*Ecological Applications*, 19(1), 2009, pp. 181–197


# Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data

Steven J. Phillips,[1,8] Miroslav Dudík,[2] Jane Elith,[3] Catherine H. Graham,[4] Anthony Lehmann,[5] John Leathwick,[6] and Simon Ferrier[7]

[1]*AT&T Labs—Research, 180 Park Avenue, Florham Park, New Jersey 07932 USA*
[2]*Computer Science Department, Princeton University, 35 Olden Street, Princeton, New Jersey 08544 USA*
[3]*School of Botany, University of Melbourne, Parkville, Victoria 3010 Australia*
[4]*Department of Ecology and Evolution, 650 Life Sciences Building, Stony Brook University, New York 11794 USA*
[5]*Climatic Change and Climate Impacts, University of Geneva, 7 Route de Drize, 1227 Carouge, Switzerland*
[6]*NIWA, Hamilton, New Zealand*
[7]*New South Wales Department of Environment and Climate Change, P.O. Box 402, Armidale 2350 Australia*

*Abstract*. Most methods for modeling species distributions from occurrence records require additional data representing the range of environmental conditions in the modeled region. These data, called background or pseudo-absence data, are usually drawn at random from the entire region, whereas occurrence collection is often spatially biased toward easily accessed areas. Since the spatial bias generally results in environmental bias, the difference between occurrence collection and background sampling may lead to inaccurate models. To correct the estimation, we propose choosing background data with the same bias as occurrence data. We investigate theoretical and practical implications of this approach. Accurate information about spatial bias is usually lacking, so explicit biased sampling of background sites may not be possible. However, it is likely that an entire target group of species observed by similar methods will share similar bias. We therefore explore the use of all occurrences within a target group as biased background data. We compare model performance using target-group background and randomly sampled background on a comprehensive collection of data for 226 species from diverse regions of the world. We find that target-group background improves average performance for all the modeling methods we consider, with the choice of background data having as large an effect on predictive performance as the choice of modeling method. The performance improvement due to target-group background is greatest when there is strong bias in the target-group presence records. Our approach applies to regression-based modeling methods that have been adapted for use with occurrence data, such as generalized linear or additive models and boosted regression trees, and to Maxent, a probability density estimation method. We argue that increased awareness of the implications of spatial bias in surveys, and possible modeling remedies, will substantially improve predictions of species distributions.

*Key words: background data; presence-only distribution models; niche modeling; pseudo-absence; sample selection bias; species distribution modeling; target group.*

## Introduction

Species distribution modeling (SDM) is an important tool for both conservation planning and theoretical research on ecological and evolutionary processes (Loiselle et al. 2003, Kozak et al. 2008). Given sufficient resources, SDM can be based on data gathered according to rigorously defined sampling designs, where both presence and absence of species is recorded at an environmentally and spatially representative selection of sites (Cawsey et al. 2002). However, for most areas of the world and most species, resources are too limited to gather large sets of data including both presences and absences, and furthermore, many species have been extirpated from much of their original range. For these reasons, SDM relies heavily on presence-only data such as occurrence records from museums and herbaria (Ponder et al. 2001, Graham et al. 2004, Suarez and Tsutsui 2004). These occurrence data often exhibit strong spatial bias in survey effort (Dennis and Thomas 2000, Reddy and Dávalos 2003, Schulman et al. 2007), meaning simply that some sites are more likely to be surveyed than others; such bias is typically spatially autocorrelated, but this paper allows for arbitrary spatial bias. This bias, referred to as sample selection bias or survey bias, can severely impact model quality; however, the effect of such bias has received little attention in the SDM literature. We present a theoretical analysis of sample selection bias for several presence-only SDM methods. We also describe a general

Manuscript received 31 December 2008; revised 14 May 2008; accepted 21 May 2008. Corresponding Editor: D. F. Parkhurst.
[8] E-mail: phillips@research.att.com

approach for coping with biased occurrence data, and empirically test its efficacy.

The range of model types for fitting presence-only data has expanded rapidly over the last decade. In ecology, the most common methods for these data were originally those that fitted envelopes or measured point-to-point similarities in environmental coordinates (Busby 1991, Carpenter et al. 1993). These methods use only occurrence data, ignoring the set of environmental conditions available to species in the region. More recent methods achieve better discrimination by modeling suitability relative to the available environment. Information on the available environment is provided by a sample of points from the study region. We refer to these points as background or pseudo-absence data. Examples of specialized programs include Hirzel's ecological niche factor analysis ("ENFA" or "Biomapper"; Hirzel et al. 2002) and Stockwell and Peterson's genetic algorithm for rule-set prediction ("GARP"; Stockwell and Peters 1999, Peterson and Kluza 2003). More generally, a broad range of logistic regression methods can be adapted to this situation, either in an approximation (modeling presences against background rather than against absences) or with more rigorous statistical procedures that correct for the possibility of true presences appearing in the background data (Keating and Cherry 2004; Ward et al., *in press*). Because the regression-related methods and other newer initiatives show generally higher predictive performance than other approaches (e.g., Elith et al. 2006, Hernandez et al. 2006), we focus here on a subset of more successful, widely used methods: boosted regression trees (BRT; Leathwick et al. 2006, De'ath 2007), maximum entropy (Maxent; Phillips et al. 2006), multivariate adaptive regression splines (MARS; Leathwick et al. 2005), and generalized additive models (GAM; Yee and Mitchell 1991, Ferrier et al. 2002).

These methods all require information about the range of environmental conditions in the modeled region, given by background samples. Some modelers think of the background samples as implied absences: partly because the word "pseudo-absences" gives that impression. However, the intention in providing a background sample is not to pretend that the species is absent at the selected sites, but to provide a sample of the set of conditions available to it in the region. The critical step in selection of background data is to develop a clear understanding of the factors shaping the geographic distribution of presence records. Two key elements are the actual distribution of the species and the distribution of survey effort. Potentially, the latter can be spatially biased, i.e., there may be sample selection bias. Most SDMs are fitted in environmental space without consideration of geographic space, so the importance of spatial bias is that it often causes environmental bias in the data. If a spatially biased sample proportionately covered the full range of environments in the region, then it would cause no problem in a model based on environmental data. However, this is usually not the case. If the bias is not accounted for, a fitted model might be closer to a model of survey effort than to a model of the true distribution of the species. For example, a species with a broad geographic distribution might only have been recorded in incidental surveys close to towns and beside roads. Background samples are commonly chosen uniformly at random from the study region; this characterizes the range of environments in the region well, but fails to indicate sample selection bias. If the roadsides and towns are not a random sample of the environment, applying any of the above modeling techniques to these data will produce a model that best describes the differences in the distribution of the presence sites compared to the background data. For example, if roads in this region happen to follow ridges, and if towns happen to be associated with the most fertile soils, then a model will find that ridges and fertile soils are positively correlated with the distribution of the species, whereas in reality they best describe the distribution of roads and towns, and hence survey effort.

The most straightforward approach to address this problem would be to manipulate the occurrence data in order to remove the bias, for example by discarding or down-weighting records in over-sampled regions (e.g., the de-biasing averages approach of Dudík et al. [2005]) or by surveying under-represented regions. However, such manipulations are hampered by incomplete information about the distribution of survey effort. In addition, the paucity of presence records for many species of interest makes discarding records unpalatable, and resources may not be available to conduct new surveys. The data may also be biased in a way that cannot be "fixed" by collecting new data: if many forested areas have been cleared, new surveys will not provide presence records of forest-dependent species in cleared areas. In the same way, less arid, more fertile areas are more likely to have been transformed by human activity, so new surveys would result in occurrence data that are biased toward arid or infertile areas. In these cases, the sample selection bias is an inherent part of the realized, current distribution of the species.

An alternative approach is to manipulate the background data. While some studies explore this idea (e.g., Zaniewski et al. 2002, Engler et al. 2004, Lütolf et al. 2006), the ecological literature lacks a coherent theoretical exploration, and the proposed solutions seem to represent different and probably incompatible reasoning. The approach we propose is to design the selection of background data so they reflect the same sample selection bias as the occurrence data. This aims to achieve the same environmental bias in both data sets. For example, if presence data are only taken from easily surveyed portions of the study region, then background data should be taken from the same areas (Ferrier et al. 2002). The hope is that a model based on biased

presence data and background data with the same bias will not focus on the sample selection bias, but will focus on any differentiation between the distribution of the occurrences and that of the background. In other words, if the species occupies particular habitats within the sampled space, the model will highlight these habitats, rather than just areas that are more heavily sampled. This has been justified theoretically for Maxent (Dudík et al. 2005; summarized here in *Maxent models for biased samples*). In the regression case, we could find no clear treatment of how to understand and interpret models using presence–pseudo-absence data, particularly with varying biases in the underlying data, so we present that here. We first investigate how to interpret models produced with random background, using the theory of use–availability sampling in habitat-selection studies (Keating and Cherry 2004). We extend the analysis to biased data, and show that under reasonable conditions, models created using background data with the same sample selection bias as the presence data can be interpreted in the same way as models produced with completely unbiased data.

It can be difficult to create background data with the same bias as presence data since we seldom know the sample selection distribution exactly. As an alternative, if presence records are derived from natural history collections, records for a broad set of species could be used to estimate survey effort. The set of species should be chosen so as to represent the specimen collection or observation activities of collectors of the target species. In general, the groups should contain species that are all collected or observed using the same methods or equipment; such groups of species are called target groups (Ponder et al. 2001, Anderson 2003). Broad biological groups (birds, vascular plants, and so on) are likely to be suitable. The sites for all records from all species in the target group then make up the full set of available information on survey effort and can be used as background data; we call such a set of sites target-group background.

