ECOGRAPHY

Research

Bunching up the background betters bias in species distribution models

Julien Vollering, Rune Halvorsen, Inger Auestad and Knut Rydgren

J. Vollering (https://orcid.org/0000-0002-7409-2898) ✉ (julien.vollering@hvl.no), I. Auestad (http://orcid.org/0000-0001-6321-0433) and K. Rydgren (http://orcid.org/0000-0001-8910-2465), Dept of Environmental Sciences, Western Norway Univ. of Applied Sciences, Sogndal, Norway. – R. Halvorsen (http://orcid.org/0000-0002-6859-7726) and JV, Dept of Research and Collections, Natural History Museum, Univ. of Oslo, Oslo, Norway.

Ecography
42: 1–11, 2019
doi: 10.1111/ecog.04503

Subject Editor: Cory Merow
Editor-in-Chief: Miguel Araújo
Accepted 18 June 2019

Sets of presence records used to model species' distributions typically consist of observations collected opportunistically rather than systematically. As a result, sampling probability is geographically uneven, which may confound the model's characterization of the species' distribution. Modelers frequently address sampling bias by manipulating training data: either subsampling presence data or creating a similar spatial bias in non-presence background data. We tested a new method, which we call 'background thickening', in the latter category. Background thickening entails concentrating background locations around presence locations in proportion to presence location density. We compared background thickening to two established sampling bias correction methods – target group background selection and presence thinning – using simulated data and data from a case study. In the case study, background thickening and presence thinning performed similarly well, both producing better model discrimination than target group background selection, and better model calibration than models without correction. In the simulation, background thickening performed better than presence thinning when the number of simulated presence locations was low, and vice versa. We discuss drawbacks to target group background selection, why background thickening and presence thinning are conservative but robust sampling bias correction methods, and why background thickening is better than presence thinning for small sample sizes. Particularly, background thickening is advantageous for treating sampling bias when data are scarce because it avoids discarding presence records.

Keywords: bias correction, Maxent, presence-background modeling, presence thinning, sampling bias, species distribution model, target group background selection, virtual species

Introduction

Opportunistically collected presence data harbor vast information about species' distributions, but distribution models based on these data risk mischaracterizing occurrence–environment relationships (Guisan and Zimmermann 2000, Ponder et al. 2001). In particular, georeferenced presence records available from museum collections

NORDIC SOCIETY OIKOS

www.ecography.org

and other compiled sources come with the important caveat that they are sampled nonrandomly – from well-known hotspots, protected areas, and easily accessible locations near cities and roads (Hijmans et al. 2008, Loiselle et al. 2008). Therefore, presence data typically used for distribution modeling are the product of two probability distributions: the true presence probability of the species, and the conditional probability of the species having been sampled (surveyed, detected and recorded) given its presence (Merow et al. 2013). As a result, any covariation between sampling probability and explanatory variables used to model presence probability may give rise to effects of sampling bias in models (Yackulic et al. 2013). At the same time, alternatives to using only biased presence data, such as collecting occurrence data under an appropriate sampling design (Guisan and Zimmermann 2000) or incorporating expert-drawn range maps into the model (Merow et al. 2017), are often unfeasible.

Presence-background distribution models estimate relative presence probability by comparing presence locations (hereafter: 'presences') to a background that consists of all locations in the study area: locations where the species is present as well as 'uninformed background' locations where its occurrence is unknown (Phillips and Elith 2013, Halvorsen et al. 2015). These models are especially vulnerable to the effects of sampling bias, and usually require correction (Phillips et al. 2009). One correction strategy is to unconfound model predictions formally, by factoring out an approximation of the sampling probability distribution, estimated either from presences of the species itself (Warton et al. 2013), from presences of other species (Stolar and Nielsen 2015, Merow et al. 2016), or from surveyed presence and absence of other species (Fithian et al. 2015). For example, estimated sampling probabilities supplied to the popular Maxent software as a 'bias file' are factored out of predictions formally (Phillips et al. 2006, Merow et al. 2013 Appendix 5, Merow et al. 2016). Another frequently employed strategy is to reduce the effects of sampling bias informally, by adjusting the training data (e.g. selecting presences or uninformed background locations under a specific scheme). Adjusting training data unconfounds model predictions indirectly, through changed parameter estimates and selected variables.

Two informal bias correction methods prevalent in presence-background models are target group background selection (Ponder et al. 2001) and presence thinning (Pearson et al. 2007, Veloz 2009). Target group background selection restricts the background to locations with recorded presence of a particular group, usually a higher-rank taxon, that includes the modeled species. This method is motivated by the assumption that the presence records of the target group reflect the sampling probability distribution that led to the presence records of the modeled species. If this holds true, the comparison of presences against the target group background nullifies the effect of sampling bias, such that the resulting distribution model yields true relative presence probability (Phillips et al. 2009). Presence thinning (aka 'filtering') subsamples presences to obtain a reduced set more uniformly distributed in space. A more uniform distribution can be achieved by spatially stratified subsampling or by setting a minimum nearest-neighbor distance (Aiello-Lammens et al. 2015). In effect, presence thinning reduces the amount of clustering in the presence data, under the assumption that the clustering is due more to uneven sampling probability than to uneven presence probability. It should be noted that thinning of presences in environmental space rather than geographic space has also been explored, but less commonly (Varela et al. 2014). We use the term 'presence thinning' for its geographic variant only.

Theoretically, another approach to informal bias correction – which, to our knowledge, has never been tried – is to increase the density of uninformed background locations around presences. We refer to this proposed method as 'background thickening,' since the operation can be thought of as inverse to presence thinning; instead of reducing the clustering of presences, the clustering of the uninformed background is increased. Background thickening is similar to target group background selection insomuch as it aims to nullify bias in the presences by creating a similar bias in the background (Phillips et al. 2009). The rationale is more similar to presence thinning though, since the training data are manipulated based only on spatial proximity. For models that capture ratios of explanatory variable probability density between presence and background locations (Aarts et al. 2012, Merow et al. 2013), the effect of background thickening is expected to be similar to that of presence thinning; presence thinning decreases the numerator of the ratio, while background thickening increases the denominator. A key difference between these two approaches is that, by omitting a fraction of presences, presence thinning discards potentially useful occurrence information. In principle, the information content of a larger sample should allow more accurate estimation of environment–occurrence relationships, given that the effects of sampling bias are accounted for (Fourcade et al. 2014).

