Methods in Ecology and Evolution British Ecological Society

Methods in Ecology and Evolution doi: 10.1111/j.2041-210X.2011.00182.x

Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions

J. Andrew Royle*, Richard B. Chandler, Charles Yackulic and James D. Nichols

U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, MD 20708, USA

Summary

1. Understanding the factors affecting species occurrence is a pre-eminent focus of applied ecological research. However, direct information about species occurrence is lacking for many species. Instead, researchers sometimes have to rely on so-called presence-only data (i.e. when no direct information about absences is available), which often results from opportunistic, unstructured sampling. MAXENT is a widely used software program designed to model and map species distribution using presence-only data.

2. We provide a critical review of MAXENT as applied to species distribution modelling and discuss how it can lead to inferential errors. A chief concern is that MAXENT produces a number of poorly defined indices that are not directly related to the actual parameter of interest – the probability of occurrence (ψ). This focus on an index was motivated by the belief that it is not possible to estimate ψ from presence-only data; however, we demonstrate that ψ is identifiable using conventional likelihood methods under the assumptions of random sampling and constant probability of species detection.

3. The model is implemented in a convenient R package which we use to apply the model to simulated data and data from the North American Breeding Bird Survey. We demonstrate that MAXENT produces extreme under-predictions when compared to estimates produced by logistic regression which uses the full (presence/absence) data set. We note that MAXENT predictions are extremely sensitive to specification of the background prevalence, which is not objectively estimated using the MAXENT method.

4. As with MAXENT, formal model-based inference requires a random sample of presence locations. Many presence-only data sets, such as those based on museum records and herbarium collections, may not satisfy this assumption. However, when sampling is random, we believe that inference should be based on formal methods that facilitate inference about interpretable ecological quantities instead of vaguely defined indices.

Key-words: Bayes' rule, detection probability, logistic regression, MAXENT, occupancy model, occurrence probability, presence-only data, species distribution model

Introduction

Species distribution is naturally characterized by the probability of occurrence of a species, say $\psi(x) = \Pr(y(x) = 1)$ where $y(x)$ is the true occurrence state of a species at some location (pixel) x (Kéry 2011). Inference about $\psi(x)$ can be achieved directly from presence–absence data using logistic regression and related models (MacKenzie *et al.* 2002). However, ecologists are not always fortunate enough to have presence–absence data, and many data sets exist which only contain locations of species presence – so-called presence-only data.

MAXENT (e.g. Phillips *et al.* 2006) is a popular software package for producing 'species distribution' maps from presence-only data. Interestingly, MAXENT does not produce estimates of occurrence probability but, instead, produces estimates of an ill-defined 'suitability index' (Elith *et al.* 2011). Because MAXENT does not correspond to an explicit model of species occurrence, it is not suitable for making explicit predictions of an actual state variable or testing hypotheses about factors that influence occurrence probability. Support for producing indices of species distribution from presence-only data, as opposed to

*Correspondence author. E-mail: aroyle@usgs.gov
Correspondence site: http://www.respond2articles.com/MEE/

estimates of occurrence probability, has been justified in the literature based on the *incorrect* assertion that occurrence probability ψ (sometimes referred to as 'prevalence' or occupancy) cannot be estimated from presence-only data.

The principle aim of our paper is to show that occurrence probability *can* be estimated from presence-only data. We consider a formal model-based approach to analysis of presence-only data. We emphasize the critical assumption required for statistical inference about species occurrence probability from presence-only data, which is random sampling of space as a basis for accumulating presence-only observations. In addition, the estimator we devise here is most relevant only when species detection probability is constant. We conclude that, under these assumptions, inference about occurrence probability can be achieved directly from presence-only data using conventional likelihood methods (e.g. Lancaster and Imbens 1996). We suspect that this is surprising to many users of MAXENT and related species distribution modelling tools in the light of repeated statements to the contrary in the literature (e.g. Phillips and Dudik 2008; Elith *et al.* 2011; Kéry 2011), asserting that probability of occurrence is not identifiable. For example, Elith *et al.* (2010) state that

> Formally, we say that prevalence is not identifiable from presence-only data (Ward *et al.* 2009). This means that it cannot be exactly determined, regardless of the sample size; this is a fundamental limitation of presence-only data.

In fact, Ward *et al.* (2009) do not make such a definitive claim. Their precise claim is

> [...occurrence probability...] is identifiable only if we make unrealistic assumptions about the structure of [...the relationship between occurrence probability and covariates....] such as in logistic regression....

In that context, it seems that subsequent references to Ward *et al.* (2009) misconstrue their result. In our view, logistic regression (or other binary regression models) is hardly unrealistic. Indeed, such models are the most common approach to modelling binary variables in ecology (and probably all of statistics), especially in the context of modelling species occurrence (MacKenzie *et al.* 2002; Tyre *et al.* 2003; Kéry *et al.* 2010). Even more generally, the logistic function is the canonical link of the binomial GLM (McCullagh and Nelder 1989, p. 38) and, as such, it is customarily adopted and widely used, and even books have been written about it (Hosmer and Lemeshow 2000).

We demonstrate the application of the formal model-based framework for estimating occurrence probability from presence-only data using a data set derived from the North American Breeding Bird Survey, and we provide an R package for producing estimates of species distribution model parameters from presence-only data.

Before proceeding, we note that the statistical principle of maximum entropy (Jaynes 1957, 1963; Jaynes and Bretthorst 2003) is widely applied to problems in statistics and other disciplines, and our development here is not critical of these ideas. Rather, we are critical of the routine application of the software package MAXENT as applied to species distribution modelling. We specifically object to the pervasive views in the MAXENT user community that one should avoid characterizing species distribution by occurrence probability, that occurrence probability is not identifiable and that one should instead obtain indices of species occurrence probability by using MAXENT.

