Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

Logistic Methods for Resource Selection Functions and Presence-Only Species Distribution Models

Steven J. Phillips
AT&T Labs − Research
180 Park Avenue, Florham Park, NJ 07932

Jane Elith
School of Botany, The University of Melbourne
Parkville 3010, Australia

Abstract

In order to better protect and conserve biodiversity, ecologists use machine learning and statistics to understand how species respond to their environment and to predict how they will respond to future climate change, habitat loss and other threats. A fundamental modeling task is to estimate the probability that a given species is present in (or uses) a site, conditional on environmental variables such as precipitation and temperature. For a limited number of species, survey data consisting of both presence and absence records are available, and can be used to fit a variety of conventional classification and regression models. For most species, however, the available data consist only of occurrence records — locations where the species has been observed. In two closely-related but separate bodies of ecological literature, diverse special-purpose models have been developed that contrast occurrence data with a random sample of available environmental conditions. The most widespread statistical approaches involve either fitting an exponential model of species' conditional probability of presence, or fitting a naive logistic model in which the random sample of available conditions is treated as absence data; both approaches have well-known drawbacks, and do not necessarily produce valid probabilities. After summarizing existing methods, we overcome their drawbacks by introducing a new scaled binomial loss function for estimating an underlying logistic model of species presence/absence. Like the Expectation-Maximization approach of Ward et al. and the method of Steinberg and Cardell, our approach requires an estimate of population prevalence, $\Pr(y = 1)$, since prevalence is not identifiable from occurrence data alone. In contrast to the latter two methods, our loss function is straightforward to integrate into a variety of existing modeling frameworks such as generalized linear and additive models and boosted regression trees. We also demonstrate that approaches by Lele and Keim and by Lancaster and Imbens that surmount the identifiability issue by making parametric data assumptions do not typically produce valid probability estimates.

Introduction

We study a modeling task that is central to two related but largely separate bodies of ecological literature. Ecologists investigating resource selection by animals seek to characterize those areas within a region of interest that are "used" by a particular species or individual animals (Manly et al. 2002), while ecologists studying a broad range of sessile or mobile species wish to predict the suitability of sites for occupation or persistence of the species (Franklin 2010). In both cases, the available data frequently consist of a collection of geographic locations with evidence of use by (or presence of) the species together with data on environmental covariates in the region of interest, termed available (or background) data. The most desirable output is the probability of use (resp., probability of presence) conditional on environmental covariates; the shape of the response to the covariates is also important for understanding how the species relates to its environment. Methods for estimating probability of use/presence and related indices are important – they have been used extensively for a variety of applications in ecology and conservation, and according to Google Scholar, a seminal resource selection text (Manly et al. 2002) has been cited 1237 times while an influential SDM paper (Elith et al. 2006) has received 970 citations.

While there is shared agonizing over data and model interpretation in the two bodies of ecological research — what defines "presence" rather than a transitory or chance visit, how to determine absolute "absence", what defines the background area from which the species is selecting sites, what ecological insight can be derived from model outputs (Lele and Keim 2006; Pulliam 2000; Desrochers et al. 2010; Johnson et al. 2006; Franklin 2010) — we focus here on the underlying statistical questions rather than ecological interpretation. In particular, we study maximum-likelihood logistic models of probability of presence and introduce a new method for estimating such models. For brevity, we will use only the "presence–background" terminology from here on.

In practice, exponential models are most often used, fitted using logistic regression (Manly et al. 2002) or maximum entropy (Phillips, Anderson, and Schapire 2006), in both cases providing a maximum-likelihood estimate of relative probability of presence. Exponential models have the drawback of being unbounded above, so that when estimates are scaled to estimate true (rather than relative) probability of

presence, estimates greater than 1 may be obtained (Keating and Cherry 2004; Ward et al. 2009). Therefore, logistic models are preferable simply for giving values bounded by $[0, 1]$. Maxent estimates are typically rescaled to remain within $[0, 1]$ (Phillips and Dudík 2008), but the result is no longer a maximum likelihood model. "Naive" logistic models are sometimes simply fitted to presence-background data, treating the background as if it were absence (Ferrier et al. 2002; Elith et al. 2006), but the resulting models are not proportional to probability of presence (Phillips et al. 2009). (They are monotonically related, which is sufficient for some applications, such as when model outputs are thresholded to yield binary values rather than probabilities.)