Procedurally, background thickening could be implemented by sampling an equal number of uninformed background locations from a discrete area around each presence, or by sampling all uninformed background locations in proportion to a continuous distribution of interpolated presence density. We note that previous studies have used presence density as an estimate of the sampling probability distribution (Elith et al. 2010, Clements et al. 2012, Kramer-Schadt et al. 2013, Fourcade et al. 2014), but all of these studies factored this distribution out of predictions formally, using Maxent's bias file, rather than using it to guide background selection. We return to this important distinction between formal bias correction and background thickening in our discussion.

The aim of this paper is to explore background thickening as a new, informal sampling bias correction method for presence-background distribution models. We examine the effects of background thickening compared to no bias correction and two established bias correction

methods: target group background selection and presence thinning. We do not provide a comprehensive evaluation of the strictly predictive performance of these different methods, but rather utilize evidence from simulated and real data, together with a priori rationale, to make our assessment. Simulation is an important and powerful strategy for evaluating distribution modeling methods because it allows rigorous assessment against a known truth (Zurell et al. 2010, Guisan et al. 2017). On the other hand, simulated patterns and processes are by definition less realistic than real ones, so we complement our simulation with a case study of bias correction methods in distribution models of Sitka spruce *Picea sitchensis*. Sitka spruce occurs natively in a narrow coastal band along the northern Pacific coast of North America, with its broadest range and strongest development in British Columbia and southeast Alaska (Fig. 1; Harris 1984, Peterson et al. 1997). Sitka spruce is a good test case for sampling bias correction for two reasons. First, because of its economic importance and high detectability, Sitka spruce's native distribution is relatively well known (Peterson et al. 1997), and can serve as a point of reference for model predictions. Second, Sitka spruce's strong gradient in occurrence probability from coast to inland (Harris 1984, Peterson et al. 1997) runs perpendicular in space to what we expect is a strong gradient in collection intensity from California to Alaska. Therefore, we can ask how well correction methods reproduce the distribution along the north–south gradient of uneven sampling, without blotting out the coastal–inland gradient that represents true differences in presence probability.

Using scenarios that simulate modeling a species' distribution from a spatially biased sample, and a case study of distribution modeling with real occurrence data, we ask: 1) whether background thickening is effective compared to other commonly employed sampling bias correction methods, and 2) which circumstances affect the applicability of these methods.

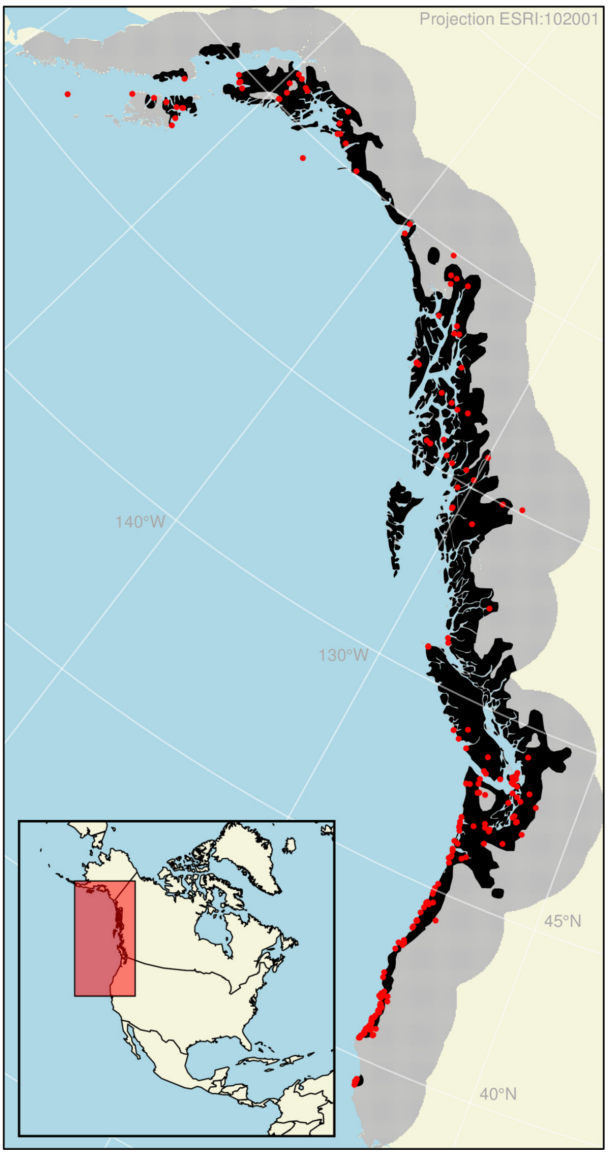

Figure 1. The 243 presences used for modeling the distribution of Sitka spruce (red dots), and an estimate of Sitka spruce's true native distribution (black) adapted from Harris (1984). The study area used for modeling (grey) is plotted underneath the estimated distribution. Note that presences which are close together show as partially overlapping at this scale of representation.

Methods

Virtual species simulation

Simulation inputs

We defined the true presence probability of a virtual species as a logistic function of four BIOCLIM variables, across an area spanning the northern Pacific coast of North America (Supplementary material Appendix A). By sampling the Bernoulli distribution, we transformed this probability distribution into a binary, realized occurrence distribution. We also defined a distribution of relative sampling probability across the same area, based on real human population density and proximity to major roads. Explanatory variables comprised 15 BIOCLIM variables not used to define true presence probability, because we consider it unlikely to have the exact drivers of a distribution available in real modeling applications.