To measure the effectiveness of target-group background, we compared it to random background using several modeling methods and the same data set as a recent comprehensive comparison of modeling methods (Elith et al. 2006). The data set covers 226 species from diverse regions of the world, with a wide range of sample sizes (2 to 5822, with a median of 57). The regions exhibit varying amounts of sample selection bias, with Ontario, Canada showing the most striking bias, toward the more populous south. A crucial aspect of this data set is that it contains independent, well-structured presence–absence test data. The test data were collected independently of the training data, using rigorous surveys in which the species' presence or absence was recorded at a collection of test sites. This allows us to evaluate model performance in a way that is largely unaffected by sample selection bias since the predictive performance of the models is evaluated on this test data, rather than the presence-only training data. We focus on average performance across broad groups of species rather than detailed expert evaluation of individual species models, and compare several of the better-performing methods from the study of Elith et al. (2006). This allows us to determine how sample selection bias impacts performance of presence-only species distribution models on typical data sets, and whether target-group background can effectively counteract sample selection bias on such data sets. Whilst the effect of background sample selection has been mentioned in relation to individual modeling methods (e.g., Lütolf et al. 2006, Elith and Leathwick 2007, Phillips and Dudík 2008), this paper focuses on the general problem and on its relevance across a range of species, environments, and modeling methods.

### *The dangers of sample selection bias: an example*

When presence–absence data are available, there are a number of modeling methods that are known to be resilient to sample selection bias (Zadrozny 2004). However, bias can have a powerful effect on models derived from presence-background data; to demonstrate this dichotomy, we briefly consider a synthetic species in Ontario, Canada, and use the continuous environmental variables described in Elith et al. (2006). The probability of presence for the species (Fig. 1) is defined to be 1 for any location which is within the middle 40% of the range of all environmental variables. For each variable outside of the middle 40% of its range, the probability of presence is multiplied by a factor ranging linearly from 0.7 (at the extremes of the variable's range) to 1.0 (at the 30th and 70th percentiles). The particular constants used here were chosen for illustrative purposes only, to create a synthetic species with a broad preference for mid-range conditions in all variables.

Occurrence data are often biased toward human population centers and roads (Reddy and Dávalos 2003). Therefore, roughly following the human population and road density of Ontario, we modeled sample selection bias with a sampling distribution that is uniform in the southern 25% of Ontario, uniform with $b$ times lower intensity in the northern 50% of the province, and a linear transition of sampling intensity in between; we varied $b$ between 1 (unbiased sampling) and 100 (strongly biased sampling). Several predictor variables for Ontario have a strong north–south trend, so this spatial bias will translate into a bias in predictor space. Samples were generated by repeatedly picking a site according to this sampling distribution and then randomly labeling the site either as a presence (with probability equal to the species' probability of presence there) or absence (with the remaining probability). Sampling continued until there were exactly 200 presences. Thus a full data set for each value of $b$ contained 200 presences and a variable number of absences, depending on how many were selected in creating the set of 200 presences. Two boosted

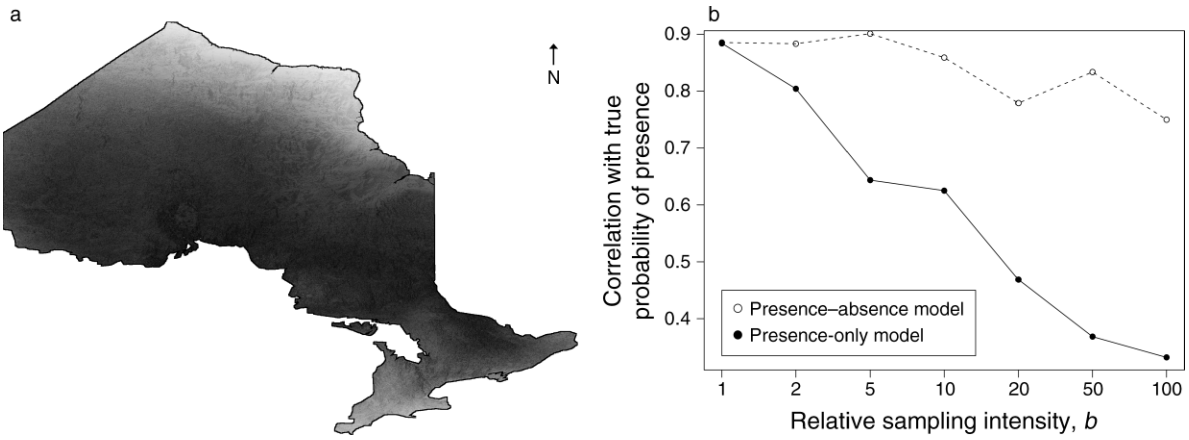


FIG. 1. Effect of sample selection bias on predictive accuracy for an artificial species in Ontario, Canada. (a) Probability of presence for the species, with darker shades indicating higher probabilities. (b) Correlation between model output and true probability of presence, measured across the whole region ($y$-axis), for various degrees of sample selection bias. Bias was introduced by sampling uniformly in the southern 25% of the region and uniformly $b$ times lower in the northern 50% of the region, with a linear transition in between; the $x$-axis shows values of $b$. Models were made using boosted regression trees with no interactions, fitted using fivefold cross-validation.

regression tree models (see *Modeling methods*) were then created: one with the set of presences and absences, and a second with the 200 presences together with 10 000 background samples chosen uniformly at random from the region, and weighted so that presence and background have equal weight, as in Elith et al. (2006). We used 10 000 samples as this is large enough to accurately represent the range of environmental conditions in the study region; more background samples do not improve model performance (Phillips and Dudík 2008).

The presence–absence models are highly correlated with true probability of presence, even under severe sample selection bias ($b = 100$). This happens because BRT is a "local" learner (Zadrozny 2004), so the model generated with biased training data converges asymptotically to the unbiased model (for large sample sizes) as long as two conditions hold: sampling probability is non-zero in the whole region, and sampling is conditionally independent of species presence given the environmental conditions. In contrast, for the presence-only models, correlation with true probability of presence quickly drops as sample selection bias increases (Fig. 1). For $b = 50$, the presence–absence model is visibly similar to true probability of presence, while the presence-only model appears only weakly related (Fig. 2). We note that the strong sample selection bias depicted in Fig. 2 may actually be very moderate compared to true occurrence data, where sampling intensity can vary by a factor of tens of thousands (Schulman et al. 2007: Fig. 4).

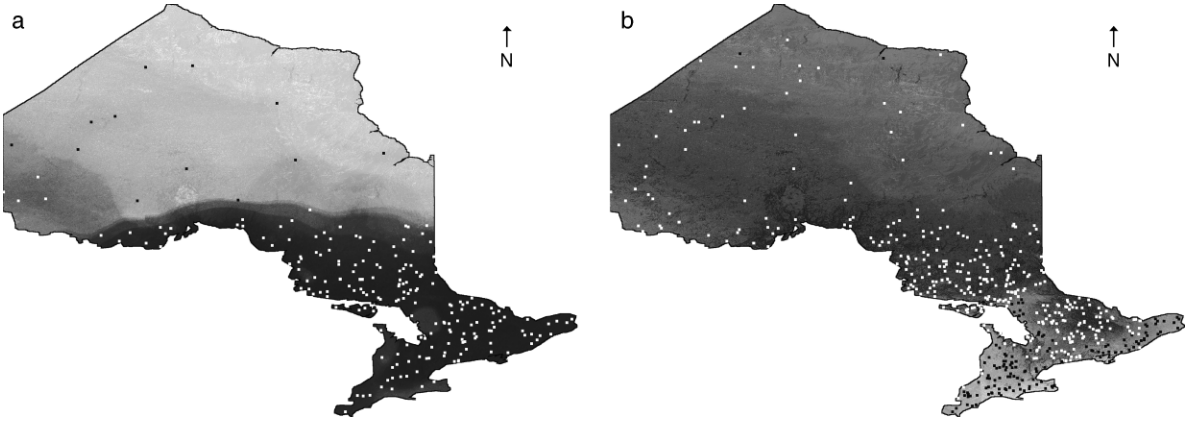


FIG. 2. Predicted probability of presence modeled from (a) biased presence-only data and (b) biased presence–absence data. Both models were generated using boosted single-node regression trees, fitted with fivefold cross-validation. Black and white dots show sampled locations used for model building. Sampling intensity in the southern 25% of the region was 50 times higher than in the northern 50% of the region, with a linear transition in between. The presence-only model is strongly influenced by the bias, whereas the presence–absence model is not: compare with the true probability of presence in Fig. 1.

## Models and Analysis

### *Preliminaries*

In the analyses that follow, we consider an area with a total of $N$ sites. For each site $t$, there are $v$ known covariates (measured environmental variables) denoted by $\mathbf{x} = (x_1, \ldots, x_v)$. An observation $(t, y)$ records whether at a particular time the species is present ($y = 1$) or absent ($y = 0$) at the site $t$. This treatment allows for the possibility that a species is present at a given site during one observation and absent in the next, as may happen for vagile species. The probability that the species is present at a site $t$, denoted $P(y = 1 \mid t)$, may therefore lie somewhere between 0 and 1. Formally, observations are taken from a distribution over a sample space consisting of pairs $(t, y)$, where $t$ is a site and $y$ is the response variable. We will use $P$ to denote probability under spatially unbiased sampling from this sample space, i.e., each site has equal probability $(1/N)$ of being sampled. For example, the prevalence of the species, denoted $P(y = 1)$, is the fraction of sites at which the species is present (for perfectly detectable non-vagile species), or the probability of observing the species at a randomly chosen site (for perfectly detectable vagile species).

A collection of observations is unbiased in environmental space if it samples each combination of environmental covariates proportionately to the amount of the study area that has those covariate values. Therefore, observations that are spatially unbiased are also environmentally unbiased, though the converse is not always true.

### *Modeling methods*

The modeling methods considered here use two distinct approaches for presence-only modeling. The first approach is derived from regression techniques, which are normally applied to presence–absence modeling. These methods estimate probability of presence from training data consisting of presences and absences for a given species. They have been adapted for use with presence-only data by treating the background data as if it were absence data. They are all logistic methods, modeling probability of presence as $P(y = 1 \mid \mathbf{x}) = \exp[f(\mathbf{x})]/(1 + \exp[f(\mathbf{x})])$ for some function $f$ of the environmental variables, and they differ mainly in the form of the function $f$. We used the following presence–absence methods:

1) Generalized additive models (GAM) use nonparametric, data-defined smoothers to fit nonlinear functions (Hastie and Tibshirani 1990, Yee and Mitchell 1991).