Genesis of presence-only data

The original motivation for the development of MAXENT was to estimate and model the distribution of a species using presence-only data (Dudik *et al.* 2004; Phillips *et al.* 2004). Species distribution is naturally characterized by occurrence probability, which provides a quantitative description of the probability of the focal species occurring at a location *and* a mechanism for generating explicit predictions of occurrence and testing hypotheses related to factors that influence occurrence. MAXENT attempts to approximate the probability of occurrence by using a logistic transformation of its suitability index (Phillips & Dudik 2008). Before explaining the details of this indirect method, we first consider a model to describe the genesis of presence-only data, and the common approach of estimating the probability of occurrence using standard sampling methods.

OCCUPANCY OR PRESENCE/ABSENCE SAMPLING EXPERIMENT

We imagine that presence-only data arise by randomly sampling spatial units, say x (e.g. corresponding to a pixel), and then observing a random variable y which we will assert is true presence or absence at x. The state space of potential values of x will be denoted by $\mathcal{X}$, and we assume henceforth that $\mathcal{X}$ is prescribed by the investigator. The data resulting from random sampling are $y(x)$ for each x in the sample, say $x_1,\ldots,x_N$. Naturally (under random sampling), we assume

$$y(x) \sim \text{Bernoulli}(\psi(x)) \qquad \text{eqn 1}$$

where $\psi(x) = \Pr(y(x) = 1)$ is the probability that the species is present at pixel x – or the probability that pixel x is 'occupied'. To be explicit, we note that $\psi(x)$ is the conditional probability of *y given x*, which we will subsequently denote by $\psi(y|x)$. In practical applications, attention is usually focused on developing covariate models on the logit-transform of $\psi(y|x)$, for example,

$$\text{logit}(\psi(y|x)) = \beta_0 + \beta_1 z(x)$$

where $z(x)$ is some landscape or habitat covariate. For example, elevation, forest cover or annual precipitation was measured at pixel x. This is a standard logistic regression model (e.g. Hosmer and Lemeshow 2000) and corresponds also to the state model underlying occupancy models (MacKenzie *et al.* 2006). The model is widely used to model occurrence, range and distribution of species. We note that while the logistic model is the most widely used because it is the canonical link function of the binomial GLM (McCullagh and Nelder 1989), many other link functions are available, although seldom are alternatives considered in ecological applications.

PRESENCE-ONLY SAMPLING EXPERIMENT

We adopt the view here that presence-only data, that is, a sample of locations for which $y = 1$, arise by discarding the $y = 0$ observations from a data set that arose by random sampling as described earlier. That is, we sample pixels randomly and obtain $x_1,\ldots,x_N$ and record $y(x_1),\ldots,y(x_N)$. Then, we consider only those sites $x_1,\ldots x_n$ for which $y(x) = 1$. The corresponding subset of locations constitutes our data set, which we will label here $x_1,\ldots,x_n$. We use 'n' here instead of 'N' as above and recognize that the presence-only x's are a reordered version of a subset of the initial sample.

Likelihood analysis

The basic characteristic of presence-only data is that the variable y is no longer random in our sample, that is, because $y = 1$ with probability 1 for all observations. Instead, x is the random quantity, and the set of n locations $x_1,\ldots,x_n$ are the data upon which inference is based. Importantly, the specific values of x that appear in the sample represent a biased selection from all possible values $\mathcal{X}$, favouring those for which $y = 1$. To clarify the nature of the induced bias in our sample of x, we invoke Bayes rule. In the remainder, we use $\pi()$ to represent the probability distributions of x, and $\psi()$ to represent probability distributions of y.

The central statistical problem in the analysis of presence-only data is to identify the likelihood of the observations x in the presence-only sample. To find the likelihood, we need to identify the conditional probability distribution $\pi(x|y = 1)$, that is, that of x for presence-only ($y = 1$) pixels. We can compute $\pi(x|y = 1)$ by an application of Bayes' rule:

$$\pi(x|y = 1) = \frac{\psi(y = 1|x)\pi(x)}{\psi(y = 1)} \qquad \text{eqn 2}$$

This might appear to be an awkward invocation of Bayes rule because we often do not think of spatial location as the outcome of a stochastic process in most contexts. It is somewhat more natural in the context of environmental covariates (Lancaster and Imbens 1996; Lele and Keim 2006), but they are equivalent formulations. We proceed with this development in terms of x here because this is pervasive in the MAXENT literature.

The probability distribution $\pi(x)$ is that describing the possible outcomes of the random variable x – 'pixel identity'. We regard the state space of x as discrete here, having M unique elements, and therefore $\pi(x) = 1/M$. $\psi(y = 1|x)$ is the probability that $y = 1$ conditional on x – what we refer to as 'occurrence probability'. Then, $\psi(y = 1)$ is the marginal probability that a pixel is occupied, which is, by definition, the integral of $\psi(y = 1|x)$ over $\mathcal{X}$or, in the case of a discrete landscape, the sum over all elements of $x \in \mathcal{X}$:

$$\psi(y = 1) = \sum_{x\in\mathcal{X}} \psi(y = 1|x)\pi(x)$$

We see that this is the 'spatially averaged' occurrence probability and has also been called prevalence in the literature (e.g. Ward *et al.* 2009).

Equation 2 makes it clear that the x's for which $y = 1$ are not a representative sample of all x's. Intuitively, the presence-only sample (i.e. x's for which $y = 1$) will favour pixels for which $\psi(y = 1|x)$ is large relative to $\psi(y = 1)$.

