The Annals of Applied Statistics
2013, Vol. 7, No. 4, 1917–1939
DOI: 10.1214/13-AOAS667
© Institute of Mathematical Statistics, 2013

FINITE-SAMPLE EQUIVALENCE IN STATISTICAL MODELS FOR PRESENCE-ONLY DATA

By William Fithian[1] and Trevor Hastie[2]

Stanford University

Statistical modeling of presence-only data has attracted much recent attention in the ecological literature, leading to a proliferation of methods, including the inhomogeneous Poisson process (IPP) model, maximum entropy (Maxent) modeling of species distributions and logistic regression models. Several recent articles have shown the close relationships between these methods. We explain why the IPP intensity function is a more natural object of inference in presence-only studies than occurrence probability (which is only defined with reference to quadrat size), and why presence-only data only allows estimation of relative, and not absolute intensity of species occurrence.

All three of the above techniques amount to parametric density estimation under the same exponential family model (in the case of the IPP, the fitted density is multiplied by the number of presence records to obtain a fitted intensity). We show that IPP and Maxent give the exact same estimate for this density, but logistic regression in general yields a different estimate in finite samples. When the model is misspecified—as it practically always is—logistic regression and the IPP may have substantially different asymptotic limits with large data sets. We propose "infinitely weighted logistic regression," which is exactly equivalent to the IPP in finite samples. Consequently, many already-implemented methods extending logistic regression can also extend the Maxent and IPP models in directly analogous ways using this technique.

1. Introduction. In recent years ecologists have devoted significant attention to the problem of estimating the geographic distribution of a species of interest from records of where it has been found in the past. There are many motivations for solving this problem, including planning wildlife management actions, monitoring endangered or invasive species, and understand-

Received October 2012; revised June 2013.

[1] Supported by NSF VIGRE Grant DMS-0502385.

[2] Supported in part by NSF Grant DMS-10-07719 and NIH Grant RO1-EB001988-15.

Key words and phrases. Presence-only data, logistic regression, maximum entropy, Poisson process models, species modeling, case-control sampling.

ing species' response to different habitats. A great variety of experimental designs and statistical methods exist for tackling this problem, and can be found in the literature on resource-selection functions [Manly et al. (2002), Lele and Keim (2006)], case-augmented designs [Lee, Scott and Wild (2006), Dorazio (2012)] and site occupancy modeling [MacKenzie (2006)].

Ecologists have proposed many statistical methods for modeling such data, including the inhomogeneous Poisson process (IPP) model [Warton and Shepherd (2010)], maximum entropy (Maxent) modeling of species distributions [Phillips, Dudík and Schapire (2004), Phillips, Anderson and Schapire (2006), Phillips and Dudík (2008)] and the logistic regression model along with its various generalizations such as GAM, MARS and boosted regression trees [Hastie, Tibshirani and Friedman (2009)]. See Elith et al. (2006) for discussion and comparison of these and other methods in common use.

In recent years several articles have emerged detailing connections between the three modeling methods above. Each method takes as its input a presence-only data set along with a set of background points consisting of a regular grid or random sample of locations in some geographic region of interest. Warton and Shepherd (2010) showed that logistic regression estimates converge to the IPP estimate when the size of the presence-only data set is fixed and the background sample grows infinitely large. Aarts, Fieberg and Matthiopoulos (2012) additionally described a variety of models for presence-only and other data sets whose likelihoods may all be derived from the IPP likelihood. Renner and Warton (2013) further explore the connection between Maxent and the IPP, taking up the important issue of how we might check the IPPs modeling assumptions.

Our primary aim in writing this paper is to provide additional clarity to this topic, recapitulating and deriving the results in a unified framework and extending them in several directions. We view all three major methods as solutions to the same parametric density estimation problem.

1.1. *Presence-only data.*

Modeling of species distributions is simplest and most convincing when the observations of species presence are collected systematically. In a typical design, a surveyor visits a one-square-kilometer patch of land for one hour and records how many specimens she discovers in that interval. The records of unsuccessful surveys are called absence records, a mild misnomer since ecologists recognize that specimens could be present but go undetected. A data set reflecting presence or absence of a species in each sampling unit is called presence–absence data.

