Presence-only data and the EM algorithm

Gill Ward

Department of Statistics, Stanford University, CA 94305, U.S.A.

email: gward@stanford.edu

and

Trevor Hastie

Department of Statistics, Stanford University, CA 94305, U.S.A.

email: hastie@stanford.edu

and

Simon Barry

Australian Government Department of Agriculture, Fisheries and Forestry,

GPO Box 858, Canberra ACT 2601, Australia

and

Jane Elith

School of Botany, The University of Melbourne, Parkville, Victoria, Australia

and

John R. Leathwick

National Institute of Water and Atmospheric Research, P O Box 11115, Hamilton, New Zealand

SUMMARY:

In ecological modeling of the habitat of a species, it can be prohibitively expensive to determine species absence. Presence-only data consists of a sample of locations with observed presences and a separate group of locations sampled from the full landscape, with unknown presences. We propose an EM algorithm to estimate the underlying presence-absence logistic model for presence-only data. This algorithm can be used with any off-the-shelf logistic model. For models with stepwise fitting procedures, such as boosted trees, the fitting process can be accelerated by interleaving expectation

steps within the procedure. Preliminary analyses based on sampling from presence-absence records of fish in New Zealand rivers illustrates that this new procedure can reduce both deviance and the shrinkage of marginal effect estimates that occur in the naive model often used in practice. Finally, it is shown that the population prevalence of a species is only identifiable when there is some unrealistic constraint on the structure of the logistic model. In practice, it is strongly recommended that an estimate of population prevalence be provided.

KEY WORDS: Presence-only data; Use-availability data; EM algorithm; Logistic model; Boosted trees

1. Introduction

Modeling wildlife habitat selection is important for effective ecological management of a species. Logistic models are typically used to estimate species distribution, ideally based on records of presences or absences collected at randomly sampled locations. However, obtaining such *presence-absence data* may be difficult or expensive, and often records of species presence are easier to obtain than definitive absences. For example, herbaria and museums have extensive historical occurrence records (Elith and Graham *et al.*, 2006) and radiotelemetry provides a rich source of range locations for mobile species (Frair, Nielsen, Merrill, Lele, Boyce, Munro, Stenhouse and Beyer, 2004). Environmental information can be collected for these recorded presence locations using geographical information system (GIS) technology.

Some methods use only these presences to estimate a species range; for example HABITAT (Walker and Cocks, 1991) estimates the range of a species via a convex hull defined in environmental space. However, many methods also require a *background sample* or *pseudo-absences* - a random sample of locations taken from the region or *landscape* of interest. Although the presence or absence of a species at the background locations is unknown, they provide a summary of the landscape, against which we can compare the observed presences. Indeed, using these data can considerably increase prediction accuracy (Elith and Graham *et al.*, 2006). This combination of a sample of locations with observed presences, and a background sample of locations from the landscape is what we will refer to as *presence-only* data.

The model construction in this paper is based on strict assumptions about the sampling mechanisms. In particular, we assume that the observed presences in the presence-only sample are taken at random from all locations where the species is present. For mobile species, where presence or absence may change over time, we assume that this random

sampling is at a rate proportional to the proportion of time the species is present at that location. Additionally, we assume that the background sample is sampled at random from the full landscape of locations. In practice, this second assumption is often only approximately true; GIS provides an easy way to generate environmental covariates for locations chosen at random. However, the presence-only sample is often biased. Records at herbaria or museums are typically ad-hoc records of species occurrence collected by individuals, and may be biased, for example, towards more accessible locations (Reese, Wilson, Hoeting and Flather, 2005).

Such bias in observed presences may be counteracted by sampling the background data with the same bias. For example, ecologists have used for background data the presence records for different species in the same region (e.g. Dudík, Schapire and Phillips, 2006). Taking this a step further, Zaniewski, Lehmann and Overton (2002) use the recorded presences of other species to build a statistical model of sampling rate given environmental covariates. Predicted probabilities for this sampling bias model are then used as weights in sampling the background data. These same techniques can be used with the EM procedure described in this paper. However, a more thorough treatment of this bias is beyond the scope of this paper.

The modeling of presence-only data in ecology is reviewed in Keating and Cherry (2004), where it is referred to as use-availability data, and Pearce and Boyce (2006). Typically a logistic model is fit to the recorded presences and background data; we refer to this as the *naive model*. If a species is rare, the background data will resemble the true absences and the naive model will be close to the true model. However, for more common species, the naive model can be highly biased (Keating and Cherry, 2004); we explore the form of that bias later in this paper. An alternate procedure, the exponential model of Manly et al. (2002), uses a log-linear model of the presence-absence probability which is easy

to estimate via existing modeling procedures. However, this approach does not attempt to estimate the true presence-absence logistic model, and has other recorded failings (Keating and Cherry, 2004). An ad-hoc method of Engler, Guisan and Rechsteiner (2004) uses an initial model to discover background locations where there is a low predicted probability of species presence. These locations are then used as absences in a second presence-absence logistic model. In essence this is the aim of our procedure — to label background data as true presences or absences — however, we approach this in a more rigorous manner.

Lancaster and Imbens (1996) provide an exact method for estimating a logistic regression model with presence-only data. It uses direct maximization of a likelihood based on the presence-only data to estimate the linear predictor for the presence-absence logistic regression model. However, this technique is rarely used in practice as there is no easily-available implementation and convergence problems have been reported. Our method indirectly maximizes this same presence-only likelihood in a way that is more robust, can easily incorporate more flexible models and is straightforward to implement using existing statistical software.

Typical presence-absence methods attempt to model the probability of presence conditional on environmental covariates. In contrast, Maxent (maximum entropy modeling, Phillips, Anderson and Schapire, 2006) estimates the density of environmental covariates conditional on species presence. This has been shown to be equivalent to a Gibbs density with certain constraints, which implies a log probability model as used by Manly. Although Maxent has been shown to perform well in practice (Elith and Graham *et al.*, 2006), in this paper we concentrate instead on improving the logistic regression model.

Within ecology the presence-only problem arises in paradigms with subtle distinctions. In "use-availability" modeling, we are interested in the relative frequency of use of an

area by wide-ranging animals; for less mobile species, "presence-absence" is modeled. However, for both these cases it is possible to arrive at a unique definition. We define the probability of presence of a species as measured across a certain time window. For non-mobile species, this time window is irrelevant, up to the lifetime of the species. Typically, for mobile species, the window is implicitly defined by the sampling mechanism, for example by the length of time a site was observed, or by the discretization of radiotelemetry data. Any reference to a probability of presence in this paper thus implicitly refers to such a time window.