Modeling scenarios

We modeled the relative presence probability of the virtual species under six modeling scenarios, crossing two sample sizes (250 or 25 presences) with three sampling bias approaches (no correction, presence thinning or background thickening). The fourth sampling bias approach in this study, target group background selection, was not included in the simulation because there exists no target group for a truly virtual species (but see Ranc et al. 2017). We varied sample size to test our hypothesis that presence thinning would be less effective with smaller sample sizes, since this method works by discarding records. Presence samples were drawn

from the realized occurrence distribution with relative probability equal to their sampling probability. Both sample sizes were replicated 50 times to overcome idiosyncrasies of single samples, and all three sampling bias approaches were applied in parallel to each sample. Thus, a total of 300 ($2 \times 50 \times 3$) models were built for the virtual species.

In scenarios without bias correction, training data comprised unmodified presence locations and 10 000 uniformly sampled, uninformed background locations. In scenarios with presence thinning, we employed the 'spThin' R package (ver. 0.1.0; Aiello-Lammens et al. 2015) to retain the maximum number of presences possible under a minimum separation distance of 50 km, and also used 10 000 uniformly sampled, uninformed background locations. Although subjective, we chose 50 km as the minimum separation distance to produce an appreciable thinning effect in the smaller samples, which had more scattered records. In scenarios with background thickening, we left presences unmodified, but concentrated 10 000 uninformed background locations within a discrete thickening radius around presences. Specifically, we sampled the study area with relative probability equal to the number of presences within a radius length of each location; for example, a location within a radius length of two presences was twice as likely to be included in the uninformed background as a location within a radius length of only one presence. For each sample, we set the length of the thickening radius equal to the mean spatial autocorrelation range of explanatory variables selected in the uncorrected model. The spatial autocorrelation ranges of individual explanatory variables were determined from variograms built using the 'gstat' R package (ver. 1.1-6; Pebesma 2004, Gräler et al. 2016), and set equal to the maximum distance in the variogram if the curve showed no sill. In other words, we concentrated uninformed background locations within an area expected to be environmentally similar to presences, among relevant explanatory variables. Different choices of minimum separation distance and thickening radius might have improved the efficacy of respectively presence thinning and background thickening, but since we did not use additional data to optimize either choice, we believe our comparison is fair.

Model building

We built models identically from all 300 training data sets, using the default workflow and settings in the 'MIAmaxent' R package (ver. 1.0.0; Vollering et al. 2018). MIAmaxent creates models in much the same way as the popular Maxent software (Phillips et al. 2006), but replaces lasso regularization with forward stepwise selection to produce simpler models (Halvorsen et al. 2016). Like Maxent, MIAmaxent transforms each explanatory variable into a number of 'derived variables' (Halvorsen 2013, Halvorsen et al. 2015) and parameterizes models following the principle of maximum entropy (Fithian and Hastie 2013). Forward stepwise selection in MIAmaxent proceeds in two hierarchical stages, following Halvorsen et al. (2015): first, a parsimonious set of derived variables is selected from those created for each explanatory variable; second, these sets of derived variables are treated as inseparable units and a second round of selection picks among them.

Model evaluation

We evaluated predictions from each model by two metrics. First, we quantified similarity to the true presence probability distribution using Warren's I, which measures similarity between probability distributions (Warren et al. 2008, including erratum). Second, we quantified agreement with the realized occurrence distribution using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, which measures discrimination (Fielding and Bell 1997). Both metrics were calculated using all locations in the simulation area, and with both we employed Student's t-test for paired observations to infer differences between the three sampling bias approaches.

Sitka spruce case study

Presence records and explanatory variables

We obtained Sitka spruce presence records georeferenced to the United States or Canada from the global biodiversity information facility (GBIF; <http://doi.org/10.15468/dl.bsgw6c>) and reviewed the precision of their coordinates manually (Supplementary material Appendix B). Our explanatory variables comprised the first 19 BIOCLIM variables (Nix 1986, Booth 2018). We accessed two different data products that supply these 19 variables – WorldClim (ver. 2.0; Fick and Hijmans 2017) and CHELSA (ver. 1.2; Karger et al. 2017) – and employed both in parallel, to assess the effect of climate data uncertainty in the models (Beale and Lennon 2012, Morales-Barbero and Vega-Álvarez 2018). Explanatory variables were reprojected to equal-area grid cells with 1 km^2-resolution, and the study area was delineated with a 200-km buffer around presences.

Sampling bias correction methods

Models of Sitka spruce took one of four sampling bias approaches: 1) no correction, 2) target group background selection, 3) presence thinning or 4) background thickening. Models without bias correction were trained with presences and 10 000 uniformly sampled, uninformed background locations. The three remaining approaches operated with either alternative presences or alternative uninformed background locations. We implemented target group background selection using the pine family, Pinaceae, as the target group. Study area locations where GBIF recorded Pinaceae present (<https://doi.org/10.15468/dl.j6g0ip>) were used as uninformed background locations (unless they belonged to the background as Sitka spruce presences). We implemented presence thinning using the 'spThin' R package (ver. 0.1.0; Aiello-Lammens et al. 2015) with a minimum separation distance of 25 km. Without additional information, choosing a minimum separation distance is necessarily subjective, owing to the confounding nature of sampling probability (Aiello-Lammens et al. 2015). We chose 25 km, which was the longest

distance that appeared to maintain coverage of the full extent of the species' range. We implemented background thickening by the procedure used in the virtual species simulation. Specifically, uninformed background data were selected in proportion to the number of presences within a radius length of their location, with the radius defined as the mean spatial autocorrelation range of relevant explanatory variables.

Model building

We built eight distribution models of Sitka spruce, combining two climate data products and four sampling bias approaches. The model building procedure was identical for all eight training data sets, so we can attribute differences between models using the same climate data product entirely to differences in sampling bias approach. We built models using the 'MIAmaxent' package, like in the simulation (ver. 1.0.0; Vollering et al. 2018). However, in the case study, if a selected set of derived variables resulted in an unreasonable model response to a particular explanatory variable – defined as a response with one or more local minima – we repeated the first stage of derived variable selection under a stricter selection criterion (Supplementary material Appendix B). If the response remained unreasonable, we excluded the offending explanatory variable and repeated the full selection process. These adjustments were motivated by insights from gradient analysis of vegetation, which suggest that species responses to complex gradients are generally unimodal (Austin 2007, Halvorsen 2012).