Existing maximum-likelihood logistic methods include those of Steinberg and Cardell (1992), Lancaster and Imbens (1996), Lele and Keim (2006) and Ward et al. (2009), which we will refer to as the SC, LI, LK and EM models respectively. The first was brought to our attention during review of this paper; it appears to be untested in ecology so here we explain it and note its equivalence to a maximum entropy formulation. The second and third are given some attention here, because they are used by ecologists. They depend on a parametric assumption to make model parameters identifiable, and we demonstrate with a simple example that deviation from the assumption results in very poorly calibrated models. The EM method uses the Expectation-Maximization algorithm to impute presence/absence in the background data, and takes population prevalence, $\Pr(y = 1)$, as an additional parameter. This method is theoretically attractive and can be combined with a variety of model fitting methods, but has not seen significant ecological use as it requires specialized software and there is no generally available implementation. The SC method has been described only for parametric models; it takes prevalence as an input to avoid the strong assumption of LI and LK.

Our main contribution is to introduce a new scaled binomial loss function for presence-background data and demonstrate that it is easily expressed as a link function for generalized linear and additive models (GLM and GAM) or optimized by existing boosted regression tree (BRT) software, in order to give maximum likelihood logistic models of presence/absence. Our loss function is related to that of the LI model; the primary point of departure is that we take the population prevalence as a supplied parameter, thus avoiding unrealistic parametric assumptions about the data. We present experiments that suggest that our approach is at least as effective a solution as the EM and SC methods, and we argue that our loss function is simple to integrate with a variety of established model fitting methods.

Derivation of scaled binomial loss

Let L be the landscape of interest, and L_1 the subset of L where the species is present. Let U be a set of presence samples drawn from L_1 and B a set of background samples drawn from L. Following Phillips et al. (2009), we formalize a sampling model for presence-background data that simplifies the analysis. Let f_p be the fraction of presence samples in the training data, i.e., $f_p = n_p/(n_p + n_b)$ where $n_p = |U|$ and $n_b = |B|$. We regard the data as having been generated by the following process: each sample is drawn uniformly from L_1 with probability f_p (a presence sample) and uniformly from L with the remaining probability $(1-f_p)$ yielding a background sample. We use the symbol P_{UA} to describe probabilities under this sampling model, and reserve the symbol $\Pr$ to refer to probabilities over L. We use s to represent sampling stratum, with $s = 1$ for presence samples and $s = 0$ for background samples. We refer to $\Pr(y = 1)$ as the population prevalence, or simply prevalence.

When a statistical model is applied to presence-background data by treating the background data as absence (a "naive" model) it produces an estimate of $P_{\text{UA}}(s = 1|\boldsymbol{x})$, the probability that a sample is a presence sample given environmental conditions $\boldsymbol{x}$. While this is not the same as conditional probability of presence of the species, $\Pr(y = 1|\boldsymbol{x})$, the two probabilities are related by a simple formula (Phillips et al. 2009; see also Keating and Cherry 2004; Lancaster and Imbens 1996; Ward et al. 2009). If we define

$$r = \frac{(1 - f_p)}{f_p} \Pr(y = 1) \tag{1}$$

then

$$P_{\text{UA}}(s = 1|\boldsymbol{x}) = \frac{1}{1 + r/\Pr(y = 1|\boldsymbol{x})}. \tag{2}$$