2. The presence-only problem and the EM algorithm

A common aim of ecological studies is to model the probability that a species of interest is present, $y = 1$ (vs. absent, $y = 0$), conditional on some ecological covariates $\mathbf{x}$, $\mathbb{P}(y = 1|\mathbf{x})$. This probability is usually modeled via its logit

$$\text{logit}\, \mathbb{P}(y = 1|\mathbf{x}) = \eta(\mathbf{x}) \;\; \Rightarrow \;\; \mathbb{P}(y = 1|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1 + e^{\eta(\mathbf{x})}} \tag{1}$$

where $\eta(\mathbf{x})$ can be linear in $\mathbf{x}$ (as in logistic regression) or a non-linear function of $\mathbf{x}$ e.g. GAMs (Hastie and Tibshirani, 1990) or boosted trees (Friedman, 2001). In studies where the *true* presences and absences of a species are known for a random sample of locations, these models are fit using established methods. We will refer to these as *presence-absence* data and methods.

However, in the presence-only problem, we know only where the species was present, not where it was absent. Along with these *observed presences* we can generate *background data*, which are locations where the true presence or absence is unknown. Ideally this background data set should be larger than the number of observed presences and large enough to provide a good representation of the landscape. We denote these observed presences and background data by $z = 1$ and $z = 0$ respectively. Note that when $z = 1$,

we know $y = 1$. However, when $z = 0$ we do not know whether y is 0 or 1. For both our observed presences and background data, we have measurements of environmental variables or covariates $\mathbf{x}$.

Using this notation, we can clearly state the presence only problem. We wish to estimate $\mathbb{P}(y = 1|\mathbf{x})$, the probability of a true presence y, given covariates $\mathbf{x}$. However, the observed presences and background data are generated by $\mathbb{P}(\mathbf{x}|z = 1)$ and $\mathbb{P}(\mathbf{x}) = \mathbb{P}(x|z = 0)$ respectively. Using a case-control approach (McCullagh and Nelder, 1989, p111), we can turn these probabilities around in order to obtain $\mathbb{P}(z = 1|\mathbf{x}, s = 1)$, the conditional probability of an observed presence $z = 1$. The notation $s = 1$ is a construct of case-control modeling and indicates that this observation is in our presence-only data sample.

What we refer to as the *naive model* attempts to fit the logistic model (1) directly to the observed z:

$$\text{logit}\, \mathbb{P}(z = 1|s = 1, \mathbf{x}) = \eta_{\text{naive}}(\mathbf{x}). \tag{2}$$

The primary difference between the presence-absence (1) and naive presence-only (2) models is that the background data $z = 0$ include some true presences $y = 1$ as well as true absences $y = 0$. This creates bias in the naive presence-only model, which we illustrate through a simple example in Section 2.2 and further investigate in Section 3.2.

The aim of this paper is to estimate the presence-absence model (1) using presence-only data. We start by giving an overview of our new modeling procedure (Section 2.1), then fill in the details later (Section 3). We also provide a simple example which illustrates the naive and new models, and the differences between them (Section 2.2). Throughout these sections we assume that we know the overall population prevalence of the species $\pi = \mathbb{P}(y = 1)$. Later we look at an example of sensitivity analysis of the uncertainty in π (Section 4), and illustrate under what conditions we can also estimate π directly from these data (Section 5).

2.1 *The EM procedure*

In fitting a model to the presence-only data, we consider two possibilities: where the background data $z = 0$ are assumed to have either known or unknown y:

(1) If we knew the true y for the background data $z = 0$, then we could estimate the presence-only logistic model (1) by fitting a model directly to the y given $\mathbf{x}$. As the true $y = 0, 1$ in these data occur in different proportions to the overall landscape, we also make a simple "case-control" adjustment to the intercept of the fitted model.

(2) As we do not know the true y for the background data $z = 0$, we can think of these y as missing data. We use an iterative procedure that tries to estimate or impute the unknown y at each iteration and then fits a model using these imputed y. At each iteration we apply the following two steps, until subsequent iterations result in the same model:

   (a) We replace the unknown ys with our best estimate $\hat{y} = \hat{\mathbb{P}}(y = 1|\mathbf{x})$ estimated from the model fit in the previous iteration.

   (b) We assume that we "know" these ys and thus apply the procedure described in (1).

The procedure in (2) is an implementation of the Expectation-Maximization (EM) algorithm of Dempster, Laird and Rubin (1977).

[Figure 1 about here.]

We provide a more mathematical description of the algorithm in Figure 1. The initial estimate of y for $z = 0$ is $\pi = \mathbb{P}(y = 1)$ because this is our best guess prior to fitting any model. We provide the exact form for the intercept adjustment of the fitted models, where n_p is the number of observed presences ($z = 1$) and n_u is the number of background data ($z = 0$). For logistic procedures used in the maximization step that cannot handle non-

integer responses, a work-around is illustrated in Web Appendix B. Details and derivation of this algorithm are given in Section 3

2.2 *A simple example*

[Figure 2 about here.]

To illustrate the bias in the naive model and the mechanism of the EM procedure, we present a simple example where the probability of true presence y depends only on elevation:

$$\text{logit}\left(\mathbb{P}(y=1|\mathbf{x})\right) = \beta_0 + \beta_1 \times \text{ elevation.}$$

We take β_1 to be positive, so the species prefers higher elevations, and set $\pi = 0.4$. There are 10 data points with $z = 1$ and 20 data points with $z = 0$. If we knew the true y (Figure 2a) we could obtain a good estimate of the underlying logistic regression model. However, if we fit the same model to the observed z (Figure 2b) then the slope $\hat{\beta}_1$ of the estimated model is too shallow.

[Figure 3 about here.]

The EM algorithm for this example uses a logistic regression fit at each iteration; the outcomes for these fits are 1 for $z = 1$ but for $z = 0$ they depend on the fitted model from the previous iteration. In the first iteration, the outcome is set to π for each $z = 0$ (Figure 3a). In the second iteration (Figure 3b) outcomes for $z = 0$ are set to the predicted values from the first iteration. This is repeated for subsequent iterations and the EM algorithm terminates when the models estimated in the previous and current iterations are the same (Figure 3d). In this example, the resulting fitted model is very close to that obtained using the true y (Figure 3a).

3. Details and development of the algorithm

To estimate η for the presence-absence model (1) we wish to maximize the likelihood for the presence-only data, with respect to η. We first derive the form of this *observed likelihood*, which is based on $\mathbb{P}(z|s=1,\mathbf{x})$. We can then derive the steps of the EM algorithm applied to this likelihood. In order to do this we first calculate a slightly different *full likelihood* based on $\mathbb{P}(y,z|s=1,\mathbf{x})$. We provide the derivation of the observed likelihood in this section, as we believe it can help the reader understand the structure of this presence-only problem. All other propositions are proved in Web Appendix A.