In this expression of Bayes rule, the variable x is 'pixel identity' for which $\pi(x)$ is constant. It is not an indicator of whether pixel x appears in the sample. In the latter case, $\Pr(y = 1|x)$ would be the probability of occupancy conditional on pixel x being in the sample which has no clearly useful meaning. That said, random sampling is important for the invocation of Bayes rule – by imagining that the presence-only sample arises by first random sampling pixels for presence–absence and then discarding the $y = 0$ pixels (See Appendix). Alternatively, it can be justified by sampling randomly the sample frame consisting of all $y = 1$ observations. Under random sampling, either with or without replacement, the probability that a sample unit appears in the sample is constant and thus has no effect on Eqn 2.

THE LIKELIHOOD

We note that this expression of Bayes rule appears in a large number of species distribution modelling papers that involve the development or application of MAXENT. However, these papers never provide further analysis of the result, instead making the (incorrect) claim that direct analysis of $\pi(y = 1|x)$ is intractable because the background prevalence ($\psi(y = 1)$ here) is not identifiable. In fact, this is incorrect, as has been noted in other contexts that produce similarly biased data (e.g. case–control studies, Lancaster and Imbens 1996, and 'resource selection probability functions', RSPF; Manly *et al.* 2002; Lele and Keim 2006). To clarify this, we use the previous application of Bayes rule to describe the likelihood for the presence-only data.

We emphasize that $\psi(y = 1|x)$ here is precisely the 'probability of occurrence' or occupancy probability, that is, $\Pr(y = 1|x)$, as in MacKenzie *et al.* (2002, 2006) and Tyre *et al.* (2003). In practice, these probabilities would usually depend on parameters, say β, which we write $\psi(y|x;\beta)$. For example, occurrence probability might vary according to a polynomial response over space, on an appropriate scale. Therefore, $\pi(x|y = 1)$ is

$$\pi(x_i|y_i = 1) = \frac{\psi(y_i = 1|x_i;\beta)\pi(x_i)}{\sum_{x\in\mathcal{X}} \psi(y = 1|x;\beta)\pi(x)}.$$

and with $\pi(x)$ constant, it cancels from the numerator and denominator and so

$$\pi(x_i|y_i = 1) = \frac{\psi(y_i = 1|x_i;\beta)}{\sum_{x\in\mathcal{X}} \psi(y = 1|x;\beta)}. \qquad \text{eqn 3}$$

The likelihood for an observation x_i in our presence-only data set is based on $\pi(x|y = 1;\beta)$ regarded as a function of the parameters β. Therefore, for the presence-only sample $x_1,\ldots,x_n$, the joint likelihood is

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} \frac{\psi(y_i = 1|x_i;\beta)}{\sum_{x\in\mathscr{X}} \psi(y = 1|x;\beta)}. \qquad \text{eqn 4}$$

Parameters β can be estimated by maximizing the likelihood using standard methods. Further, inferences in the form of hypothesis tests, confidence intervals or model selection can be achieved using conventional statistical ideas.

As noted earlier, the denominator is the marginal probability of occurrence over the landscape, and it is computed by summing over all elements of $x \in \mathscr{X}$ where $\mathscr{X}$ is the state space of x, that is, the landscape as defined by the analyst. Clearly, this marginal probability could be estimated by evaluating $\psi(y = 1|x)$ at a random sample of points x independent of y (Lele and Keim 2006). Sometimes, a sample of $\mathscr{X}$ chosen independent of y is referred to as the 'background' in species distribution modelling or, in the context of case–control models, 'contaminated controls' (Lancaster and Imbens 1996).

IMPERFECT DETECTION OF SPECIES

In practice, we expect bias in observing species presence, such that the probability of detecting a species given that it is present should be less than 1. A standard model of this phenomenon (MacKenzie *et al.* 2002; Tyre *et al.* 2003) is constructed as follows: Let y_{obs} be the observed species presence and then define the probability of detection as $\Pr(y_{\text{obs}} = 1|y = 1\ , x) = p$. If p is constant spatially, the marginal probability of the contaminated observations y_{obs} is $p\psi(y|x)$, and we see that the constant p cancels from the Eqn 4 and inferences about ψ are unaffected.

GEOGRAPHIC VS. ENVIRONMENTAL SPACE

MAXENT is now always discussed in terms of environmental covariates, say z (typically vector-valued), whereas we have developed the problem thus far in terms of space, x. Elith *et al.* (2011; Appendix) make the first attempt we know of to reconcile what they call the 'geographic space' (in terms of x) and the 'environmental space' (in terms of z) formulations. Earlier papers seem to simply substitute z for x without much (if any) discussion of that. Conceptually, there is no reason to regard x and z differently with regard to the application of Bayes rule, and so simply substituting z for x is reasonable, if we assume that z is randomly sampled instead of x. In discrete space, this is assured because x in that case is merely an index to unique elements of z, so that $z(x)$ is effectively a 1-to-1 transformation of x, that is, the elements of the sample frame of $z(x)$ are associated with unique elements of x. It is then somewhat more natural to view the application of Bayes rule in terms of how occupancy status relates to environmental covariates because the view of z as a random variable is standard. When y only depends on x through z, this is often expressed as:

$$\pi(z(x)|y = 1) = \frac{\psi(y = 1|z(x))\pi(z(x))}{\psi(y = 1)}$$

By the law of total probability, the marginal probability $\psi(y = 1|z(x))$ can be computed directly or estimated if a random sample of $z(x)$ independent of y is available (Lele and Keim 2006).

We introduce covariates into the model by modelling the relationship between $\psi(y = 1|z(x))$ and those covariates, for example, using a logit link:

$$\text{logit}(\psi(y = 1|z(x))) = \beta_0 + \beta_1 z_1(x) + \ldots\beta_J z_J(x) \qquad \text{eqn 5}$$

where β_0 is an intercept and the other β's are coefficients associated with each of J covariates.