We can transform a naive model into a model of $\Pr(y = 1|\boldsymbol{x})$ by simply inverting Eqn. 2:

$$\Pr(y = 1|\boldsymbol{x}) = \frac{rP_{\text{UA}}(s = 1|\boldsymbol{x})}{1 - P_{\text{UA}}(s = 1|\boldsymbol{x})} \tag{3}$$

In other words, $\Pr(y = 1|\boldsymbol{x})$ is r times the odds of $P_{\text{UA}}(s = 1|\boldsymbol{x})$. This is why exponential models are widely used in resource selection studies: the odds of a logistic model $(1/(1 + e^{-\eta(\boldsymbol{x})}))$ are exponential $(e^{\eta(\boldsymbol{x})})$. That is, a naive logistic model for presence-background data yields an exponential model of probability of presence.

Lancaster and Imbens (1996) took a different approach: they specified a logistic model of $\Pr(y = 1|\boldsymbol{x})$, i.e.,

$$\Pr(y = 1|\boldsymbol{x}) = \frac{1}{1 + \exp(-\eta(\boldsymbol{x}))} \tag{4}$$

and they assumed the response $\eta(\boldsymbol{x}) = \boldsymbol{\beta} \cdot \boldsymbol{x}$ is a linear function of the environmental predictors. Substituting into Eqn. 2, we obtain

$$P_{\text{UA}}(s = 1|\boldsymbol{x}) = \frac{c}{1 + \exp(-\eta'(\boldsymbol{x}))}. \tag{5}$$

where $c = 1/(1 + r)$ and

$$\eta'(\boldsymbol{x}) = \eta(\boldsymbol{x}) - \ln(r/(1 + r)). \tag{6}$$

We can therefore obtain η by fitting η' (in Eqn. 2) and then offsetting it by $\ln(r/(1 + r))$. The model in Eqn. 5 is not quite a logistic model, because of the scaling parameter c. Lancaster and Imbens (1996) proposed a maximum likelihood method for estimating prevalence (and hence c) and

```
scaledlogit <- function(c)
{
 linkfun <- function(mu) qlogis(mu/c)
 linkinv <- function(eta) c*plogis(eta)
 mu.eta <- function(eta) c*.Call("logit_mu_eta",
               eta, PACKAGE = "stats")
 valideta <- function(eta) TRUE
 validmu <- function(mu) all(mu>0) && all(mu<c)
 link <- paste("scaledlogit(", c, ")", sep="")
 structure(list(
  linkfun=linkfun, linkinv=linkinv, mu.eta=mu.eta,
  valideta=valideta, name=link), class="link-glm")
}

c <- 1/(1+prevalence)
glm(s ~ x,
  family=binomial(link=scaledlogit(c)),
  weights=ifelse(s==1,1,sum(s==1)/sum(s==0))
  start=c(0,0))
```

Figure 1: R code to fit a univariate model to $P_{\text{UA}}(s = 1|x)$ with scaled binomial loss. Predicted values of $P_{\text{UA}}(s = 1|x)$ or its logit can be converted to logistic estimates of $\Pr(y = 1|x)$ using Eqn. 3 or Eqn. 6. Sample weights equalize the total weight of presence and background data, so r equals the (supplied) estimate of the species' prevalence in Eqn.1.

the parameters of η'. [1] However, the population prevalence $\Pr(y = 1)$ is not identifiable in practice from presence-background data (Ward et al. 2009) – it is only identifiable under the assumption that the logit of the true probability of presence is *exactly* linear in the predictors. Ward et al. (2009) strongly advise against making such an unrealistic assumption – after all, linearity is always at best an approximation to the complexity of natural phenomena (further, assuming an approximately linear response is often not justified ecologically; Austin 2002).