Unfortunately, dedicated surveys recording sampling effort are expensive, especially for rare or elusive species. For many species of interest, the only data available are museum or herbarium records of locations where a specimen was found and reported, for instance, by a motorist or hiker. Typically

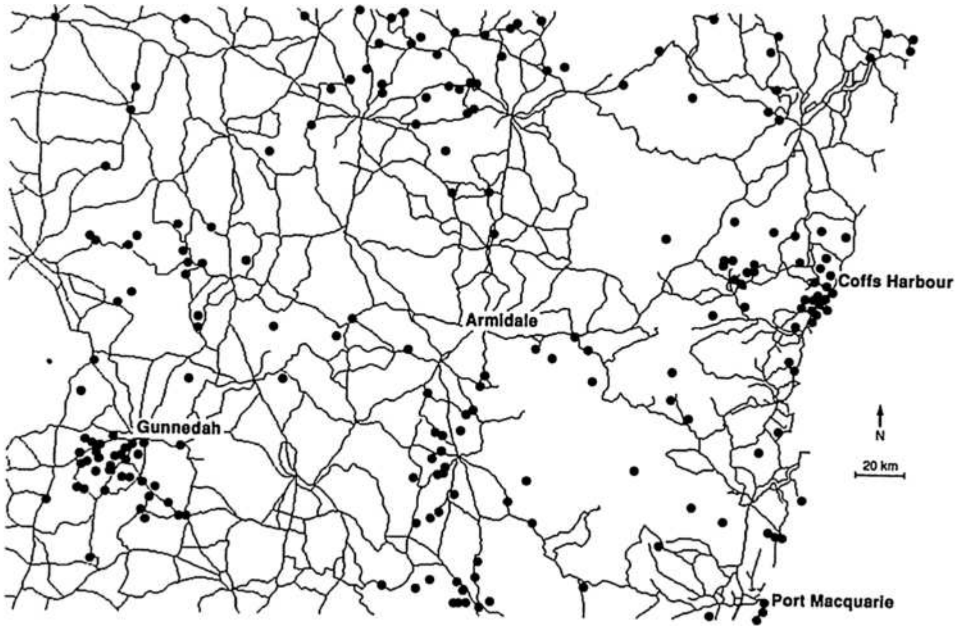

FIG. 1. *Sampling bias in presence-only data for koalas. Taken from Margules et al. (1994).*

these presence-only records are collected haphazardly and frequently suffer from unknown sampling bias such as that illustrated in Figure 1. The clustering of koala sightings near roads and cities probably has more to do with the behavior of people than of koalas.

In recent years many such presence-only data sets have become available electronically, and geographic information systems (GIS) enable ecologists to remotely measure a variety of geographic covariates without having to visit the actual locations of the observations. As a result, presence-only data has become a popular object of study in ecology [Elith et al. (2006)].

1.2. *What should we estimate*? Before we can sensibly decide how to model presence-only data, we must address the issue of what it is we are modeling in the first place. How should we think of "species occurrence," the scientific phenomenon nominally under study? This issue arises with presence-only and presence–absence data alike.

1.2.1. *Occurrence probability.* Figure 2 is a typical "heat-map" output of a study of the willow tit in Switzerland using count data [Royle, Nichols and Kéry (2005)]. The map reveals which locations are more or less favored by the species (in this case, high elevation and moderate forest cover appear to be the bird's habitat of choice). The legend tells us that the color of a region reflects the local probability of "occurrence."

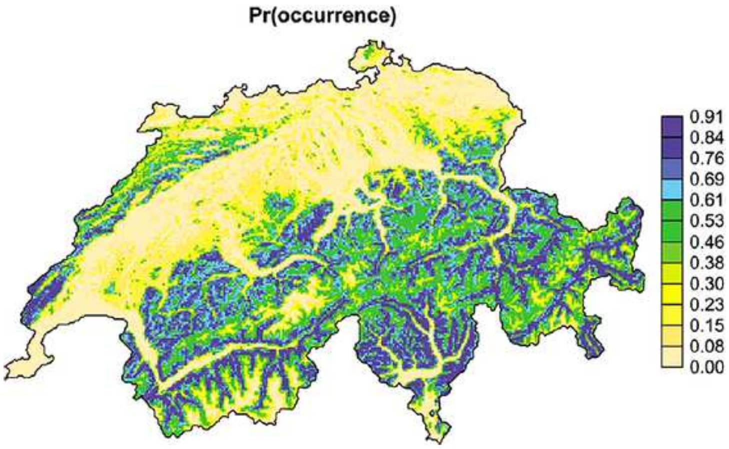

Fig. 2. *Heat map of occurrence probabilities. Taken from Royle, Nichols and Kéry (2005).*

But precisely what event has this probability? Reading the paper, we discover that occurrence means that there is at least one willow tit present on a survey path through a 1 km × 1 km quadrat of land. In this case, the authors analyze a presence–absence data set using a hierarchical model that explicitly accounts for the possibility that a bird was present but not detected at the time of the survey.

Because the survey path length varies across sampling units, the authors use it in their model as a predictor of presence probability. It is not specified which value of this predictor is used in generating the heat map, which makes the map difficult to interpret.

