Ecology, 94(6), 2013, pp. 1409–1419

On estimating probability of presence from use–availability or presence–background data

Steven J. Phillips[1,3] and Jane Elith[2]

[1]*AT&T Labs–Research, 180 Park Avenue, Florham Park, New Jersey 07932 USA*
[2]*School of Botany, University of Melbourne, Parkville 3010 Australia*

Abstract. A fundamental ecological modeling task is to estimate the probability that a species is present in (or uses) a site, conditional on environmental variables. For many species, available data consist of "presence" data (locations where the species [or evidence of it] has been observed), together with "background" data, a random sample of available environmental conditions.

Recently published papers disagree on whether probability of presence is identifiable from such presence–background data alone. This paper aims to resolve the disagreement, demonstrating that additional information is required.

We defined seven simulated species representing various simple shapes of response to environmental variables (constant, linear, convex, unimodal, S-shaped) and ran five logistic model-fitting methods using 1000 presence samples and 10 000 background samples; the simulations were repeated 100 times. The experiment revealed a stark contrast between two groups of methods: those based on a strong assumption that species' true probability of presence exactly matches a given parametric form had highly variable predictions and much larger RMS error than methods that take population prevalence (the fraction of sites in which the species is present) as an additional parameter. For six species, the former group grossly under- or overestimated probability of presence. The cause was not model structure or choice of link function, because all methods were logistic with linear and, where necessary, quadratic terms. Rather, the experiment demonstrates that an estimate of prevalence is not just helpful, but is necessary (except in special cases) for identifying probability of presence. We therefore advise against use of methods that rely on the strong assumption, due to Lele and Keim (recently advocated by Royle et al.) and Lancaster and Imbens. The methods are fragile, and their strong assumption is unlikely to be true in practice. We emphasize, however, that we are not arguing against standard statistical methods such as logistic regression, generalized linear models, and so forth, none of which requires the strong assumption.

If probability of presence is required for a given application, there is no panacea for lack of data. Presence–background data must be augmented with an additional datum, e.g., species' prevalence, to reliably estimate absolute (rather than relative) probability of presence.

Key words: availability; background; identifiability; logistic; measuring use vs. non-use; presence–background; prevalence; resource selection; species distribution model.

Introduction

We study a modeling task that is central to two related bodies of ecological literature. Ecologists studying a broad range of species wish to map species' distributions or predict the suitability of sites for occupation or persistence of the species (Franklin 2010), while ecologists investigating resource selection by animals seek to characterize those areas within a region of interest that are "used" by a particular species or individual animals (Manly et al. 2002). In both cases, the data at hand frequently consist of a collection of geographic locations with evidence of presence of (or use by) the species, together with data on environmental covariates in the region of interest, termed background (or available or sometimes pseudo-absence) data (Aarts et al. 2012). In this paper we investigate statistical methods that estimate the probability that the species is present at a site (respectively, the probability that it uses the site) conditional on environmental covariates. Methods for estimating probability of presence and related indices are important; they have been used extensively for a variety of applications in ecology and conservation, and according to Google Scholar, a seminal resource selection text (Manly et al. 2002) has been cited 2170 times while an influential species distribution modeling paper (Elith et al. 2006) has received 1926 citations at the time of writing.

Although there is shared concern over terminology and model interpretation in the two bodies of ecological research—what defines use rather than a transitory or

Manuscript received 6 September 2012; revised 2 January 2013; accepted 4 January 2013. Corresponding Editor: B. D. Inouye.

[3] E-mail: phillips@research.att.com

chance visit, how to define and determine absence or non-use, what delimits the available area from which the species is selecting sites, what ecological interpretation can be given to model outputs (Pulliam 2000, Johnson et al. 2006, Lele and Keim 2006, Jiménez-Valverde et al. 2008, Beyer et al. 2010, Desrochers et al. 2010, Franklin 2010)—we focus here on the underlying statistical questions rather than the difficulties of ecological interpretation. In particular, we study logistic models of probability of presence or use. For brevity, we will use only the "presence" and "background" terminology from here on. We also restrict our attention to the case that presence samples are independent, as opposed to spatially and temporally autocorrelated samples derived from wildlife telemetry (Aarts et al. 2008), and free from spatial bias in sampling effort (Phillips et al. 2009). Furthermore, we assume a discretized sample space in which the study area is partitioned into a set of sites (e.g., derived from a rectangular grid), and each presence sample consists of a site drawn uniformly from the subset of sites that is occupied by the species. While other sampling models have been developed for independent presence data (for example, regarding presence samples as dimensionless points in point-process models; Chakraborty et al. 2011, Aarts et al. 2012), the discrete-space model is most natural for the modeling methods and the statistical question that we study here. We emphasize that probability of presence of a species depends not only on the distribution of the species itself, but also on how a site is defined: a species is more likely to be present in a larger site than a smaller site with the same conditions.

In practice, exponential models are most often used for presence–background data, fitted using logistic regression (Manly et al. 2002) or the maximum entropy method (Phillips et al. 2006), in both cases providing a maximum-likelihood estimate of relative probability of presence (also called a resource selection function, or RSF). The output of these methods is proportional to absolute probability of presence, but the constant of proportionality is not generally known. Exponential models also have the drawback of being unbounded above, so that if their estimates are rescaled in an attempt to estimate absolute (rather than relative) probabilities, estimates greater than 1 may be obtained (Keating and Cherry 2004, Ward et al. 2009). Models of absolute probability of presence would therefore be preferable, especially logistic models, which are naturally bounded within [0, 1]. Maxent estimates are typically post-processed to produce logistic models (Phillips and Dudík 2008), but the result is no longer a maximum-likelihood model, and it will not generally give a good estimate of absolute probability of presence unless additional information is available to inform a parameter (τ) used in the post-processing (Elith et al. 2011). Exponential models are not the only models used: in the species distribution modeling literature, "naive" logistic models are sometimes simply fitted to presence–background data, treating the background as if it were absence (Ferrier et al. 2002, Elith et al. 2006). The resulting models are not proportional to probability of presence, but they are monotonically related (Phillips et al. 2009), which is sufficient for some applications, such as when model outputs are thresholded to yield binary values (although some threshold rules are sensitive to the amount of background data), or when only rankings are of interest. Li et al. (2011) proposed converting a naive model to a model of absolute probability of presence using the assumption that species' probability of presence reaches 1 at some "prototypical" sites; we have critiqued this approach elsewhere (Phillips 2012).