Model evaluation

The presence-only data used to train the models are inappropriate for evaluating the results of bias correction, because they themselves are the source of bias (Veloz 2009, Halvorsen 2012). Spatially stratified cross validation does not solve this problem, because it neutralizes the effect of biases that are specific to data partitions, but not the effect of biases that are uniformly distributed across all data (Radosavljevic and Anderson 2014). Lacking surveyed presence–absence data, we evaluated models first by comparing their mapped predictions to documented estimates of Sitka spruce's distribution (Harris 1984, Peterson et al. 1997), assessing how well the models discriminated between estimated presence and absence. Second, we used systematically collected plot data from the United States' Forest Inventory and Analysis (FIA) program to assess how well model predictions were calibrated. Specifically, we measured how closely model predictions corresponded to relative presence probability in the FIA survey. FIA survey plots are spaced on average 5 km apart, and provide reliable observations of Sitka spruce presence (Bechtold and Patterson 2005). However, in the publicly available data, plot georeferences are intentionally inaccurate by up to 1.6 km, and up to 20% of georeferences are also swapped between plots in the same county (Burill et al. 2018). To keep this spatial muddling from impairing our analysis, we quantified relative presence probability at the county level. We discarded counties whose plot density in the database deviated from the norm, because estimating relative presence probability from presences alone requires that sampling intensity is uniform. Among the remaining counties in the study area, 1831 plots recorded Sitka spruce presence. We used the relative sums of model predictions within each of these counties to calculate how many of the 1831 presences the model predicted to occur there, and compared these predicted frequencies with the empirical frequencies to assess overprediction and underprediction in space.

Software and data deposition

Analyses were performed in R, ver. 3.5.1 (<www.r-project.org>), unless stated otherwise. All data and code necessary to reproduce results is deposited in the Dryad Digital Repository: <https://doi.org/10.5061/dryad.bb6f284> (Vollering et al. 2019).

Results

Virtual species simulation

The mean number of presences remaining after presence thinning was 76 for samples of 250, and 18 for samples of 25. The lengths of the background thickening radii ranged from 600 to 1300 km, but the effect of background thickening did not covary strongly with radius length (Supplementary material Appendix C).

With samples comprising 250 presences, both metrics of performance were highest for models employing presence thinning, followed by models using background thickening, and then models without bias correction (Fig. 2). With samples comprising 25 presences, predictions from models using background thickening clearly outperformed the other two sampling bias approaches, whose results were similar.

Sitka spruce distribution models

Background thickening radii (800, 850 km) were longer than the buffer radius used to delineate the study area (200 km), so background thickening altered the density of uninformed background locations but did not restrict their spatial extent (Fig. 3, Supplementary material Appendix D). The four sampling bias approaches resulted in conspicuously different variable selections; among directly comparable models (those using the same climate data), only one explanatory variable was selected in all four cases, and a number of explanatory variables were selected under a single sampling bias approach only (Table 1).

Models using target group background selection deviated strongly from the rest by predicting high relative presence probability for inland parts of the northern half of the study area (Fig. 4). Predictions from the other models were high predominantly in coastal regions. Models without bias correction differed from those applying presence thinning or background thickening by predicting much higher relative

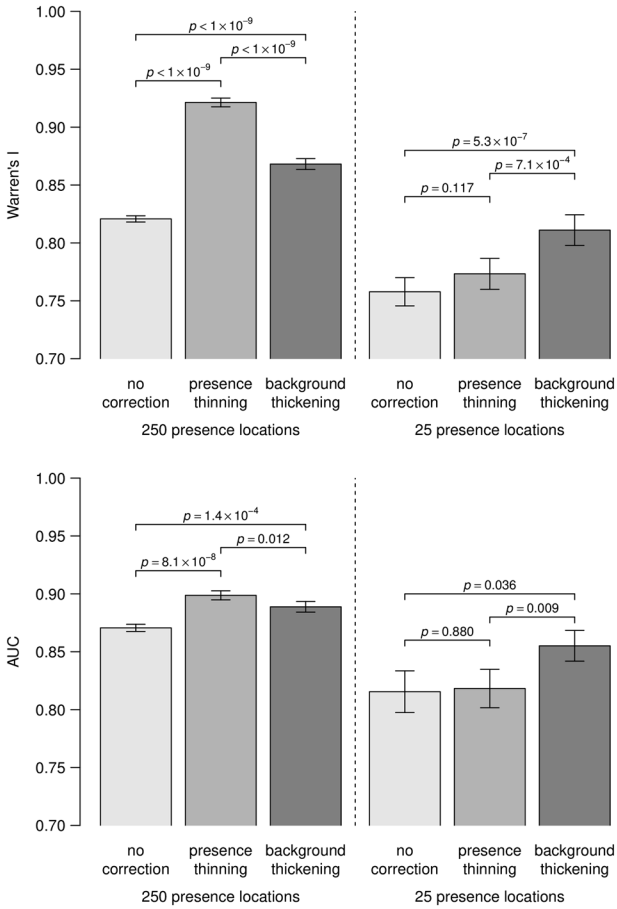

Figure 2. Similarity of model predictions to the true probability distribution (Warren's I; top panel), and their ability to discriminate between realized presences and absences (AUC; bottom panel), under different combinations of sampling bias approach (no correction, presence thinning, background thickening) and sample size (250 presences, 25 presences). Each bar represents the mean of fifty samples and error bars show the standard error of the mean. Brackets show the p-value for the null hypothesis that two population means are equal, from Student's t-test for paired samples.