Even if we could interpret the heat map as the probability of a bird being present anywhere in the quadrat (not just along a path of unspecified length), this probability would still be larger in a 2 km × 2 km sampling unit and smaller in a 100 m × 100 m one. Therefore, the very definition of "occurrence probability" in a presence–absence study depends crucially on the specific sampling scheme used to collect the presence–absence data. Consequently, interpreting the legend of such a heat map can only make sense in the context of a specific quadrat size, namely, whatever size was used in the study. We would recommend that this information always be displayed alongside the plot to avoid conveying the false impression (suggested by a heat map) that occurrence probability is an intrinsic property of the land, when it is really an extrinsic property.

Though the choice of quadrat size used to define occurrence probability is ecologically arbitrary, it can in principle yield estimates with meaningful interpretations. By contrast, estimating occurrence probability in a presence-only study is a murkier proposition. Any method purporting to do so without reference to quadrat size would be predicting the same occurrence probability within a large or small quadrat, which cannot make sense.

1.2.2. *Occurrence rate.* Since occurrence probability is only meaningful with reference to a specific quadrat size, it is a somewhat awkward quantity to model in a presence-only study. In this context it is more natural to estimate an occurrence *rate* or intensity: that is, a quantity with units of inverse area (e.g., $1/\mathrm{km}^2$) corresponding to the expected number of specimens *per unit area*. Under some simple stochastic models for species occurrence, including the Poisson process model considered here, specifying the occurrence rate is equivalent to specifying occurrence probability simultaneously for all quadrat sizes.

Unfortunately, a presence-only data set only affords us direct knowledge of the expected number of specimen *sightings* per unit area. The absolute sightings rate is reflected in the number of records in our data set, but, at best, this rate is only proportional to the occurrence rate discussed above, which typically is the real estimand of interest. We must assume that our sightings only constitute a small fraction of the species' population over our study region, possibly with repeated sightings of the same specimen. Without other data or assumptions we would have no way of knowing what this constant of proportionality might be.

In other words, the absolute *sightings* rate is observable but usually not of direct interest, while the absolute *occurrence* rate is interesting but not observable without another source of information. Using presence-only data alone, we can at best hope to estimate a relative, not absolute, occurrence rate. Even assuming that the sightings and occurrence rates are proportional is optimistic, since it rules out sampling bias like that in Figure 1, an issue we take up again in Section 2.5.

1.3. *Notation.* We now introduce notation we will use for the remainder of the article. We begin with some geographic domain of interest $\mathcal{D}$, typically a bounded subset of $\mathbb{R}^2$. If the time of an observation is an important variable, we might alternatively take $\mathcal{D} \subseteq \mathbb{R}^3$, so that our observations have both space and time coordinates. Associated with each geographic location $z \in \mathcal{D}$ is a vector $x(z)$ of measured features.

Our presence-only data set consists of n_1 locations of sightings $z_i \in \mathcal{D}$ for $i = 1, 2, \ldots, n_1$, accompanied by n_0 "background" observations z_i for $i = n_1 + 1, \ldots, n_1 + n_0$ (typically a regular grid or uniformly random sample from $\mathcal{D}$). Finally, let $x_i = x(z_i)$ be the features for observation i, and y_i be a 0/1 indicator that i is a presence sample. Our treatment of these data as random or fixed will vary throughout the article.

1.4. *Outline.* The rest of the paper is organized as follows. In Section 2 we define the log-linear inhomogeneous Poisson process (IPP) model and its application to presence-only data, with special focus on interpreting its parameters and their maximum likelihood estimates. In particular, the estimate of the intercept α reflects nothing more than the total number of presence samples and, as such, is typically not of scientific relevance for the reasons discussed in Section 1.2.2. In fact, IPP model estimation amounts to parametric density estimation in an exponential family model, followed by multiplication of the fitted density by n_1. The density thus obtained reflects the relative rate of sightings as a function of geographic coordinates z.

Aarts, Fieberg and Matthiopoulos (2012) showed that many methods in species distribution modeling can be motivated by the IPP model. We review these connections and generalize them for several illuminating examples. In Section 3 we consider a particularly important example, showing that the popular Maxent method of Phillips, Dudík and Schapire (2004) follows immediately from partially maximizing the IPP log-likelihood with respect to α, a result which is explored further in Renner and Warton (2013). Hence, given any set of presence and background points, the Maxent and IPP methods obtain identical estimates for the slope $\hat{\beta}$ and for the density.