LIKELIHOOD ANALYSIS IN R

We provide an example here involving a landscape comprised of 10 000 pixels. We consider a single covariate z, which for our purposes we define by simulating 10 000 values from a standard normal distribution. The probability of occurrence is defined as

$$\text{logit}(\psi(x)) = -1 - 1 * z(x).$$

We generated y (true occurrence) for each pixel on the landscape as a Bernoulli trial with probabilities $\psi(x)$ and sampled 2000 occupied pixels for which we used the resulting values of $z(x)$ as our data set. The likelihood function definition requires about a half-dozen lines of R code. Every line of **R** code for simulating data, defining the likelihood and obtaining the MLEs is given in the following box:

```
z<- rnorm(10000,0,1) # simulate a covariate
lpsi<- -1 -1* z # define the linear predictor
      # occurrence probability
psi<-exp(lpsi)/(1+exp(lpsi))
      # generate presence-absence data
y<-rbinom(10000,1,psi)
      # keep the presence-only data
data<- sample(z[ y==1] ,2000)
      # define the neg log-likelihood

lik<-function(parm){
  beta0<-parm[1]
  beta1<-parm[2]
 gridpsi<-
  exp(beta0+beta1* z)/(1+exp(beta0+beta1* z))
 datapsi<-
  exp(beta0 + beta1* data)/(1+exp(beta0+
    beta1* data))
 -1* sum(log(datapsi/(sum(gridpsi))))
}
  # minimize it
out<-nlm(lik,c(0,0),hessian=TRUE)
  # produce the estimates
out$estimate
```

We conducted 5000 simulations under the model described above and found that the MLEs were unbiased (Fig. 1).

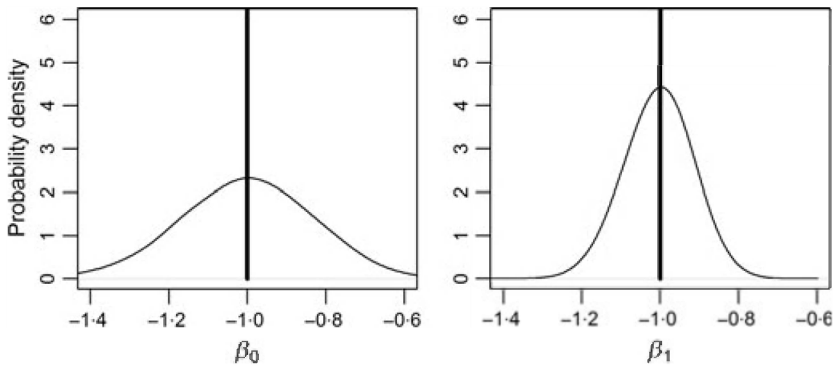

Fig. 1. Distributions of the maximum likelihood estimates obtained by fitting our model to 5000 simulated data sets. β_0 and β_1 are the intercept and slope parameters of the linear model of occurrence probability ($\psi(y = 1|z)$). The data-generating values are indicated by vertical lines. Kernel density estimators were used to represent the distributions.

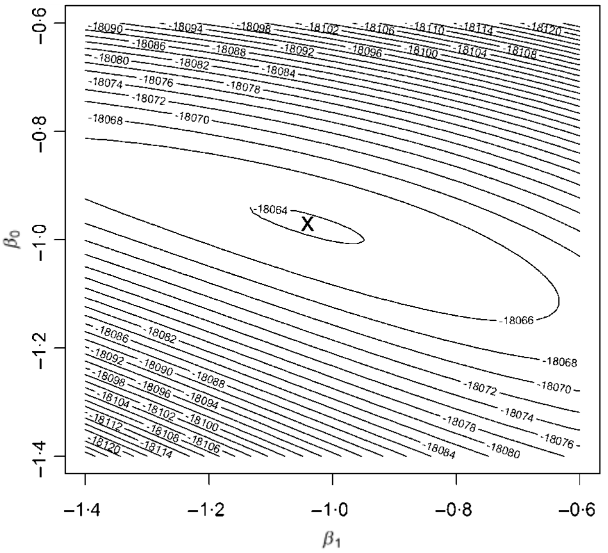

Fig. 2. The log-likelihood surface of the MAXLIKE model for a data set simulated using logit($\psi(y = 1|z)) = -1-1*z$. The 'X' indicates the maximum. Parameters of the model are identifiable, but there exists a prominent ridge in the likelihood.

Furthermore, the log-likelihood has a distinct mode (Fig. 2), indicating that occurrence probability, $\psi(y = 1|z)$, is identifiable in the situations we examined, under random sampling. We do note, however, that there is a prominent ridge in the likelihood, highlighting the low information content inherent in presence-only data. We developed an R package 'MAXLIKE', which implements the likelihood analysis in some generality.

MAXENT analysis

The definition of the likelihood for an observed sample of x from presence-only sites is straightforward using Bayes Rule as we demonstrated earlier. MAXENT is not using the likelihood as a basis for estimation. Instead, MAXENT is operating on what the various authors refer to as the 'maximum entropy distribution'. Using a single covariate, $z(x)$, as an example, Phillips & Dudik (2008; and others) define the 'maximum entropy distribution' as:

$$q(x_i) = \frac{\exp(\beta z(x_i))}{\sum_x \exp(\beta z(x))} \qquad \text{eqn 6}$$

As in the development of Eqn 4, the summation in the denominator is taken over the background or, if available, all pixels $x \in \mathscr{X}$. In addition, MAXENT obtains β by maximizing a *penalized* version of Eqn 6.

