant records from the GBIF database were assessed (October 2014) and merged with the available literature records. Bat records from Egypt were mostly taken from the expert-revised BioMAP (Biodiversity Monitoring and Assessment Project) database (Basuony et al. 2010). Only species with enough presence locations that are located in at least 5 larger spatial blocks were included in this study (allowing model evaluation on spatial-block cross-validation; see below). This was fulfilled by 21 species, each occupying > 20 unique cells at resolution of 5 × 5 km^2 (Table 1). The geographical range of records was assessed (based on the literature and IUCN), and spatial outliers were excluded. The coverage of all available records shows obvious signs of spatial bias towards western Europe and only sparse sampling in Africa and western Asia (Supplementary material Appendix 1 Fig. A2). Presence locations were purposefully split into training and testing data, using only records from outside Egypt's boundaries for training, thereby keeping the Egyptian presence data as entirely independent evaluation data. For sampling of background points, however, Egypt was not excluded from the study area, and hence background points can be sampled from Egypt as well (see Fig. 1 for a flowchart of the methods applied).

The determination of the study area is critical, especially for presence-only models (Pearce and Boyce 2006). We decided against a single fixed large study area that covers presence locations of all 21 species to keep only areas of potential accessibility to the bats (Barve et al. 2011) and avoid inflating the discrimination ability of the model (e.g. higher AUC; Barve et al. 2011). Instead, for each species, the study area was determined based on the geographical extent of the records: a rectangular bounding box containing a 1000 km buffer around the species extent of occurrence (see Supplementary material Appendix 1 Fig. A3 for an example). We used a buffer of 1000 km, as we found it suitable for the study species 'bats', which have, on average, high dispersal ability and large home range. Potential covariates (and species presences) were projected into Mollweide equal-area projection and rasterised to a resolution of 5 × 5 km^2. We assessed potential environmental variables for multi-collinearity, and assembled a final list of covariates with a maximum generalized variance inflation factor value less than 3 (for details see Supplementary material Appendix 3 and Supplementary material Appendix 1 Fig. A4).

Block cross-validation

Cross-validation is commonly used for evaluating model performance when no independent data are available. Random splitting does not guarantee spatial independence due to spatial autocorrelation (both training and testing data will be spatially adjacent), and so may overestimate model performance (Bahn and McGill 2013, Radosavljevic and Anderson 2014). To maintain independence between folds and improve transferability of models, a spatial form of cross-validation was used by splitting the study area into coarse checkerboard blocks, and then distribute blocks randomly into folds (Fithian et al. 2015). We performed 5-fold spatial-block cross-validation (hereafter: cross-validation). The larger the block

Table 1. List of Egyptian bat species used in this study, with total number of records available, number of unique records (in parentheses: the number of records used to train the model on cross-validation/those kept aside for independent evaluation, i.e. 'Egyptian records'), number of occupied grid cells at the resolution of 5 × 5 km (number of cells outside/inside Egypt), and best estimated number of background points used to run the DWPR-GLM and elastic-net models. See main text and Supplementary material Appendix 1 Fig. A5 for more details.

	Species	No. records (total)	No. records (unique)	No. occupied cells	No. background points (×1000)
1	*Asellia tridens* (trident leaf-nosed bat)	414	**339** (238/101)	**285** (209/76)	300
2	*Barbastella leucomelas* (Sinai barbastelle)	50	**41** (32/9)	**34** (28/6)	75
3	*Eptesicus bottae* (Botta's serotine bat)	99	**80** (64/16)	**74** (61/13)	300
4	*Hypsugo ariel* (fairy pipistrelle)	58	**54** (21/33)	**45** (21/24)	300
5	*Nycteris thebaica* (Egyptian slit-faced bat)	939	**840** (789/51)	**741** (704/37)	300
6	*Nycticeinops schlieffeni* (Schlieffen's bat)	122	**109** (107/2)	**103** (101/2)	300
7	*Otonycteris hemprichii* (Hemprich's long-eared bat)	222	**166** (127/39)	**149** (115/34)	150
8	*Pipistrellus deserti* (desert pipistrelle)	35	**21** (15/6)	**20** (15/5)	150
9	*Pipistrellus kuhlii* (Kuhl's pipistrelle)	2074	**1840** (1751/89)	**1646** (1574/72)	200
10	*Pipistrellus rueppellii* (Rueppell's pipistrelle)	113	**94** (86/8)	**83** (75/8)	100
11	*Plecotus christii* (desert long-eared bat)	110	**88** (24/64)	**72** (23/49)	75
12	*Rhinolophus clivosus* (Arabian horseshoe bat)	150	**103** (53/50)	**85** (48/37)	300
13	*Rhinolophus hipposideros* (lesser horseshoe bat)	5313	**5288** (5274/14)	**3395** (3383/12)	300
14	*Rhinolophus mehelyi* (Mehely's horseshoe bat)	424	**413** (404/9)	**375** (368/7)	150
15	*Rhinopoma cystops* (lesser mouse-tailed bat)	132	**96** (52/44)	**80** (50/30)	200
16	*Rhinopoma microphyllum* (greater mouse-tailed bat)	283	**237** (214/23)	**212** (194/18)	200
17	*Rousettus aegyptiacus* (Egyptian fruit bat)	1046	**874** (766/108)	**702** (633/69)	200
18	*Tadarida aegyptiaca* (Egyptian free-tailed bat)	150	**141** (131/10)	**135** (126/9)	300
19	*Tadarida teniotis* (European free-tailed bat)	1335	**1308** (1284/24)	**1151** (1131/20)	150
20	*Taphozous nudiventris* (naked-bellied tomb bat)	278	**216** (170/46)	**193** (158/35)	100
21	*Taphozous perforatus* (tomb bat)	226	**203** (161/42)	**186** (150/36)	200