In Section 4 we discuss so-called "naive" logistic regression and its connections to the IPP model. We derive its likelihood as a conditional form of the IPP likelihood, but show that if the log-linear model is misspecified this convergence may not occur until the background sample is quite large. The need for a large background sample is due not only to variance, but also to bias that persists until the proportion n_1/n_0 becomes negligibly small. We show, however, that if we upweight all the background samples by large weight $W \gg 1$, we can use logistic regression to recover the IPP estimate $\hat{\beta}$ precisely with any finite presence and background sample. This procedure, which we call "infinitely weighted logistic regression," is a device for using GLM software to maximize the IPP log-likelihood. Section 5 recapitulates the relationships and contains discussion.

2. The inhomogeneous Poisson process model.

The IPP is a simple model for a random set of points $\mathbf{Z}$ falling in some domain $\mathcal{D}$. Both the number and locations of points are random. It can be defined by its intensity function

$$\lambda : \mathcal{D} \longrightarrow [0, \infty), \tag{1}$$

which indexes the likelihood that a point falls at or near z. For $A \subseteq \mathcal{D}$, write

$$\Lambda(A) = \int_A \lambda(z)\, dz \tag{2}$$

and assume $\Lambda(\mathcal{D}) < \infty$.

There are two main ways to formally characterize an IPP with intensity λ. One simple definition is that the total number of points is a Poisson random variable with mean $\Lambda(\mathcal{D})$ and, conditionally on the number of points, their locations are independent and identically distributed with density $p_\lambda(z) = \lambda(z)/\Lambda(\mathcal{D})$. That is, an IPP is an i.i.d. sample from p_λ whose size is itself random.[3]

Alternatively, we can think of an IPP as a continuous limit of an independent Poisson count model for ever-finer discretizations of $\mathcal{D}$. If $N(A) = \#(\mathbf{Z} \cap A)$, the number of points falling in set A, then

$$N(A) \sim \text{Poisson}(\Lambda(A)) \tag{3}$$

with $N(A)$ and $N(B)$ independent for disjoint sets A and B. For more on the IPP and other point process models, see Gaetan and Guyon (2009) or Cressie (1993).

In the case of a finite discrete domain $\mathcal{D} = \{z_1, z_2, \dots, z_m\}$, the IPP model reduces to a discrete Poisson model, with $N(z_i) \sim \text{Poisson}(\lambda(z_i))$. In this sense, the IPP model may be seen as a limit of finer and finer discretizations of $\mathcal{D}$. We discuss this connection further in Section 2.4.

Warton and Shepherd (2010) proposed modeling species sightings $z_1, \dots, z_{n_1}$ as arising from an IPP whose intensity is log-linear in the features $x(z)$:

$$\lambda(z) = e^{\alpha + \beta' x(z)}. \tag{4}$$

The formal linearity assumption is less restrictive than it seems, since our features $x(z)$ could include polynomial terms, interactions, splines or other basis expansions, which substantially broaden the space of possible $\lambda(z)$.

Interpreting the IPP as an i.i.d. sample with random size, we see that α and β play very different roles. Since α only multiplies $\lambda(z)$ by a constant, it has no effect on $p_\lambda(z) = \lambda(z)/\Lambda(\mathcal{D})$. The "slope" parameters β completely determine p_λ, while α scales the intensity up or down to determine the expected sample size $\Lambda(\mathcal{D})$.

2.1. *Geographic space and feature space.* In the context of logistic regression, it can be more natural to think of the x_i as a sample of points in "feature space" [i.e., the range of $x(z)$] rather than as the features corresponding to a sample in the geographic domain $\mathcal{D}$. There is no real distinction between these two viewpoints, so long as we adjust for the fact that some values of x are more common in $\mathcal{D}$ than others.

Let $A_x = \{z : x(z) = x\}$ and $h(x) = \int_{A_x} 1\, dz$. Then if the set $\mathbf{Z}$ is an IPP with intensity $\lambda(x(z))$, the corresponding set $x(\mathbf{Z})$ is an IPP with intensity $\lambda_x(x) = \lambda(x) \cdot h(x)$ and, conditionally on n_1, their distribution is $p_x(x) \propto p_\lambda(x) \cdot h(x)$. For more detailed discussion see Elith et al. (2011) and Johnson et al. (2006).

[3] Cressie (1993) and Aarts, Fieberg and Matthiopoulos (2012) refer to an IPP conditioned on n_1 as a "Conditional IPP"; this is exactly an i.i.d. sample of size n_1 from the density $p_\lambda(z)$.