2) Multivariate adaptive regression splines (MARS) provide an alternative regression-based technique for fitting nonlinear responses. MARS uses piecewise linear fits rather than smooth functions and a fitting procedure that makes it much faster to implement than GAM (Friedman 1991, Elith and Leathwick 2007).

3) Boosted regression trees (BRT), also known as stochastic gradient boosting (Friedman 2001, Leathwick et al. 2006), use a form of forward stage-wise regression to construct a sum of regression trees. Each stage consists of a gradient-descent step, in which a regression tree is fitted to the derivatives of the loss function. Cross-validation is used to avoid overfitting by halting model growth based on predictive accuracy on withheld portions of the data.

The second approach is probability density estimation, where the presence data are assumed to be drawn from some probability distribution over the study region. The task is to estimate that distribution. This approach is represented here by a single method, called Maxent (Phillips et al. 2006, Dudík et al. 2007), described in *Maxent models with unbiased samples*. Whenever we present examples, we use either BRT or Maxent, since these are the two methods out of those considered here that performed best in the comparison of methods by Elith et al. (2006). The settings used for BRT have been improved over those used previously and we use a recent version of Maxent (version 3.0) with default settings. For both methods, therefore, the statistical performance we report for random background is improved over that presented by Elith et al. (2006).

### *Presence–absence models with random background*

Before we analyze the use of presence–absence models (such as BRT, GAM, and MARS) on presence–background data under bias, we must first understand the use of these methods on unbiased data. Using unbiased presence data and random background gives a sample model known in habitat-selection studies as a use–availability sampling design (Keating and Cherry 2004) and defined as follows. The full set of training data consists of a set of samples, each obtained either by randomly choosing a sample with $y = 1$ to get a presence sample (a fraction $p$ of the whole set), or randomly choosing a sample from the full set of $N$ sites to get a background sample (the remaining fraction, $1 - p$). This sampling model suffers from two complications. First, the set of background samples typically includes both sites with $y = 1$ and sites with $y = 0$, a problem referred to as contaminated controls (Lancaster and Imbens 1996). Second, the sampling intensity (probability that a given data point will be chosen as a sample) may differ between presence and background samples, which makes it a case-control sampling design. The relative sampling intensity is determined by the parameter $p$. Our goal in this section is to understand the effect of these two complications, and in particular, to determine exactly what quantity is being estimated when a model is fitted to use–availability data.

For mathematical simplicity in our analyses, we use two steps to model the process by which each training sample is derived. The first step is a random decision about whether the current sample will be presence (probability $p$) or background (probability $1 - p$). The second step is a random draw either from the population of presences or from the full set of available sites, according to the outcome of the first step.

We will use $P_{\mathrm{UA}}$ to denote probability under this sampling model. $P_{\mathrm{UA}}$ is formally defined as a joint probability model over triples $(t, y, s)$ where $s$ is an auxiliary variable representing sampling stratum: $s = 1$ for presence samples and $s = 0$ for background samples. Therefore, $P_{\mathrm{UA}}(s = 1) = p$ and $P_{\mathrm{UA}}(s = 0) = 1 - p$, and by definition,

$$P_{\mathrm{UA}}(\mathbf{x}|s = 1) = P(\mathbf{x}|y = 1)$$

and

$$P_{\mathrm{UA}}(\mathbf{x}|s = 0) = P(\mathbf{x}). \quad (1)$$

When a presence–absence model is applied to use–availability data, the response variable being modeled is $s$, not $y$, so we obtain an estimate of $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$, i.e., the probability that a site will be chosen as a presence sample rather than a background sample, conditioned on the environmental variables. It is crucial to note that this is not the same as $P(y = 1 \mid \mathbf{x})$, the probability of occurrence conditioned on the environmental variables. Indeed, if we define

$$r = \frac{(1 - p)}{p} P(y = 1)$$

then we obtain the following relationship, similar to Eq. 11 of Keating and Cherry (2004), but without their large-sample assumption:

$$P_{\mathrm{UA}}(s = 1|\mathbf{x}) = \frac{1}{1 + r/P(y = 1|\mathbf{x})}. \quad (2)$$

This relationship is proved as follows:

$$\begin{aligned}
&P_{\mathrm{UA}}(s = 1|\mathbf{x}) \\
&\quad = P_{\mathrm{UA}}(\mathbf{x}|s = 1)P_{\mathrm{UA}}(s = 1)/P_{\mathrm{UA}}(\mathbf{x}) && \text{[Bayes' rule]} \\
&\quad = \frac{[P_{\mathrm{UA}}(\mathbf{x}|s = 1)P_{\mathrm{UA}}(s = 1)]}{[P_{\mathrm{UA}}(\mathbf{x}|s = 1)P_{\mathrm{UA}}(s = 1) + P_{\mathrm{UA}}(\mathbf{x}|s = 0)P_{\mathrm{UA}}(s = 0)]} && \text{[since } s = 0 \text{ or } 1\text{]} \\
&\quad = \frac{pP_{\mathrm{UA}}(\mathbf{x}|s = 1)}{pP_{\mathrm{UA}}(\mathbf{x}|s = 1) + (1 - p)P_{\mathrm{UA}}(\mathbf{x}|s = 0)} && \text{[definition of } p\text{]} \\
&\quad = 1/(1 + a) && \text{[dividing through by } pP_{\mathrm{UA}}(\mathbf{x}|s = 1)\text{]}
\end{aligned}$$

where $a$ satisfies

$$\begin{aligned}
a &= \frac{(1 - p)}{p} \cdot \frac{P_{\mathrm{UA}}(\mathbf{x}|s = 0)}{P_{\mathrm{UA}}(\mathbf{x}|s = 1)} \\
&= \frac{(1 - p)}{p} \cdot \frac{P(\mathbf{x})}{P(\mathbf{x}|y = 1)} && \text{[by Eq. 1]} \\
&= \frac{(1 - p)}{p} \cdot \frac{P(y = 1)}{P(y = 1|\mathbf{x})} && \text{[Bayes' rule]} \\
&= \frac{r}{P(y = 1|\mathbf{x})}.
\end{aligned}$$

This has strong implications for interpretation of any model fitted to presence–background data using a presence–absence method, as the quantity being approximated is not equal to, or even proportional to, probability of presence. Despite these problems, this sampling model and the resulting estimate of $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ have been extensively used in SDM (Ferrier et al. 2002, Zaniewski et al. 2002, Elith et al. 2006). Using an estimate of $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ for species modeling is reasonable as long as care is taken in the interpretation of model values. While $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ is not proportional to probability of presence, it is a monotone increasing function of probability of presence, i.e., it correctly ranks probability of presence. In particular, this means that any binary prediction made by thresholding $P(y = 1 \mid \mathbf{x})$ (i.e., predicting presence only for sites with $P(y = 1 \mid \mathbf{x})$ above some threshold) can be obtained by thresholding $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$, and vice versa, although the required thresholds will differ. When measuring model performance, measures that depend only on ranking of test data (such as the area under the receiver operating characteristic curve) might therefore be insensitive to the distinction between modeling $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ or $P(y = 1 \mid \mathbf{x})$, although the two approaches will likely yield different models.

In habitat-selection studies using resource selection functions, the emphasis is on deriving $P(y = 1 \mid \mathbf{x})$ from $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ by inverting Eq. 2. If $P(y = 1 \mid \mathbf{x})$ is assumed to be an exponential function, then $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ is logistic. A logistic model fitted to $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ can thus be used to infer parameters of an exponential model for $P(y = 1 \mid \mathbf{x})$ (Boyce et al. 2002, Manly et al. 2002). However, this approach is controversial in the habitat-selection literature (Keating and Cherry 2004). An alternative way of estimating $P(y = 1 \mid \mathbf{x})$ from presence-only data involves using the expectation–maximization (EM) algorithm to iteratively infer probability of occurrence for the background sites (estimation) and feed the results back into maximum likelihood parameter estimation (maximization; Ward et al., *in press*). Whilst this approach has strong theoretical justification, it requires knowledge of $P(y = 1)$, and the implementation is not yet widely available, so we do not use it here. In summary, modeling $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x})$ is the best currently available way to apply presence–absence models to presence-only data, and is therefore the approach we take here.

### *Presence–absence models with biased background*

We have argued that sample selection bias is widespread in species occurrence data. We would therefore like to be able to correct for this bias. As in the unbiased case we cannot estimate $P(y = 1 \mid \mathbf{x})$ without further knowledge of the prevalence $P(y = 1)$. Instead, we prove under a mild assumption that if the background data have the same bias as the occurrence data, the resulting model is monotonically related to $P(y = 1 \mid \mathbf{x})$, as in the unbiased case. We therefore assume

that both background and presence samples are selected nonuniformly using the same sample selection distribution. A practical example could be that presence records are collected by driving along roads while stopping at random sites and walking up to 100 m from the road to record sightings of the species. This sample selection is biased toward roadsides, which in turn are likely to be biased away from gullies or particular rough terrain. To generate background data with the same bias, we randomly select sites within a distance of 100 m from any road (note that these might coincide with presence points). For this example, the sample selection distribution is uniform over sites whose distance from the road is at most 100 m and zero elsewhere.

We introduce an additional auxiliary variable $b$ to represent potentially biased selection of samples: samples are now drawn from a distribution over triples ($t$, $y$, $b$), and only samples with $b = 1$ are used for model training. Analogously to the unbiased case, a presence–absence model fitted to a biased use–availability sample gives an estimate of $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x}, b = 1)$. The derivation of Eq. 2 is still valid if we condition all probabilities on $b = 1$, so Eq. 2 generalizes to

$$P_{\mathrm{UA}}(s = 1|\mathbf{x}, b = 1) = \frac{1}{1 + r'/P(y = 1|\mathbf{x}, b = 1)} \qquad (3)$$

where

$$r' = \frac{(1 - p)}{p} P(y = 1|b = 1)$$

which is a constant independent of $\mathbf{x}$.