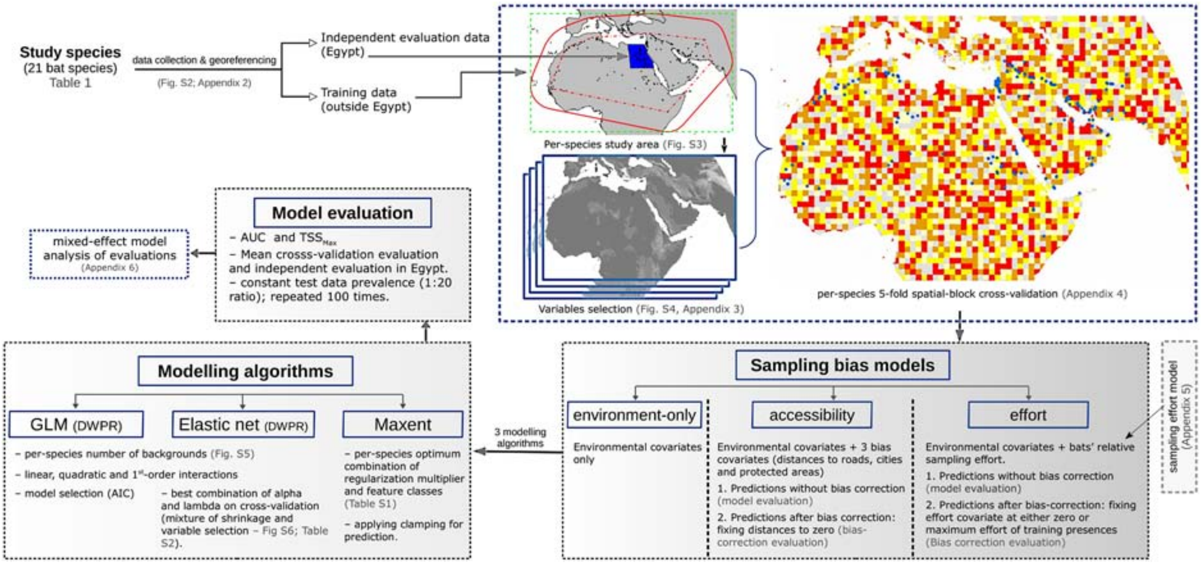
Study species
(21 bat species)
Table 1
data collection & georeferencing
(Fig. S2; Appendix 2)
Independent evaluation data
(Egypt)
Training data
(outside Egypt)
Per-species study area (Fig. S3)
Variables selection (Fig. S4, Appendix 3)
per-species 5-fold spatial-block cross-validation (Appendix 4)
Model evaluation
– AUC and TSS$_{Max}$
– Mean crosss-validation evaluation and independent evaluation in Egypt.
– constant test data prevalence (1:20 ratio); repeated 100 times.
mixed-effect model analysis of evaluations
(Appendix 6)
Modelling algorithms
GLM (DWPR)
Elastic net (DWPR)
Maxent
– per-species number of backgrounds (Fig. S5)
– linear, quadratic and 1st-order interactions
– model selection (AIC)
– best combination of alpha and lambda on cross-validation (mixture of shrinkage and variable selection – Fig S6; Table S2).
– per-species optimum combination of regularization multiplier and feature classes (Table S1)
– applying clamping for prediction.
3 modelling algorithms
Sampling bias models
environment-only
accessibility
effort
Environmental covariates only
Environmental covariates + 3 bias covariates (distances to roads, cities and protected areas)
1. Predictions without bias correction (model evaluation)
2. Predictions after bias correction: fixing distances to zero (bias-correction evaluation)
Environmental covariates + bats' relative sampling effort.
1. Predictions without bias correction (model evaluation)
2. Predictions after bias-correction: fixing effort covariate at either zero or maximum effort of training presences (Bias correction evaluation)
sampling effort model
(Appendix 5)

Figure 1. Flowchart of analyses in this study, illustrated with data for *Asellia tridens*. Sampling-bias models and modelling algorithms were combined factorially. Only results for validation with AUC are presented in the manuscript, while TSS-results are given in the Supplementary material.

size, the higher is the need for more data to be able to run models on spatial-block cross-validation. We used blocks of 100 × 100 km^2 (20 × 20 cells), which is not very strong, as we find this more appropriate for the available data. Presence and background locations within each block were used together for model training or testing: for the three modelling algorithms, potential background locations of the left-out cross-validation fold were not used (masked) during model calibration and were used exclusively for evaluation (similar to 'masked geographically structured approach' of Radosavljevic and Anderson 2014, see also Fig. 7 in Fithian et al. 2015). For each species, we used a different blocking structure, balancing the number of available presences between folds and avoiding extrapolation in environmental space (for more details see Supplementary material Appendix 4).

Sampling-bias models

We compared models without bias correction (environment-only model, our reference) with two methods of addressing bias. Firstly, we considered human accessibility as the main source of sampling bias and ran SDMs with additional bias covariates describing 'distances to nearest cities, roads and protected areas' (the 'accessibility model'). For bias-free predictions, we set all distances to zero; thus, a bias-free prediction can be interpreted as the relative intensity of the target species records if all locations across the study area had perfect accessibility (Warton et al. 2013). Second, assuming relative effort is the main source of sampling bias, we incorporated a single bias covariate describing the 'relative intensity of sightings of all bat species' (the 'effort model'). This sampling-effort bias covariate was actually the prediction from a (different) all-bats-model, which predicted the number of all records as a function of non-environmental variables (terrain roughness, distance to cities, distance to main roads, human population density, protected area – details in Supplementary material Appendix 5 and Fig. 1 bottom-right). For bias-adjusted predictions, the sampling-effort covariate needed to be set to a common level, for which there is no obvious choice. We therefore adjusted it to either of two values of sampling effort: the maximum fitted value at the target species' training presence locations, and at zero. In the first case, the bias-adjusted prediction can be interpreted as the relative intensity of species reporting if all the study area receives the sampling effort of the best presence location of the target species. In the second case, the bias-adjusted prediction is independent of effort, and the effort covariate is used only to correct the coefficients of the environmental variables. Although the latter value seems simpler, the predicted values lack intuitive interpretation. As the results were very similar, we here only use the maximum fitted value at training presences.

Modelling algorithms

For each sampling bias model (environment-only, accessibility and effort), we employed three modelling algorithms under the PPM framework with cross-validation and adjustment for model complexity: GLM, elastic net, and Maxent. Both GLM and elastic net model the number of records as a Poisson regression, with elastic net including a mixture of lasso and ridge regularisation (shrinkage; L1- and L2-regularisation, respectively; Friedman et al. 2010). Maxent (ver. 3.3.3k; Phillips and Dudík 2008) is a machine-learning algorithm for presence-only data, effectively and mathematically akin to a Poisson GLM, but with different functional forms for the predictors; it also applies an ad-hoc form of lasso regularisation (Renner and Warton 2013). We used GLM and elastic net to implement a 'down-weighted

Poisson regression' (DWPR, following Renner et al. 2015). This approach uses weights to make the number of presences a negligible proportion of all data and scales the data to the actual area. As a consequence, DWPR estimates model parameters including the intercept. A small weight (10^{-6}) was assigned to presence locations, while background points were given a higher weight equal to the area of the study region divided by the number of background points used.