Beyond these, existing maximum-likelihood logistic methods for presence–background data include those of Steinberg and Cardell (1992), Lancaster and Imbens (1996), Lele and Keim (2006), the Expectation-Maximization approach of Ward et al. (2009) and the scaled binomial loss of Phillips and Elith (2011), which we will refer to as the SC, LI, LK, EM, and SB methods, respectively. Lee et al. (2006) proposed an approach that is closely related to the LI method. The LK method recently was strongly advocated by Royle et al. (2012), who proposed maximizing the same likelihood function (Royle et al. 2012: Eq. 4, Lele and Keim 2006: Eq. 1).

Lancaster and Imbens (1996) and subsequent authors have noted that with enough presence–background data, we can determine *relative* probability of presence of the species conditional on environmental conditions. This means that the species' probability of presence can be determined up to a multiplicative constant of proportionality, but identifiability of the constant of proportionality is a concern. (Note that we can discuss the constant of proportionality without assuming any particular model structure, and in particular, we are not talking about the intercept of a logistic model.) If we knew the population prevalence of the species (the fraction of sites in the study area in which the species is present), this would serve as an additional constraint that would determine the constant of proportionality. However, Ward et al. (2009) formally proved that prevalence is not identifiable from presence–background data if we make no assumptions about the structure of the true probability of presence. The idea of the proof is simple: if we imagine two species for which one has exactly half the probability of presence of the other (i.e., exactly the same covariates affect both species in the same way, but one is less common than the other), then presence records for the two species are identically distributed and therefore cannot be used to distinguish between the two species. Probability of presence is identifiable if we make a strong assumption about the structure of the species' probability of presence, but Ward et al. (2009) suggest that such an assumption is unrealistic. Our primary purpose in this paper is to explain this assumption and why it is

unrealistic, and demonstrate that it can result in very unreliable models.

The crux of the identifiability issue is that if two estimates of a species' probability of presence derived from presence–background data are exactly proportional (i.e., one is a constant times the other), then their likelihoods (e.g., as defined by Eq. 1 of Lele and Keim 2006) are equal. Therefore, each fits the data exactly as well as the other, so there is no way to choose between them. The LI and LK methods circumvent this problem by severely restricting the available set of models, so that no two can be proportional. They do this by making the strong assumption that the true conditional probability of presence of the species is *exactly* logit-linear in the predictor variables. This addresses the problem, because no two logit-linear functions are exactly proportional, although with some previously ignored exceptions, e.g., constant functions. Lele and Keim (2006) and Royle et al. (2012) consider other link functions in addition to the commonly used logit link, and while our analysis applies equally in that case, we will for simplicity restrict our discussion to the logit link.

Royle et al. (2012:545) made the strongest claims about the LI/LK approach: their abstract claims "we demonstrate that [probability of presence] is identifiable using conventional likelihood methods under the assumptions of random sampling and constant probability of species detection," without mention of any other assumptions. Indeed, they claim that "lack of identifiability of occurrence probability is not a general feature of presence-only data" (Royle et al. 2012:550) and that the problem of identifiability can be solved simply by choosing a suitable parametric model (e.g., logistic) as long as not all covariates are categorical, citing Lele and Keim (2006). However, these claims are incorrect, as we will demonstrate with a collection of examples. Simply put, the choices made by the modeler cannot overcome an inherent limitation of the available data. Similarly, Dorazio (2012:1304) describes the LK method and a variant of the LI method (Lee et al. 2006) as "approaches for estimating [probability of occurrence] that do not require knowledge of species prevalence," without mention of any restrictive assumptions. In addition, Dorazio (2012:1304) credits the point-process model of Warton and Shepherd (2010) with providing "an estimate of species abundance, and therefore occurrence, at any location" using presence-only data. However, the point-process approach models the density of presence records, not species abundance (W. Fithian and T. Hastie, *unpublished manuscript*), and therefore does not provide an estimate of probability of presence.

The strong parametric assumption of the LI and LK methods makes these methods very different from standard logistic regression. The assumption is not justified by any ecological theory; there is no reason to expect that the logit of the probability of presence of any species should be *exactly* linear in any predictor variable. In contrast, with careful choice of predictors, we can often expect that linearity is a good *approximation* of the truth, as is required for standard logistic regression. Although this distinction may seem slight, it is important: this paper demonstrates the risks of the strong parametric assumption with a collection of simple examples in which even a small deviation from the strong assumption results in very poorly calibrated models.

The risky assumption can be avoided, for example, by taking an estimate of population prevalence as an additional parameter, as is done by the SC, EM, and SB methods. The estimate could derive from limited presence–absence surveys (which may be difficult and expensive to obtain, especially for large sites and/or cryptic species) or from expert opinion. In either case, the estimate may involve substantial uncertainty; nevertheless, in our simple examples, the EM, SC, and SB methods strongly outperform the LI and LK methods even when there is a moderate amount of uncertainty in the prevalence estimate. However, we emphasize that comparing LK/LI to SC, EM, and SB is not an "apples to apples" comparison. We are not placing them on a level playing field, because the latter methods are given an additional piece of data. The purpose of the comparison is to demonstrate that the additional datum is not just helpful (as would be obvious), but required. Without the additional datum, the LK and LI methods produce models that are proportional to probability of presence, but they do not correctly estimate the constant of proportionality except in special cases or by good fortune.

We note that some of the ideas that we further develop here first appeared in published conference proceedings, along with Fig. 4 (Phillips and Elith 2011).

Five Logistic Methods for Presence–Background Data

Here we present an overview of five available maximum-likelihood-based methods for deriving logistic models from presence–background data. These methods are all compared in the experimental comparison that follows. Let L be the landscape of interest, and L_1 and L_0 be the subsets of L in which the species is present or absent, respectively. Let P be a set of presence samples (drawn uniformly from L_1) and B a set of background samples (drawn uniformly from L). We use y to represent the presence ($y = 1$) or absence ($y = 0$) of the species, and s to represent sampling stratum: $s = 1$ for samples in P and $s = 0$ for samples in B. We are concerned with parametric logistic models, i.e., our models take the following form:

$$\Pr(y = 1 \mid \mathbf{x}; \boldsymbol{\beta}) = \frac{1}{1 + \exp[-\eta(x)]}$$

where $\eta(x)$ is a function of the set of predictor variables, $\mathbf{x}$, described by the set of parameters, $\boldsymbol{\beta}$; we will assume for simplicity of this presentation that η is just a linear

function with coefficients $\boldsymbol{\beta}$. Given these definitions, we now describe the existing methods for determining the parameters $\boldsymbol{\beta}$. Three of the methods (EM, SC, and SB) require as an additional input an estimate of the species' prevalence $\Pr(y = 1)$, which we denote by π.