In many cases we can make the assumption that $P(y = 1 \mid \mathbf{x}, b = 1) = P(y = 1 \mid \mathbf{x})$, i.e., that sampling effort and presence of the species are conditionally independent given $\mathbf{x}$. Under this assumption, the right-hand side of Eq. 3 simplifies to $1/[1 + r'P(y = 1 \mid \mathbf{x})]$. Thus, the function we are fitting, $P_{\mathrm{UA}}(s = 1 \mid \mathbf{x}, b = 1)$, is monotonically related to what we are truly interested in, $P(y = 1 \mid \mathbf{x})$. A simple case for which the conditional independence assumption is true is when all variables that affect presence of the species are included among the covariates. Similarly, we obtain conditional independence if all variables that affect sample selection are included among the covariates (Zadrozny 2004). In general, though, conditional independence may not hold. For example, a pioneer plant species that is correlated with disturbance may be more common than climatic conditions would suggest near roads and towns, exactly where sample selection bias is higher. Unless disturbance level is used as a predictor variable, the conditional independence assumption would be incorrect.

### *Maxent models with unbiased samples*

Maxent is a general technique for estimating a probability distribution from incomplete information (Jaynes 1957). It has been applied to species distribution modeling by assuming that the presence data have been drawn from some probability distribution $\pi$ over the study region, and using the presence records for a species to determine a set of constraints that are likely to be satisfied by $\pi$ (Phillips et al. 2006, Dudík et al. 2007). Maxent then produces as output the distribution of maximum entropy among all distributions satisfying those constraints; note that the distribution is over sites in the study region, not over environmental conditions. The constraints require that the expected value of each environmental variable (or some functions thereof, referred to as features) under this estimated distribution closely match its empirical average. Maximizing entropy is desirable, as doing otherwise would be equivalent to imposing additional (unfounded) constraints on the output distribution. Maximizing entropy also has the useful property that it results in a distribution with a simple mathematical description: under the Maxent distribution, the probability of a site is an exponential function of the features.

The Maxent distribution can be related to conditional probability of presence as follows. The probability $\pi(t)$ is the probability of the site $t$ conditioned on the species being present, i.e., the conditional probability $P(t \mid y = 1)$. We define

$$f(\mathbf{x}) = \frac{P(\mathbf{x}|y = 1)}{NP(\mathbf{x})}$$

i.e., $f(\mathbf{x})$ is the average of $\pi(t)$ over sites with $\mathbf{x}(t) = \mathbf{x}$. This gives

$$P(y = 1|\mathbf{x}) = \frac{P(y = 1)}{P(\mathbf{x})} P(\mathbf{x}|y = 1) \qquad \text{[Bayes' rule]}$$

$$= Nf(\mathbf{x})P(y = 1) \qquad \text{[definition of } f\text{]}.$$

The function $f(\mathbf{x})$ is therefore proportional to probability of presence, and the exponential function describing the Maxent distribution is an estimate of $f(\mathbf{x})$. Note, however, that with presence-only data we typically do not know the constant of proportionality $P(y = 1)$, i.e., the prevalence of the species, since $P(y = 1)$ is not estimable from presence-only data alone (Ward et al., *in press*).

### *Maxent models for biased samples*

Maxent has been available now for five years as a stand-alone program that enables the spatial modeling of presence-only data. Because such data are often biased, the authors have worked on methods for dealing with sample bias, one of which, called FactorBiasOut, we briefly describe here (for technical details, see Dudík et al. [2005]). To describe the impact of sample selection bias on density estimation, we introduce the notation $p_1p_2$ for the site-wise product of two probability distributions normalized over the study region, i.e., $p_1p_2(t) = p_1(t)p_2(t)/\Sigma_{t'} p_1(t')p_2(t')$. As opposed to the case of unbiased estimation, we now assume that the presence sites for a species are biased by a sample selection distribution $\sigma$, in other words, the presence

sites are recorded by observers who pick locations randomly according to $\sigma$, rather than uniformly at random (in the notation of *Presence–absence models with biased background*, $\sigma(t) = P(t \mid b = 1)$). The presence sites are therefore samples from the distribution $\sigma\pi$ rather than from the true species distribution $\pi$.

The FactorBiasOut method estimates $\sigma\pi$, then factors out the bias $\sigma$. It does this by outputting the distribution that minimizes the relative entropy $\mathrm{RE}(\sigma q \,||\, \sigma)$ among all choices of the probability distribution $q$, subject to the constraints mentioned in *Maxent models with unbiased samples*, with the constraints now applying to $\sigma q$, since that is the distribution from which we have samples. Relative entropy, also known as Kullback-Liebler (KL) divergence, measures how different two probability distributions are. It makes sense to seek to minimize the difference from $\sigma$, since a null model would have the species distribution being uniform, so the presence data would simply be drawn from $\sigma$.

In the special case that there is no sample selection bias, i.e., $\sigma$ is the uniform distribution, FactorBiasOut is just the standard Maxent, since minimizing entropy relative to the uniform distribution is the same as maximizing entropy. Under reasonable conditions, the output of FactorBiasOut converges, with increasing sample size, to the distribution $q_\sigma$ that minimizes $\mathrm{RE}(\sigma\pi \,||\, \sigma q)$ among the class of Gibbs (i.e., exponential) distributions. This generalizes the result for the unbiased case, that the output of Maxent converges to the Gibbs distribution that minimizes $\mathrm{RE}(\pi \,||\, q)$ (Dudík et al. 2007). In other words, the output of FactorBiasOut converges to a distribution that is close, in a strict sense and as in the unbiased case, to the true distribution $\pi$, so bias has been removed from the prediction.

As described so far, the FactorBiasOut method requires knowledge of the sampling distribution $\sigma$. However, it is enough to have a set $S$ of independent samples from $\sigma$. We can use $S$ as background data for fitting a Maxent distribution and then apply the resulting model to obtain a distribution over the entire study area. For large $|S|$, the resulting distribution converges to the same distribution $q_\sigma$. To summarize, we have shown that, as with the regression models, using background data with the same sample selection bias as the occurrence data yields a Maxent model with theoretical properties that are analogous to the unbiased case.

## Experimental Methods

### Data sources

We used data for 226 species from six regions of the world: the Australian Wet Tropics (AWT), Ontario, Canada (CAN), northeast New South Wales, Australia (NSW), New Zealand (NZ), South America (SA), and Switzerland (SWI). The species represent a range of geographic distributions, habitat specialization, and biological groups/life forms. Similarly, there is a wide range in the amount of training data per species (2–5822 occurrence records, median 57). In the independent evaluation data, the presence or absence of each species is described at between 102 and 19 120 sites. There are 11–13 environmental data layers per region, and the layers are typical of what is used for SDM. Environmental data varied in functional relevance to the species and spatial resolution. Data for three regions (NSW, NZ, SWI) had more direct links to species' ecology at the local scale than the climate-dominated variables from AWT, CAN, and SA (Elith et al. 2006, Guisan et al. 2007). Layers from AWT, NSW, NZ, and SWI had grid cell sizes of around 100 m and those from CAN and SA were 1 km. More details on the species and environmental data layers can be found in Elith et al. (2006).

### Background treatments

Two sets of background data were used. First, we used 10 000 sites selected uniformly at random from each region (as in Elith et al. [2006], and referred to as random background). Second, and uniquely for this study, for each of the 226 species we generated a set of background data consisting of the presence localities for all species in the same target group (referred to as target-group background). The target groups were birds or herpetofauna for AWT; birds for CAN, plants, birds, mammals or reptiles for NSW; and plants for NZ, SA, and SWI (Table 1).

### Evaluation statistics

The modeled distributions were evaluated for predictive performance using the independent presence–absence sites described above. We used the area under the receiver operating-characteristic curve (AUC) to assess the agreement between the presence–absence sites and the model predictions (Fielding and Bell 1997). The AUC is the probability that the model correctly ranks a random presence site vs. a random absence site, i.e., the probability that it scores the presence site higher than the absence site. It is thus dependent only on the ranking of test data by the model. It provides an indication of the usefulness of a model for prioritizing areas in terms of their relative importance as habitat for a particular species. AUC ranges from 0 to 1, where a score of 1 indicates perfect discrimination, a score of 0.5 implies random predictive discrimination, and values less than 0.5 indicate performance worse than random.

When we are working with presence-only data, we can define the AUC of a model on a set of presence sites relative to random background as the probability that the model scores a random presence site higher than a random site from the study area. The resulting AUC measures the model's ability to distinguish test sites from random, but the value of the AUC is harder to interpret than in the presence–absence case. While a score of 0.5 still indicates discrimination that is no better than random, the maximum value attainable is typically less than 1 (Wiley et al. 2003, Phillips et al. 2006).

TABLE 1. Target groups and measures of training and testing bias.

| Target group | Region | No. species | $AUC_{TG}$ | $AUC_{eval}$ |
|---|---|---|---|---|
| AWT-bird | Australian wet tropics | 20 | 0.8337 | 0.7887 |
| AWT-plant | Australian wet tropics | 20 | 0.841 | 0.5649 |
| CAN | Ontario, Canada | 20 | 0.9473 | 0.9216 |
| NSW-bird | New South Wales | 10 | 0.8789 | 0.877 |
| NSW-mammal | New South Wales | 7 | 0.9341 | 0.8402 |
| NSW-plant | New South Wales | 29 | 0.7054 | 0.6303 |
| NSW-reptile | New South Wales | 8 | 0.9219 | 0.8539 |
| NZ | New Zealand | 52 | 0.7443 | 0.7619 |
| SA | South America | 30 | 0.7502 | 0.7667 |
| SWI | Switzerland | 30 | 0.8564 | 0.8256 |

*Notes:* For each target group, $AUC_{TG}$ is the area under the receiver operating characteristic curve (AUC) of training presence sites vs. random background for a Maxent model trained on all presence sites for the target group. $AUC_{eval}$ is the AUC of the same model evaluated using the set of test sites for that target group vs. random background. A high value of $AUC_{TG}$ indicates that the training sites are highly biased and that sample-selection bias can be predicted well as a function of environmental conditions. A high value of $AUC_{eval}$ indicates that the test sites and training sites have similar strong biases.