To estimate the appropriate number of background points (for GLM and elastic net), 25 repeated series of DWPR-GLMs were run, each one progressively increasing the number of randomly sampled background points. We used the number of background points at which the log-likelihood asymptoted (Supplementary material Appendix 1 Fig. A5; Renner et al. 2015). For both GLM and elastic-net models, 1) we included linear, quadratic and 1st-order interactions between variables; 2) no interactions were allowed between the environmental and bias covariates (Warton et al. 2013); 3) we standardised all covariates to a mean of zero and standard deviation of one. For GLMs, a random sample of background points (number estimated from the asymptoted log-likelihood curve) was used to run an initial model. The initial model was then simplified using AIC-informed backward stepwise selection and the remaining variables were used to cross-validate the final models.

Running elastic net (R package glmnet; Friedman et al. 2010) requires tuning of two parameters, α (describes the balance of ridge and lasso) and λ (degree of regularisation; for more details see, Hastie et al. 2009). For each species and bias manipulation, we estimated the best combination of α and λ by 5-fold cross-validation of 11 models (α ranging from zero to one with an increment of 0.1). For each model, the optimal λ value was determined by fitting a series of cross-validated models to a range of λ values ('regularisation path'; by default 100 values estimated from the data) and the λ value that showed the minimum mean cross-validated error was used for predictions. Similarly, the α value with the lowest error (Poisson deviance) was selected (Supplementary material Appendix 1 Fig. A6). To report the performance of the elastic net (and for comparisons with the results of other techniques), the selected values of α and λ were used to run the cross-validation models manually. Due to the computational limitation of explicitly using a user-defined λ during model training (stated by glmnet help page), models were fitted without providing a λ-value, allowing the fit of many models over the regularisation path. For prediction we used the optimal λ-value estimated from cross-validation. The best-estimated values of α are shown in Supplementary material Appendix 1 Table A2. Lasso ($\alpha = 1$) rather than ridge was chosen for almost half of the species in environment-only and accessibility models. For the effort model, only three species had ridge models ($\alpha = 0$) and all others had α-values < 1.

Maxent default settings were adapted according to advice in the literature (Merow et al. 2013, Radosavljevic and Anderson 2014). Maxent, by default, uses combinations of feature classes (transformations of covariates: linear 'L', quadratic 'Q', hinge 'H', threshold 'T', and product 'P') depending on the number of presence locations available, allowing for complex species–environment relationships (Phillips and Dudík 2008). We adapted functions from the 'ENMeval' package (Muscarella et al. 2014) to run Maxent models on cross-validation at different complexity levels and feature class combinations (48 models = 8 regularisation multiplier values ranging from 0.5 to 4 with increment of 0.5 × 6 feature class combinations [L/LQ/H/LQH/LQHP/LQHPT]). The default number of background points used by Maxent (10 000) is insufficient to represent the environmental variability in large study areas as used here (Renner and Warton 2013), and hence all potential grids were considered as background points. We used clamping while predicting to the left-out fold, meaning that if any value of a covariate is beyond its training range, these values will be replaced by the closest value during training (Anderson and Raza 2010). For each species and bias manipulation, the combination of regularisation multiplier and feature class that shows the highest mean testing-AUC (on cross-validation) were selected for predictions. The optimum combinations of Maxent's feature classes and regularisation multiplier are shown in Supplementary material Appendix 1 Table A1. The selected feature classes deviated from Maxent's default: 10 species always had simple features (L/LQ) for all bias manipulations, with the effort model showing more complex features in some cases. Moreover, the best-estimated regularisation multiplier also deviated from the default value of one, with many species having values > 1, indicating little overfitting.

Model evaluation

We evaluated the models using threshold-independent (the area under the ROC curve, AUC) and threshold-dependent (true skill statistic, TSS) metrics. We tried to avoid some of the metrics' known caveats through 1) the use of species-specific study areas; 2) block cross-validation minimising environmental extrapolation; and 3) using the same blocking structure to run different modelling techniques and bias manipulations of the same species.

In presence-only SDMs, presences comprise only a tiny fraction of the background. During evaluation, the use of too many background points with equal weights for commission and omission error can distort the evaluation (Lobo et al. 2008). To be able to compare AUCs (computed using the 'dismo' R package) among models for the same species, we set the numbers of test presences and random test background (test data prevalence) to a ratio of 1:20, and repeated this 100 times to average stochastic effects. TSS was calculated using the threshold that maximises the sum of sensitivity and specificity (Liu et al. 2013), again using a constant ratio of presences and background. AUC and TSS were calculated for each combination of species, algorithm and bias manipulation.

We summarised the overall effect of different bias manipulations in two ways, using linear mixed-effect models (R-package lme4; Bates et al. 2015). Firstly, to evaluate the effect of incorporating bias covariates into the model, we compared evaluation of the environment-only models (without any bias correction) to those of bias-accounted models (accessibility and effort) without conditioning on a particular value of the bias covariate during prediction (modelling evaluation). Second, to explore the effect of bias correction on model evaluation, we performed similar comparisons, with bias-free predictions of the bias-accounted models instead (bias correction evaluation; for details see Supplementary material Appendix 6). We expect models incorporating the impact of bias to outperform (signified by a higher AUC or TSS) those that do not. The problem is that we do not have bias-free data, and therefore our test data exhibit the same bias as the data used to create the model. Under these circumstances, models allowing for bias may be worse than those that do not. In either case, the difference between bias-corrected and control models is a measure of the impact of bias. In the mixed models, we used model evaluation as response variable, species as random effect, and model type, bias correction (and their interactions), total number of training presences, and total number of pixels at per-species study area (range size) as fixed effects (for more details, see Supplementary material Appendix 6).