PROPOSITION 1: Given the usual logistic model (1), we can use a case-control style adjustment to show

$$\text{logit}\left(\mathbb{P}(y=1|\mathbf{x},s=1)\right)=\eta(\mathbf{x})+\log\left(\frac{n_p+\pi n_u}{\pi n_u}\right) \tag{3}$$

where n_p and n_u are the number of observed presences $z=1$ and background data $z=0$ respectively.

PROPOSITION 2: The *observed likelihood* for the presence-only data is given by:

$$\begin{aligned} L(\eta|\mathbf{z},X)&=\prod_i \mathbb{P}(z_i|s_i=1,\mathbf{x}_i)\\ &=\prod_i\left(\frac{\frac{n_p}{\pi n_u}e^{\eta(\mathbf{x}_i)}}{1+\left(1+\frac{n_p}{\pi n_u}\right)e^{\eta(\mathbf{x}_i)}}\right)^{z_i}\left(\frac{1+e^{\eta(\mathbf{x}_i)}}{1+\left(1+\frac{n_p}{\pi n_u}\right)e^{\eta(\mathbf{x}_i)}}\right)^{1-z_i}. \end{aligned} \tag{4}$$

Proof. We will prove the form of the probability $\mathbb{P}(z=1|s=1,\mathbf{x})$; the construction of the observed likelihood from these probabilities follows immediately. We start with a total-probability argument across $y=0,1$:

$$\begin{aligned} \mathbb{P}(z=1|s=1,\mathbf{x})=&\mathbb{P}(z=1|y=1,s=1,\mathbf{x})\mathbb{P}(y=1|s=1,\mathbf{x})\\ &+\mathbb{P}(z=1|y=0,s=1,\mathbf{x})\mathbb{P}(y=0|s=1,\mathbf{x}). \end{aligned} \tag{5}$$

Note that $\mathbb{P}(z=1|y=1,s=1,\mathbf{x})$ is a sampling probability of whether true presences $(y=1)$ in our presence-only data were observed $(z=1)$ or not $(z=0)$. The random

sampling assumptions outlined in the introduction imply that the sampling of the observed presences is independent of $\mathbf{x}$, given the true presence or absence. Thus we use an application of Bayes rule to show

$$\mathbb{P}(z=1|y=1,s=1,\mathbf{x}) = \mathbb{P}(z=1|y=1,s=1) = \frac{\mathbb{P}(z=1,y=1|s=1)}{\mathbb{P}(y=1|s=1)} \tag{6}$$

The expected number of true presences ($y=1$) in our data is $n_p + \pi n_u$ — all data with $z=1$ plus a proportion π of $z=0$ — and so $\mathbb{P}(y=1|s=1) = (n_p+\pi n_u)/(n_p+n_u)$. Also, by definition of z and y, $\mathbb{P}(z=1,y=1|s=1) = \mathbb{P}(z=1|s=1) = n_p/(n_p+n_u)$. Plugging these into (6) we get

$$\mathbb{P}(z=1|y=1,s=1,\mathbf{x}) = \frac{n_p}{n_p+\pi n_u} \tag{7}$$

Further, $\mathbb{P}(z=1|y=0,s=1)=0$ because all $y=0$ in the data must occur for $z=0$.

From Proposition 1, we know $\mathbb{P}(y=1|s=1,\mathbf{x}) = e^{\eta^\star(\mathbf{x})}/(1+e^{\eta^\star(\mathbf{x})})$ where $\eta^\star(\mathbf{x}) = \eta(\mathbf{x}) + \log\big((n_p+\pi n_u)/(\pi n_u)\big)$. Substituting all the probabilities in (5) we have

$$\mathbb{P}(z=1|s=1,\mathbf{x}) = \frac{n_p}{n_p+\pi n_u}\,\frac{e^{\eta^*(\mathbf{x})}}{1+e^{\eta^*(\mathbf{x})}} + 0 \tag{8}$$

$$= \frac{\frac{n_p}{\pi n_u} e^{\eta(\mathbf{x})}}{1+\left(1+\frac{n_p}{\pi n_u}\right)e^{\eta(\mathbf{x})}} \tag{9}$$

after some manipulation. ∎

Direct maximization of this likelihood is difficult, though Lancaster and Imbens (1996) have implemented this for η linear in $\mathbf{x}$. We continue by providing the full likelihood, which also depends on the unknown y.

PROPOSITION 3: The *full likelihood* for the presence-only data, in terms of both z and y is:

$$\begin{aligned} L(\eta|\mathbf{y},\mathbf{z},X) &\propto \prod_i \mathbb{P}(y_i|s_i=1,\mathbf{x}_i) \\ &= \prod_i \left(\frac{e^{\eta^\star(\mathbf{x}_i)}}{1+e^{\eta^\star(\mathbf{x}_i)}}\right)^{y_i} \left(\frac{1}{1+e^{\eta^\star(\mathbf{x}_i)}}\right)^{1-y_i} \end{aligned} \tag{10}$$

where $\eta^\star(\mathbf{x}_i) = \eta(\mathbf{x}) + \log\big((n_p+\pi n_u)/(\pi n_u)\big)$.

Note that in (4) and (10) we assume that π is known. Note also that with y known, the full likelihood (10) for η does not depend on z. In the next proposition we use this full likelihood in applying the Expectation-Maximization algorithm (EM algorithm, Dempster, Laird and Rubin, 1977) to the observed likelihood.

PROPOSITION 4: The observed likelihood (4) can be maximized using the EM algorithm. The algorithm uses alternate maximization and expectation steps as outlined in Figure 1. In the E step, the unknown ys are estimated using the model from the previous fit. The M step involves maximization of the full likelihood (10) with the estimated ys imputed.

Thus we have shown that the EM algorithm for presence-only data (Figure 1) maximizes the observed likelihood (4). In other words, this new approach is estimating the underlying logistic model $\eta(\mathbf{x})$ of the probability of interest $\mathbb{P}(y = 1|\mathbf{x})$.