The correlation, COR, between a prediction and 0–1 observations in the presence–absence test data set is known as the point biserial correlation, and can be calculated as a Pearson correlation coefficient (Zheng and Agresti 2000). It differs from AUC in that, rather than depending only on rank, it measures the degree to which prediction varies linearly with the observation. Because it depends on the prediction values rather than simply on their order, it is likely to be sensitive to the effect of varying relative sampling intensity in the training data (Eq. 2).

To assess whether there is a monotone relationship between two variables, we use Spearman's rank correlation coefficient ($\rho$), which is a nonparametric measure of correlation. We use $\rho$ rather than Pearson's product-moment correlation ($r$) to avoid two assumptions required by the latter: that the relationship between the two variables is linear, and that the data are drawn from normal distributions.

### *Measuring bias*

In order to measure the effect of bias on predictions, it is useful to be able to measure the amount of bias in a set of presence-only samples. Specifically, we would like to measure the amount of bias for each target group. We do this by estimating how well we can discriminate target-group sites from the background, by using Maxent to make a model of target-group sites and using the AUC of the target-group sites vs. background as a measure of discrimination. We refer to this value as $AUC_{TG}$. If $AUC_{TG}$ is high, it means that the environmental variables can be used to distinguish the spatial distribution of target-group presences from random background, and therefore target-group presences sample environmental space in very different proportions from the proportions present in the study area, i.e., the target-group presences are biased both in environmental and geographic space. We therefore use $AUC_{TG}$ as an estimate of sample selection bias for the target group, but with the following two reservations. First, spatial bias will only be picked up by $AUC_{TG}$ if it results in bias in environmental space, i.e., if some environmental conditions are more strongly represented in the target-group presence data than we would expect based on the proportion of sites with those conditions. Any spatial bias that is independent of the environmental variables will not be picked up by $AUC_{TG}$. However, such spatial bias is less problematic than the bias measured by $AUC_{TG}$, since a species distribution model cannot use it to distinguish presences from background. Second, the target group may truly occupy only part of the environmental space represented in the study area, in which case $AUC_{TG}$ may be higher than 0.5 even if there is no sample selection bias, i.e., even if the presence records were gathered with uniform survey effort across the study area. For these reasons, $AUC_{TG}$ should be interpreted carefully only as an estimate of bias. Note also that the use of Maxent models here is not essential; any of the methods used in this paper would have sufficed.

Once we have an estimate of bias in the training data, it is possible to measure how well this bias estimate predicts sampling effort in the evaluation data. A simple systematic design for evaluation data would uniformly sample the study region, and therefore have no bias. However, bias may arise, for example if the evaluation data derive from a survey of only part of the region, such as all uncleared, forested areas. If the sample selection and evaluation biases are similar, we might expect it would help us in constructing better-performing models. We measure the similarity of the biases using the value $AUC_{eval}$, defined as the AUC of the Maxent model of training group sites, with the AUC evaluated using test sites (both presences and absences) vs. random background. A high value of $AUC_{eval}$ indicates that environmental conditions at the test sites are very similar to those at the training sites, and different from most of the study region. The amount of bias varied considerably between regions and target groups (Table 1), with the strongest bias and the highest value of $AUC_{eval}$ occurring in Canada (Fig. 3). AWT-plant training data

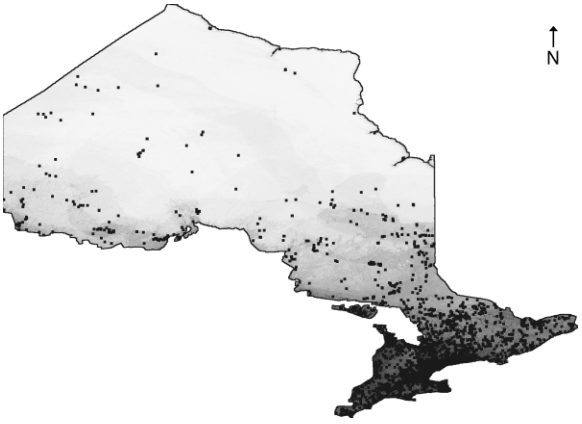


FIG. 3. Bias in the Canada training data used in Elith et al. (2006). Training sites for all species combined are shown as black dots and exhibit a strong bias toward the south of the region. Test sites exhibit a very similar pattern of bias (not shown). The region is shaded to indicate strength of prediction of a maximum entropy (Maxent) model trained on these training sites, with dark shades indicating stronger prediction. Note that the bias is stronger than the bias shown for the artificial species in Fig. 2.

were least effective at predicting test sites ($AUC_{eval} = 0.5649$).

## RESULTS

The average AUC and COR values improved for all methods when using target-group background (Table 2). The improvement in each statistic was highly significant for all methods ($P < 0.001$, two-tailed Wilcoxon signed rank test, paired by species). According to an analysis of variance, the three factors affecting AUC and COR (species, background, and algorithm) are all highly significant ($P < 1 \times 10^{-14}$, $F$ test), with the strongest effect being for species. The effect of background is slightly greater than that of algorithm for both AUC and COR (Table 3). With target-group background, the best methods achieved average AUC values above 0.7 in all regions (Fig. 4). The improvement in AUC scores depended strongly on the estimated amount of bias in training data for the target group (Fig. 5) and with the degree to which the distribution of training data can be used to predict test sites (Fig. 6). For all four methods, there was a strong monotone dependence of improvement in AUC on both estimates of bias as measured by Spearman's rank correlation coefficient (Table 4), with a high level of statistical significance in all cases.

Using target-group background has a visually marked effect on some predictions. The greatest improvement in AUC was for a Canadian species, the Golden-crowned Kinglet, a generalist species that is widely distributed across Ontario and that favors old conifer stands. For this species, the AUC rose from 0.3379 to 0.8412 for Maxent and from 0.2920 to 0.8648 for BRT; the predictions with and without target-group background are very different (Fig. 7). The model with target-group background is much more widespread, excluding mostly the southernmost tip of Ontario, which is the only part of the province that is predominantly deciduous. The map produced with target-group background is much closer visually to maps of breeding evidence and relative abundance for this species (Cadman et al. 2008), differing mainly by strongly predicting the far northeast of the province, where there is little current evidence of breeding.

## DISCUSSION

For all the algorithms we consider here, using target-group background gave a substantial improvement in model performance, measured by both AUC and COR (Table 2). To evaluate the extent of the improvement, we would like to know how it compares with the differences between modeling methods. Elith et al. (2006) found that presence-only modeling methods fell into three distinct groups. The lower group consisted largely of methods that do not use background data, such as BIOCLIM (Busby 1991). The middle group contained traditional regression-based methods such as GAM and MARS among others, while the top group included Maxent and BRT. The improvement due to target-group background (Table 2) is similar to the difference between groups in Elith et al. (2006). In fact, an analysis of variance shows the effect of background type as being larger than the effect of modeling method (Table 3). We conclude that appropriate choice of background data affects model performance for the four methods presented here as much as the choice of modeling method. Since all tested methods benefit from appropriate background, we recommend both well-informed selection of method and careful choice of background samples.

The improvement varied considerably between target groups, with the largest gains seen for target groups with the most biased training data (Fig. 5). This addresses an anomaly from Elith et al. (2006), where BIOCLIM was one of the worst methods in all regions except Canada, where it was one of the best. With target-group

TABLE 2. Area under the receiver operating characteristic curve (AUC) and correlation between predictions and 0–1 test data (COR) for the methods considered; values shown are averages over all 226 species.

| | Random background | | Target-group background | |
|---|---|---|---|---|
| Model | AUC | COR | AUC | COR |
| BRT | 0.7275 | 0.2130 | 0.7544 | 0.2435 |
| Maxent | 0.7276 | 0.2100 | 0.7569 | 0.2446 |
| MARS | 0.6964 | 0.1787 | 0.7260 | 0.2145 |
| GAM | 0.6993 | 0.1765 | 0.7368 | 0.2196 |

*Notes:* For random-background models, background data were chosen uniformly at random from the study area. For target-group background, background data are the sites with presence records for any species from the same target group. Models are boosted regression trees (BRT), maximum entropy (Maxent), multivariate adaptive regression splines (MARS), and generalized additive models (GAM).

TABLE 3. Coefficients for an analysis of variance for AUC and COR evaluated on independent presence–absence test data for models of 226 species.

| | Algorithm | | | | Background | | Effect SE | | |
|---|---|---|---|---|---|---|---|---|---|
| Measure | BRT | GAM | MARS | Maxent | Random | Target group | Species | Algorithm | Background |
| AUC | 0.0128 | −0.0101 | −0.0169 | 0.0141 | −0.0154 | 0.0154 | 0.0228 | 0.0030 | 0.0021 |
| COR | 0.0157 | −0.0146 | −0.0160 | 0.0149 | −0.0180 | 0.0180 | 0.0241 | 0.0032 | 0.0023 |

*Note:* Factors were species (per-species effects not shown), algorithm used to make the model (BRT, GAM, MARS, or Maxent), and background data used for the model (random or target group).

background, all the methods considered in this paper perform better than BIOCLIM in all regions. This confirms that the previous anomalous results in Canada were due to a strong bias in the occurrence data impacting the performance of any method that used background data. With target-group background, performance of the methods that use background data is now consistent across regions (Fig. 4; compare with Fig. 5 of Elith et al. [2006]).

The effect of target-group background varies species by species, and one might expect that it would be systematically affected by characteristics of a species distribution, in particular the species' prevalence in the study area. We investigated this question, measuring the prevalence of a species as the fraction of test sites in which the species is present. However, we found no clear patterns. For BRT, the improvement in AUC is slightly larger for generalist species (those with high prevalence), while the improvement in COR is slightly larger for specialists (with low prevalence). In contrast, for Maxent, the improvement in AUC was unaffected by prevalence, while COR values improved more for generalists. Details are omitted, since the results were inconclusive.