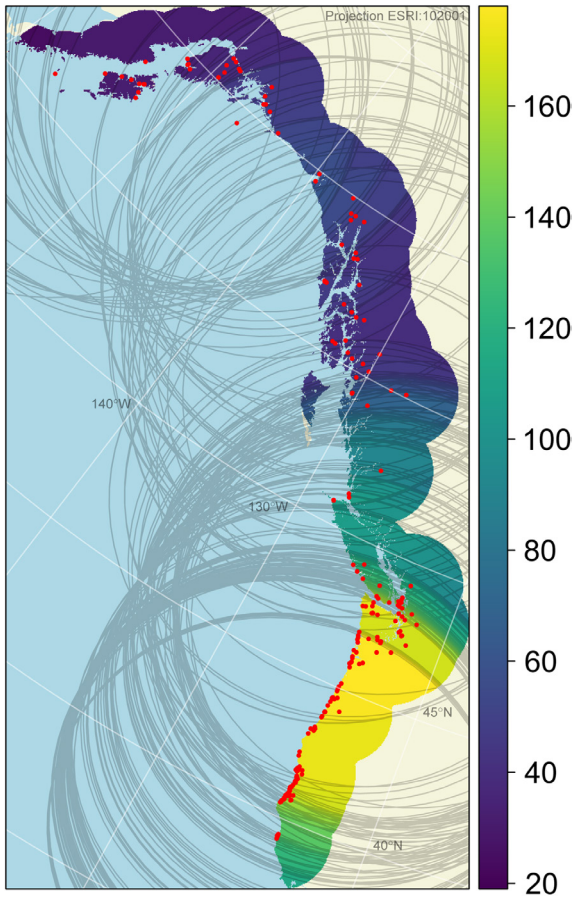

Figure 3. The study area's relative probability of inclusion in the thickened background, for the model of Sitka spruce using CHELSA climate data. The 850 km background thickening radius (grey ellipse) around each Sitka spruce presence (red dot) is shown.

presence probability in the southernmost part of the range than in any other part. Accordingly, models without bias correction showed stronger overprediction in the south and stronger underprediction in the north than those applying presence thinning or background thickening (Fig. 5). Whether the models used WorldClim or CHELSA data made no difference to these patterns (Supplementary material Appendix E).

Discussion

Does background thickening work?

Our results show that background thickening improves the predictive performance of distribution models trained on biased samples. Background thickening causes models to recover a simulated distribution with greater fidelity and reproduce occurrence patterns of Sitka spruce with greater realism, compared to models without correction. Particularly, Sitka spruce models were better calibrated across regions of varying collection intensity than models without correction. The fact that AUC calculated from the biased presence sample was higher for models without correction than models using background thickening (Table 1) reinforces the argument that without unbiased, independent data, AUC is a misleading metric of model quality (Veloz 2009, Jiménez-Valverde 2012, Fourcade et al. 2018).

Background thickening is superior to target group background selection for models of Sitka spruce, because the latter results in predictions that accord very poorly with Sitka spruce's known distribution. These correction methods both aim to neutralize bias in presences by reproducing the same bias in the background, but target group background selection will usually represent a much stronger departure from uniform background sampling than background thickening, because the environmental bias of target group presences can be arbitrarily strong (Støa et al. 2018), while the environmental bias of a thickened background is constrained by the strength of spatial autocorrelation. Therefore, target group

Table 1. Comparison of Sitka spruce distribution models, differing in sampling bias approach and climate data product. 'Derived variable selection threshold' refers to the p-value used as the threshold for selection during forward stepwise selection of derived variables. 'Excluded explanatory variables' refers to sets of derived variables that were eliminated during variable selection because they resulted in unreasonable model responses. 'No. derived variables' shows the total number of derived variables in the model, while 'Explanatory variables' shows which explanatory variables these derived variables represent. 'Training AUC' measures threshold-independent discrimination, and was calculated using all 243 cleaned presences and all uninformed locations in the study area, to allow direct comparison between models (Lobo et al. 2008).

Sampling bias approach	Climate data product	No. presences	No. background	Derived variable selection threshold	Excluded explanatory variables	No. derived variables	Explanatory variables	Training AUC
No correction	WorldClim	243	10 243	0.01	–	8	bio04 bio15 bio08 bio10	0.918
No correction	CHELSA	243	10 243	0.01	bio03 bio15 bio08	12	bio06 bio02 bio10 bio12 bio05	0.931
Target group background selection	WorldClim	243	3983	0.001	bio03	6	bio07 bio15 bio08 bio10	0.807
Target group background selection	CHELSA	243	4155	0.001	–	6	bio02 bio15 bio06 bio13	0.794
Presence thinning	WorldClim	97	10 097	0.01	–	5	bio07 bio10 bio19	0.919
Presence thinning	CHELSA	97	10 097	0.01	–	5	bio07 bio10 bio02	0.925
Background thickening	WorldClim	243	10 243	0.001	–	7	bio07 bio08 bio15 bio10	0.904
Background thickening	CHELSA	243	10 243	0.01	bio03 bio15	10	bio06 bio02 bio04 bio08 bio12	0.918

background selection may be a better counterweight to very strong environmental bias, but it is also more likely to overcorrect. Indeed, in our case study, target group background selection eliminated overprediction in the south compared to no correction, but simultaneously obliterated the pattern of coastal affinity in the northern part of the range.

Several authors have attributed the shortcomings of target group background selection to spatially uneven species richness in the target group – species-poor environmental conditions are underrepresented in the background, so presence probability tends to be overestimated there, and vice versa (Warton et al. 2013, Ranc et al. 2017). But the confounding