3.1 *Stepwise procedures and the EM method*

Recent work in ecological modeling has seen a move toward more flexible models, such as generalized additive models (GAM) and boosted trees (e.g. Leathwick, Rowe, Richardson, Elith and Hastie, 2005; Leathwick, Elith, Francis, Hastie and Taylor, 2006). Both GAM and boosted trees can be the logistic model of choice in the maximization step of this EM algorithm (Figure 1). However the stepwise nature of the boosted trees model suggests an improvement; interleaving expectation steps with every few boosting steps will decrease the time to convergence of the EM algorithm. In practice this decrease can be an order of magnitude. In Web Appendix C we show how to implement this interleaving using the boosted trees package GBM in R (Ridgeway, 2004). Although we use boosted trees as an example, this could be implemented for any stepwise procedure.

3.2 *The Naive Logistic Model*

In recent literature (e.g. Ferrier, Drielsma, Manion and Watson, 2002) one approach to the presence-only problem has been to fit a logistic model directly to the observed presences and background data. We refer to this as a naive logistic model and use η_{naive} to refer to the linear predictor of this model, as defined in (9).

The naive logistic model is doubly problematic. Assuming that the background data are all absences ignores the "contamination" with true but unknown presences. Further, the true (but in part unknown) presences and absences are not sampled proportionally to their prevalence in the landscape. We can improve the naive model by using a case-control adjustment for η_{naive} to account for the differing sampling rates (e.g. Mace, Waller, Manley, Lyon and Zuuring, 1996). We then show how this adjusted η_{naive} is related to the true η. All proofs for these propositions are provided in Web Appendix A.

PROPOSITION 5: The case control adjusted naive model $\eta^{\text{adj}}_{\text{naive}}(\mathbf{x})$ is

$$\eta^{\text{adj}}_{\text{naive}}(\mathbf{x}) = \eta_{\text{naive}}(\mathbf{x}) - \log\left(\frac{(1-\pi)n_p}{\pi n_u}\right).$$

PROPOSITION 6: The case control adjusted naive model $\eta^{\text{adj}}_{\text{naive}}(\mathbf{x})$ can be written in terms of the presence-absence model $\eta(\mathbf{x})$ as follows:

$$\eta^{\text{adj}}_{\text{naive}}(\mathbf{x}) = -\log(1-\pi) - \log\left(e^{-\eta(\mathbf{x})}+1\right). \tag{11}$$

[Figure 4 about here.]

Note that when π is small, most η are large and negative, thus $-\log\left(e^{-\eta(\mathbf{x})}+1\right) \approx \eta(\mathbf{x})$ and the naive model is similar to the true model. More generally, the adjusted naive model $\eta^{\text{adj}}_{\text{naive}}$ is increasing in the true η, but nonlinearly (Figure 4). Although the true and naive predictors are similar up to a constant for $\eta \ll 0$, when $\eta > 0$ the naive model considerably underestimates the rate at which η is increasing. In particular, the naive linear predictor

is bounded above: $\eta_{\text{naive}} \leqslant -\log(1-\pi)$. Hence the estimated probabilities for the naive model will be underestimated for locations with higher probabilities of presence.

[Figure 5 about here.]

In practice, this bias has considerable effect on logistic regression estimates: if the true η is linear in $\mathbf{x}$, i.e. $\eta(\mathbf{x}) = \mathbf{x}^T\beta$, then any naive logistic regression model must be biased. In particular, the estimates for the naive model β_{naive} will tend to underestimate the slopes β. This is illustrated in simulations of presence-only data (Figure 5). Although the estimates for the EM procedure are approximately unbiased, the naive model estimates are shrunk to zero, with more shrinkage for larger π. However, note that even for $\pi = 0.1$, the naive estimates are noticeably different from the truth. We will see that this under-estimation of effect size also occurs for more complex models (Figure 6). This shrinkage of effect size occurs because of the contamination of true presences $y = 1$ in the background data $z = 0$. Because of this contamination, the $z = 1$ and $z = 0$ are more similar to each other than the $y = 1$ and $y = 0$. It is easy to see how this shrinkage may impact variable selection, e.g. important variables may not appear statistically significant.

For flexible modeling procedures it may seem tempting to use the relationship between the true and naive models (11) to make a post-hoc transformation of the naive model. However, from (11) we see there is an implicit constraint of $\eta^{\text{adj}}_{\text{naive}} < -\log(1-\pi)$; exceeding this value results in undefined η. In practice this constraint is not observed, nor is there an easy way to enforce this, and the transformed naive model may be undefined for many locations in the landscape.

4. Example: The Longfin Eel

Assessing models based on presence-only data is difficult, because there is typically no validation data with true presences and absences. Although simulated data are a useful

tool, they typically do not reflect the noise and complex structure inherent in ecological data. To overcome these issues, we have generated presence-only data sets from presence-absence data of diadromous fish in the rivers of New Zealand (more details are given in Leathwick, Rowe, Richardson, Elith and Hastie, 2005, and in the acknowledgements). In particular we looked at the Longfin Eel *Anguilla dieffenbachii* which has a high prevalence, occurring at 51.3% of all locations sampled. To reduce spurious effects, we repeated the presence-only data generation 10 times, with different random seeds; the results provided are amalgamated across these repetitions. There are 21 environmental covariates describing conditions at the sampled site as well as up- and downstream. This includes some variables omitted from Leathwick, Rowe, Richardson, Elith and Hastie (2005), in particular the presence of a dam downstream, as this provides a clear illustration of the difference in the naive model and EM procedure. In the full data set, eels were present in 20% of locations with a dam downstream.

[Figure 6 about here.]

The presence-only samples were generated according to the sampling assumptions set out in this paper. Ideally the background data should be a random sample from all river locations in New Zealand; here we assume that the fish data set consists of such a sample, so we can compare performances of different models. Hence the sample of background data was generated by sampling randomly from all available data. Then the naive presences sample was generated by sampling from the remaining presence locations. A random sample of one quarter of the data were set aside as a validation set and the rest were used to train the models. These training sets contained around 4,400 and 1,300 background data and observed presences respectively. The validation set contained one third of these numbers, of which around 1,200 were true presences and 700 were true absences.

We fitted a boosted trees model to these data, using the EM procedure (Figure 1), which we will call the *EM model*, with $\pi = 0.513$. This is compared to the *naive model*, fitting boosted trees directly to the presence-only data with a case-control adjustment to enable a fair comparison with the EM model. Additionally, as a measure of an optimal result, we calculated the *full model*, a boosted trees model calculated on the same data, but using the full knowledge of the true presences and absences. A total of 10,000 trees were fitted for each model, each with a depth of 4, allowing four-way interaction terms. In practice the optimal number of trees for a model would be chosen by minimizing a prediction set deviance using the observed likelihood (4). However, here the deviance was calculated using the full likelihood (10) of the true presences, so that the EM and naive models could be compared with the full model. As the validation set was not a random sample from all locations, a further case-control adjustment was made in calculating deviances. In general, the full and observed likelihoods generate similar shaped curves that are minimized at a similar number of trees. The naive model required the fewest trees (Figure 7), while the EM models were optimized at around 3,000 trees, similar to the optimal number of trees for the model based on the true presences.