Note that target-group background substantially improved predictions in Switzerland (Fig. 5), and the improvement is statistically significant for all methods ($P < 0.001$, two-tailed Wilcoxon signed rank test, paired by species). This is initially surprising, since the presence-only training data set is extensive and of high quality. However, the sites only sample a subset of the country (forested areas) and therefore they do not represent areas that could support forest but are not currently forested. This means that use of random pseudo-absences misled the models to some extent. The only region where target-group background reduced

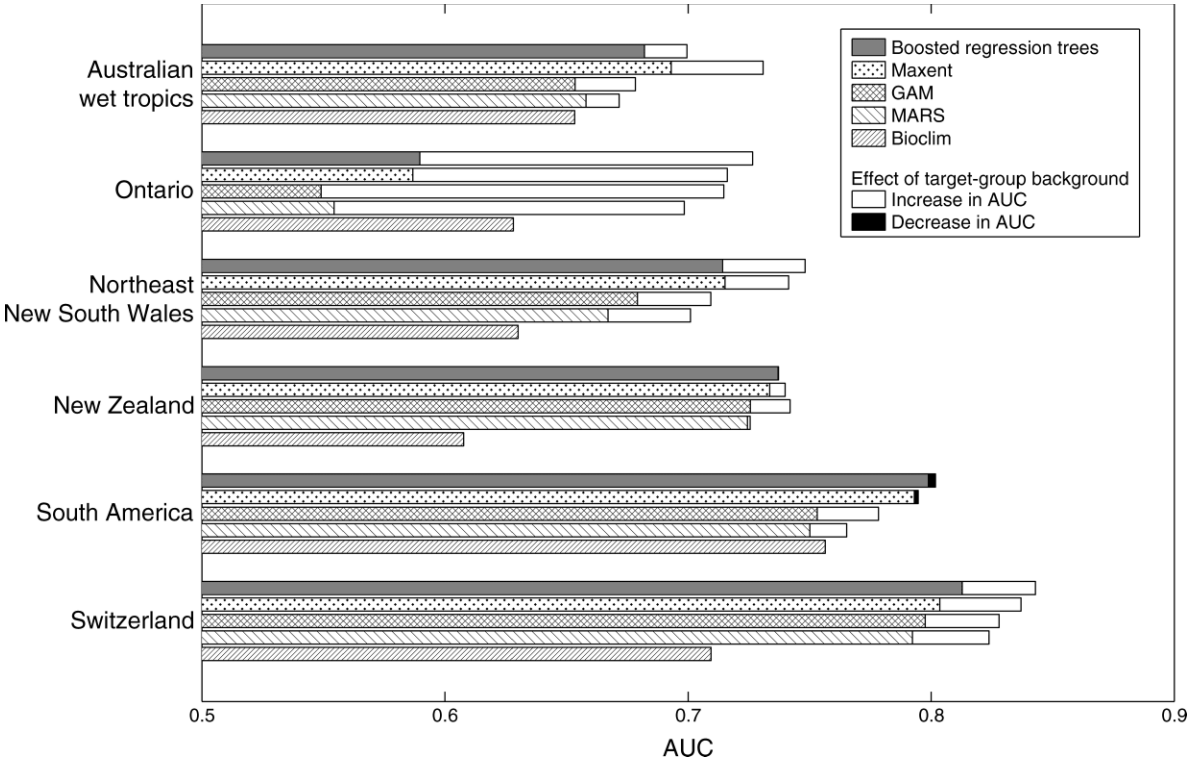


FIG. 4. Performance using target-group background of methods in each of the modeled regions, measured using area under the receiver operating characteristic curve (AUC) on independent presence–absence test data.

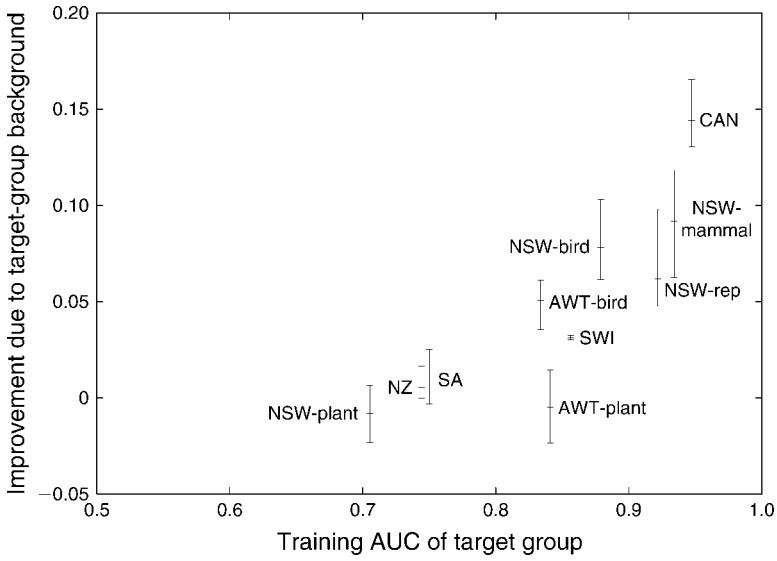


FIG. 5. Plot of improvement in AUC on independent presence–absence test data when using target-group background instead of random background. Models were created using four methods (boosted regression trees [BRT], maximum entropy [Maxent], multivariate adaptive regression splines [MARS], and generalized additive models [GAM]), and minimum, mean, and maximum improvement in AUC across methods are shown for each target group (endpoints of bars are minimum and maximum values). The *x*-axis is a measure of the amount of bias in training data for the target group. It is obtained by training a Maxent model using all presence sites for the target group, and measuring the AUC of the training sites relative to random background. The abbreviations are: AWT, Australian Wet Tropics; CAN, Canada; NSW, New South Wales; NZ, New Zealand; SA, South America; SWI, Switzerland.

average performance was South America, for BRT and Maxent, but the decrease is small and not statistically significant ($P > 0.65$ for BRT, $P > 0.84$ for Maxent, two-tailed Wilcoxon signed rank test, paired by species).

When using random background, all the modeling methods we consider will make predictions that are biased toward areas that have been more intensively sampled. In comparison, target-group background removes some of this bias, spreading predictions into unsampled areas with similar environmental conditions to sampled areas where the species is present. The test sites for most of our target groups exhibit similar spatial

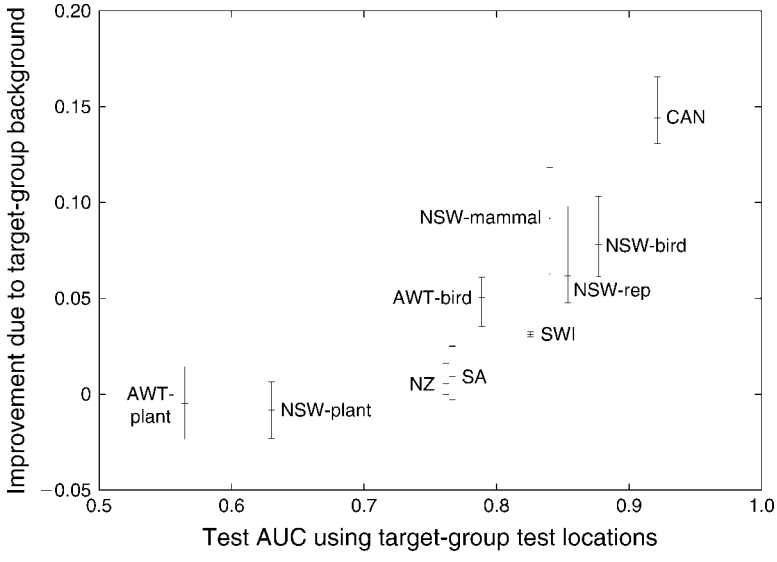


FIG. 6. Scatter plot of improvement in AUC on independent presence–absence test data when using target-group background instead of random background. The *x*-axis is a measure of how well target-group background predicts the distribution of test sites, namely, the AUC of a Maxent model trained on all presence sites for the target group and tested using all test sites for that group versus random background sites. Models were created using four methods (GAM, MARS, BRT, Maxent), and minimum, mean, and maximum improvement in AUC across methods are shown for each target group.

distributions to the training sites, and therefore target-group background will cause prediction strength (i.e., model output values) to decrease at test sites relative to less-sampled areas, compared with random background. Thus, it is crucial that our test data are presence–absence data, so that we are measuring discrimination at test sites, rather than comparing them to random background. If the test data were presence-only, environmental bias in conditions at test sites would strongly influence test results. For example, the Maxent models trained with target-group background have much lower AUC (0.7168) than models trained with random background (0.8201) if the AUC in both cases is measured using presences at test sites relative to random background, rather than relative to absences at test sites. The use of presence-only evaluation data may explain why Lütolf et al. (2006) found that an approach similar to target-group background decreased GLM model performance.

One concern with using target-group background is that we are focusing only on parts of geographic (and thus environmental) space that contain presence samples. Predictions to unsampled areas could therefore be less reliable. This effect is not evident in our statistical results: the average AUC for the groups NSW-plant and AWT-plant, whose test sites are not well predicted by the distribution of training sites, barely changes when using target-group background (Fig. 6). Nevertheless, predictions into unsampled areas, especially those with conditions outside the range observed in sampled areas, should be treated with strong caution. We also note that a critical assumption of the target-group approach is that the data for all species in the group were collected using the same methods, so that the target-group occurrences represent an estimate of sampling effort that is applicable for each member of the group. The set of species in the target group should be chosen with this in mind.

TABLE 4. Spearman rank correlations of improvement in AUC when using target-group background instead of random background.

| Model | Correlation with training bias | | Correlation with test bias | |
|---|---|---|---|---|
| | Spearman's ρ | *P* | Spearman's ρ | *P* |
| Maxent | 0.87 | 0.002 | 0.81 | 0.008 |
| GAM | 0.90 | <0.001 | 0.93 | <0.001 |
| BRT | 0.75 | 0.017 | 0.87 | 0.002 |
| MARS | 0.84 | 0.004 | 0.95 | <0.001 |

*Notes:* The improvement is correlated against the degree of bias in the training data for each target group ("training bias") and a measure of how well the training data for each target group predict the test sites ("test bias"). In each case, we give Spearman's rank correlation coefficient (ρ) and the two-sided *P* value for the null hypotheses that $\rho = 0$.