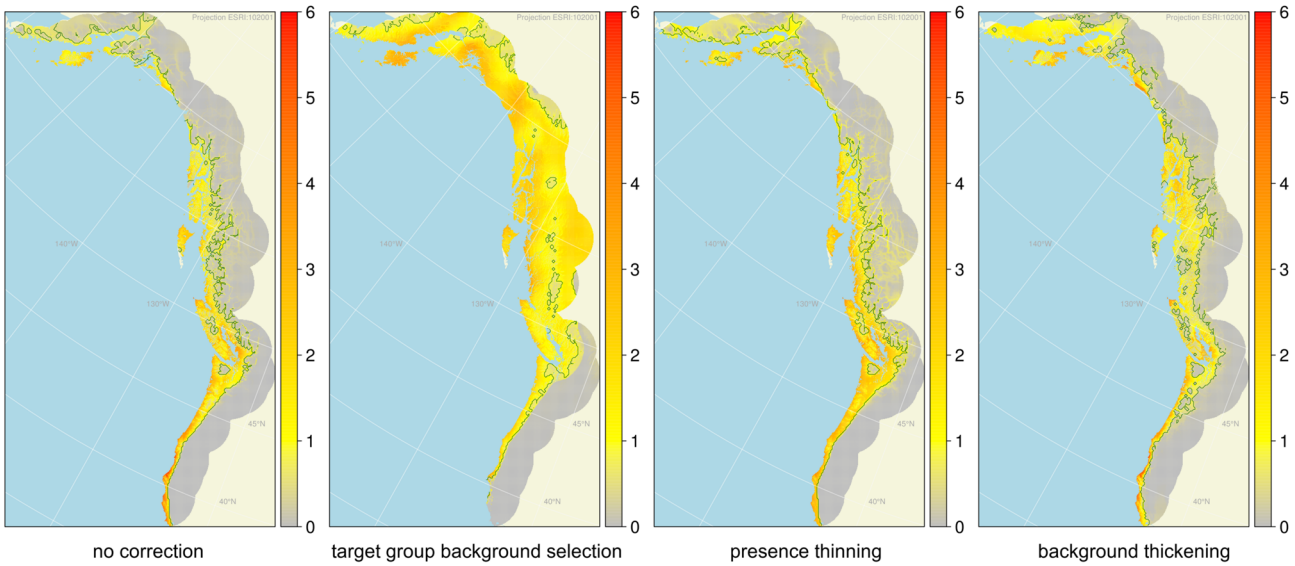

Figure 4. Relative presence probability of Sitka spruce, predicted by models using CHELSA climate data and four different sampling bias approaches. Predictions are given in probability ratio output (PRO) units, which means the value of one is the expected relative presence probability in a randomly drawn training data location – i.e. an 'average' training data location has PRO = 1 (Halvorsen 2013). To visualize differences between large or small values, the color scale represents log2(PRO + 1). The dark green lines show the 5% omission threshold in model predictions – i.e. the value above which 95% of presences occur.

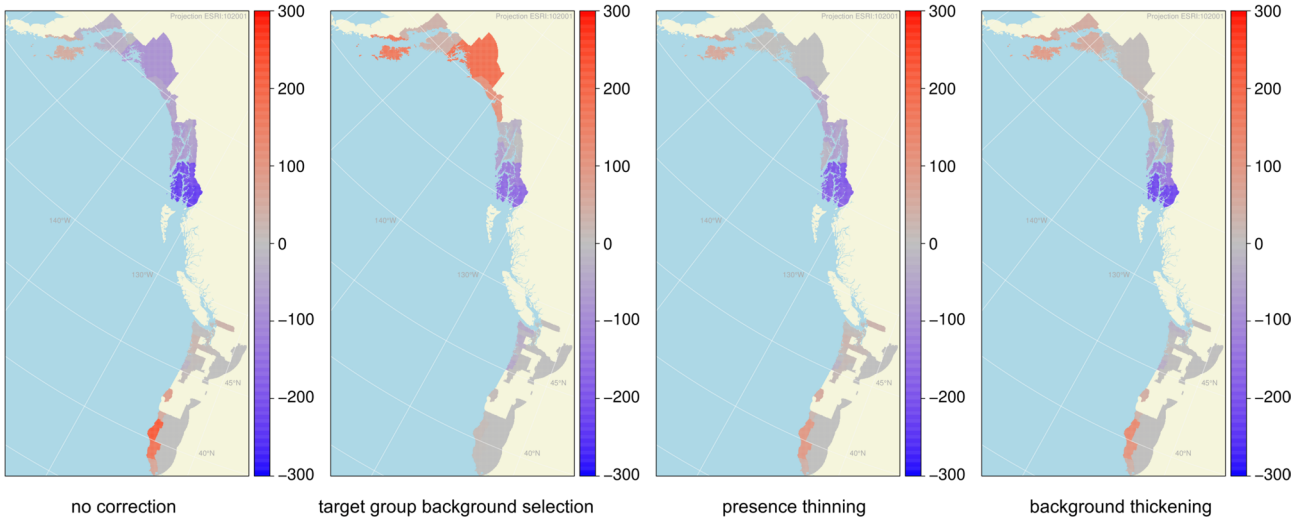

Figure 5. Overprediction and underprediction of relative presence probability of Sitka spruce, compared to surveyed presence in the forest inventory and analysis (FIA) program. Predictions are from models using CHELSA climate data and four different sampling bias approaches. Plotted values represent the difference between the actual number of surveyed Sitka spruce presences in each county (sum = 1831), and the fraction of 1831 presences expected to occur in each county based on model predictions (red shows overprediction while blue shows underprediction). The number of surveyed presences is considered only in U.S. counties with equal plot density in the FIA database.

effect of species richness actually indicates a more general problem: that the target group as a whole has an environmental tendency. Unless the combined presence probability of the target group is uniform across conditions, the target group's presence records will reflect presence probability as well as sampling probability, rather than sampling probability alone. Under these circumstances, the target group background introduces bias. For example, if the target group as a whole has an affiliation for continental climate, but the modeled species is neutral with respect to continentality, the target group background will cause the modeled species to show an affiliation for oceanic climate (Støa et al. 2018). Moreover, even if target group occurrence is environmentally neutral, there is another potential drawback: the restricted number of background observations might result in too coarse characterization of the background (Renner et al. 2015, Støa et al. 2018) or unwanted extrapolation beyond training data (Elith et al. 2010; Supplementary material Appendix F).