We therefore assume that an estimate of population prevalence is provided as an input. The estimate could derive from limited presence-absence surveys or from expert opinion, and sensitivity to the estimate can be explored by running multiple models. With the parameter c now given, Eqn. 5 describes a link function which can be used in GLMs or GAMs. There is widely-available software for fitting such models, and in particular, the `glm` method and the `gam` package in R allow specification of the link function (Fig. 1). Eqn. 5 similarly defines a loss function — "scaled binomial loss" — that can be minimized with BRT methods, including `gbm` (Friedman 2001) which has been recognized as a powerful method for modeling species distributions (Elith et al. 2006; Elith, Leathwick, and Hastie 2008). We defined a scaled logit link family (similar to Fig. 1) for the `mboost` R package which has a generic link family mechanism (Bühlmann and Hothorn 2007). We also implemented a scaled binomial loss function in C++ to add to the most widely-used implementation of BRT, the `gbm` package for R (Ridgeway 2007). This package uses hard-coded loss functions so that terminal node estimates can be custom-designed for each error distribution. For our scaled binomial loss implementation, we used the weighted average gradient of loss for terminal node estimates, forgoing the single step of Newton-Raphson used for Bernoulli loss in `gbm`, as the latter might diverge as scaled binomial loss is not convex.

[1] Lancaster and Imbens (1996) also describe an approach for estimating the linear predictor η if the prevalence is known, based on the generalized method of moments. The approach is difficult and has not been applied in ecology due to lack of available software. In this paper we use "LI" only to refer to the maximum likelihood approach of Lancaster and Imbens (1996) for estimating parameters when the prevalence is unknown.

Summary of existing methods

We present the LI method above. The LK method defines the log likelihood of the presence samples as $\sum_{\boldsymbol{x}\in U} \ln \Pr(\boldsymbol{x}|y = 1; \boldsymbol{\beta})$. Applying Bayes' rule and dropping terms that do not depend on $\boldsymbol{\beta}$ yields:

$$\sum_{\boldsymbol{x}\in U} \ln \Pr(y = 1|\boldsymbol{x}; \boldsymbol{\beta}) - |U| \ln \Pr(y = 1; \boldsymbol{\beta}) \qquad (7)$$

An empirical estimate of the second term is given by

$$\frac{|U|}{|B|} \sum_{x\in B} \Pr(y = 1|x; \boldsymbol{\beta}). \qquad (8)$$

Standard numerical optimization techniques (Lele and Keim 2006) or more involved optimization methods (Lele 2009) are used to maximize Eqn. 7 and hence obtain a maximum likelihood estimate of $\boldsymbol{\beta}$. The parameters are identifiable only under the parametric assumption described above.

The SC method breaks the log likelihood of the entire landscape L given parameters $\boldsymbol{\beta}$ into terms that depend either on L_1 or all of L. These terms can then be empirically estimated using the samples U and B (as in Eqn. 8). Combining terms (using an estimate of prevalence) yields a pseudo-likelihood that approximates the log likelihood of L:

$$\sum_{\boldsymbol{x}\in B} -\ln(1 + \exp(\eta(\boldsymbol{x}))) + p\,|B|\,\eta(\boldsymbol{\mu}) \qquad (9)$$

where p is the given prevalence and $\boldsymbol{\mu}$ is the vector of means of predictors over U. Eqn. 9 can be maximized using standard numerical optimization techniques. An alternative derivation of this pseudo-likelihood can be obtained using a maximum-entropy based approach called "class-robust logistic regression" (Dudík and Phillips 2008), replacing the class-robustness by the constraint that modeled prevalence on B should equal p. The SC method has the advantage of being the only one (out of those considered here) with a convex loss function.

EM is a general-purpose algorithm for estimating missing data (Dempster, Laird, and Rubin 1977). For the current application, the missing data is the value of y for background samples. EM repeatedly fits a model (e.g., a GLM, GAM or BRT) using the current estimates of the missing data (a "Maximization" step), then uses that model to update the estimates of the missing data (an "Expectation" step). This iterative process is repeated until convergence, and can be very slow for complex fitting procedures (e.g. BRT).