[Figure 7 about here.]

The two most important predictors in modeling the true presences were the summer temperature and the presence of a dam downstream. Figure 6 illustrates that the naive model tends to underestimate the range of the effect that each of these predictors has on η. This is particularly noticeable in the binary downstream dam variable, and echoes the pattern in Figure 5 that the naive logistic regression model shrinks the parameter estimates toward zero. This leads to the effect seen in Figure 7, that the predicted η for the naive model tend to be shrunk to zero in comparison to the model based on the true presences.

The performance of the EM model lies somewhere between that of the naive and true models. There is some shrinkage of the marginal effects and estimated ηs, though it is not as pronounced as for the naive model. This occurs because the presence probabilities imputed at the E-steps are estimates of the true model, thus introducing an averaging effect on the estimated presences/absences for the background data. Figure 7 illustrates that the EM model also has lower deviance than the naive model (calculated on a validation set where the true presences and absences are known). In this case, with $\pi = 0.513$, the EM model considerably reduces the excess in deviance over the model based on the true presences, compared to the naive model. Although these results are exploratory, these effects were seen across other species of fish with $\pi > 0.2$, but with less of an effect for smaller π. It should be noted, however, that these results were for data that were sampled according to the assumptions set out in this paper; in practice, sampling of presences may be very ad-hoc.

[Figure 8 about here.]

Finally, we ran a sensitivity analysis of the model, for one of the 10 sampling repetitions, to determine the effect of differing prior beliefs of π. The EM model was fitted using a range of different πs between 0.3 and 0.8, in increments of 0.05. The validation set deviance was calculated for each model, using the observed likelihood (4), and the minimum deviance recorded. Figure 8(a) illustrates that these minimum deviances are themselves minimized around $\pi = 0.6$, with low minimum deviances for values of π between 0.5 and 0.7. Thus there is only a small change in predictive ability when increasing π slightly. However, if the prior belief was that the true π was smaller than 0.513, then the choice of π would influence the prediction deviance considerably. It should be noted here that although Figure 8(a) suggests we could estimate π as that which minimizes the

deviance, this would be incorrect; in the next section we show that π is not estimable when fitting a boosted trees model.

The marginal effects on η of all variables follow a pattern similar to that in Figure 8(b). Across different values of π, the shape of the marginal effect is relatively constant, with the magnitude of the effect increasing with π. Unsurprisingly, as π gets smaller the effect size tends to resemble that of the naive model (not shown), as we are assuming that there are few presences in the background data. Thus, in our example, the shapes of the marginal effects are not sensitive to changes in π, but the magnitudes of the effects are.

5. Estimating π

Ideally we would like π to be *identifiable*; in other words, that we can estimate π, as well as η, from presence-only data. However, in the following proposition we show that this is not feasible in practice. Proofs of the first two parts of this proposition are given in Web Appendix A; the third part is illustrated using a simple example.

PROPOSITION 7: Identifiability of π can be summarized as follows:

(a) π is *not* identifiable if we make no assumptions about the structure of η.

(b) π is identifiable only if we make unrealistic assumptions about the structure of $\eta(\mathbf{x})$ such as in logistic regression where $\eta(\mathbf{x})$ linear in $\mathbf{x}$: $\eta(\mathbf{x}) = \mathbf{x}^T\beta$.

(c) Even when π is identifiable, the estimate is highly variable.

[Figure 9 about here.]

We illustrate the high variability in estimated π using an example where data are simulated from a simple logistic model (Figure 9). (Details of this estimation procedure are given in Web Appendix D.) Even in this simple example, where the assumed structure of η is correct (and linear), the 95% confidence interval for π is large (0.15 to 0.75) and

π is highly correlated with the intercept α. Where the true model deviates from the assumed structure, the estimates are highly unstable; Lancaster and Imbens (1996) report failure of convergence of the direct maximization of the observed likelihood in several examples. These results strongly contra-indicate estimating π from presence-only data in any situation.

6. Conclusions

We have proposed an application of the EM algorithm that provides a flexible method of estimating the underlying presence-absence model from presence-only data. It can work as a wrapper to (almost) any logistic regression modeling procedure. At each iteration the probabilities of presence are calculated, given the current model, for the background data (E-step) and the case-control logistic model is re-estimated using these new weighted or non-binary outcomes (M-step). For stagewise logistic modeling procedures, such as boosted trees, E-steps may be interleaved within the procedure to save computation time.

This EM model gives approximately unbiased estimates in a simple linear logistic regression simulation study. In comparison, there is considerable shrinkage toward zero in the estimates from the naive logistic regression model, fitted directly to the presence-only data. In our more complex example, a boosted trees EM model outperforms the naive boosted trees model; there is less shrinkage of the marginal effects and the prediction deviance for the presence-absence validation set is smaller for the EM model. Unsurprisingly, the EM model still has higher prediction deviance than the boosted trees model fitted using the true presence-absence data.

Previous work in estimating the presence-absence model from presence-only data has attempted to simultaneously estimate the population prevalence π. However, we have shown that π is not identifiable when no assumptions are made about the structure of η, such as in boosted trees models. In addition, even when unrealistic assumptions are

made about the structure of η, e.g. linear, the resulting estimate of π is highly variable and heavily dependent on that assumption. We recommend obtaining an estimate of π from some other source and using sensitivity analysis to assess the dependence of the results on this estimate. If no estimate of π is available, then the naive model is the only logistic model available.

Alternatively, a Bayesian approach could be considered in situations when π is not precisely known. As π is not realistically identifiable in the presence-only framework, an informative prior should be used to summarize the strength of the knowledge of π. Standard errors of the fitted model resulting from such an analysis would be larger than in the EM approach described in this paper; this increase would reflect the uncertainty inherent in the prior for π.

This application of the EM algorithm provides a flexible way of estimating species distribution from the extensive records of species presence in herbaria and museums, and from radiotelemetry studies. Because of its simplicity, we believe it can be easily adopted by anyone working in this field.

Supplementary Material

The Web Appendices referenced in Sections 2.1 and 5 is available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowledgements

Fish distribution data were taken from the New Zealand Freshwater Fish Database, which is curated by New Zealand's National Institute of Freshwater and Atmospheric Research.

The authors are very grateful to Holger Höfling for identifying the naive likelihood as the density of a curved exponential family. We would also like to thank Rob Tibshirani,

Art Owen and the Hastie-Tibshirani Research Group of Stanford University for their helpful feedback on this work.