The LK method

The LK method (Lele and Keim 2006, Lele 2009) defines the log likelihood of the presence samples as

$$\sum_{\mathbf{x}\in P} \ln \Pr(\mathbf{x} \mid y = 1; \boldsymbol{\beta}).$$

Applying Bayes' rule and dropping terms that do not depend on $\boldsymbol{\beta}$ yields the objective

$$\sum_{\mathbf{x}\in P} \ln \frac{\Pr(y = 1 \mid \mathbf{x}; \boldsymbol{\beta})}{\Pr(y = 1; \boldsymbol{\beta})} \tag{1}$$

which can be rewritten as:

$$\sum_{\mathbf{x}\in P} \ln \Pr(y = 1 \mid \mathbf{x}; \beta) - |P| \ln \Pr(y = 1; \beta).$$

The background data are used to give an empirical estimate of the second term, resulting in the objective:

$$\sum_{\mathbf{x}\in P} \ln \Pr(y = 1 \mid \mathbf{x}; \boldsymbol{\beta}) - |P| \ln \frac{\sum_{\mathbf{x}\in B} \Pr(y = 1 \mid \mathbf{x}; \boldsymbol{\beta})}{|B|}. \tag{2}$$

Here $|P|$ is the number of presence samples and $|B|$ is the number of background samples, and the species' prevalence (the average probability of presence over the whole landscape) has been approximated by the average probability over the background samples. Standard numerical optimization techniques (Lele and Keim 2006) or more involved methods (Lele 2009) are used to find the coefficients $\boldsymbol{\beta}$ that maximize Eq. 2.

Eq. 1 can be thought of as a relative likelihood; it describes the probability of the presence records (in the numerator) relative to the probability averaged over all sites (i.e., the prevalence, in the denominator). If we multiply all probabilities by a constant, that constant will factor out of the ratio. Therefore, if two models are proportional, they have exactly the same log likelihood. With enough data, the LK method therefore finds a model that is as close as possible to being *proportional* to the species' true probability of presence. In other words, although the LK method aims to estimate *absolute* probability of presence, the likelihood it uses (Eq. 1) measures only *relative* probability of presence. Therefore it may not yield a good approximation of absolute probability of presence, because its predictions can be off by a constant factor. The method has only been proven to estimate absolute probabilities in special cases, e.g., if the true species probability of presence and the fitted model are both exactly logit–linear and not all predictors are categorical (Lele and Keim 2006). However, we note the need for an additional condition in our experiment, namely that the species' response must be nonconstant.

The EM method

Expectation-Maximization (EM) is a general-purpose algorithm for estimating missing data during model-fitting (Dempster et al. 1977). It is naturally applied to presence–background data by regarding the the value of y for background samples as missing data (Ward et al. 2009). EM works in a sequence of iterations, each time improving its estimate of the missing data. To start, the value of y at each background point is initialized to π. A "Maximization" step is then performed, which fits a maximum-likelihood logistic model to the current values of y, with a case-control adjustment to account for unequal sampling rates of presences and absences. An "Expectation" step then applies the model to update the estimates of the missing data, i.e., the values of y at the background points. This process is repeated until convergence. The species' prevalence π is needed both in the initialization step and in the case-control adjustment.

The SC method

The SC method (Steinberg and Cardell 1992) begins by considering, as a thought experiment, the log likelihood of the entire landscape L. This log likelihood is most naturally written as two sums, over sites in L_0 and L_1, respectively. The primary insight of the SC method is that the log likelihood can be rewritten in a way that the sums are either over all of L or over L_1:

$$\begin{aligned}\sum_{\mathbf{x}\in L_0} \ln[1 - \Pr(y = 1 \mid \mathbf{x}, \boldsymbol{\beta})] + \sum_{\mathbf{x}\in L_1} \ln \Pr(y = 1 \mid \mathbf{x}, \boldsymbol{\beta}) \\ = \sum_{\mathbf{x}\in L} \ln[1 - \Pr(y = 1 \mid \mathbf{x}, \boldsymbol{\beta})] \\ + \sum_{\mathbf{x}\in L_1} \Big(\ln \Pr(y = 1 \mid \mathbf{x}, \boldsymbol{\beta}) - \ln[1 - \Pr(y = 1 \mid \mathbf{x}, \boldsymbol{\beta})]\Big).\end{aligned}$$

These two sums then can be estimated empirically using the samples B and P, respectively (similarly to Eq. 2). Combining terms and simplifying yields a pseudo-likelihood that approximates the log likelihood of L

$$\sum_{\mathbf{x}\in B} -\ln(1 + \exp[\eta(\mathbf{x})]) + \pi |B| \eta(\boldsymbol{\mu}) \tag{3}$$

where $\boldsymbol{\mu}$ is the vector of means of predictor variables over P. Eq. 3 can be maximized using standard numerical optimization techniques.

The SB method

The scaled binomial loss (SB) method applies existing logistic model-fitting methods, using a modification to the standard binomial loss function in order to deal with presence–background data (Phillips and Elith 2011). Let f_p be the fraction of presence samples in the training data, i.e., $f_p = |P|/(|P| + |B|)$. Following Lancaster and Imbens (1996), we regard the presence and background data as all having been generated together by the following process: each sample is drawn uniformly from L_1 with probability f_p (a presence

sample) and uniformly from L with the remaining probability $(1 - f_p)$ yielding a background sample. If we use P_{UA} to describe probabilities under this use–availability (UA) sampling model, then $P_{\text{UA}}(s = 1 \mid \boldsymbol{x})$ can be simply expressed in terms of η:

$$P_{\text{UA}}(s = 1 \mid \mathbf{x}) = \frac{1}{1 + r + \exp[-\eta(\mathbf{x}) + \ln r]} \quad (4)$$

where

$$r = \frac{(1 - f_p)}{f_p}\pi. \quad (5)$$

The SB method finds the parameters of η that maximize the likelihood of the presence and background data as described by Eq. 4; this can be done with some existing logistic regression packages by describing Eq. 4 as a user-supplied link function. The name of the method derives from the observation that, although Eq. 4 is not the standard binomial loss function, it can written as a constant scaling factor $(1/(1 + r))$ times a binomial loss function.