2.2. *Maximum likelihood for the IPP.* The score equations for the log-linear IPP are simple and enlightening. The IPP log-likelihood in terms of the presence samples is

$$\ell(\alpha,\beta)=\sum_{i:y_i=1}(\alpha+\beta' x_i)-\int_{\mathcal{D}} e^{\alpha+\beta' x(z)}\,dz-\log n_1!. \tag{5}$$

Differentiating with respect to α, we obtain the score equation

$$n_1=\int_{\mathcal{D}} e^{\alpha+\beta' x(z)}\,dz=\Lambda(\mathcal{D}). \tag{6}$$

That is, whatever $\hat{\beta}$ is, $\hat{\alpha}$ plays the role of a "normalizing" constant guaranteeing that $\lambda(z)$ integrates to n_1, the number of total presence records. Hence, if n_1 is not of scientific interest, then neither is $\hat{\alpha}$.

Solving for α in (6) and ignoring constants, we obtain the partially maximized log-likelihood

$$\ell^*(\beta)=\sum_{i:y_i=1}\left(\beta' x_i-\log\int_{\mathcal{D}} e^{\beta' x(z)}\,dz\right)=\sum_{i:y_i=1}\log p_\lambda(z_i), \tag{7}$$

which is the same log-likelihood we would obtain by conditioning on n_1 and treating the z_i as a random sample with density $p_\lambda(z)=\frac{e^{\beta' x(z)}}{\int_{\mathcal{D}} e^{\beta' x(z)}\,dz}$.

Finally, differentiating (7) with respect to β and dividing by n_1 gives the remaining score equations:

$$\frac{1}{n_1}\sum_{i:y_i=1} x_i=\frac{\int_{\mathcal{D}} e^{\beta' x(z)}x(z)\,dz}{\int_{\mathcal{D}} e^{\beta' x(z)}\,dz}=\mathbb{E}_{p_\lambda}x(z). \tag{8}$$

Solving (8) amounts to finding β for which the expectation of $x(z)$ under $p_\lambda(z)$ matches the empirical mean over the presence samples.

Hence, maximum likelihood for a log-linear IPP may be thought of as an algorithm with two discrete steps:

1. Estimate the density p_λ: find $\hat{\beta}$ for which $\mathbb{E}_{\hat{p}_\lambda}x(z)$ matches the empirical means of the presence sample x_i.
2. Multiply $\hat{p}_\lambda$ by n_1: find $\hat{\alpha}$ for which $\hat{\lambda}(z)=n_1\cdot\hat{p}_\lambda(z)$.

Unless n_1 is meaningful, then, the IPP is essentially density estimation. In our view, it is rare that n_1 merits much scientific interest, but there are important cases where it might. For instance, if we are comparing multiple species, study areas or periods of study, and if we believe that sampling effort is comparable across the different studies, then comparing the n_1 from each data set may teach us something.

Note, however, that in each of these cases our inference target can be viewed as a relative intensity *across* the different data sets. If we wish to make such comparisons, the right approach may simply be to expand the survey area $\mathcal{D}$ to include multiple regions or time periods and add region identity or species identity as a feature, then perform a combined analysis. n_1 for the combined analysis (the total number of sightings across all the different data sets) would then typically not be of much interest.

2.3. *Numerical evaluation of the integral.* When we cannot evaluate the integrals in equations (5)–(8) analytically, we replace them with numerical integrals based on the background samples. Hence, (5) becomes

$$\ell(\alpha, \beta) = \sum_{i:y_i=1} \alpha + \beta' x_i - \frac{|\mathcal{D}|}{n_0} \sum_{i:y_i=0} e^{\alpha+\beta' x_i} - \log n_1!, \tag{9}$$

where $|\mathcal{D}| = \int_{\mathcal{D}} 1\, dz$ represents the total area of the region.

The background points may be either a uniform sample from $\mathcal{D}$ or a regular grid. Quadrature weights may also be assigned to the background points to approximate the integral with a weighted sum, instead of the unweighted sum represented above.

We could repeat the derivation of Section 2.2 to obtain the criteria

$$\frac{|\mathcal{D}|}{n_0} \sum_{i:y_i=0} e^{\alpha+\beta' x_i} = n_1, \qquad \frac{\sum_{i:y_i=0} e^{\beta' x_i} x_i}{\sum_{i:y_i=0} e^{\beta' x_i}} = \frac{1}{n_1} \sum_{i:y_i=1} x_i. \tag{10}$$

Throughout, we will refer to (9) as the numerical IPP log-likelihood to distinguish it from (5). In practice, fitting the IPP means solving (10) for some background sample.