Results

Validation by block cross-validation vs by Egypt hold-out

The overall mean AUC is 0.88 ± 0.08 on cross-validation (0.75 ± 0.13 in Egypt's evaluation; Fig. 2–3). Because results for AUC and TSS were similar, we present only results for AUC (Supplementary material Appendix 1 Fig. A11 and supplement for TSS). Model evaluations in Egypt were, on average, poorer than on cross-validation. However, both types of evaluation (using 'bias-corrected' predictions) show positive correlation (Kendall's tau; $n = 21$, Supplementary material Appendix 1 Fig. A7), with highest consistency for the accessibility model. As expected, species with fewer occupied cells tend to have higher evaluation variability (uncertainty), and hence provide less robust analyses (Supplementary material Appendix 1 Fig. A9–A10).

Effect of bias corrections: mixed-effect model analysis across species and model types

Sampling-bias model explained most of the variation in model validation, followed by the modelling algorithm, and their interaction (Fig. 2 and Supplementary material Appendix 6.1). The total number of training presence (positive effect) and the range size (negative effect) were much less important. On average, the accessibility model performed much better than environment-only and effort models (Fig. 2 and Supplementary material Appendix 6.1). For mean cross-validations, effort models also had relatively higher AUCs than the environment-only model (the difference is much smaller for elastic net and GLM; Fig. 2a). However, for validation in Egypt, effort models were similar to the environment-only model (Fig. 2b, see also Supplementary material Appendix 6.1).

We can also evaluate the performance of the sampling-bias models when their bias covariates are set to a fixed value, i.e. when predicting to bias-free data. However, since we have no reference survey data to compare them to, these predictions reveal the impact of bias-correction on the model, rather than assess the model's performance. These evaluations are presented in Fig. 3 and Supplementary material Appendix 6.2.

Effect of model type

The predictive performance of the three modelling algorithms showed fair to moderately high correlation coefficient (most $r > 0.6$, all significant; Fig. 4). For mean cross-validation evaluations, GLM performed best, followed by elastic net and Maxent. The order is reversed for independent validations in Egypt: Maxent has highest AUC-values, followed by elastic net, then GLM (Fig. 2–4, and Supplementary material Appendix 6).

Discussion

Sampling bias, if not corrected for effectively, can substantially affect the predicted intensity and model evaluation of SDMs. Our results suggest that accessibility bias covariates describe the bias in our focal species' training data well, compared to sampling efforts, and their use led to higher validation scores. Evaluated on data that are themselves spatially biased, we find that removing sampling bias did not improve the predictive performance on cross-validation or in Egypt (Fig. 3), highlighting the limitation of evaluating bias correction without independent bias-free presence–absence data. Collectively, the three modelling algorithms performed similarly well in cross-validation, though predicting to new sites (Egypt) gave Maxent a slight advantage (Fig. 2b and 3b).

Sampling-bias correction using presence-only data

Few approaches exist to correct for sampling bias, some not applicable when data are sparse. Spatial filtering (i.e. aggregation to single records within some larger buffer: Anderson and Raza 2010, Boria et al. 2014) might be wasteful when only few presences exist. Removal of clumped data will reduce training sample size and may remove some of the environmental conditions occupied by the species (depending on the heterogeneity of the landscape, selected distance, and pixel size).

Target-group background selection strictly assumes that related species were collected with the same method and equipment (Phillips et al. 2009), bias is similar across species (Warton et al. 2013), and species have the same chance of being recorded in all locations (Yackulic et al.