The LI method

The LI method (Lancaster and Imbens 1996) is also based on Eq. 4, but regards the species' prevalence π as a parameter to be estimated, rather than a user-supplied value; a similar approach is described by Lee et al. (2006). General-purpose nonlinear optimization software can be used to simultaneously estimate values for π and $\boldsymbol{\beta}$ that maximize the likelihood according to Eq. 4. We note that Lancaster and Imbens (1996) also describe an approach for estimating η if the species' prevalence is known, based on the generalized method of moments. The approach is difficult and, to our knowledge, it has not been applied in ecology due to lack of available software; we therefore do not consider it further here.

Methods

Our experimental evaluation considers the simple case of a single predictor variable x whose values range uniformly from 0 to 1 across the landscape. We study seven simulated species whose probability of presence is defined by the seven functions in Table 1. These functions all range smoothly across the landscape, and because they remain bounded between 0 and 1, they define probabilities. They are chosen to represent a variety of shapes of the response of a species to its environment (constant, linear, convex, unimodal, s-shaped), rather than particular functional forms. It is important to note that these are simulated examples of different species, but that the underlying model structure that we will use is the same in all cases: logistic models with linear and potentially quadratic terms.

Any practical method for modeling $\Pr(y = 1 \mid x)$ needs to be robust enough to produce reasonable approximations across the likely range of species–environment relationships, and our simulations fall within and explore some of that range. Furthermore, species respond to and interact with their environment in complex ways, so the smooth response curves used here are likely to be simpler than for real species. The functions considered here can therefore be thought of as simple cases that should be handled with ease by any practical modeling method.

Table 1. Probability of presence for seven simulated species used in the experimental evaluation.

Simulated species	Probability of presence
Constant	$\Pr(y = 1 \mid x) = 0.3$
Linear	$\Pr(y = 1 \mid x) = 0.05 + 0.2x$
Quadratic	$\Pr(y = 1 \mid x) = 0.5 - 1.333(x - 0.5)^2$
Gaussian	$\Pr(y = 1 \mid x) = 0.75 \exp[-(4x - 2)^2]$
Semi-Logistic	$\Pr(y = 1 \mid x) = 8/(1 + \exp[4 - 2x])$
Logistic-1	$\Pr(y = 1 \mid x) = 1/(1 + \exp[4 - 2x])$
Logistic-2	$\Pr(y = 1 \mid x) = 1/(1 + \exp[4 - 8x])$

Note: The probability of presence of each species is a function of a single environmental predictor, x, whose value ranges uniformly from 0 to 1 across the landscape.

Some readers may note that outside of our simulated landscape (with x in the range [0, 1]), some of these functions fall outside the range [0, 1]. This is not a concern, as the functions can be extended by defining them to be either 0 or 1 for values of x outside of [0, 1]. Readers may also be concerned about the variable x being bounded in our simulations; however, in any finite landscape all variables will necessarily be bounded, and our choice of 0 and 1 for the bounds is arbitrary and does not affect our analysis.

We ran the five modeling methods on randomly drawn data for each species, with 1000 presence samples drawn according to the species' probability of presence and 10 000 background samples chosen uniformly from [0, 1]. The number of presence samples is greater than for most presence–background data sets, and the samples do not suffer from any of the biases that plague presence data in practice (Reddy and Dávalos 2003, Phillips et al. 2009), so any practical modeling method should at least make reasonable models from such a large quantity of high-quality data.

For the "Quadratic" and "Gaussian" species, the true response is unimodal, so quadratic terms were included for all the modeling methods (i.e., we are allowing all methods a reasonable chance of performing well). The simulations were repeated 100 times and the model output was compared both visually against the true probability of presence, and statistically using root mean square (RMS) error of fitted values compared against true probability of presence. We used RMS because it measures how well true probabilities are being estimated. We did not use AUC, as the latter is rank-based and therefore insensitive to predictions all being incorrect by a constant factor.

Some of the methods (SC, EM, and SB) require an estimate of the species' prevalence, for example, derived from limited presence/absence surveys or expert opinion, and there is likely to be some level of error in such

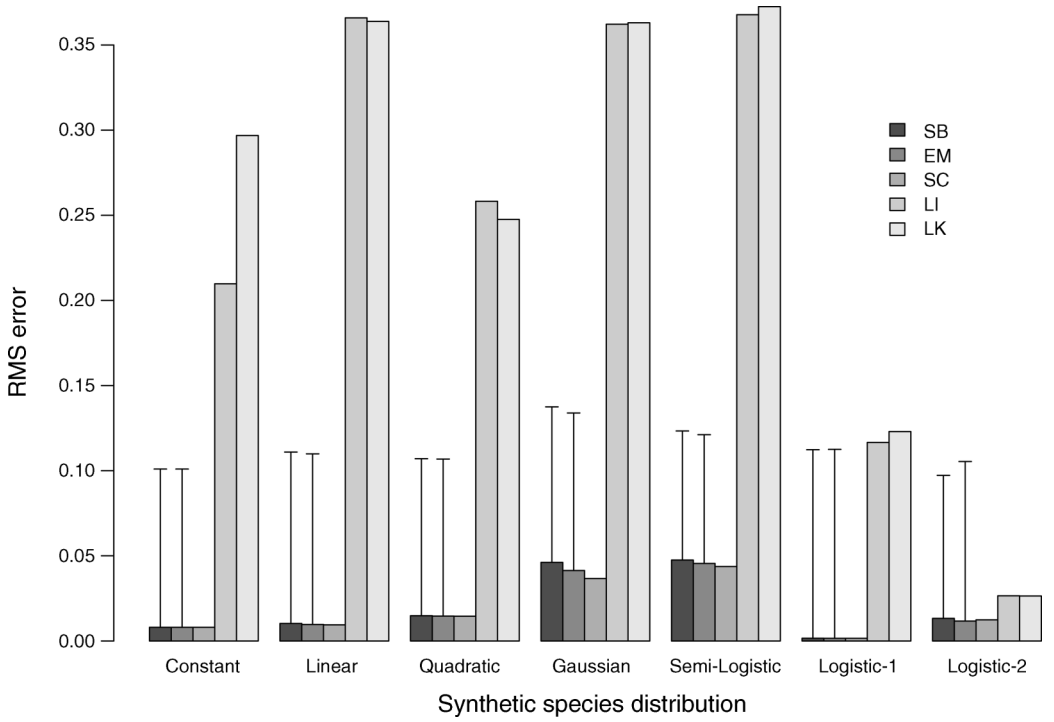

FIG. 1. Root mean square (RMS) error of logistic models fitted to synthetic species distributions, each of whose probability of presence is a function of one variable (details in Table 1). Training data consisted of 1000 presence samples and 10 000 background samples chosen from a landscape with the predictor uniform in [0, 1]. Results shown are the mean of 100 simulations. Model abbreviations are: SB, scaled binomial loss of Phillips and Elith (2011); EM, expectation-maximization approach of Ward et al. (2009); SC, Steinberg and Cardell (1992); LI, Lancaster and Imbens (1996); LK, Lele and Keim (2006). Whiskers on the SB and EM bars indicate the increase in RMS error when there is an additive error of 0.1 in the provided prevalence parameter. SB and EM use glm for model fitting; other methods use the nlm function in R.