Comparison with existing methods

First we explore the implication of the parametric assumption in both LI and LK. We generated models for a simple simulated species (Fig. 2) for which the logit of $\Pr(y = 1|x)$ is not exactly linear in x. A logistic model that assumes a linear logit therefore cannot be a perfect estimate of $\Pr(y = 1|x)$, but despite this, we could hope to get a reasonable estimate. Any practical method for modeling $\Pr(y = 1|x)$ needs to be robust enough to deal with the likely range of species-environment relationships, and this simulation falls within that range. We present results primarily for the LI method (Fig. 2 a to c); results were very similar for LK (Fig. 2 d). Model predictions varied wildly between simulations and were unsuccessful at estimating true probability of presence (Fig. 2 a). Furthermore, although the methods are guaranteed to converge to the maximum-likelihood model parameters given enough data, 1000 presence samples and 10000 background samples were not sufficient (Fig. 2 c,d). These results support previous findings (Ward et al. 2009) that LI parameter estimates are highly variable, especially when true probabilities deviate from the assumed parametric form. In contrast, combining GLM with either scaled binomial loss or EM gave a close approximation of true probability of presence (Fig. 2 a), with minimal variation between simulations.

The LI and LK methods do yield accurate resource selection functions (RSF; Manly et al. 2002), i.e., the modeled response is approximately proportional to true probability of presence (Fig. 2 b). They do not yield resource selection probability functions (RSPF) – i.e., correct probabilities – since the constant of proportionality is not close to 1.

In our second experiment, we applied a GLM with scaled binomial loss (Fig. 1) to data generated from a bivariate logistic model, with details of the experiment matching those of Ward et al. (2009) to allow comparison of results. Parameter estimates using scaled binomial loss were approximately unbiased, unlike for naive models (Fig. 4). Variance of the parameter estimates increased with higher prevalence, and appears comparable to those obtained by Ward et al. (2009; Fig. 3). Because these artificial data have linear logit, the LK method also produced approximately unbiased parameter estimates, but with higher variance. In contrast to the LK method though, scaled binomial loss is also valid in the general case that the data are not exactly parametric — for example, it gives a good approximation to the simulated distribution of Fig. 2 (green line).

For our third experiment (Fig. 3), we used the same test data as for Fig. 4 and fit both naive models and models with scaled binomial loss using two implementations of BRT, `gbm` (Ridgeway 2007) and `mboost` (Bühlmann and Hothorn 2007). GAMs with scaled binomial loss were also included. The naive models all under-estimated the slope of the response to the predictor variables (Fig. 3 bottom), and this bias was essentially corrected by use of scaled binomial loss (Fig. 3 top). The GAM and `gbm` models were broadly similar, though `mboost` exhibited more bias in naive simulations and greater variance and some residual bias in simulations with scaled binomial loss.