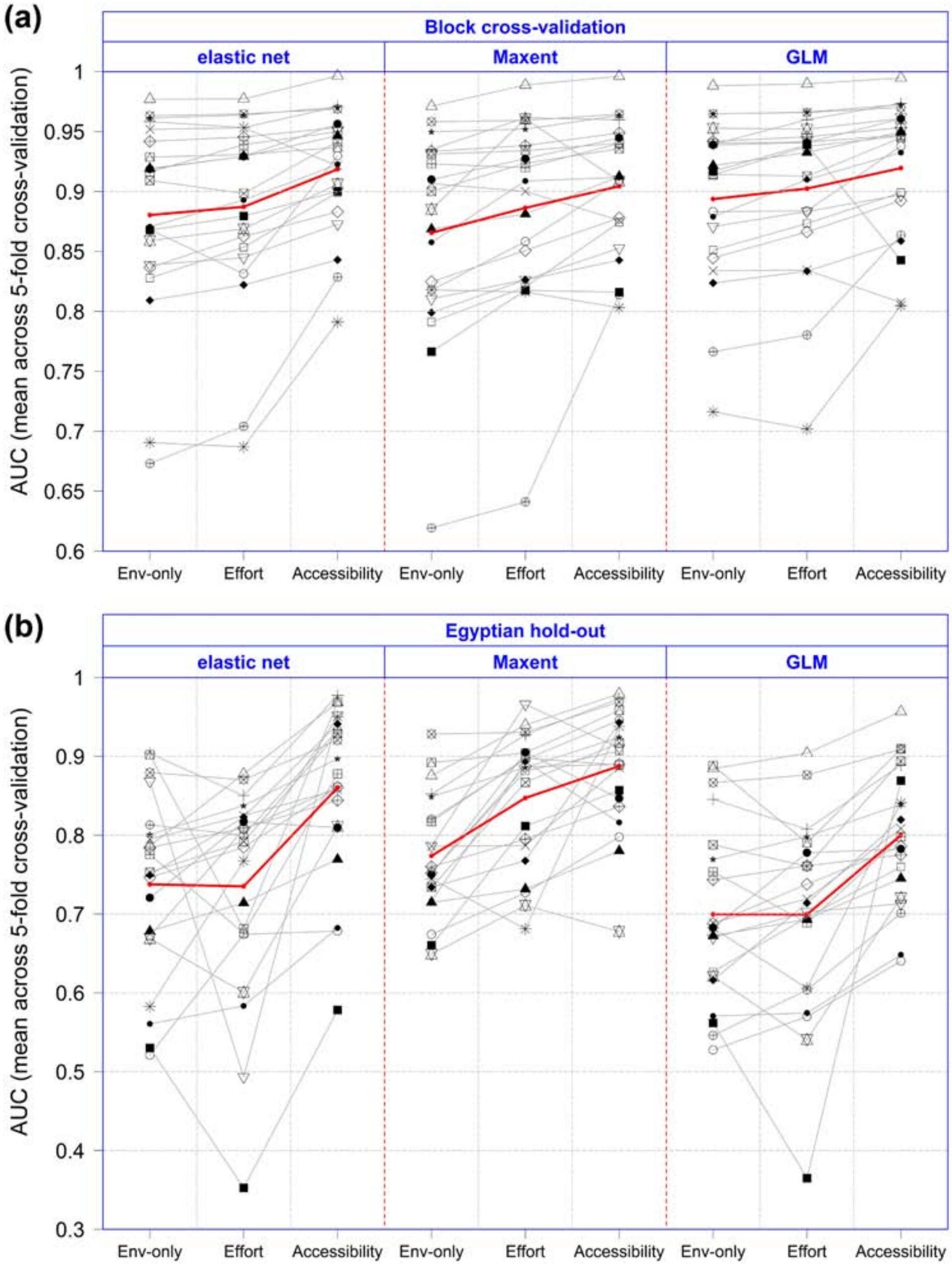

Figure 2. Mean AUC of each species calculated either on 5-fold spatial block cross-validation (a) or in Egypt (b). Each plot compares evaluation of environment-only models to bias-accounting models (effort and accessibility), without correcting for sampling bias (Modelling evaluation; Supplementary material Appendix 5.1). Each species is represented by different symbols (see Supplementary material Appendix 1 Fig. A7 for species names). Red lines indicate overall mean AUC for each modelling algorithm and bias manipulation applied. For evaluations of bias-free prediction, see Fig. 3.

2013). Moreover, it replaces the observer bias with a (spatial) species-richness bias and can be understood as modelling the likelihood of encountering the target species rather than a non-target species (Warton et al. 2013, Warren et al. 2014). Also, it does not distinguish between areas unsuitable for any of the species and areas of low accessibility (no records, but potentially suitable: Warren et al. 2014, Fithian et al. 2015). The target-group background cannot be used when data on the related species are similarly limited, as this increases the risk of extrapolating in environmental space (Mateo et al. 2010, Merow et al. 2013).

Bias layers attempt to describe with which biases presence locations were recorded (Phillips et al. 2009, Warton et al. 2013). Our external sampling-effort model (Fig. 1 bottom right) is such an attempt to derive a bias layer (Elith et al. 2010). The bias layer is currently only implemented in Maxent, and we are not aware of any study that applied a

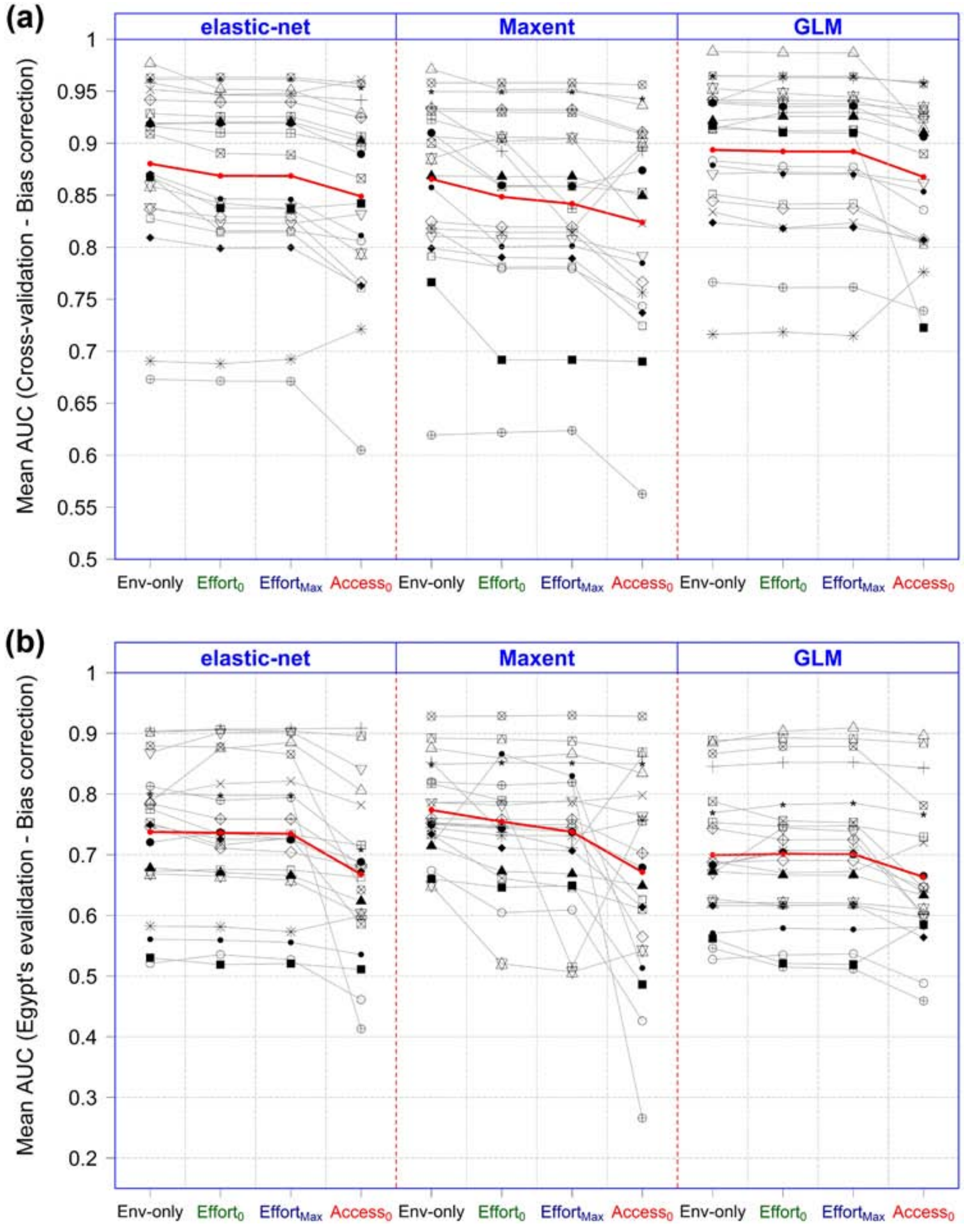

Figure 3. Species mean AUC calculated either on cross-validation (a) or in Egypt (b), after sampling bias correction (using bias-free prediction; for details, see Supplementary material Appendix 6.2). Each species is represented by different symbols (see Supplementary material Appendix 1 Fig. A7 for species names). Red lines indicate the overall mean AUC at each modelling algorithm and bias models applied.

similar approach using other modeling algorithms. To maintain consistency of the analyses across different modelling algorithms we did not consider using a bias layer here. Both bias layer and target-group are used to sample background with the same bias as presence locations, which is a sensitive tuning step that strongly influences model evaluation (Chefaoui and Lobo 2008). We think that model-based bias correction, as applied here, is more plausible in data deficient situations and if applied efficiently, it frees us from artificially manipulating presences or background points, and rather focus on sampling bias correction. The effectiveness of bias correction depends on whether bias covariates are actually able to describe the bias in the available data.