an estimate. We therefore ran these methods first using the true prevalence, then with the true value ± 0.1, in order to assess the sensitivity of the methods to errors in the prevalence estimate. Note that this additive error corresponds to a large relative error in prevalence. For example, for a species that is present in 20% of the landscape, our sensitivity analysis uses prevalence estimates of 0.1 and 0.3, respectively, corresponding to relative errors of $\pm 50\%$, i.e., assuming the species is present in 50% more sites than it really occupies, or half as many sites.

We implemented all five model-fitting methods in R. Their implementations are included in the Supplement, along with R code for our experimental simulations.

RESULTS

Overall, the statistical comparison revealed a stark contrast between two groups of methods: the methods that make a strong parametric assumption (LI and LK) and those that take the species' prevalence as a parameter (SC, EM, and SB). The LI and LK methods had RMS errors that were greater than those of the other methods for all simulated species, and approximately 10 times greater for all but one of the simulated species (Fig. 1). Differences in performance within groups were minor. When prevalence was mis-specified by ± 0.1, the RMS error of the EM and SB methods increased accordingly to ~ 0.1, which was much lower than for the LI and LK methods in five of seven cases (whiskers on EM and SB bars in Fig. 1). For SC, mis-specifying the prevalence sometimes caused the pseudo-likelihood to be unbounded above, and therefore maximized by infinite coefficient values and yielding predicted probabilities of exactly 0 or 1. Similar behavior has been noted by Lancaster and Imbens (1996), even for correctly specified prevalence. This behavior would have to be carefully addressed in a practical implementation of the method; here we simply did not plot whiskers for SC in Fig. 1.

A visual inspection of model predictions also indicated a stark difference between the two groups when the true probability of presence is not exactly logit-linear (Fig. 2). LI models (left column; gray lines) and LK models (right column; gray lines) do not appear close to the true probability of presence (black lines). In contrast, a single run of the SB method (black dots) closely approximated the true probability of presence in all cases. Mis-specification of prevalence resulted in limited, but consistent, over- or under-prediction in the SB method (black dashed lines). Results for the EM method, and for the SC method with correctly specified prevalence, were similar to SB.

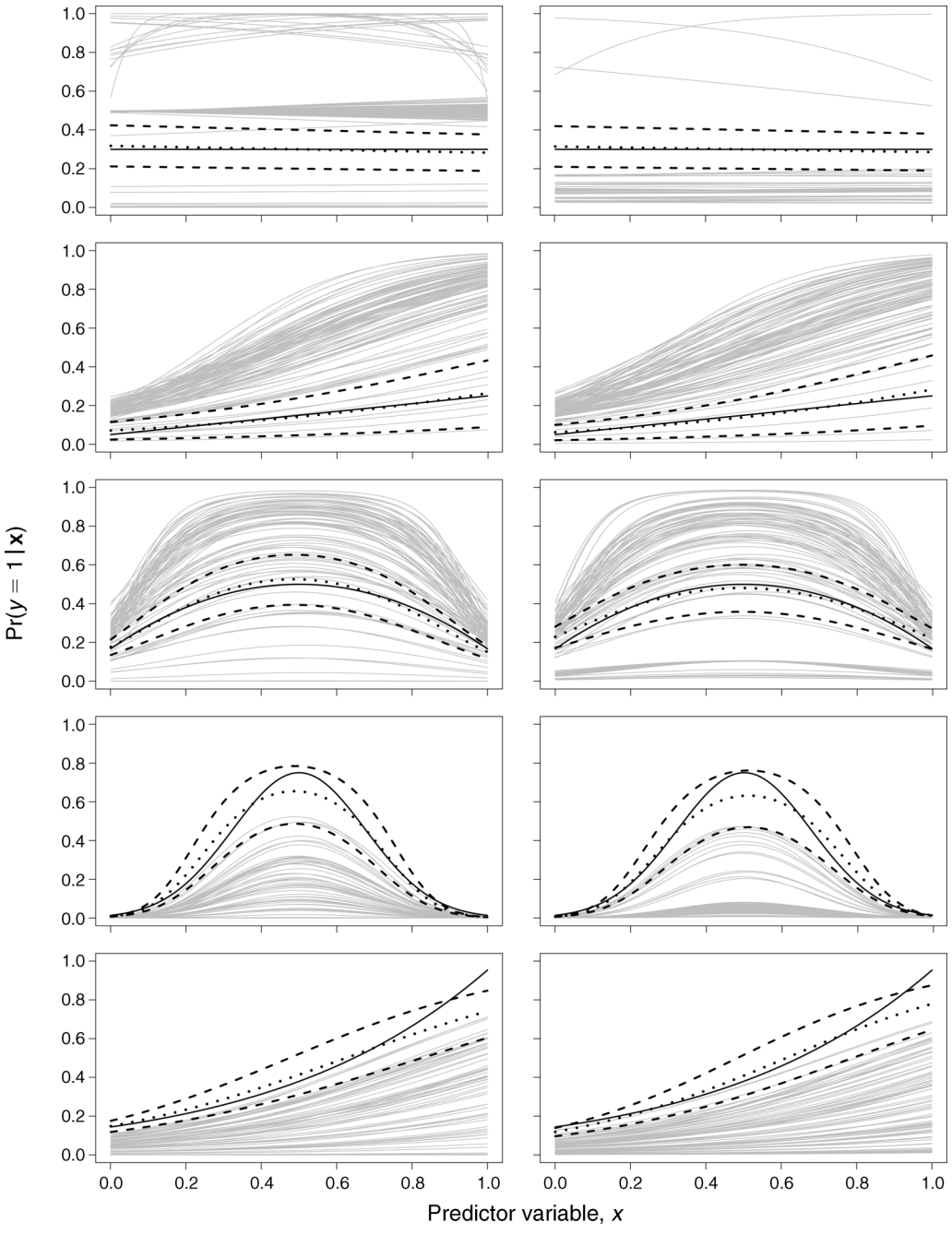

FIG. 2. Lele-Keim (left) and Lancaster-Imbens (right) models (gray lines represent 100 simulations) from simulated data (black line). Models were fit to each data set using the nlm function in R. For comparison, three scaled binomial (SB) models were fit using true prevalence (black dots) and true prevalence $\pm 10\%$ (black dashed lines). The functions used for true probability of presence do not have exactly linear or quadratic logit; the functions are (from top) Constant, Linear, Quadratic, Gaussian and Semi-Logistic. See Table 1 for mathematical definitions of the functions. Quadratic terms were included in models for the Quadratic and Gaussian cases. For each simulation, 1000 presence samples and 10 000 background samples were chosen from a landscape with the predictor variable x uniform in [0, 1].