The performance of background thickening compared to presence thinning varies importantly depending on the data at hand; background thickening is better for small samples of the virtual species while presence thinning is better for large samples of the virtual species. Both methods result in similarly realistic predictions of Sitka spruce's distribution. These results support our hypothesis that retaining all presences gives background thickening an advantage over presence thinning in certain modeling situations. Specifically, when the number of presences is sufficiently small, the bias correction benefit obtained by discarding presences is apparently outweighed by an accompanying decline in the accuracy of parameter estimation. This tradeoff is implicit in all applications of presence thinning, regardless of the strength of the subsampling; that is, stronger thinning of presences may counteract bias more effectively, but it simultaneously increases model variance. At some point along the gradient towards fewer presences, the net effect becomes detrimental. Among studies that demonstrate improvement to model accuracy by presence thinning (Kramer-Schadt et al. 2013, Verbruggen et al. 2013, Beck et al. 2014, Boria et al. 2014), those with comparatively fewer available records show only small improvements (Verbruggen et al. 2013, Boria et al. 2014). Having tested only two sample sizes here, it might be tempting in future work to search for the tipping point in sample size below which background thickening surpasses presence thinning. However, other factors such as sample redundancy or niche breadth likely affect this balance, so aiming for a universal recommendation for choosing between these two methods based on sample size is probably misguided. Notably, in our case study, the choice between WorldClim and CHELSA climate data caused greater variation in model predictions than the choice between background thickening and presence thinning (Supplementary material Appendix G), which suggests that these two correction methods sometimes have quite similar effects, compared to other sources of model variation.

To build on our proof of concept, further research should 1) evaluate the performance of background thickening across a more comprehensive range of real and simulated modeling scenarios and 2) compare different implementations of background thickening.

How does background thickening work?

Ostensibly, background thickening resembles using presence density to correct bias formally – as others have done with Maxent's bias file (Elith et al. 2010, Clements et al. 2012, Kramer-Schadt et al. 2013, Fourcade et al. 2014). Indeed,

background thickening is like sampling background locations from an estimate of sampling probability based exclusively on density of training presences. However, background thickening is actually fundamentally different from Maxent's bias file method. Merow et al. (2013) clarify the distinction as 'biased background' versus 'biased prior' methods. The bias file prompts Maxent to find the distribution most similar to the sampling probability distribution (the prior), subject to constraints dictated by presences, and then to divide by the sampling probability distribution (Merow et al. 2016). Thus, model predictions indicate how much the presence pattern deviates from the estimated sampling probability distribution. Because sampling probability is factored out formally by division, model predictions are directly dependent on this estimate, and very high or low values can lead to illogical results. For example, if their estimated sampling probability is sufficiently low, conditions entirely without presence may be assigned larger presence probability than conditions saturated with presence (Merow et al. 2013 Appendix 5). The same outcome is unlikely to occur with background thickening, because informal methods avoid dividing by an estimate of sampling probability. Under background thickening, conditions without presence could be assigned high presence probability only if the ratio of presences to uninformed background were high at similar conditions. Conversely, conditions saturated with presence could be assigned low presence probability only if similar conditions were overrepresented in the uninformed background. In either case, the modeled outcome is easily justified.

Essentially, background thickening emphasizes the comparison between a presence and its surroundings; a presence with conditions which are exceptional compared to its surroundings will inform model predictions more than a presence with conditions which are commonplace compared to its surroundings. This emphasis matches our intuition: without additional information, finding a species on the only mountain in a lowland area does more to convince us that its presence probability is high on mountains than finding it on a single peak surrounded by many other mountains. In the former case we know the empirical presence probability on mountains to be one, while in the latter it could easily be much lower than one. The stronger emphasis on proximal comparison under background thickening means that a difference in the number of presences between two geographically and environmentally distinct regions will have relatively little effect on a model, unless the characteristics that differentiate the regions also differentiate presences from background within both regions. It is worth noting that such regional differences may arise not only from sampling bias but also from disequilibrium in the species' distribution (Elith et al. 2010).

When does background thickening work?

Correcting sampling bias in opportunistically collected presence data is inherently heuristic, because the true sampling probability distribution is unknown. Without added, reliable information about the true sampling or presence probability distributions, the extent to which a correction reduces existing bias or introduces new bias always remains ambiguous (Yackulic et al. 2013). Target group background selection and formal methods like Maxent's bias file make strong assumptions about the true sampling probability distribution. If these assumptions are not underpinned by strong justifications – such as surveyed presence–absence data from similar species (Fithian et al. 2015), or a mechanistic understanding of the sampling process (Ponder et al. 2001) – we contend that background thickening and presence thinning are preferable methods, since they treat the presences themselves as the only reliable information. Presence thinning and background thickening are blunt instruments, but they are less likely than target group background selection and formal correction methods to overcorrect. Applying these two methods in parallel, given the heuristic nature of bias correction, can improve confidence (Fourcade et al. 2014).

The limitations of presence thinning and background thickening may be understood by considering our working definition of sampling bias as covariation between sampling probability and explanatory variables (Yackulic et al. 2013). Covariation may stem from sampling probability being: 1) spatially clustered (i.e. positively spatially autocorrelated), 2) unclustered but correlated to explanatory variables or 3) a combination of the two. Background thickening and presence thinning are expected to counteract effects of bias arising by spatial clustering only, because they operate on spatial proximity, while bias of the second kind is inherently difficult to tease out without additional information (Fithian et al. 2015). To illustrate: suppose presences are evenly spaced along a uniform road network, roads track low elevations, and true presence probability is independent of elevation. If elevation is used as an explanatory variable, the resultant low-elevation bias is remedied neither by presence thinning nor by background thickening. In fact, any correction method not dependent on additional occurrence data will struggle to alleviate this type of bias. The difficulty is that as the strength of the bias increases with the strength of the road-elevation correlation, so too does the number of presences necessary to disentangle the effect of roads from the effect of elevation (Fithian and Hastie 2013, Fithian et al. 2015). Thus, only correction methods using additional occurrence data are able to address this kind of bias, to the extent that the additional data characterize sampling probability accurately.