Our use of either accessibility or sampling effort bias covariates did not lead to notably different conclusions (Fig. 5). Using sampling-effort bias covariate led to, on average, little to moderate evaluation changes (Fig. 2–3). It is in fact a model-based version of the target-group background approach, without strict selection of the background data points. The assumption that closely related species have similar bias may not hold. Surprisingly, most of the available presences of our focal species were located in

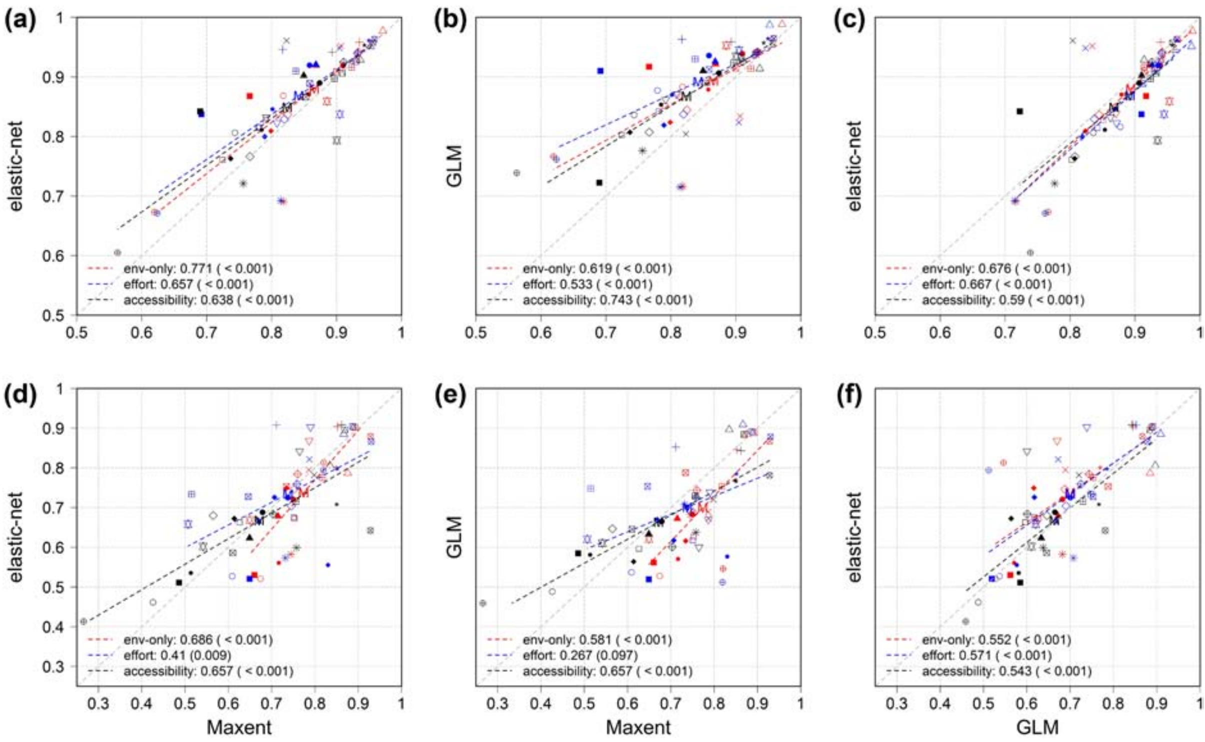

Figure 4. Kendall's correlation of the per-species mean AUC between pairs of modelling algorithms. Each species is represented by different symbols (see Supplementary material Appendix 1 Fig. A7 for species names) with colours referring to bias models (using predictions of environment-only model and bias-free prediction of accessibility and effort model). 'M' indicates the overall mean evaluation. Top row panels are mean evaluation using spatial block cross-validation, while those in the bottom row are independent evaluations in Egypt.

areas of low to moderate estimated efforts (Supplementary material Appendix 1 Fig. A13), but still strongly biased towards roads and cities (and, to less extent, protected areas; Supplementary material Appendix 1 Fig. A13).

Ideally, evaluating bias correction requires independent 'bias-free' presence–absence testing data (Phillips et al. 2009). Such data are typically not available, especially in developing countries, and removing bias from test data is very difficult (Dudík et al. 2005, Smith 2013). On average, validation of bias-corrected predictions using well-structured, independent presence–absence data from rigorous surveys led to improved predictions (Elith and Leathwick 2007, Phillips et al. 2009, Mateo et al. 2010, Syfert et al. 2013, Warton et al. 2013, Boria et al. 2014). However, this improvement is not happening in all situations and depends on the modelling conditions, species prevalence, validity of assumptions, and how effective bias covariates are in describing bias in training presences (Phillips et al. 2009, Warton et al. 2013, Fourcade et al. 2014). As covariates that affect sampling may also affect the distribution of a species (e.g. avoiding deserts), no method can fully correct for sampling bias in presence-only data without affecting the niche model (Merow et al. 2014, Guillera-Arroita et al. 2015).

Correction of sampling bias leads to larger areas of suitable habitats due to higher suitability estimates in low-accessible sites (Phillips et al. 2009, Warton et al. 2013; Fig. 5). Predictions at such sites are of lower reliability and should be interpreted with caution (Supplementary material Appendix 1 Fig. A16 for an example). They can be used to guide future surveys and conservation planning, but not for taking serious conservation decisions (Guisan et al. 2006).