The last two synthetic species in Table 1 (Logistic-1 and Logistic-2) conform to the assumptions of the LI and LK models (i.e., they are logit-linear). For these species, results were mixed (Figs. 1 and 3). All models had reasonable statistical performance for the "Logistic-2" species, but the LI and LK methods failed to produce a good approximation of probability of presence for "Logistic-1." Note that the probability of presence of the "Logistic-1" species is exactly proportional to that of the "Semi-Logistic" species, and therefore presence–

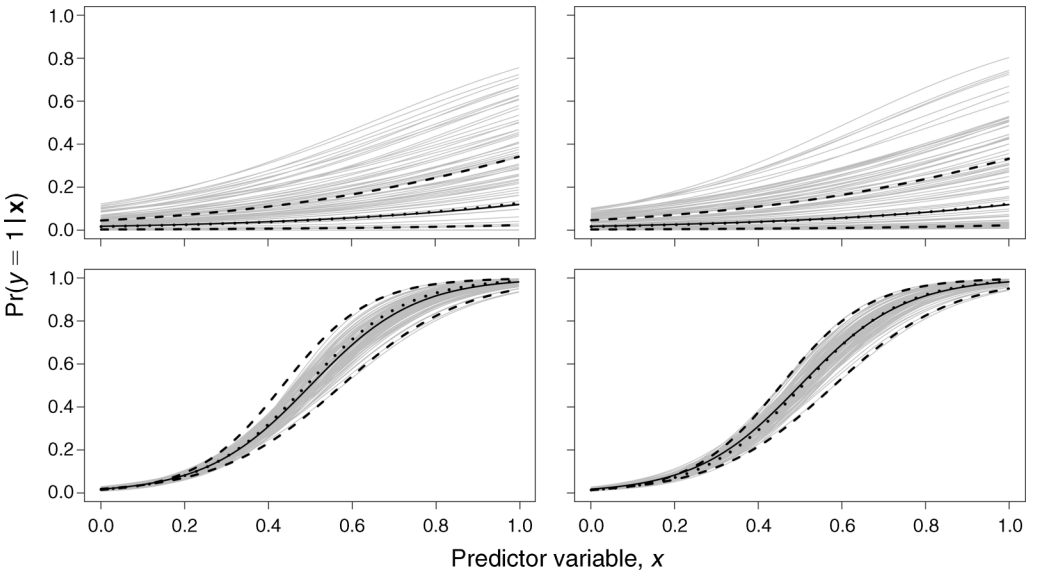

FIG. 3. Lele-Keim (left) and Lancaster-Imbens (right) models (gray lines) from simulated data (black line). Simulations were performed as for Fig. 2; scaled binomial (SB) models with true prevalence (black dots) and true prevalence $\pm 10\%$ (black dashed lines) are also shown. The functions are (from top) Logistic-1 and Logistic-2; in both cases, the true probability of presence has linear logit. See Table 1 for mathematical definitions of the functions.

background data for these two species are indistinguishable, as the distribution $\Pr(x \mid y = 1)$ is the same for both species. The LI and LK models for these two species are the same, except for random variation between simulations (i.e., they identify no difference between these "species").

The high RMS errors of LI and LK can be attributed in some cases to a large variation in model predictions between simulations. As an example, the LI and LK methods on the "Semi-Logistic" (or equivalently, the "Logistic-1") simulated species produced a wide spread of estimated coefficient values (Fig. 4c, d), which resulted in a broad range of estimated probability of presence (Fig. 4a). Although the LI and LK methods will converge to a stable combination of model parameters, given enough data (black square in Fig. 4c, d; black dashed lines in Fig. 4a), 1000 presence samples and 10 000 background samples were not sufficient to achieve convergence. The LI and LK methods did, in contrast, produce accurate estimates of *relative* probability of presence, i.e., they yielded accurate resource selection functions (RSF), as can be seen when rescaling the models so that the prediction at $x = 1$ matches the true probability of presence (Fig. 4b). They did not yield good resource selection probability functions (RSPF) (i.e., correct probabilities) because the constant of proportionality is not close to 1.

DISCUSSION

We surveyed maximum-likelihood-based methods for logistic modeling of species' probability of presence (or probability of use) from presence–background (or use–availability) data. The methods fall into two camps: the LI (Lancaster and Imbens 1996) and LK (Lele and Keim 2006, Royle et al. 2012) methods use a strong parametric assumption to make probability of presence identifiable, whereas the EM (Ward et al. 2009), SC (Steinberg and Cardell 1992), and SB (Phillips and Elith 2011) methods require the user to supply an estimate of the species' population prevalence.

Our experiments show that for many reasonable responses of a species to its environment, the methods that use a strong parametric assumption (LI and LK) fail to adequately estimate the species' true probability of presence from presence–background data. Our comparison is not on a level playing field, as the LI and LK methods use less data than the others. The important finding of our experiments is not the obvious fact that the additional datum used by SC, EM, and SB is *helpful*, but that it is *required* (except in special cases, which we will discuss), in contrast to claims by Royle et al. (2012).

Real species interact with their environment in complex ways, so the smooth response curves that we have considered here should be thought of as straightforward examples that any practical model should handle with ease. We therefore join Ward et al. (2009) in strongly recommending against methods that rely on the strong parametric assumption. The only methods that we can recommend for making maximum-likelihood-based logistic models of species' probability of presence from presence-only data are the EM, SC, or SB methods, all of which require a user-supplied estimate of prevalence.