The evaluation data we have used here measure model performance according to the ability to predict the realized distribution of a species, as represented by presence–absence data at test sites. We note that many applications of species distribution models depend on predicting potential distributions, rather than realized distributions (Peterson et al. 1999). A species may have failed to disperse due to geographic barriers, or be excluded from an area due to competition. In the current evaluation, prediction into such areas would be penalized; however we note that it is usually not possible, with either occurrence or presence–absence data, to test ability to predict potential distribution. It is possible that some of the species considered here are absent from significant portions of their potential distribution, so our conclusions refer to the ability of models to predict realized distributions. We note also that the present study concerns the ability to derive accurate models in a single geographic area under fixed climatic conditions. Therefore, our conclusions do not necessarily apply to uses of species distribution models involving extrapola-

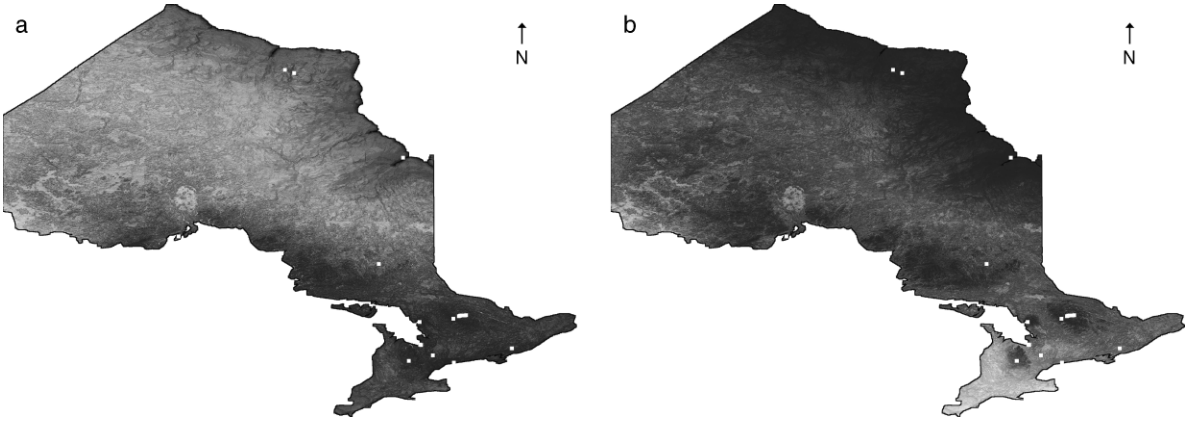


FIG. 7. Maxent predictions in Ontario, Canada for the Golden-crowned Kinglet, a widely distributed generalist species, created (a) without and (b) with use of target-group background. Dark shades indicate stronger prediction, while white or black dots are presence sites used in training. Without target-group background, the prediction is similar to the model of sampling effort (Fig. 3). Target-group background results in stronger prediction in less-sampled areas, reducing dependency of sampling effort.

tion, i.e., producing a model using one set of environmental variables and then applying it to another set with the same names, but describing conditions for a different time or geographic area. Examples of such extrapolations involve future climate conditions (Thomas et al. 2004) or areas at risk for species invasions (Thuiller et al. 2005).

### *Alternate explanations*

We have assumed so far that the improvement in performance due to target-group background is due to properly accounting for sample selection bias in the training data. Here we consider other explanations for the performance improvement.

*Factoring in the test site bias.*—When modeling a species distribution, we may be more interested in model performance under some conditions than others, in particular, under conditions that are broadly suitable for the species or target group. For example, if we want a model to predict the specific niche of a montane species within an alpine area, in a broad region that includes a lot of lowland, we should make sure that all different montane conditions are represented in the evaluation data. However, if we were to include a number of lowland sites in proportion to lowland area, our evaluation statistics would not tell us much about the quality of prediction in the alpine area, since a high AUC value can be obtained by simply ranking montane areas higher than lowlands. In general, evaluation data should be chosen in a way that is relevant to the required output and use of the models, and so may focus on restricted areas.

In the case that evaluation data are biased toward areas representing only a subset of environmental conditions, we expect better performance if training data have the same bias, so that model development is focused on the environmental conditions that will be examined during model evaluation. This can be done formally, for example by transductive learning where unlabeled test data are used to reweight training data (Huang et al. 2007). It is possible, therefore, that the reason that target-group background improves model performance is that it focuses training on the most important areas of the region, which are also the areas with the most test data.

For presence-only modeling, training sites for a target group will be drawn from broadly suitable areas for the group. The distributions of target-group sites and test sites may therefore be similar, in which case using target-group background brings the spatial distribution of the full complement of training data (presences plus background) closer to that of the test data. To see formally why this is advantageous, consider the case of Maxent. Assume the true species distribution is $\pi$ and the sampling distribution is $\sigma$. When using FactorBiasOut, the output converges to the distribution $q^*_\sigma$, which minimizes $\mathrm{RE}(\sigma\pi \,||\, \sigma q)$ among Gibbs distributions $q$ (see *Maxent models for biased samples*). We can expect that $q^*_\sigma$ is close to $q^*$, the distribution that minimizes $\mathrm{RE}(\pi \,||\, q)$, but it is not always true that $q^*_\sigma = q^*$ (Dudík et al. 2005). To obtain the best test results, we would like the Maxent distribution to approximate $\pi$ with respect to the distribution of test data, i.e., we should find $q^*_{\mathrm{test}}$ that minimizes $\mathrm{RE}(\sigma_{\mathrm{test}}\pi \,||\, \sigma_{\mathrm{test}}q)$ as a function of $q$. If $\sigma = \sigma_{\mathrm{test}}$, this is exactly what FactorBiasOut does, and what target-group background approximates. Otherwise, we must rely on the assumption that $q^*_\sigma$ and $q^*_{\mathrm{test}}$ are similar.

For the presence–absence methods, the reasoning is similar. If test sites are chosen according to the distribution $\sigma_{\mathrm{test}}$, then we are evaluating how well our predictions model probability of occurrence under $\sigma_{\mathrm{test}}$, i.e., $P_{\sigma_{\mathrm{test}}}(y = 1 \,|\, x)$. From *Presence–absence models with biased background*, we know that presence–absence methods applied to presence-only data and background data with the same bias are approximating a monotonic function of $P_\sigma(y = 1 \,|\, x)$. Therefore, the best we can hope for is $\sigma = \sigma_{\mathrm{test}}$; otherwise we must rely on the assumption that $P_\sigma(y = 1 \,|\, x)$ and $P_{\sigma_{\mathrm{test}}}(y = 1 \,|\, x)$ are similar.

Testing on similar conditions to those encountered during training has the potential to increase estimates of model performance, in addition to the improvement given by properly accounting for sample selection bias in the training data. Indeed, this seems to be the case for the regression-based methods (BRT, GAM, and MARS): note the higher correlation of performance with test bias than with training bias in Table 4. In contrast, for Maxent the correlation decreases somewhat, and we conclude that for this data set, properly dealing with training bias is a sufficient explanation of the performance improvement for Maxent given by target-group background.

*Target-group data suggest true absences.*—In some situations, target-group sites without records for a particular species can be interpreted as true absences. For example, in presence-only data collections, including some of those used here, many sites are research stations or other well-known sites that have been visited multiple times and have multiple recorded species constituting an inventory of species present there. Therefore, species that are not recorded at such sites are likely to be absent. If most target-group sites are well inventoried, then absence records can be derived by selecting sites that have a record from the target group but not for the species being modeled.

On the other hand, a lot of herbarium and museum records are there because a collector has noticed a species in an odd place (e.g., it might be considered a range expansion), because the collector has a primary interest in that species, or because the species is rare and all occurrences are recorded. In such cases, the collector will not be recording all species from the target group.

In all experiments, we used all target-group records as background. We call this approach overlapping background, because the background data include presences

of the modeled species (as it belongs to the target group). However, if target-group sites where the modeled species was not observed are true absences, then we expect better results if we treat them as such. To test this hypothesis, we removed the sites where the modeled species was recorded from the target-group background, resulting in what we call nonoverlapping background. This removes the problem of contaminated controls (see *Presence–absence models with random background*) and results in a case-control sampling model. If the selection of survey sites is biased according to a distribution $\sigma$, then it results in a case-control sampling model for $P_\sigma(y = 1 \mid x)$, which may be assumed to be equal to $P(y = 1 \mid x)$ (but see *Presence–absence models with biased background*). A presence–absence model fitted using nonoverlapping background data can then be used to index probability of occurrence; if the species prevalence under $\sigma$ is known, then a case-control adjustment can be made in order to estimate probability of occurrence (Keating and Cherry 2004).

We tried this alternative approach (without a case-control adjustment, as species prevalence cannot be derived from our data set) for the presence–absence methods in our study (Table 5). We observed very little difference in performance between the two background formulations. The biggest difference is a slight improvement in performance for GAM with overlapping background. Thus, for our data set at least, there is no benefit to interpreting missing records from target-group sites as true absences.

TABLE 5. Performance of presence–absence methods using target-group background when presences for the modeled species are included in the background (overlap) or excluded (interspersed).

| | Overlap background | | Interspersed background | |
|---|---|---|---|---|
| Model | AUC | COR | AUC | COR |
| BRT | 0.7544 | 0.2435 | 0.7544 | 0.2442 |
| GAM | 0.7368 | 0.2196 | 0.7315 | 0.2092 |
| MARS | 0.7260 | 0.2145 | 0.7222 | 0.2102 |

### *Related approaches*

A related option is to use target-group background data to directly model survey effort (Zaniewski et al. 2002). The surveyed sites are modeled against a random background sample from the region. The resulting model of survey effort can be used to make a weighted selection of background data, with higher probability sites being selected most often, for use in species distribution modeling. The advantage is that a large amount of biased background data can be produced, even if the target-group background data are limited. The danger is that the extra step of modeling introduces an extra source of error on top of the variability in model output caused by varying survey effort. The present study arose from a comparison of this method (which we term modeled target-group background) against target-group background and random background, using a subset of the species modeled by Elith et al. (2006). The preliminary results (not shown here) suggested that target-group background clearly outperforms modeled target-group background. The size of the improvement of target-group background over random background suggested that a larger study was warranted, resulting in the present paper.