It is natural to wonder how $q(x)$ relates to the likelihood given above in Eqn 4. It is clear from the development (e.g. Phillips and Dudik 2008 and others) that MAXENT implies the following strict equivalence

$$\pi(x|y = 1) \equiv q(x)$$

For example, Phillips and Dudik (2008) and subsequent papers state that MAXENT is 'estimating $\pi(x)$', where $\pi(x)$ is their notation for what we labelled $\pi(x|y = 1)$ above. To be precise, the Phillips and Dudik (2008) state: 'However, if we have only occurrence data, we cannot determine the species' prevalence (Phillips *et al.* 2006; Ward *et al.* 2009). Therefore, instead of estimating $P(y = 1|x)$ directly, we estimate the distribution π.'

We therefore are led to ask: In what sense is MAXENT 'estimating π'? The only clear interpretation of $\pi(x|y = 1) \equiv q(x)$ is that MAXENT is estimating a *specific* version of $\pi(x|y = 1)$ in which $\Pr(y = 1|z(x)) = \psi(y = 1|z(x)) = \exp(\beta z(x))$ (i.e. occurrence probability is modelled as an exponential function) and, furthermore, a penalized form of that specific $\pi(x|y = 1)$. Clear advantages to either of these two methodological choices ($\psi(y = 1|z(x)) = \exp(\beta z(x))$ and the penalty) have not been established. In particular, modelling probabilities by a simple exponential function does not appear to be customary, or even very natural, as it does not have bounded support on [0,1] as $\psi(y = 1|z(x))$ must.

IDENTIFIABILITY OF β_0 OR 'SPECIES PREVALENCE'

There is a widespread and incorrect belief (see Phillips and Dudik quote above) that species prevalence (Elith *et al.* 2011 also use the term 'proportion of occupied sites') cannot be determined from presence-only data (Phillips and Dudik 2008; Elith *et al.* 2011; Kéry 2011), and this is widely used as justification for producing vaguely defined 'suitability indices'. While this is repeatedly asserted, there is never any specific discussion or argument as to why this is the case. In fact, it is in direct contradiction to existing literature (Lancaster and Imbens 1996; Lele and Keim 2006).

As we demonstrated, lack of identifiability of occurrence probability is not a general feature of presence-only data. We can clearly estimate the intercept term in Eqn 5 by maximum likelihood, if a suitable parametric form of $\psi(y = 1|z)$ is assumed (Lancaster and Imbens 1996) and a continuous covariate is present (Lele and Keim 2006). Lacking a particular parametric form, one must know $\psi(y = 1)$ (Lancaster and Imbens 1996) and, if covariates are only nominal categorical, then only relative probabilities of occurrence are achievable (Lele and Keim 2006). Conversely, it is clear that, no matter the composition of covariates, with the choice $\psi(y_i = 1|z) = \exp(\beta z_i)$, the intercept term is *not* identifiable, as the intercept cancels from Eqn 6 (Lele and Keim 2006). Thus, the inability to estimate occurrence probability is a feature of the specific model used by MAXENT and not a feature of presence-only data. As such, we do not see an advantage to using the exponential form for $\psi(y = 1|z)$ over the more conventional logistic occurrence probability model.

Logistic output from maxent

In relying on the 'MAXENT distribution', instead of adopting a direct focus on occurrence probability, the ability to estimate the intercept is lost. Despite this, MAXENT provides an *ad hoc* approach to producing 'logistic output' (Phillips and Dudik 2008; Elith *et al.* 2011, Appendix). This procedure amounts to prescribing a particular value of β_0, so that the output of MAXENT can be interpreted as a probability. Indeed, graphical output from MAXENT explicitly and misleadingly labels such output 'probability of presence'. Phillips and Dudik (2008) imply some objective basis for this procedure, by stating (Phillips and Dudik 2008, abstract) 'we describe a new logistic output format that gives an *estimate* of the probability of presence' (our emphasis). The implication here is that one is able to *estimate* 'probability of presence' using MAXENT. In fact, as the rest of the paper makes clear, the user is required to prescribe a specific value for the intercept. Phillips and Dudik provide the following equation:

$$\psi = \frac{\exp(\beta_0)q}{1 + \exp(\beta_0)q}$$

or, equivalently,

$$\text{logit}(\psi) = \beta_0 + \log(q)$$

where they recommend setting $\beta_0 \equiv H$ where H is the mean $\log(q)$ for the observed data. They provided an argument that one might as well set $\beta = 0$ so that the procedure '...assigns typical presence sites probability of presence close to 0.5'. (Phillips and Dudik 2008, p. 165). This is questionable because β_0 for 'typical presence sites' could be any real number, and thus, the bias of MAXENT will depend on how different the true value is from the prescribed value of β_0. Thus, Phillips and Dudik (2008) fail to provide an objective approach to estimating this value.

'REGULARIZATION' IN MAXENT

In MAXENT, the specific objective function maximized is *not* the likelihood given in Eqn 4 earlier. Rather, it is the exponential function along with a penalty term that has the effect of penalizing the maximum entropy distribution for large covariate effects. This penalization is termed 'regularization' in the MAXENT terminology. The effect of the penalty is clear – it shrinks the regression coefficients to 0 – and it is a standard concept in smoothing methods and other contexts (Green and Silverman 1994; Tibshirani 1996). General motivation for the need of a penalty term in the context of species distribution modelling is not clear, and we have not seen specific justification given in the literature other than the suggestions that it may prevent over-fitting or save the user time by avoiding the need to formally compare competing hypotheses (Phillips *et al.* 2004). As an alternative to regularization, we recommend that exercising restraint in the creation of covariate data sets and maintaining a focus on developing *a priori* models can also prevent over-fitting.

The practical problem in using the penalized objective function is that it will generally lead to *biased* estimators of the important β parameters, and we believe that this should be considered and understood by users of MAXENT prior to analysis. In our view, this penalty is not necessary in developing occupancy models from presence-only data. Moreover, the way in which it is handled by MAXENT seems *ad hoc* – which is to say, the smoothing parameters are fixed a priori based on heuristics.