The motivation for the methods with the strong parametric assumption is the desire to estimate absolute

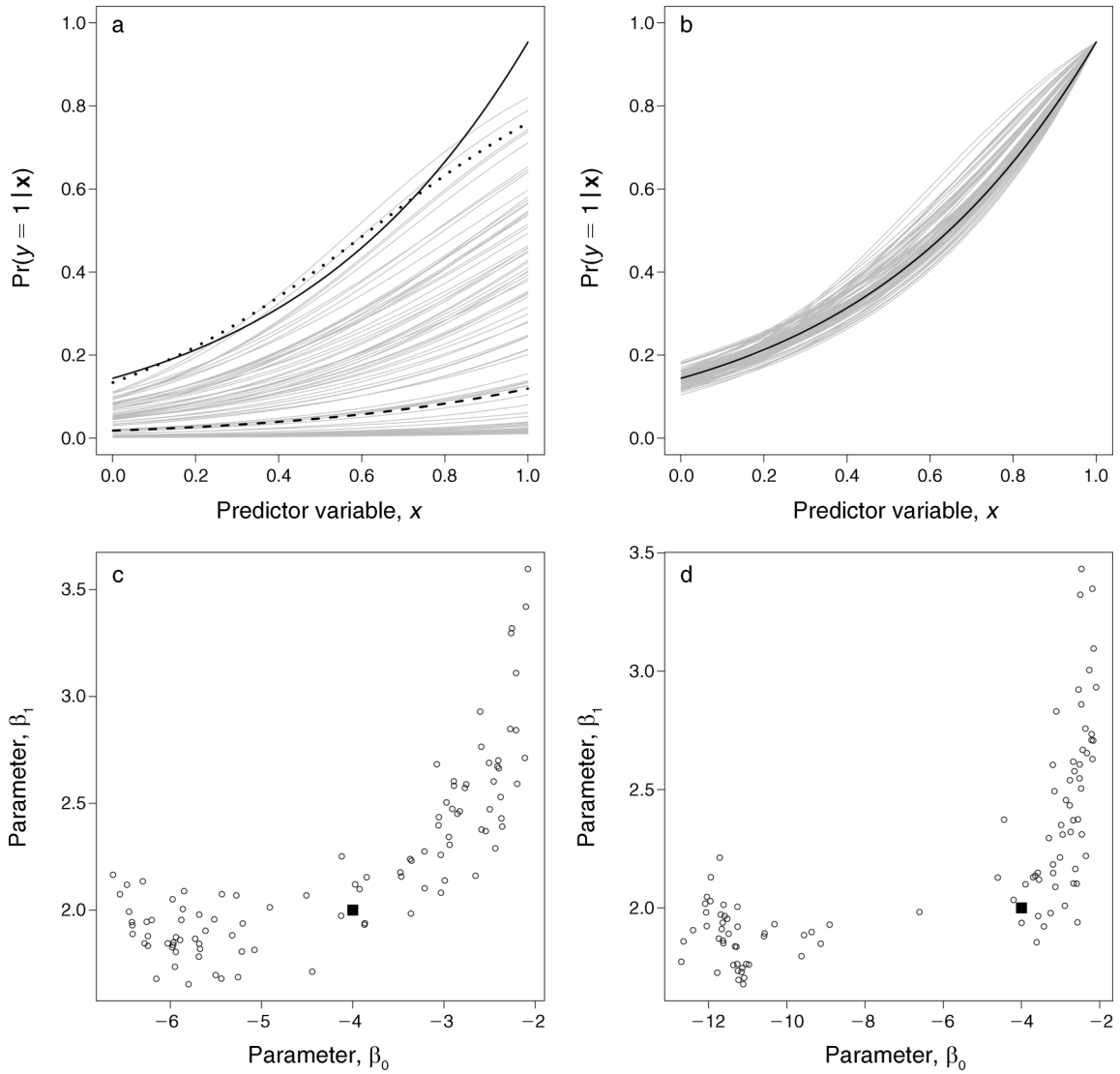

FIG. 4. Lancaster-Imbens and Lele-Keim models for the Semi-Logistic simulated species with $\Pr(y = 1 \mid \mathbf{x}) = 8/(1 + \exp[4 - 2x])$ (black line), made using 1000 presence samples and 10 000 background samples from a landscape with the predictor variable x uniform in [0, 1]. Models were fit for 100 simulations using the nlm function in R. (a) One hundred LI models (gray), and the maximum-likelihood estimate that LI would produce given unlimited data ($\Pr(y = 1 \mid \mathbf{x}) = 1/(1 + \exp[4 - 2x])$, black dashes). An EM model (black dots) is shown for comparison (only one simulation shown; others simulations were similar). (b) LI models after rescaling so that predictions at $x = 1$ match the true probability of presence. (c) LI fitted parameters β_0 and β_1 (open circles) and true parameter value (black square). (d) LK fitted parameters β_0 and β_1 (open circles) and true parameter value (black square).

probability of presence from presence-only data without the extra fieldwork required to estimate prevalence. However, we show that these methods are very fragile; they can give very bad estimates, even when the experimental or empirical data deviate only slightly from the strong assumption. Moreover, the strong assumption will generally be false for real species data: there is no reason to expect the probability of presence of any species to *exactly* match any particular model structure, as all models are approximations. Although it is obviously possible to make the LI and LK models more complex by use of quadratics, interactions, splines, different link functions, and so forth, their problem is not lack of complexity. Additional complexity was not necessary in our experiment, as evidenced by the good fit of alternative methods (SC, EM, and SB) using the same simple model structure (logistic models with linear and potentially quadratic terms). Our message is that there is no panacea for lack of data, and if one really wants absolute (rather than relative) probabilities, there is generally no alternative to collecting some data. Most of our simulated species have probability of presence whose logit is not exactly linear (or quadratic) in the predictors, which is reasonable because linearity is

always only an approximation to the complexity of ecological phenomena (Bio et al. 1998). However, in all cases, the true probabilities are well approximated by the EM, SB, and SC logistic methods, demonstrating that the logit is reasonably close to linear (or quadratic) in the predictors, yet the LI and LK methods failed to give useful estimates of the species' probability of presence. When the logit of probability of presence is exactly linear (and not constant), the LK and LI methods will converge to the true probability of presence, given enough data (Lancaster and Imbens 1996, Lele and Keim 2006; see also section *The LK method*). Nevertheless, 1000 presence samples were insufficient for convergence in one of our two simulated species.