Evaluations using spatial-block cross-validation vs Egyptian hold-out

We used spatial-block cross-validation to avoid overestimating model performance and underestimating predictive errors (Bahn and McGill 2013, Renner et al. 2015, Roberts et al. 2017). Block cross-valuations suggested a better model performance than evaluation on Egyptian data. This can be explained by differences in sample size and by environmental variability: on average, our cross-validation models had a larger mean number of testing presences and higher environmental variability compared to evaluations at smaller scale (Table 1). However, evaluations at both scales were positively correlated, which supports the idea that evaluations on cross-validation (larger extent) can be indicative of performance at the local scale.

Evaluation metrics and predictive performance

Using AUC and TSS led to consistent conclusions. The overall evaluation scores are fair to high, taking into account

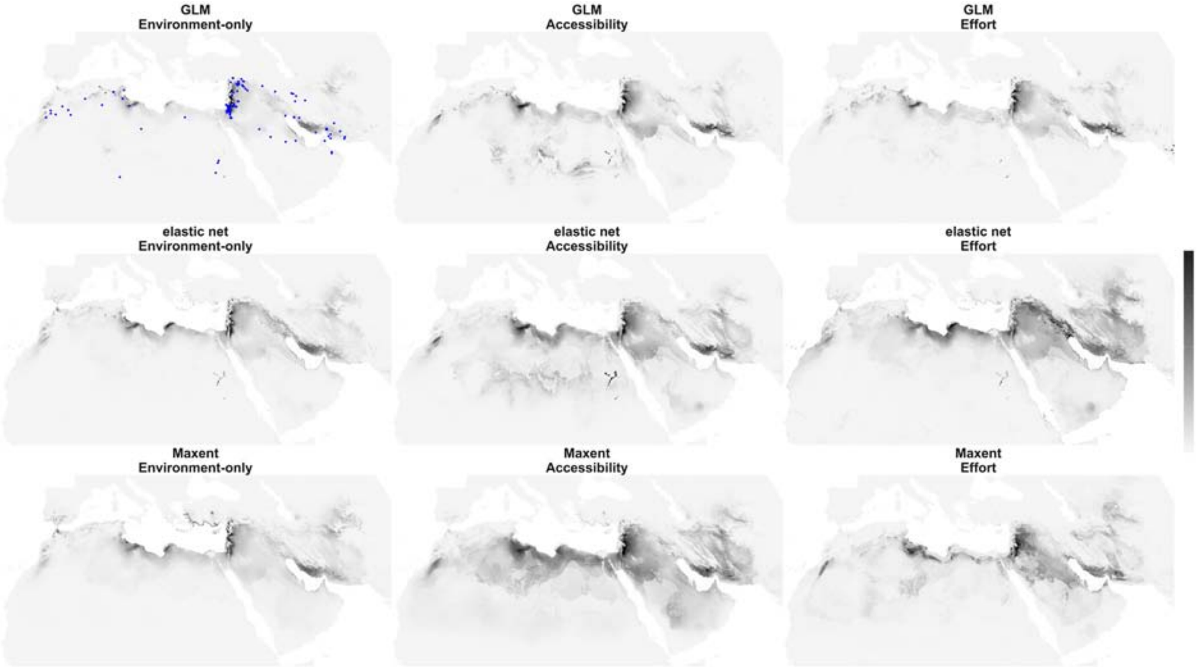

Figure 5. Mean cross-validated predicted distribution of *Otonycteris hemprichii*, of different modelling algorithms (rows) and bias models (columns). Maps were rescaled to relative intensities between zero and one, as different modelling algorithms do not have the same scale. Darker colour indicates higher predicted relative intensity. Blue points (in the top left panel) represents available records used for cross-evaluation model training. For visualisation, extreme values (> 0.9995 quantile of predicted values) were replaced with their next smaller value, as GLM and elastic net are subject to some few extreme predictions (see main text, and Supplementary material Appendix 1 Fig. A15B for a comparison with unlimited predictions). For predicted maps in Egypt see Supplementary material Appendix 1 Fig. A15A.

that spatially independent evaluation yields lower scores compared to commonly used random split (Radosavljevic and Anderson 2014). In presence-only SDMs, the use of a particular value for defining good models become unreliable (Yackulic et al. 2013), due to higher uncertainty of estimates at background points and as the number of background points used is not fixed. In this study, we maintained constant test-data prevalence (1:20) across all comparisons, which gave very similar results, across all species, to the standard default approach of computing AUC (Supplementary material Appendix 1 Fig. A14). For some species, the spread over the 100 repetitions was very noticeable, however, making our approach somewhat more robust. The improved predictive performance reported by Syfert et al. (2013), for example, may be in part attributable to much higher testing prevalence. Also in our study, the larger the number of available presences and the smaller the study area was, the higher were the predictions scores.

Comparison of modelling algorithms

The three modelling algorithms applied did not lead to different conclusions and their evaluations were highly correlated. Elastic-net models showed intermediate performance in all situations. GLM (more specifically: the down-weighted Poisson regression (DWPR) with variable selection) had the highest evaluation on cross-validation, which may suggest that GLM-DWPR is more powerful with a larger number of testing presences. However, GLMs predictions in Egypt were less nuanced compared to other two modelling algorithms (Supplementary material Appendix 1 Fig. A15A, for example), which explains why it ranked lowest on the Egyptian hold-out.

Maxent had the lowest prediction error on the Egyptian data, which suggests its transferability to situations of low extrapolation. However, Maxent had the lowest evaluation on cross-validation, possibly due to clamping. We applied 'clamping' to constrain the response beyond the training range, which changes some of the predicted values. Clamping can affect model evaluation (the ranking of the predicted values) and, while its effect has not been well-explored in the literature, it seems to depend on the shape of the response curve (at both ends), the importance of the covariate, and how much environmental extrapolation occurs. We expect the effect of clamping to be small in the Egyptian hold-out compared to the cross-validation, due to less extrapolation.