References

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

Dudík, M., Schapire, R., and Phillips, S. (2006). Correcting sample selection bias in maximum entropy density estimation. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 323–330. MIT Press, Cambridge, MA.

Elith, J. and Graham *et al.*, C. H. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29,** 129–151.

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

Ferrier, S., Drielsma, M., Manion, G., and Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. ii. community-level modelling. *Biodiversity and Conservation* **11,** 2309–2338.

Frair, J. L., Nielsen, S. E., Merrill, E. H., Lele, S. R., Boyce, M. S., Munro, R. H. M., Stenhouse, G. B., and Beyer, H. L. (2004). Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology* **41,** 210–212.

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

Hastie, T. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

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

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

Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., and Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series.* .

Leathwick, J. R., Rowe, D., Richardson, J., Elith, J., and Hastie, T. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology* **50,** 2034–2051.

Mace, R. D., Waller, J. S., Manley, T. L., Lyon, L. J., and Zuuring, H. (1996). Relationships among grizzly bears, roads and habitat in the Swan Mountains, Montana. *The Journal of Applied Ecology* **33,** 1395–1404.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L., and Erickson, W. (2002). *Resource selection by animals: statistical design and analysis for field studies*. Chapman & Hall, New York, second edition.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

Pearce, J. L. and Boyce, M. S. (2006). Modeling distribution and abundance with presence-only data. *Journal of Applied Ecology* **43,** 405–412.

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

Reese, G. C., Wilson, K. R., Hoeting, J. A., and Flather, C. H. (2005). Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications* **15,** 554–564.

Ridgeway, G. (2004). gbm: Generalized boosted regression models. R package, version 1.3-5. `http://www.i-pensieri.com/gregr/gbm.shtml`.

Walker, P. A. and Cocks, K. D. (1991). HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters* **1,** 108–118.

Zaniewski, A. E., Lehmann, A., and Overton, J. M. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* **157,** 261–280.

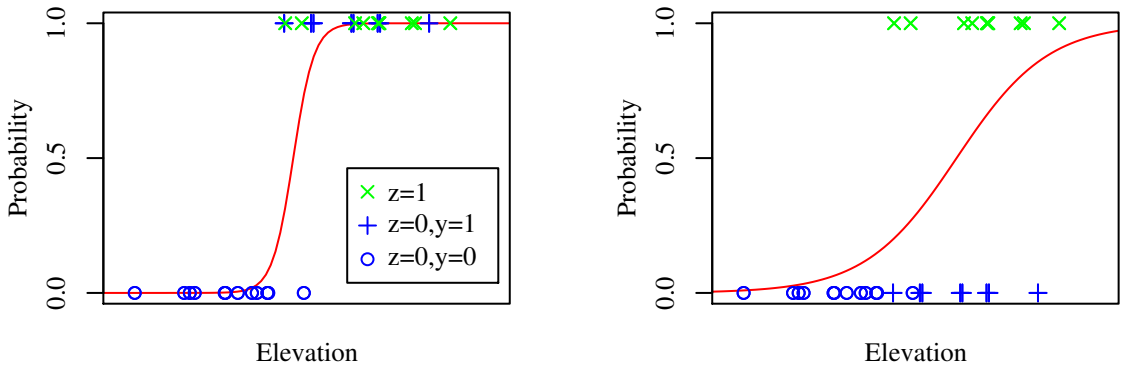
(1) Chose initial estimates: $\hat{y}_i^{(0)} = \pi \quad \text{for} \quad z_i = 0.$

(2) Repeat until convergence:

- *Maximization step:*
  - Calculate $\hat{\eta}^{*(k)}$ by fitting a logistic model of $\hat{\mathbf{y}}^{(k-1)}$ given X.
  - Calculate $\hat{\eta}^{(k)} = \hat{\eta}^{*(k)} - \log\left(\frac{n_p + \pi n_u}{\pi n_u}\right)$.
- *Expectation step:*

$$\hat{y}_i^{(k)} = \frac{e^{\hat{\eta}^{(k)}}}{1 + e^{\hat{\eta}^{(k)}}} \quad \text{for } z_i = 0 \qquad \text{and} \qquad \hat{y}_i^{(k)} = 1 \quad \text{for } z_i = 1$$

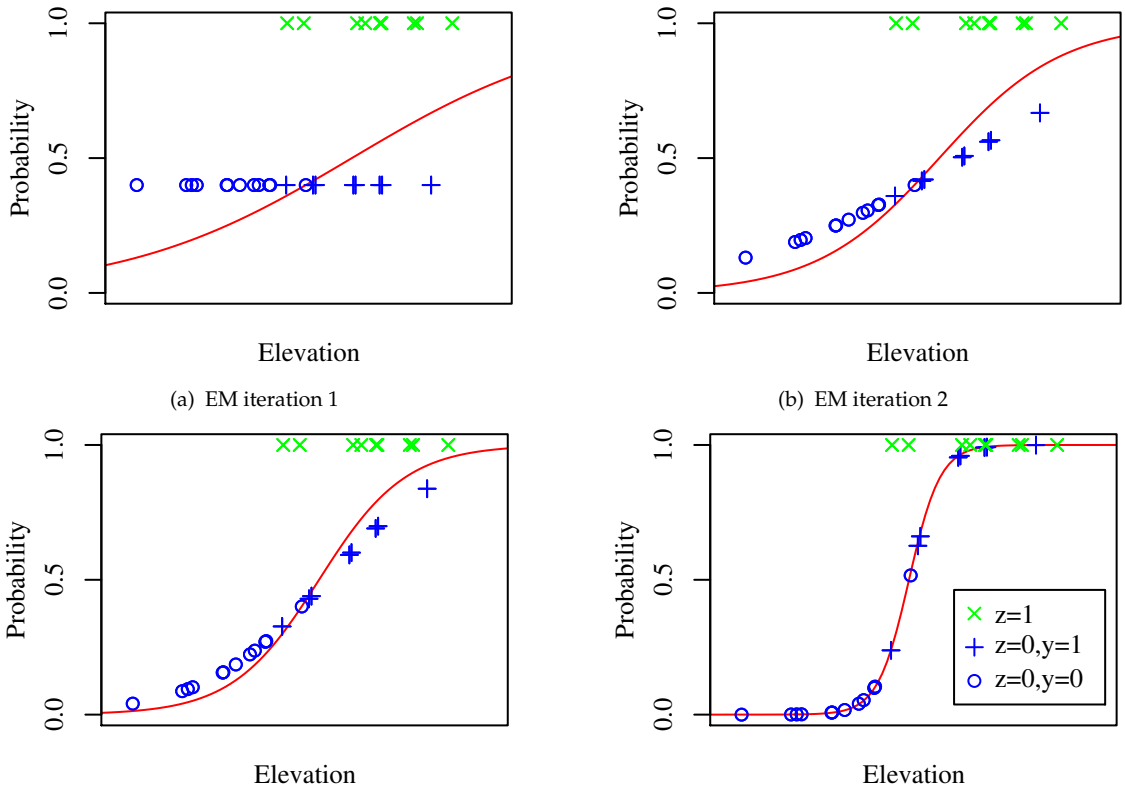
Figure 1: The EM algorithm for presence-only data.