The strong assumption of the LI and LK methods makes them very different from standard logistic (or other parametric) methods for presence–absence data: they are not "conventional likelihood methods" (Royle et al. 2012). With presence–absence data, logistic regression is appropriate and can be expected to give useful predictions whenever the predictors and their transformations (sensu Elith et al. 2011) are chosen so that the logit of the true probability of presence is approximately linear. If the true partial response to a predictor is unimodal, useful predictions can be made by adding a quadratic term to a logistic regression model, even if the true response is not exactly quadratic. In contrast, the LI and LK methods rely on the response *exactly* matching the form of the parametric model (e.g., see Lele and Keim 2006:3023), rather than approximately so, in order to identify prevalence, given enough data. However, our Logistic-1 species shows that even thousands of presence records might not suffice. Although the distinction between "exactly matching" and "approximately matching" might seem slight, we have shown that it has important consequences; when the strong assumption is false, the resulting estimates of probability of presence can be far wrong. We argue that the poor performance of these methods largely derives from the fact that any model is necessarily only an approximation to the true complexity of a species' response to its environment. Furthermore, as noted by Ward et al. (2009) and Royle et al. (2012), and further demonstrated here, the LI and LK methods can be unstable, requiring large amounts of data to converge to the optimal parameters, even when the strong assumption is true. If the species' prevalence is low (below 0.2), the LI method may not converge to a solution (Lancaster and Imbens 1996).

The EM, SC, and SB methods generally perform well in our experiments, given an estimate of prevalence, and they therefore warrant further experimentation on more realistic species data. When and how should these methods be used? The reason for, and advantage of, using these methods is to obtain an approximately unbiased estimate of absolute probability of presence, in contrast to established methods for estimating relative suitability such as Maxent or a RSF (Manly et al. 2002, Elith et al. 2011). Without true probabilities, it is difficult to compare across species in a given region for conservation planning, for example, or to know how likely it is that a species will occur at a site. Relative values can be useful but are often harder to work with in practice. The SB and EM approaches build on existing model-fitting methods, so all the usual features of the model in which they are included (ability to plot fitted functions, predict to mapped data, and so on) are available. They are useful as long as there is a reasonable possibility of making a reasonable estimate of prevalence. Such an estimate may be difficult to obtain. For example, if the species is cryptic and/or the sites are large grid cells, determining whether the species is present in even a small set of random sites may be prohibitively difficult or expensive. In that case, it is better to use methods like MaxEnt or resource selection functions and to work within the limitations of relative probabilities of presence. Using a substantially wrong estimate of prevalence (e.g., prevalence of 0.6 for a species that is, in reality, rare and restricted to one small part of the landscape), would result in EM and SB models that either broadly over-predict or under-predict the true probabilities, and can cause the SC method to yield infinite parameter values.

Acknowledgements

We thank Cory Merow for comments on a draft of this paper. J. Elith gratefully acknowledges Australian Research Council grant FT0991640.

Literature Cited

Aarts, G., J. Fieberg, and J. Matthiopoulos. 2012. Comparative interpretation of count, presence–absence and point methods for species distribution models. Methods in Ecology and Evolution 3:177–187.

Aarts, G., M. MacKenzie, B. McConnell, M. Fedak, and J. Matthiopoulos. 2008. Estimating space-use and habitat preference from wildlife telemetry data. Ecography 31:140–160.

Beyer, H. L., D. T. Haydon, J. M. Morales, J. L. Frair, M. Hebblewhite, M. Mitchell, and J. Matthiopoulos. 2010. The interpretation of habitat preference metrics under use–availability designs. Philosophical Transactions of the Royal Society B 365:2245–2254.

Bio, A., R. Alkemande, and A. Barendregt. 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. Journal of Vegetation Science 9:5–16.

Chakraborty, A., A. E. Gelfand, A. M. Wilson, A. M. Latimer, and J. A. Silander. 2011. Point pattern modelling for degraded presence-only data over large regions. Applied Statistics 60:757–776.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39:1–38.

Desrochers, A., E. J. McIntire, S. G. Cumming, J. Nowak, and S. Sharma. 2010. False negatives—a false problem in studies of habitat selection? Ideas in Ecology and Evolution 3:20–25.

Dorazio, R. 2012. Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68:1303–1312.

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

Elith, J., S. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43–57.

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.

Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, UK.

Jiménez-Valverde, A., J. M. Lobo, and J. Hortal. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions 14:885–890.

Johnson, C. J., S. E. Nielsen, E. H. Merrill, T. L. McDonald, and M. S. Boyce. 2006. Resource selection functions based on use–availability data: Theoretical motivation and evaluation methods. Journal of Wildlife Management 70:347–357.

Keating, K. A., and S. Cherry. 2004. Use and interpretation of logistic regression in habitat-selection studies. Journal of Wildlife Management 68:774–789.

Lancaster, T., and G. Imbens. 1996. Case-control studies with contaminated controls. Journal of Econometrics 71:145–160.

Lee, A., A. Scott, and C. Wild. 2006. Fitting binary regression models with case-augmented samples. Biometrika 93:385–397.

Lele, S. R. 2009. A new method for estimation of resource selection probability function. Journal of Wildlife Management 73:122–127.

Lele, S. R., and J. L. Keim. 2006. Weighted distributions and estimation of resource selection probability functions. Ecology 87:3021–3028.

Li, W., Q. Guo, and C. Elkan. 2011. Can we model the probability of presence of species without absence data? Ecography 34:1096–1105.

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, New York, New York, USA.

Phillips, S. 2012. Inferring prevalence from presence-only data: a response to 'Can we model the probability of presence of species without absence data?' Ecography 35:385–387.

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. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175.

Phillips, S., M. Dudík, J. Elith, C. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181–197.

Phillips, S., and J. Elith. 2011. Logistic methods for resource selection functions and presence-only species distribution models. Pages 1384–1389 *in* Proceedings of the 25th AAAI Conference on Artificial Intelligence, San Francisco, California, USA.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters 3:349–361.

Reddy, S., and L. M. Dávalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30:1719–1727.

Royle, J. A., R. B. Chandler, C. Yackulic, and J. D. Nichols. 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods in Ecology and Evolution 3(3):545–554.

Steinberg, D., and N. S. Cardell. 1992. Estimating logistic regression models when the dependent variable has no variance. Communications in Statistics—Theory and Methods 21(2):423–450.

Ward, G., T. Hastie, S. Barry, J. Elith, and J. Leathwick. 2009. Presence-only data and the EM algorithm. Biometrics 65: 554–563.

Warton, D., and L. Shepherd. 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Annals of Applied Statistics 4:1383–1402.

Supplemental Material

Supplement

R script files to perform the simulations of Figs. 1–3 (*Ecological Archives* E094-125-S1).