2.4. *Connection to Poisson log-linear model.* If the background z_i comprise a regular grid, we can discretize $\mathcal{D}$ into n_0 pixels A_i, each of roughly the same size $\frac{|\mathcal{D}|}{n_0}$ and centered at z_i. If $x(z)$ is continuous, then

$$\Lambda(A_i) = \int_{A_i} e^{\alpha+\beta' x(z)}\, dz \approx \frac{|\mathcal{D}|}{n_0} e^{\alpha+\beta' x_i}. \tag{11}$$

The IPP model implies that the counts $N(A_i)$ arise independently via

$$N(A_i) \sim \text{Poisson}(\Lambda(A_i)) \approx \text{Poisson}\left(\frac{|\mathcal{D}|}{n_0} e^{\alpha+\beta' x_i}\right). \tag{12}$$

Hence, the approximate log-likelihood is

$$\begin{aligned} \tilde{\ell}(\alpha, \beta) = \sum_{i:y_i=0} N(A_i)(\alpha + \beta' x_i) - \frac{|\mathcal{D}|}{n_0} \sum_{i:y_i=0} e^{\alpha+\beta' x_i} \\ - \sum_{i:y_i=0} \log N(A_i)!. \end{aligned} \tag{13}$$

Let $S_i = \{k : z_k \in A_i, y_k = 1\}$ contain the presence samples in pixel i. Then

$$\sum_{i:y_i=0} N(A_i)(\alpha + \beta' x_i) \approx \sum_{i:y_i=0} \sum_{k \in S_i} \alpha + \beta' x_k = \sum_{k:y_k=1} \alpha + \beta' x_k. \tag{14}$$

Hence, the only difference between (9) and (13) is that in the latter we also discretize the location of each presence sample to match its nearest background point.

Berman and Turner (1992) proposed using this approximation to fit the IPP model using Poisson GLM software, and Baddeley and Turner (2000) show how to generalize it to other point-process models including generalized additive models. This device provides a simple means of accessing the modeling flexibility of GLM methods at a cost of some loss of data, since it effectively replaces the covariate vector x_i for each presence sample with that of its nearest background sample.

Baddeley et al. (2010) discuss the bias incurred by the discretization, showing in particular that it vanishes in the small-pixel limit. They also propose a strategy for improving the bias, which splits pixels into subpixels whose covariates are closer to constant.

As we will see later, this discretization is not really necessary. In Section 4 we propose a different procedure, infinitely weighted logistic regression, that also allows us to fit an IPP model using GLM software but produces exactly the same estimates we would obtain by maximizing (9) on the original presence and background data.

2.5. *Identifiability and sampling bias.* Sampling bias poses a serious challenge to valid inference in presence-only studies. Scientifically, we are interested in the *occurrence process* consisting of all specimens of the species of interest. However, our data set consists of what we might call the *sightings process*, consisting only of the occurrences observed and reported by people.

We can model the sightings process as an occurrence process *thinned* by incomplete observation, as proposed by Chakraborty et al. (2011) and Renner and Warton (2013). That is, suppose that specimens occur with intensity $\tilde{\lambda}(z)$, but that most occurrences go unobserved. Each occurrence is observed with probability $s(z)$, which may depend on features of the geographic location z (e.g., proximity to the road network). If detection is independent across occurrences, then the observation process is an IPP with intensity

$$\lambda(z) = \tilde{\lambda}(z) \cdot s(z). \tag{15}$$

The trouble is that our presence-only data set only directly reflects λ, the intensity of sightings, and not $\tilde{\lambda}$.

Optimistically, we might assume that s is constant (no sampling bias). In that case, by estimating $\lambda(z)$ we are also estimating $\tilde{\lambda}(z)$ up to an unknown constant of proportionality s, so $p_{\tilde{\lambda}} = p_\lambda$ but $\tilde{\lambda} \neq \lambda$. Even in this optimistic scenario we can only estimate relative, not absolute, occurrence intensities. Phillips and Elith (2013) also elaborate the same point in the context of logistic regression models.

Slightly less optimistically, we might assume that s is an unknown function of z, but that s and $\tilde{\lambda}$ are known to depend on z through two disjoint feature sets. For instance, we could model $\tilde{\lambda}$ and s as log-linear in features $x_1(z)$ and $x_2(z)$, respectively:

$$\lambda(z) = \tilde{\lambda}(z)s(z) \tag{16}$$

$$= e^{\tilde{\alpha}+\tilde{\beta}' x_1(z)} e^{\gamma+\delta' x_2(z)}. \tag{17}$$

Then the sightings process follows the log-linear model $\lambda(z) = e^{\alpha+\beta' x(z)}$ with $\alpha = \tilde{\alpha} + \gamma$, $x = \binom{x_1}{x_2}$ and $\beta = \binom{\tilde{\beta}}{\delta}$. Note that $\tilde{\alpha}$ and $\tilde{\beta}$ are the quantities of primary scientific interest, whereas α and β are the parameters governing the process we actually observe. Nevertheless, $\tilde{\beta}$ is still identifiable from the data because β is.[4]

As $n_0, n_1 \to \infty$, our estimate $\hat{\beta}_1$ converges to the true value of $\tilde{\beta}$, the slope coefficients of $\tilde{\lambda}$. However, $\hat{\alpha}$ will converge not to $\tilde{\alpha}$ but rather to $\tilde{\alpha} + \gamma$. Without knowing γ, we have no way of estimating $\tilde{\alpha}$. By the same token, if some features appear both in x_1 and x_2—or if x_1 and x_2 are not linearly independent—the model is unidentifiable.