Finally, a comparison over four synthetic distributions

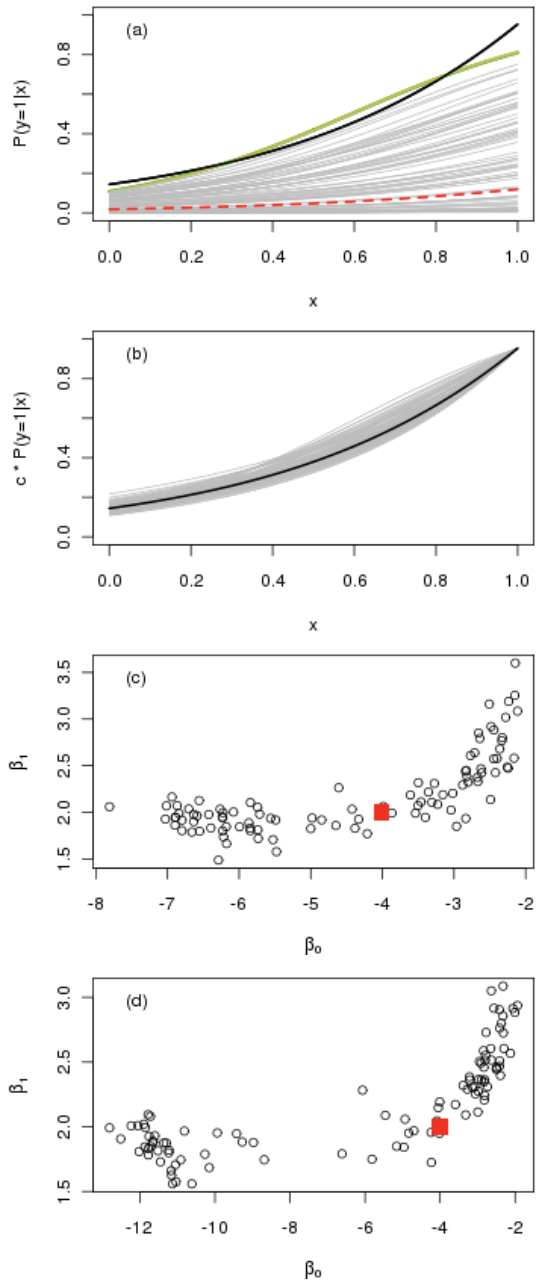

Figure 2: LI (Lancaster and Imbens 1996) and LK (Lele and Keim 2006) models from simulated data with $\Pr(y = 1|x) = 8/(1 + \exp(4 - 2x))$ (black line), with $n_p = 1000$ and $n_b = 10000$ and x uniform in $[0, 1]$. Models were fit for 100 simulations using the `nlm` function in R. (a): 100 LI models (grey), and the maximum likelihood estimate that LI would produce given unlimited data ($\Pr(y = 1|x) = 1/(1 + \exp(4 - 2x))$, dashed red line). Scaled binomial loss (green) and EM (orange) models are shown for comparison (only one simulation shown, others were similar). (b): LI models after rescaling so that predictions at $x = 1$ match the true probability of presence. (c): LI fitted parameters β_0 and β_1 (black) and true parameter values (red). (d): LK fitted parameters.