Another approach for explicitly modeling survey effort is to include it as a level in a hierarchical Bayesian framework (e.g., Gelfand et al. 2006). One advantage of this approach is that the model gives explicit estimates of uncertainty in the predictions; in contrast, for the models we have considered here, uncertainty estimates are typically obtained by bootstrapping (generating separate models for random subsets of the training data, in order to derive pointwise variance in predictions). To our knowledge the hierarchical Bayesian approach has only been applied to presence–absence data, rather than the presence-only data that are the focus of this study, so it cannot be directly compared with the target-group background approach.

Given presence records for only one species and no information on collection effort, a simple option is to define areas within the region where it is broadly possible that the species could occur. For example, if modeling a tree species in a landscape with substantial amounts of clearing for agriculture, spatial records of clearing (e.g., from remotely sensed data) could be used to define areas to be excluded from the set available for background data selection. Doing so would counteract a sample selection bias toward environmental conditions that are less suitable for agriculture, as long as the cleared areas correspond temporally with the species presence records. This is a special case of the biased background sampling approach we have described here, where the sampling intensity is zero in cleared areas, and uniform in other areas. An alternative approach to correct for this bias is to include land use as a predictor variable.

Engler et al. (2004) used a single species approach to generate weighted background points for input to GAM. They used an ecological niche factor analysis (ENFA) to create "ENFA-weighted" background points by choosing points that were within the study region but unlikely to have the species (i.e., ENFA value less than 0.3). They compared this approach to random background, and found that it improved performance according to three out of four of their evaluation measures. This approach has the aim of having background data biased in favor of areas where the species is thought to be absent. In principle, this moves the sampling design away from a use–availability design and toward being a case-control design. However, the method of Engler et al. (2004) does not address the issue of bias in the occurrence data, and the extra step of modeling in the generation of background data may

introduce spatial and environmental bias in the controls and makes models difficult to interpret.

## Conclusions

While the problem of sample selection bias has received much attention in other fields (e.g., the Nobel prize-winning econometrics work of Heckman [1979]), it has not been adequately addressed for species distribution modeling. Sample selection bias is a serious problem for species distribution models derived from presence-only data, such as occurrence records in natural history museums and herbaria. It has a much greater impact on such models than it does on models derived from presence–absence data. When the sampling distribution is known, we have shown how sample selection bias can be addressed by using background data with the same bias as the occurrence data; our analysis holds for most of the commonly-used presence-only modeling methods. Sample selection bias has been previously explicitly considered only for some individual modeling methods (Argaéz et al. 2005, Dudík et al. 2005, Schulman et al. 2007).

When the sampling distribution is not known, it can be approximated by combining occurrence records for a target group of species that are all collected or observed using the same methods. We evaluated this approach on a diverse set of 226 species and four modeling methods. For both statistical measures of model performance that we used, target-group background improved predictive performance for all modeling methods, with the amount of improvement being comparable to the difference between the best and the worst of the four modeling methods. We conclude that the choice of background data is as important as the choice of modeling method when modeling species distributions using presence-only data.

## Acknowledgments

This work was initiated by the working group on "Testing Alternative Methodologies for Modeling Species' Ecological Niches and Predicting Geographic Distributions," at the National Center for Ecological Analysis and Synthesis (Santa Barbara, California, USA). We thank all the members of the working group, as well as others who provided data used here: A. Ford, CSIRO Atherton, for AWT data; M. Peck and G. Peck, Royal Ontario Museum, and M. Cadman, Bird Studies Canada, Canadian Wildlife Service of Environment Canada, for CAN data; the National Vegetation Survey Databank and the Allan Herbarium, for NZ data; Missouri Botanical Garden, especially R. Magill and T. Consiglio, for SA data; and T. Wohlgemuth and U. Braendi from WSL Switzerland for SWI data. We thank Richard Pearson for helpful references and comments.

## Literature Cited

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

Argaéz, J. A., J. A. Christen, M. Nakamura, and J. Soberón. 2005. Prediction of potential areas of species distributions based on presence-only data. Environmental and Ecological Statistics 12(1):27–44.

Boyce, M. S., P. R. Vernier, S. E. Nielsen, and F. K. Schmiegelow. 2002. Evaluating resource selection functions. Ecological Modelling 15:281–300.

Busby, J. R. 1991. BIOCLIM: a bioclimate analysis and prediction system. Pages 64–68 *in* M. P. Austin and C. R. Margules, editors. Nature conservation: cost effective biological surveys and data analysis. CSIRO, Melbourne, Australia.

Cadman, M., D. A. Sutherland, G. G. Beck, D. Lepage, and A. R. Couturier. 2008. Atlas of the breeding birds of Ontario, 2001–2005. Ontario Nature, Toronto, Ontario, Canada.

Carpenter, G., A. N. Gillison, and J. Winter. 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants, animals. Biodiversity and Conservation 2: 667–680.

Cawsey, E. M., M. P. Austin, and B. L. Baker. 2002. Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. Biodiversity and Conservation 11:2239–2274.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. Ecology 88:243–251.

Dennis, R., and C. Thomas. 2000. Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. Journal of Insect Conservation 4:73–77.

Dudík, M., S. J. Phillips, and R. E. Schapire. 2005. Correcting sample selection bias in maximum entropy density estimation. Pages 323–330 *in* Advances in neural information processing systems 18. MIT Press, Cambridge, Massachusetts, USA.

Dudík, M., S. J. Phillips, and R. E. Schapire. 2007. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. Journal of Machine Learning Research 8:1217–1260.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2): 129–151.

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

Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41:263–274.

Ferrier, S., G. Watson, J. Pearce, and M. Drielsma. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. 1. Species-level modelling. Biodiversity and Conservation 11:2275–2307.

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

Friedman, J. 1991. Multivariate adaptive regression splines (with discussion). Annals of Statistics 19:1–141.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232.

Gelfand, A. E., J. A. Silander, Jr., S. Wuz, A. Latimer, P. O. Lewis, A. G. Rebelo, and M. Holder. 2006. Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis 1(1):41–92.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19(9):497–503.

Guisan, A., N. Zimmermann, J. Elith, C. Graham, S. Phillips, and A. Peterson. 2007. What matters for predicting spatial distributions of tree occurrences: techniques, data, or species' characteristics? Ecological Monographs 77:615–630.

Hastie, T., and R. Tibshirani. 1990. Generalized additive models. Chapman and Hall, London, UK.

Heckman, J. J. 1979. Sample selection bias as a specification error. Econometrica 47(1):153–161.
Hernandez, P., C. Graham, L. Master, and D. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773–785.
Hirzel, A. H., J. Hausser, D. Chessel, and N. Perrin. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 87:2027–2036.
Huang, J., A. J. Smola, A. Gretton, and K. M. Borgwardt. 2007. Correcting sample selection bias by unlabeled data. Pages 601–608 *in* Advances in neural information processing systems 19. MIT Press, Cambridge, Massachusetts, USA.
Jaynes, E. T. 1957. Information theory and statistical mechanics. Physics Reviews 106:620–630.
Keating, K. A., and S. Cherry. 2004. Use and interpretation of logistic regression in habitat-selection studies. Journal of Wildlife Management 68(4):774–789.
Kozak, K., C. Graham, and J. Wiens. 2008. Integrating GIS-based environmental data into evolutionary biology. Trends in Ecology and Evolution 23:141–148.
Lancaster, T., and G. Imbens. 1996. Case-control studies with contaminated controls. Journal of Econometrics 71:145–160.
Leathwick, J. R., J. Elith, M. P. Francis, T. Hastie, and P. Taylor. 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series 321: 267–281.
Leathwick, J., D. Rowe, J. Richardson, J. Elith, and T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biology 50:2034–2052.
Loiselle, B. A., C. A. Howell, C. H. Graham, J. M. Goerck, T. Brooks, K. G. Smith, and P. H. Williams. 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conservation Biology 17(6):1591–1600.
Lütolf, M., F. Kienast, and A. Guisan. 2006. The ghost of past species occurrence: improving species distribution models for presence-only data. Journal of Applied Ecology 43:802–815.
Manly, B., L. McDonald, D. Thomas, T. McDonald, and W. Erickson. 2002. Resource selection by animals: statistical design and analysis for field studies. Second edition. Kluwer Press, New York, New York, USA.
Peterson, A. T., and D. A. Kluza. 2003. New distributional modelling approaches for gap analysis. Animal Conservation 6:47–54.
Peterson, A. T., J. Soberón, and V. Sánchez-Cordero. 1999. Conservatism of ecological niches in evolutionary time. Science 285:1265–1267.
Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.
Phillips, S., and M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175.
Ponder, W. F., G. A. Carter, P. Flemons, and R. R. Chapman. 2001. Evaluation of museum collection data for use in biodiversity assessment. Conservation Biology 15:648–657.
Reddy, S., and L. M. Dávalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30:1719–1727.
Schulman, L., T. Toivonen, and K. Ruokolainen. 2007. Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. Journal of Biogeography 34(8):1388–1399.
Stockwell, D., and D. Peters. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13:143–158.
Suarez, A. V., and N. D. Tsutsui. 2004. The value of museum collections for research and society. BioScience 54(1):66–74.
Thomas, C. D., et al. 2004. Extinction risk from climate change. Nature 427:145–148.
Thuiller, W., D. M. Richardson, P. Pyšek, and G. F. Midgley. 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Global Change Biology 11:2234–2250.
Ward, G., T. Hastie, S. Barry, J. Elith, and J. Leathwick. *In press*. Presence-only data and the EM algorithm. Biometrics.
Wiley, E. O., K. M. McNyset, A. T. Peterson, C. R. Robins, and A. M. Stewart. 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16(3):120–127.
Yee, T. W., and N. D. Mitchell. 1991. Generalized additive models in plant ecology. Journal of Vegetation Science 2: 587–602.
Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. Page 114 *in* Proceedings of the Twenty-First International Conference on Machine Learning. Association of Machine Learning, New York, New York, USA.
Zaniewski, A. E., A. Lehmann, and J. M. Overton. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157:261–280.
Zheng, B., and A. Agresti. 2000. Summarizing the predictive power of a generalized linear model. Statistics in Medicine 19:1771–1781.