There could be at least two situations in which using a penalized objective function is a sensible thing to do. One is when no obvious model set can be developed *a priori* and the number of potential models is extremely large. In that case, we might wish to fit some omnibus complex model for the sake of hypothesis generation. A second possibility for using a penalty to the objective function is in the presence of sparse data, or small sample size relative to the number of predictors. In that case, some model parameters will be weakly identified, and the penalty term essentially keeps the parameter in a reasonable region of the parameter space and probably alleviates numerical errors and other pathologies that you might expect in such cases.

MAXENT SCALINGS OF $q(x)$

To implement precisely the estimation problem as implemented in MAXENT, it is not sufficient to understand the 'maximum entropy distribution'. This is because MAXENT implements various scalings to force $q(x)$ to be between 0 and 1. In particular, MAXENT uses two 'normalizing constants' that are involved in the calculation of $q(x)$: the linear normalizer (LN) and the density normalizer (DN), in addition to the penalty (λ). Thus, MAXENT minimizes the following function:

$$q(x) = -\log(\exp(\mathrm{LN} + \beta * z(x))/\mathrm{DN}) + \lambda * \beta$$

The DN is simply equal to the sum of $\exp(\mathrm{LN} + \beta z(x))$ over all x (not just the observed locations) and thus is just a function of the observed z and the coefficient β. The LN is chosen to equal -1 * maximum value of $(\beta^* z(x))$, ensuring that $\mathrm{LN} + \beta^* z(x)$ is less than zero for all x.

We see that LN itself is a redundant parameter because $\exp(\mathrm{LN} + \beta^* z(x)) = \exp(\mathrm{LN})^*\exp(\beta^* z(x))$ and LN is included in both the denominator and numerator. Furthermore, we can absorb DN directly into β, yielding the reparameterization $\beta^* = \beta/\mathrm{DN}$. As such, the scaling of q does not appear be a meaningful methodological element of the MAXENT problem.

Comparison between maxent and maxlike

SIMULATION STUDY

We compared the MAXENT suitability index to estimates of $\psi(x)$ obtained by maximizing the likelihood in Eqn 4 by fitting both models to 100 single simulated data sets. We created an artificial landscape in which each pixel had associated elevation and precipitation data. We generated presence–absence data using the following model:

$$\begin{aligned}\mathrm{logit}(\psi) = \beta_0 + \beta_1\mathrm{Elevation} + \beta_2\mathrm{Elevation}^2 \\ + \beta_3\mathrm{Precipitation} + \beta_4\mathrm{Precipitation}^2\end{aligned}$$

where $\beta = \{0{\cdot}5,2,-2,2,-1\}$ and both elevation and precipitation were standardized to have mean 0 and unit variance. We then sampled 1000 presence locations and discarded the absence locations. We estimated parameters using MAXENT and by maximum likelihood using our R package 'MAXLIKE'. Figure 3 illustrates the general bias of MAXENT predictions for estimating $\psi(x)$ and shows that the magnitude of the bias has a nonlinear form. For low values of $\psi(x)$, the index is biased high, whereas it is biased low when $\psi(x)$ is high. The fact that MAXENT's 'logistic output' is not proportional to $\psi(x)$ results from the program making a guess about species prevalence. If we were to change the default value to some other value, we would have found different results, and possibly higher degrees of bias. This nonconstant bias greatly limits the utility of the MAXENT output as an index.

NORTH AMERICAN BREEDING BIRD SURVEY

We applied the maximum likelihood estimator of occurrence probability to data from the North American Breeding Bird Survey (BBS) and compared the resulting estimates to predictions based on logistic regression and also to MAXENT's 'logistic output'. We fitted models to data on the Carolina wren (*Thryothorus ludovicianus*) using four land cover variables (per cent cover of mixed forest, deciduous forest, coniferous forest and grasslands) and latitude and longitude. We considered quadratic effects for each covariate. Fitting this model in MAXENT required that we modified the default settings, so that so-called hinge, threshold and product features were disabled. We restricted our analysis to the 2222 BBS routes surveyed in the United States during 2006, the year when the land cover variables were measured. Each BBS route is an approximately 40-km-long stretch of road consisting of 50 'stops' – points at which observers record counts of all bird species seen or heard during a 3-min period.

Traditional analyses of BBS data treat either the stop or the route as the sample unit; however, MAXENT requires data formatted as rasters (spatially referenced grids) and treats the pixel as the sample unit. Thus, we imposed a 25 km^2 grid over the study area, and for pixels with > 1 stop, we classified the pixel as being occupied ($y_i = 1$) if ≥ 1 detection was made at any of the stops in the pixel or unoccupied ($y_i = 0$) otherwise. Note that only logistic regression made use of the $y_i = 0$ data. Covariate values for each pixel in the United States were used as the 'background'.

Of the three methods, the logistic regression model makes the most use of the data and thus is expected to outperform the presence-only models. More generally, presence–absence data should always be preferred to presence-only data because observed zeros are informative about the species' range. Predicted probability maps from logistic regression have a clear interpretation as the probability that a pixel would yield

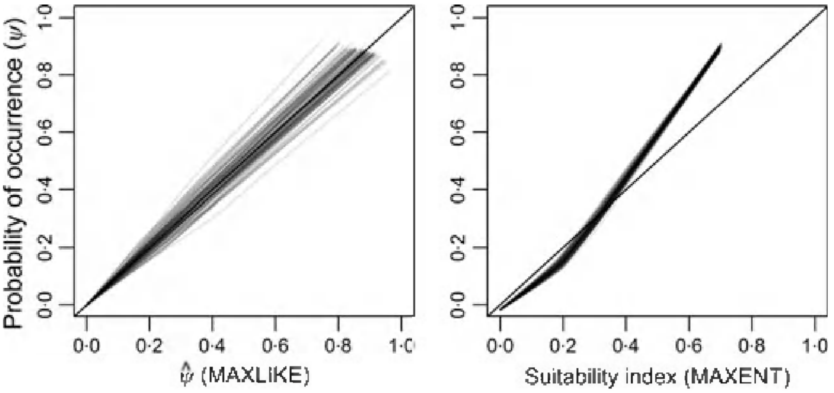

Fig. 3. Comparison between MAXLIKE and MAXENT estimates to true values of $\psi(y = 1|z(x))$. The grey lines represent the relationship between the estimate and the true value for each of 100 simulated data sets. The MAXENT index is not proportional to the probability of occurrence.