(a) Model with known y

(b) Naive model using z

Figure 2: A simple example where species presence depends only on elevation. A logistic regression model estimated using y (a) and a naive logistic regression model estimated using z (b) are illustrated by the lines. The vertical location of data points (crosses and circles) indicates the value of the outcome as used in the fitting procedure.



(a) EM iteration 1

(b) EM iteration 2

(c) EM iteration 3

(d) EM final iteration

Figure 3: Iterations of the EM algorithm using a logistic regression model as applied to the simple elevation example. The vertical location of data points (crosses and circles) indicate the value of the outcome as used in the fitting procedure and lines indicate models estimated from these data.

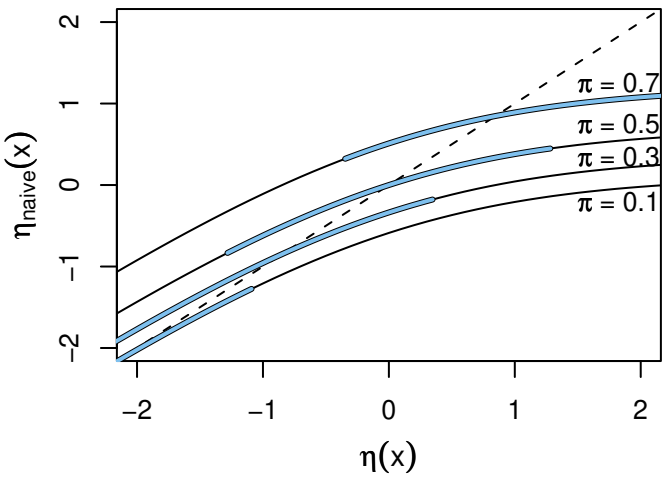
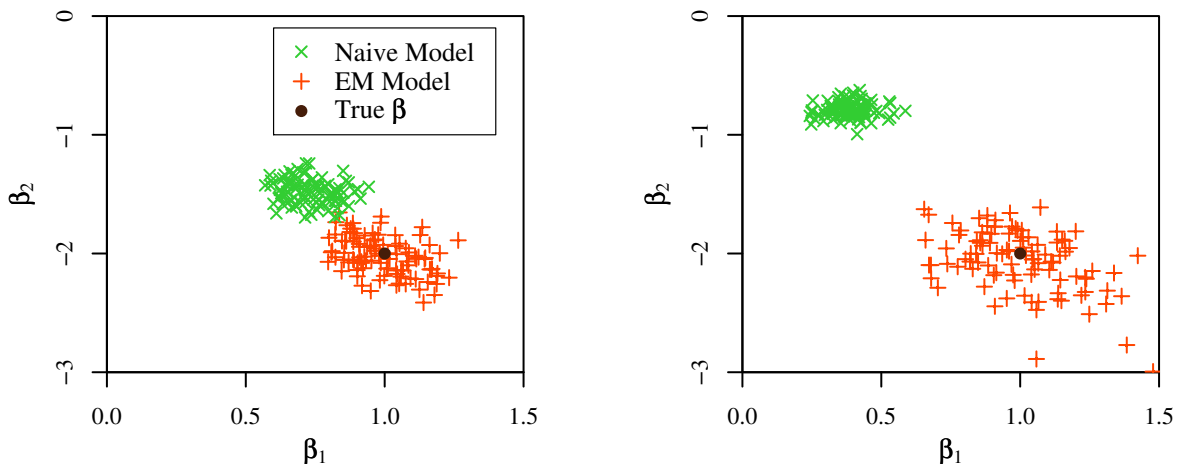
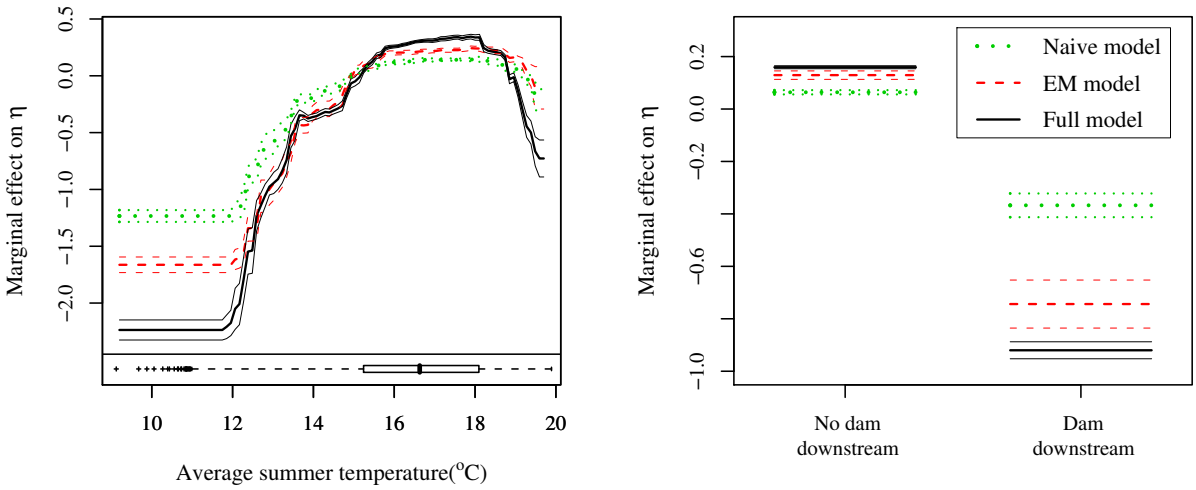


Figure 4: The case-control adjusted η_{naive} from the naive logistic regression is an increasing but non-linear function of the linear predictor from the true model of interest, η, and the population prevalence π. Thick lines indicate typical values of η for each π.