To be concrete, suppose koala occurrence is known to depend only on elevation (x_1), and that sampling bias is known to depend only on proximity to roads (x_2). Then, despite the obvious sampling bias in Figure 1, we could still estimate what elevations koalas tend to frequent, by making the correct adjustments for road proximity. By contrast, we could not estimate from presence-only data alone whether koalas tend to avoid roads, since that is confounded by sampling bias.

Whether or not s is constant, our estimate for $\alpha = \tilde{\alpha} + \gamma$ carries no real information about $\tilde{\alpha}$ unless we have independent knowledge of γ. Indeed, we have already seen that the only role $\hat{\alpha}$ plays in estimation is to make λ integrate to n_1.

The distinction between β and $\tilde{\beta}$ may be very important for some problems, but for the remainder of this article we focus on estimation of β, the slope parameters of the process we get to observe.

3. Maximum entropy.

Another popular approach to modeling presence-only data, which we will see is equivalent to the IPP, is the Maxent method proposed by Phillips, Dudík and Schapire (2004). The authors begin by

[4] As with any regression adjustment scheme, we should proceed with caution here. If our linear model is misspecified (perhaps we should have included x_2^2) and x_1 is correlated with the missing variables, even regression adjustment will not remove all bias. In perverse situations it could even make the situation worse.

assuming that the presence samples $z_1, \dots, z_{n_1}$ are a random sample from some probability distribution $p(z)$, called the species distribution.

The authors adopt the view, inspired by information theory, that our estimate $\hat{p}$ should have large entropy $H(p) = -\int_{\mathcal{D}} p(z) \log(p(z))\, dz$. Large $H(p)$ means roughly that p is close to the uniform density $1/|\mathcal{D}|$, the species distribution we would observe if the species were indifferent to all geographic features. The idea is that $\hat{p}$ should be "nearly geographically uniform," subject to constraints that make it resemble the observed data.

Phillips, Dudík and Schapire (2004) propose to choose the p which maximizes $H(p)$ subject to the constraint that the expectation of the features $x(z)$ under $\hat{p}$ matches the sample mean of those features, that is,

$$\frac{1}{n_1} \sum_{y_i=1} x_i = \int_{\mathcal{D}} x(z) \hat{p}(z)\, dz = \mathbb{E}_{\hat{p}} x(z). \tag{18}$$

They show that this criterion is equivalent to maximizing the likelihood of the parametric exponential family density:

$$p(z) = \frac{e^{\beta' x(z)}}{\int_{\mathcal{D}} e^{\beta' x(u)}\, du}. \tag{19}$$

This is exactly the form of p_λ for our log-linear IPP, and its log-likelihood is exactly the partially maximized log-likelihood $\ell^*(\beta)$, the log-likelihood for an IPP conditioned on n_1. The constraint (18) is precisely the score criterion (8) for β in an IPP, so the Maxent $\hat{\beta}$ is the same as the IPP $\hat{\beta}$. This result may also be found in Appendix A of Aarts, Fieberg and Matthiopoulos (2012).

The popular software package Maxent implements a method slightly more complex than the one originally proposed in 2004. First, it automatically generates a large basis expansion of the original features into many derived features: quadratic terms, interactions, step functions and hinge functions of the original features. Then, it fits a model by optimizing an ℓ_1-regularized version of the conditional IPP likelihood (7):

$$\sum_{y_i=1} \beta' x_i - n_1 \log\left(\int_{\mathcal{D}} e^{\beta' x(z)}\, dz \right) - \sum_j r_j |\beta_j|. \tag{20}$$

The regularization parameters r_j are chosen automatically according to rules based on an empirical study of various presence-only data sets [Phillips and Dudík (2008)].[5]

Mathematically, the basis expansion increases the dimension of $x(z)$ but changes nothing else. Moreover, the ℓ_1 regularization scheme does not constitute an essential difference with the other methods considered here. One

[5] The notation of the Maxent papers uses λ and β to denote what we call β and r, respectively.

could (and often should) regularize β when fitting an IPP model as well, especially if $x(z)$ contains many features resulting from a large basis expansion.

Penalizing the Maxent log-likelihood does not change the equivalence between the two models, so long as α is left unpenalized. If we add a penalty term $J(\beta)$ to the IPP log-likelihood (5), we still obtain (6) after differentiating with respect to α. Then, partially maximizing $\ell(\alpha, \beta) - J(\beta)$ gives us $\ell^*(\beta) - J(\beta)$, the penalized Maxent log-likelihood. This equivalence depends on our not penalizing α in (5).