When correcting for bias, no interaction should be allowed between bias and other covariates so that, for prediction, the bias can be corrected for without affecting the other covariates (Warton et al. 2013). For GLM and elastic net, we have full control of the models' interactions. However, the version of Maxent used here (3.3.3k) does not enable users to select which interactions to use, and the use of the product feature inevitably enables all pairwise interactions. In further applications, it may be recommended to disable the 'product' feature class while correcting for sampling bias. However, its use here did not lead to different conclusions compared to GLM and elastic net, suggesting that bias-suitability interactions were of limited importance. During the reviewing of this manuscript, an open-source version of Maxent (maxnet R package: Phillips et al. 2017) was released that uses the glmnet package for L1-regularization. Maxnet provides flexibility for specifying the interactions to be used, so it is possible to exclude interactions between environmental and bias variables (making them similar to our elastic-net models, but with more flexibility for other feature classes implemented in Maxent, e.g. the hinge). This early version of maxnet implements an infinitely-weighted logistic regression (IWLR; Fithian and Hastie 2013) and only L1-regularization (lasso); however, further extensions are possible, e.g. the implementation of DWPR and elastic net (similar to our elastic-net models). We implemented the down-weighted Poisson regression using both GLM and elastic net. As Poisson models, their predictions have no upper bound and may thus yield extreme predictions. We reported the existence of a few extreme predicted intensities for many species, which makes it difficult to plot their predictions on a linear scale (see Fig. 5 and Supplementary material Appendix 1 Fig. A15B for an example). In contrast, Maxent puts a constraint on the moments of the predictions, making them less subject to extreme values (Phillips et al. 2006).

Conclusion

Data-sparse regions pose challenges to modelling species distributions, exacerbated by noticeable sampling biases. We recommend the use of model-based bias correction in data-sparse situations, in which other bias corrections methods are not possible; however, the effectiveness of bias correction depends on whether bias covariates are actually able to describe the bias in the available data. Using covariates to describe site accessibility improved prediction to spatially independent hold-outs, compared to environment-only or effort models. Augmenting local records with data from across the species' range allowed us to make consistently high-quality predictions to hold-out data from an entire country (in this case Egypt). Bias-free predictions can enhance future conservation planning and target future surveys when limited resources are available to cover large study areas. However, due to possible lower certainty at unsurveyed locations, they should be used cautiously (maps including bias are of use only during model cross-validation). Without survey-based presence–absence data, no complete evaluation of the quality of bias corrections can be attempted. Down-weighted Poisson regression as well as the statistically equivalent Maxent approach led to similar results, with more flexibility in the elastic-net models (e.g. degree of shrinkage, question-led specification of non-linear effects and interactions). More important than the specific algorithm is to use the point-process modelling framework as such.

Acknowledgements – We would like to express our sincere thanks for Petr Benda (Charles Univ.) for comments on the distribution of the bat species. An earlier version of the manuscript was improved by comments of Francis Gilbert (Univ. of Nottingham). We are also grateful to Belarmain Fandohan, David R. Roberts, and Simone Ciuti for long discussions and comments on the manuscript. This manuscript benefitted from comments of the subject editor Robert P. Anderson and two anonymous reviewers.

Funding – AE-G is sponsored by the German Academic Exchange Service (DAAD) through a GERLS scholarship. This work was partially performed on the computational resource 'bwUniCluster' funded by the Ministry of Science, Research and Arts and the Univ. of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

References

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.

Bahn, V. and McGill, B. J. 2013. Testing the predictive performance of distribution models. – Oikos 122: 321–331.

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.

Basuony, M. I. et al. 2010. Mammals of Egypt: atlas, red data listing & conservation. – BioMAP and CultNat, EEAA and Bibliotheca Alexandrina, Cairo.

Bates, D. et al. 2015. Fitting linear mixed-effects models using lme4. – J. Stat. Softw. 67 doi: 10.18637/jss.v067.i01

Boria, R. A. et al. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. – Ecol. Model. 275: 73–77.

Chefaoui, R. M. and Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – Ecol. Model. 210: 478–486.

Dudík, M. et al. 2005. Correcting sample selection bias in maximum entropy density estimation. – Advances in Neural Information Processing Systems 18. The MIT Press, < http://papers.nips.cc/paper/2929-correcting-sample-selection-bias-in-maximum-entropy-density-estimation.pdf >.

Elith, J. and Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. – Divers. Distrib. 13: 265–275.

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.

Fithian, W. and Hastie, T. 2013. Finite-sample equivalence in statistical models for presence-only data. – Ann. Appl. Stat. 7: 1917–1939.

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.

Friedman, J. H. et al. 2010. Regularization paths for generalized linear models via coordinate descent. – J. Stat. Softw. 33 <www.jstatsoft.org/v33/i01/paper>.

Gaiji, S. et al. 2013. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. – Biodivers. Inform. 8: 94–122.

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. et al. 2006. Using niche-based models to improve the sampling of rare species. – Conserv. Biol. 20: 501–511.

Hastie, T. et al. 2009. The elements of statistical learning: data mining, inference, and prediction – Springer.

Liu, C. et al. 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. – J. Biogeogr. 40: 778–789.

Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.

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.

Merow, C. et al. 2014. What do we gain from simplicity versus complexity in species distribution models? – Ecography 37: 1267–1281.

Muscarella, R. et al. 2014. ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. – Methods Ecol. Evol. 5: 1198–1205.

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.

Pearce, J. L. and Boyce, M. S. 2006. Modelling distribution and abundance with presence-only data. – J. Appl. Ecol. 43: 405–412.

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. J. 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.

Phillips, S. J. et al. 2017. Opening the black box: an open-source release of Maxent. – Ecography 40: 887–893.

Ponder, W. F. et al. 2001. Evaluation of museum collection data for use in biodiversity assessment. – Conserv. Biol. 15: 648–657.

Radosavljevic, A. and Anderson, R. P. 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. – J. Biogeogr. 41: 629–643.

Renner, I. W. and Warton, D. I. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. – Biometrics 69: 274–281.

Renner, I. W. et al. 2015. Point process models for presence-only analysis. – Methods Ecol. Evol. 6: 366–379.

Roberts, D. R. et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. – Ecography 40: 913–929.

Smith, A. B. 2013. On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. – Divers. Distrib. 19: 867–872.

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.

Warren, D. L. et al. 2014. Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. – Divers. Distrib. 20: 334–343.

Warton, D. and Aarts, G. 2013. Advancing our thinking in presence-only and used-available analysis. – J. Anim. Ecol. 82: 1125–1134.

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. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? – Methods Ecol. Evol. 4: 236–243.

Supplementary material (Appendix ECOG-03149 at <www.ecography.org/appendix/ecog-03149>). Appendix 1–6.