(a) π=0.1 (b) π=0.4

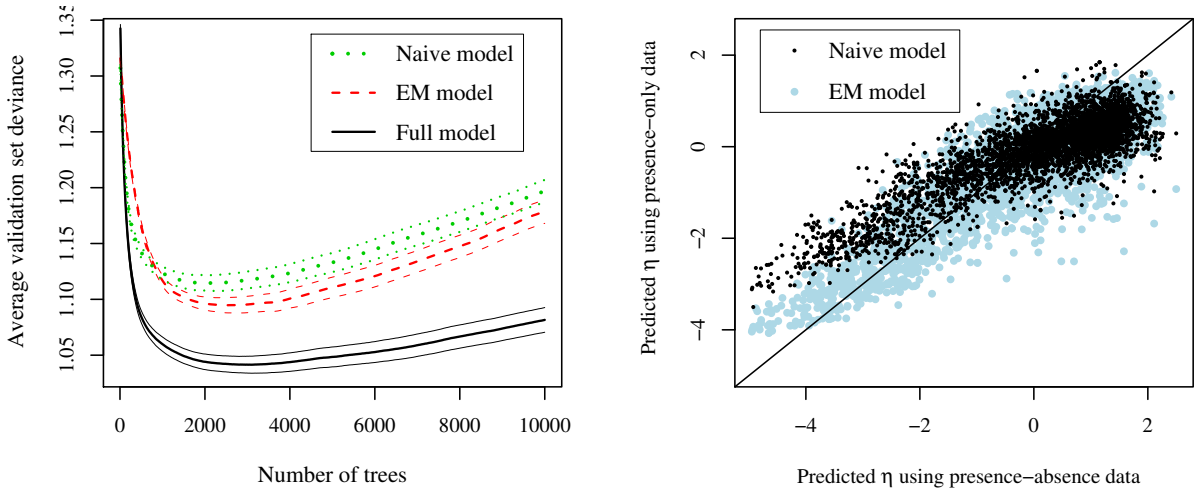
Figure 5: Parameter estimates for the naive logistic regression model are biased toward the origin, with increased bias for larger π. These estimates are from 100 simulations of presence-only data, generated from the model $\eta(\mathbf{x}) = \alpha + x_1\beta_1 + x_2\beta_2$, where $\beta_1 = 1$ and $\beta_2 = -2$. The $\mathbf{x}$ are i.i.d. standard normals and $n_p = 300$ and $n_u = 1000$. Note that the variance of the EM estimates increase with π.



(a) Average summer air temperature (^{o}C)

(b) Presence of a dam downstream

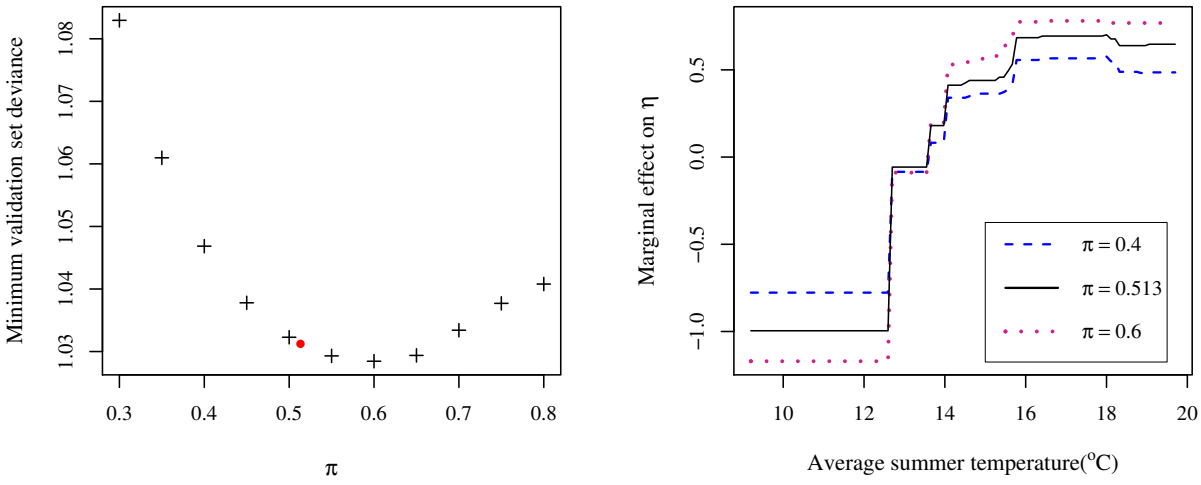
Figure 6: The marginal effect of each variable on η for the EM and naive models and for the full model based on the true presences (mean and ± 1 pointwise standard errors). The boxplot in (a) indicates the distribution of the summer temperature across all locations in the sample; 15% of these locations had a dam downstream.



(a) Validation set deviance

(b) $\hat{\eta}$ using presence-only and presence-absence data

Figure 7: The EM model has a lower validation set deviance (a) and less shrinkage in η (b) than the naive model, when predicting the presence of the Longfin Eel. The average validation set deviance is calculated from the likelihood of the true presences (mean and ± 1 pointwise standard errors). The predicted η in (b) are the best fitting EM and naive models, versus the best fitting model based on the true presences.



(a) Minimum validation set deviance

(b) Marginal effect on η

Figure 8: A sensitivity analysis for π indicates that the minimum validation set deviance for the EM model is smallest for $\pi \approx 0.6$ (a). The effect on η of average summer temperature has a consistent shape across all π, but the estimated effect magnitude increases with π (b). The validation set deviance is calculated using the presence-only likelihood, so is not comparable with Figure 7.

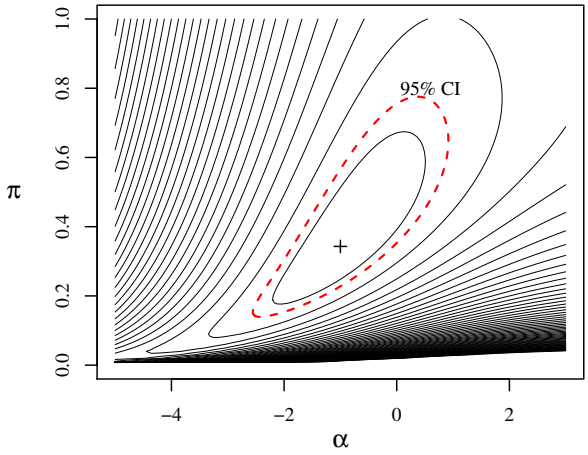


Figure 9: The likelihood surface, and a 95% confidence interval (dashed line), for π and the intercept α at the true β for a simulation of 200 observed presences and 1000 background data. The generative model is $\eta(x) = \alpha + \beta x$, where $\alpha = -1.0$, $\beta = 2$,$\pi = 0.34$ and the x are i.i.d. standard normals.