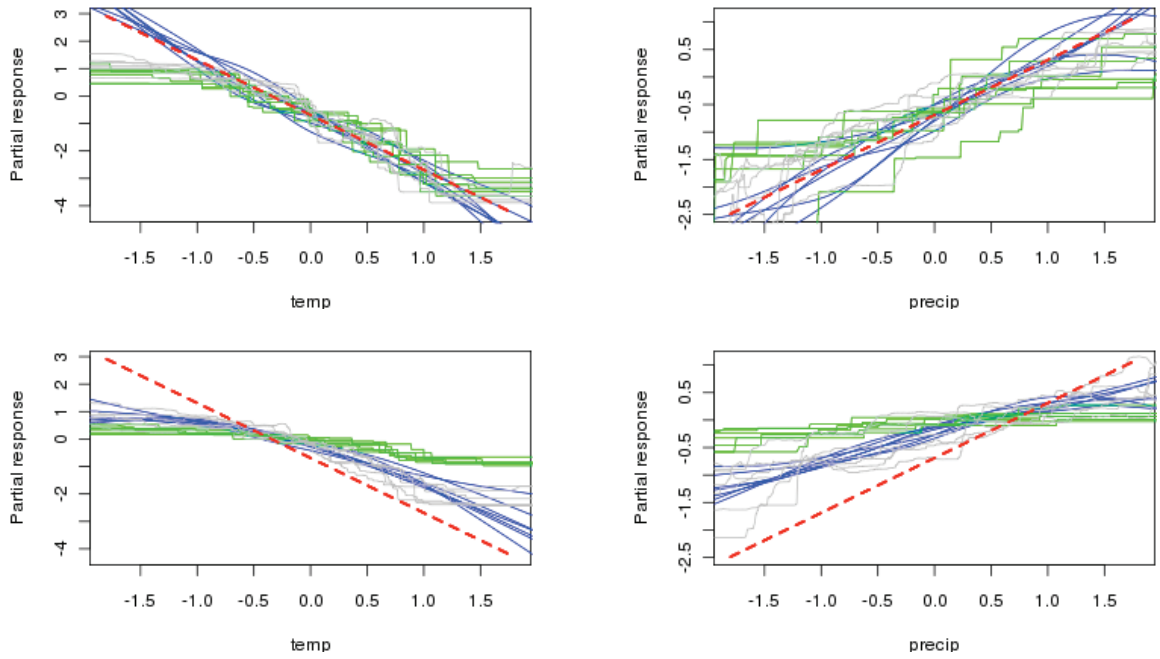

Figure 3: Naive and scaled binomial loss `gam`, `gbm` and `mboost` models of simple test data. Partial response curves are shown for the truth (dashed red line) and for 100 simulations for gbm (grey), mboost (green) and gam (blue). Data were drawn from the model $\eta = \alpha + \beta_1 \text{precip} + \beta_2 \text{temp}$, where $\alpha = -0.69$, $\beta_1 = 1$ and $\beta_2 = -2$, precip and temp were standard normals giving prevalence $= 0.4$, and $n_p = 300$ and $n_b = 1000$. All gbm and mboost models were fitted with 10000 stumps (single-node trees). Top: models with scaled binomial loss. Bottom: naive models.

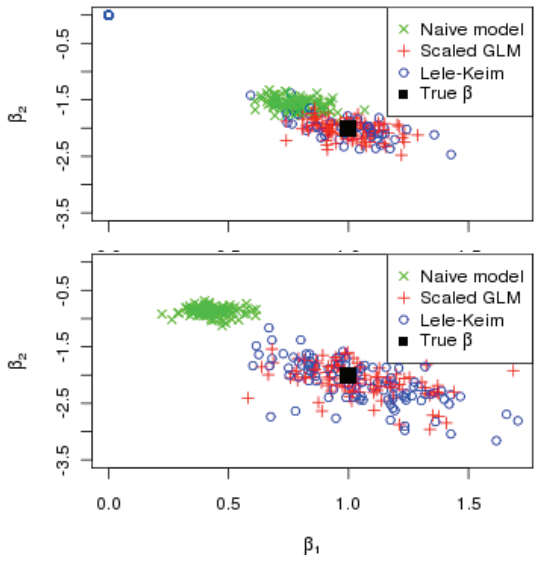

Figure 4: Parameter estimates for logistic models using presence-background data. Estimates are shown for 100 simulations generated from the model $\eta = \alpha + \beta_1 \text{precip} + \beta_2 \text{temp}$, where $\beta_1 = 1$ and $\beta_2 = -2$, precip and temp are standard normals, $n_p = 300$ and $n_b = 1000$, and α is set to give prevalence of 0.1 (top) or 0.4 (bottom).

(Table 1) showed that the scaled binomial loss, EM, and SC methods all produced good estimates of probability of presence, as measured by RMS error (Table 1), while LI and LK had poor performance. Errors in prevalence estimates of $\pm$ 10% increased RMSE equally for scaled loss and EM, but RMSE remained lower than for LI and LK in each case.

Conclusion

We surveyed existing methods for logistic modelling of species' probability of presence (or probability of use) from presence-only (or use-availability) data. Some methods use a strong parametric assumption to make model parameters (in particular, population prevalence) identifiable; based on our experiments and those of Ward et al. (2009), we strongly recommend against relying on this assumption. Instead, we recommend using an estimate of population prevalence.

We introduced a new loss function, scaled binomial loss, which can easily be used as a link function in generalized linear and generalized additive models or incorporated in established model-fitting methods such as boosted regression trees. As with the EM method of Ward et al. (2009) and the method of Steinberg and Cardell (1992), our method takes an estimate of population prevalence as input. We argue that our loss function is easier to integrate with the variety of existing model-fitting methods than either of these existing methods. Our experiments on simulated data suggest that all three can have good performance, and we feel that all three warrant further investigation for the analysis of presence-background / use-availability data in ecology.