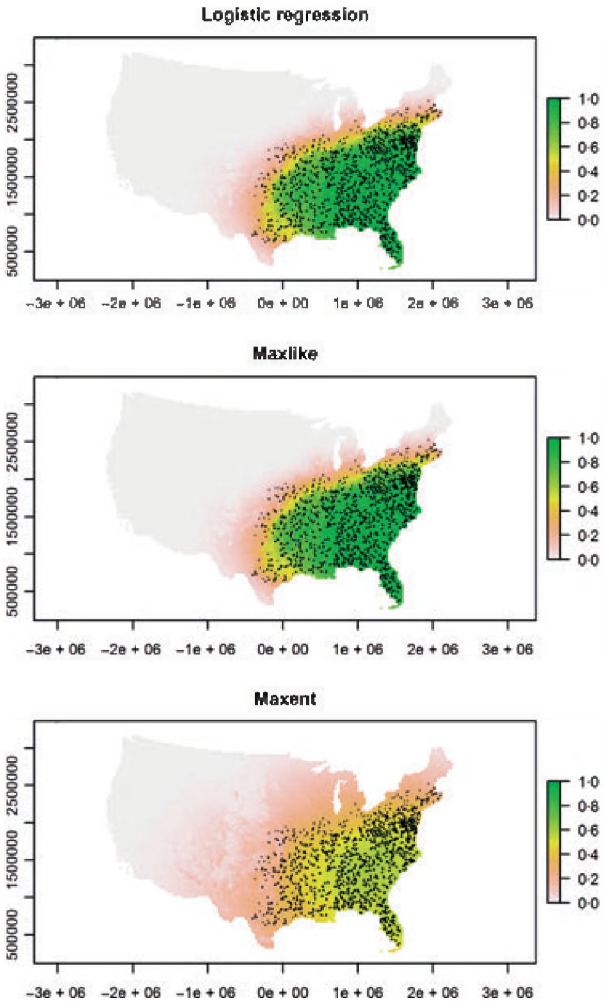

Fig. 4. Maps of the Carolina wren distribution generated using the three estimators applied to Breeding Bird Survey data.

an observation of the species in question. For these reasons, we considered the results of the logistic regression models as the standard against which we compared results of the other two estimators.

Maps of the Carolina wren distribution created using each of the 3 estimation methods are shown in Fig. 4. Salient points of the analysis are that MAXENT's logistic output is not as similar to the logistic regression predictions as those obtained by the maximum likelihood estimator and are generally inconsistent with the observed data in that sense that the resulting 'index' of species range is less defined and more geographically diffuse. From the maps, we see that MAXENT's 'logistic output' greatly underestimates the probability of occurrence throughout the core of the species' range and overestimates occurrence probability in regions where the species was never detected. The reason for this bias is the same as the bias in our simulation study – MAXENT uses a default intercept value that implies baseline prevalence of 0·50. Clearly, the MAXENT predictions will depend on this value which is *not estimated from the data using the maxent procedure*, and this subjectivity prohibits a clear interpretation of the index. Thus, while one might obtain more consistency using a different value of this parameter, there is no objective basis for setting this value.

Another important limitation of MAXENT for modelling these data relates to the difficulty of specifying the desired model of interest. For example, one cannot test for a specific interaction because the software requires that either all or none of the possible interactions are tested. Similarly, one cannot evaluate the possibility of a specific quadratic effect.

Discussion

Inference about occupancy from presence-only data has proved to be elusive. Rather than developing methods for direct inference about species occurrence, ecologists have settled for the production of ill-defined 'suitability indices' such as those issued by MAXENT. However, under random sampling, formal statistical inference about the probability of species occurrence can be achieved from presence-only data using conventional likelihood methods (Lancaster and Imbens 1996). We imagine that inference based on this likelihood should be accessible to practitioners familiar with ordinary statistical concepts.

Our simulation study using a standard logistic regression type of model indicates that occupancy probability is identifiable from presence-only data, consistent with what has been shown in related classes of models (Lancaster and Imbens 1996; Lele and Keim 2006). Some might argue that parametric assumptions are overly restrictive and, as a result, it is better to estimate something vaguely defined, which only might be proportional to occurrence probability. However, the most sensible and natural interpretation of the model underlying MAXENT is that it also assumes a parametric relationship between $\psi(y = 1|z)$ and covariates, one that is exponential. This is widely justified based solely on the incorrect assertion that the marginal probability of occurrence is not identifiable, and not based on any specific benefit of the exponential function. The lack of identifiability problem is specific to the parametric model that is implemented in MAXENT, and not a general feature of presence-only data. In our view, it does not make sense to forgo estimation of what is an eminently sensible quantity in the absence of any concrete technical or conceptual argument.

The ability to estimate occurrence probability parameters from presence-only data naturally requires larger sample sizes than from presence–absence data. Our analysis of data from the North American BBS provides sufficient data for this purpose, but such data may be unavailable in many studies. This was noted by Ward *et al.* (2009), who qualified their statement about identifiability by noting that

> Even when [...occurrence probability...] is identifiable, the estimate is highly variable.