Conclusions

We find that 'background thickening' – selecting background locations to mirror the density of presence locations – is a suitable and potentially valuable option for correcting sampling bias in presence-background models. Furthermore, we find that background thickening may be preferable to other sampling bias correction methods in data-poor modeling circumstances, when the sampling probability distribution is hard to infer and the presence sample is small. Thus, background thickening helps extract knowledge about species' distributions and occurrence–environment relationships

from confounded presence data, especially for species that we know little about. Since background thickening mitigates the negative effects of spatially autocorrelated sampling, we wonder whether it might also help presence-background models address challenges associated with endogenous spatial autocorrelation in presence records (Dormann 2007) – including spatial autocorrelation brought about by range shifts (Elith et al. 2010).

Acknowledgements – Thank you to members of the Geo-Ecology research group at the Natural History Museum, Univ. of Oslo, for comments on early results of this work.
Funding – This work was funded through a PhD grant to JV from Sogn og Fjordane University College (now Western Norway University of Applied Sciences).
Author contributions – All authors conceived and designed the study. JV carried out the analyses, JV and RH interpreted the results, JV drafted the manuscript and all authors revised the manuscript critically.

References

Aarts, G. et al. 2012. Comparative interpretation of count, presence-absence and point methods for species distribution models. – Methods Ecol. Evol. 3: 177–187.
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.
Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. – Ecol. Model. 200: 1–19.
Beale, C. M. and Lennon, J. J. 2012. Incorporating uncertainty in predictive species distribution modelling. – Phil. Trans. R. Soc. B 367: 247–258.
Bechtold, W. A. and Patterson, P. L. 2005. The enhanced forest inventory and analysis program – national sampling design and estimation procedures. – Gen. Tech. Rep. <www.fs.usda.gov/treesearch/pubs/20371>.
Beck, J. et al. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. – Ecol. Inform. 19: 10–15.
Booth, T. H. 2018. Why understanding the pioneering and continuing contributions of BIOCLIM to species distribution modelling is important. – Austral Ecol. 43: 852–860.
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.
Burill, E. A. et al. 2018. The forest inventory and analysis database: database description and user guide ver. 8.0 for phase 2. – United States Forest Service.
Clements, G. R. et al. 2012. Predicting the distribution of the Asian tapir in Peninsular Malaysia using maximum entropy modeling. – Integr. Zool. 7: 400–406.
Dormann, C. F. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. – Global Ecol. Biogeogr. 16: 129–138.
Elith, J. et al. 2010. The art of modelling range-shifting species. – Methods Ecol. Evol. 1: 330–342.
Fick, S. E. and Hijmans, R. J. 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. – Int. J. Climatol. 37: 4302–4315.
Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.
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.
Fourcade, Y. et al. 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. – Global Ecol. Biogeogr. 27: 245–256.
Gräler, B. et al. 2016. Spatio-temporal interpolation using gstat. – R J. 8: 204–218.
Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.
Guisan, A. et al. 2017. Habitat suitability and distribution models: with applications in R. – Cambridge Univ. Press.
Halvorsen, R. 2012. A gradient analytic perspective on distribution modelling. – Sommerfeltia 35: 1–165.
Halvorsen, R. 2013. A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. – Sommerfeltia 36: 1–132.
Halvorsen, R. et al. 2015. Opportunities for improved distribution modelling practice via a strict maximum likelihood interpretation of MaxEnt. – Ecography 38: 172–183.
Halvorsen, R. et al. 2016. How important are choice of model selection method and spatial autocorrelation of presence data for distribution modelling by MaxEnt? – Ecol. Model. 328: 108–118.
Harris, A. S. 1984. Sitka spruce: an American wood. – United States Dept of Agriculture.
Hijmans, R. J. et al. 2008. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. – Conserv. Biol. 14: 1755–1765.
Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. – Global Ecol. Biogeogr. 21: 498–507.
Karger, D. N. et al. 2017. Climatologies at high resolution for the earth's land surface areas. – Sci. Data 4: 170122.
Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. – Divers. Distrib. 19: 1366–1379.
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. 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? – J. Biogeogr. 35: 105–116.
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. 2016. Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. – Global Ecol. Biogeogr. 25: 1022–1036.
Merow, C. et al. 2017. Integrating occurrence data and expert maps for improved species range predictions. – Global Ecol. Biogeogr. 26: 243–258.
Morales-Barbero, J. and Vega-Álvarez, J. 2018. Input matters matter: bioclimatic consistency to map more reliable species distribution models. – Methods Ecol. Evol. 10: 212–224.
Nix, H. A. 1986. A biogeographic analysis of Australian elapid snakes. – In: Longmore, R. (ed.), Atlas of elapid snakes of Australia: Australian flora and fauna series 7. Bureau of Flora and Fauna, pp. 4–15.
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.
Pebesma, E. J. 2004. Multivariable geostatistics in S: the gstat package. – Comput. Geosci. 30: 683–691.
Peterson, E. B. et al. 1997. Ecology and management of Sitka spruce, emphasizing its natural range in British Colombia. – UBC Press.
Phillips, S. J. and Elith, J. 2013. On estimating presence probability from use-availability or presence-background data. – Ecology 94: 1409–1419.
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.
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.
Ranc, N. et al. 2017. Performance tradeoffs in target-group bias correction for species distribution models. – Ecography 40: 1076–1087.

Supplementary material (available online as Appendix ecog-04503 at <www.ecography.org/appendix/ecog-04503>). Appendix A–G.

Renner, I. W. et al. 2015. Point process models for presence-only analysis. – Methods Ecol. Evol. 6: 366–379.
Støa, B. et al. 2018. Sampling bias in presence-only data used for species distribution modelling: theory and methods for detecting sample bias and its effects on models. – Sommerfeltia 38: 1–53.
Stolar, J. and Nielsen, S. E. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. – Divers. Distrib. 21: 595–608.
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.
Vollering, J. et al. 2018. MIAmaxent: a modular, integrated approach to maximum entropy distribution modeling. – <https://CRAN.R-project.org/package=MIAmaxent>.
Vollering, J. et al. 2019. Data from: Bunching up the background betters bias in species distribution models. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.bb6f284>.
Warren, D. L. et al. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. – Evolution 62: 2868–2883.
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.
Zurell, D. et al. 2010. The virtual ecologist approach: simulating data and observers. – Oikos 119: 622–635.