Truth	Scaled logit	EM	SC	Scaled logit +	EM +	Scaled logit -	EM -	LI	LK
(x+y)/2	0.048	0.047	0.054	0.075	0.073	0.066	0.067	0.228	0.234
$1/(1+e^{3(x+y-1)})$	0.047	0.045	0.052	0.071	0.069	0.066	0.066	0.165	0.165
$8/(1+e^{4-2x})$	0.064	0.063	0.063	0.077	0.076	0.080	0.079	0.332	0.343
$1/(1+e^{4-2x})$	0.0062	0.0061	0.0061	0.0095	0.0093	0.0084	0.0083	0.1880	0.1852

Table 1: Root mean square (RMS) error of logistic models fitted to synthetic species distributions, each of which is a function of variables x and y uniform in $[0, 1]$, with $n_p = 200$ and $n_b = 600$. Results shown are the mean of 100 simulations. Models with "+" were given prevalence estimates 10% higher than truth, while "-" indicates 10% lower. Scaled logit and EM use GLM for model fitting; others use the `nlm` function in R. The synthetic responses are: (1) linear; (2) logistic with linear logit, with probabilities covering the full range from 0 to 1; (3) same as black line in Figure 2 a; (4) same as dashed red line in Figure 2 a.

Acknowledgements

We thank Miro Dudík, Trevor Hastie and three reviewers for insights and helpful commentary. JE gratefully acknowledges Australian Research Council grant FT0991640.

References

Austin, M. 2002. Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling* 157:101–118.

Bühlmann, P., and Hothorn, T. 2007. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* 22(4):477–505.

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, Series B* 39:1–38.

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

Dudík, M., and Phillips, S. J. 2008. Generative and discriminative learning with unknown labeling bias. In *Neural Information Processing Systems*, 401–408.

Elith, J.; Graham, C. H.; Anderson, R. P.; Dudík, M.; Ferrier, S.; Guisan, A.; Hijmans, R. J.; Huettmann, F.; Leathwick, J. R.; Lehmann, A.; Li, J.; Lohmann, L. G.; Loiselle, B. A.; Manion, G.; Moritz, C.; Nakamura, M.; Nakazawa, Y.; Overton, J. M.; Peterson, A. T.; Phillips, S. J.; Richardson, K.; Scachetti-Pereira, R.; Schapire, R. E.; Soberón, J.; Williams, S.; Wisz, M. S.; and Zimmerman, N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29(2):129–151.

Elith, J.; Leathwick, J.; and Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77:802–813.

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

Franklin, J. 2010. *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press.

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

Johnson, C. J.; Nielsen, S. E.; Merrill, E. H.; McDonald, T. L.; and S.Boyce, M. 2006. Resource selection functions based on use-availability data: Theoretical motivation and evaluation methods. *The Journal of Wildlife Management* 70(2):347–357.

Keating, K. A., and Cherry, S. 2004. Use and interpretation of logistic regression in habitat-selection studies. *Journal of Wildlife Management* 68(4):774–789.

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

Lele, S. R., and Keim, J. L. 2006. Weighted distributions and estimation of resource selection probability functions. *Ecology* 87(12):3021–3028.

Lele, S. R. 2009. A new method for estimation of resource selection probability function. *Journal of Wildlife Management* 73(1):122–127.

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

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

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

Phillips, S.; Dudík, M.; Elith, J.; Graham, C.; Lehmann, A.; Leathwick, J.; and Ferrier, S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* 19(1):181–197.

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

Ridgeway, G. 2007. gbm: Generalized Boosted Regression models. R package version 1.6-3, http://www.ipensieri.com/gregr/gbm.shtml.

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

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