While clearly the precision of estimates of model parameters in any specific application is a matter of sample size and complexity of the model, this does caution that sufficiently precise estimates might not be produced for all applications. We do not view this as a serious deterrent to seeking out estimates for parameters of models that are ecologically sensible independent of whether or not data are available to achieve a certain level of precision. Whether an estimator is 'highly variable' in a situation is not relevant if the estimand is the object of inference, and there are not competing (presumably more precise) estimators available for achieving that objective.

We emphasize that random sampling is critical because under this assumption, the marginal probability that a presence-only unit is included in the sample is constant, and thus, the sample inclusion probability does not affect $\pi(x|y = 1)$ constructed by invocation of Bayes rule. The issue of imperfect detection does not affect the development of the estimator, but it does affect interpretation of results. If detection probability, p, is constant spatially, then parameter estimators under the random sampling presence-only model are unaffected. Despite this, imperfect detection probability poses a number of complications, because it stands to reason that detection probability should be influenced by a whole host of things including nuisance effects such as 'effort' – which might include things such as human population or road density, such that what we obtain in many samples is a random sampling of presence-only sites from among those that are easy to access or sample. Also, detection probability might be related to ecological processes including population density of the species being studied, such that detection is more likely to occur at high-density sites and vice versa (Royle and Nichols 2003). In such cases, we expect to obtain a sample of presence-only sites that is biased toward high-density sites. Despite the importance of imperfect detection, it is possible to accommodate this in formal models for inference about species occurrence probability. So-called 'occupancy models' (MacKenzie *et al.* 2002; Tyre *et al.* 2003) generalize the logistic regression model to allow for imperfect detection, that is, false negatives, and some work has also been done to accommodate false positives within the occupancy modelling framework (Royle and Link 2006; Miller *et al.* 2011). In our view, application of any statistical procedure to presence-only data should acknowledge the core assumptions, seriously consider consequences of their violation on inferences and discuss them in the context of the specific study.

MAXENT is a popular software package for producing species distribution models, which is *not* based on a formal model for species occurrence. As such, the focus of MAXENT, as it is applied in practice, is not on formal inference about mechanisms responsible for observed species distribution pattern. We believe that it is not widely appreciated that direct inference about occurrence probability can be achieved using standard likelihood methods. We believe that the likelihood approach advanced in this paper offers a better framework for species distribution modelling because it allows users to estimate the fundamental parameter governing species distributions, the probability of occurrence. In this paper, we have also tried to provide context to certain technical facets of MAXENT that have important consequences. Principal among those are the implicit assumption equating the conditional probability $\Pr(x|y = 1)$ to the specific exponential function referred to as the 'maximum entropy distribution' and the implication that occupancy probability is modelled by an exponential model, thereby neglecting estimation of the intercept parameter and forsaking the ability to estimate occurrence probability. Second, we believe that many MAXENT users are unaware of the relevance of the penalty term that appears fundamental to MAXENT. To date, there has not been a formal justification of the need, importance or consequences of the penalty in the context of species distribution modelling, and there has been no mention of the possible bias introduced by this procedure. In our view, poorly motivated and justified technical elements of MAXENT distract from understanding the central inference problem of species distribution modelling.

Acknowledgements

We thank Peter Blank for providing habitat and weather data sets used in the BBS analysis, and members of PWRC's WMS Research Group.

References

Dudik, M., Phillips, S.J. & Schapire, R.E. (2004) Performance guarantees for regularized maximum entropy density estimation. *Proceedings of the Seventeenth Annual Conference on Computational Learning Theory*, pp. 655–662. ACM Press, New York.

Elith, J., Kearney, M. & Phillips, S.J. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 220–342.

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

Green, P.J. & Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC, New York.

Hosmer, D.W. & Lemeshow, S. (2000) *Applied Logisticregression*. Wiley-Interscience, New York.

Jaynes, E.T. (1957) Information theory and statistical mechanics. *Physical Review Series II*, **106**(4), 620–630.

Jaynes, E.T. (1963). Information theory and statistical mechanics. *Statistical Physics* (ed. K. Ford), p. 181. Benjamin, New York.

Jaynes, E.T. & Bretthorst, G.L. (2003) *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Kéry, M. (2011) Towards the modelling of true species distributions. *Journal of Biogeography*, **38**, 617–618.

Kéry, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

Lancaster, T. & Imbens, G. (1996) Case-control studies with contaminated controls. *Journal of Econometrics*, **71**, 145–160.

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

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, J.A, Royle, S. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation & Modeling: Inferring Patterns & Dynamics of Species Occurrence*. Academic Press, New York.

Manly, B.F., McDonald, L., Thomas, D.L., McDonald, T. L. & Erickson, W. (2002) *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*, 2nd edn. Kluwer Press, New York.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*. Chapman & Hall/CRC, New York.

Miller, D., Nichols, J.D., McClintock, B , Grant, E., Bailey, L., and Weir, L. (2011) Improving occupancy estimation when two types of observational error occur: non-detection & species misidentification. *Ecology*, **92**, 1422–1428, [10.1890/10-1396.1].

Phillips, S.J. & Dudik, M.. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips, S.J., Dudik, M. & Schapire, R.E. (2004) A entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, **69**, 83.

Phillips, S.J., Anderson, R.P., & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Royle, J.A. & Link, W.A. (2006) Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, **87**, 835–841.

Royle, J.A. & Nichols, J.D. (2003) Estimating abundance from repeated presence–absence data or point counts. *Ecology*, **84**, 777–790.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D. & Parris, K. & Possingham, H.P. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.

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

Received 1 August 2011; accepted 22 November 2011
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix. Random Sampling in Modeling Presence-only Data.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.