This argument generalizes immediately to a generic penalized likelihood method with any parametric form for $\log \lambda(z)$. We have established the following general proposition:

PROPOSITION 1. *Given some parametric family of real-valued functions* $\{f_\theta : \theta \in \mathbb{R}^d\}$ *with penalty function* $J(\theta)$, *consider the penalized log-likelihood* g_1 *for an IPP with intensity* $e^{\alpha + f_\theta(x(z))}$,

$$g_1(\alpha, \theta) = \left(\sum_{y_i=1} \alpha + f_\theta(x_i) \right) - \int_{\mathcal{D}} e^{\alpha + f_\theta(x(z))}\, dz - J(\theta) - \log n_1! \tag{21}$$

and the penalized log-likelihood g_2 *for a sample with density* $\propto e^{f_\theta(x(z))}$:

$$g_2(\theta) = \sum_{y_i=1} f_\theta(x_i) - n_1 \log\left(\int_{\mathcal{D}} e^{f_\theta(x(z))}\, dz \right) - J(\theta). \tag{22}$$

Then θ *maximizes* g_2 *iff* (α, θ) *maximize* g_1 *for some* α. *The same applies if we replace the integrals in (21)–(22) with sums over the background sample.*

PROOF. Partially maximize g_1 over α as in (7) to obtain g_2. □

Thus, we see that, while Maxent and the IPP appear to be different models with different motivations, they result in the exact same density estimate $\hat{p}_\lambda(z)$. In terms of the two-step algorithm we derived in Section 2.2, Maxent is identical to step 1, but it skips step 2. The IPP fit $\hat{\lambda}$ is n_1 times the Maxent fit $\hat{p}$.

4. Logistic regression. Another ostensibly different model for presence-only data is so-called "naive" logistic regression, which casts presence-only modeling as a problem of classifying points as presence $(y = 1)$ or background $(y = 0)$ on the basis of their features. The logistic regression model treats n_1, n_0 and the x_i as fixed and the y_i as random with

$$\mathbb{P}(y_i = 1 | x_i) = \frac{e^{\eta + \beta' x_i}}{1 + e^{\eta + \beta' x_i}}. \tag{23}$$

Superficially, this approach may appear ad hoc and unmotivated compared to IPP or Maxent. Nevertheless, it has enjoyed some popularity, in part because logistic regression is an extremely mature method in statistics, enjoying myriad well-understood and already-implemented extensions such as GAM, MARS, LASSO, boosted regression trees and more.

Logistic regression modeling of presence-only data has often been motivated by analogy to logistic regression for presence–absence data. Since it is not known whether the species is present at or near the background examples, these are sometimes referred to as "pseudo-absences," and the supposed naivete of the method is that it appears to treat background samples as actual absences. For instance, Ward et al. (2009) introduced latent variables coding "true" presence or absence and proposed fitting this model via the EM algorithm.

This interpretation raises once again the troublesome question of what it would mean for one of our randomly sampled background points to be a "true presence." Need there be a specimen sitting directly on the location, or is it enough for it to be within 100 m? 1 km?

Fortunately, we can sidestep these concerns, since connections between the logistic regression and IPP models yield a more straightforward interpretation.

4.1. *Case-control sampling.* Suppose the background data are a uniform random sample, and the presence data arise from a log-linear IPP. Then if we condition on n_1, the z_i are a mixture of two i.i.d. samples, one from density $e^{\alpha+\beta' x(z)}/\Lambda(\mathcal{D})$ and the other from density $1/|\mathcal{D}|$. By Bayes' rule, for a random index i,

$$\mathbb{P}(y_i = 1|z_i) = \frac{\mathbb{P}(y_i = 1)\mathbb{P}(z_i|y_i = 1)}{\mathbb{P}(y_i = 0)\mathbb{P}(z_i|y_i = 0) + \mathbb{P}(y_i = 1)\mathbb{P}(z_i|y_i = 1)} \tag{24}$$

$$= \frac{n_1 e^{\alpha+\beta' x_i}/\Lambda(\mathcal{D})}{n_0/|\mathcal{D}| + n_1 e^{\alpha+\beta' x_i}/\Lambda(\mathcal{D})} \tag{25}$$

$$= \frac{e^{\eta+\beta' x_i}}{1 + e^{\eta+\beta' x_i}}, \tag{26}$$

with $e^\eta = \frac{n_1 e^\alpha |\mathcal{D}|}{n_0 \Lambda(\mathcal{D})}$. Since $\mathbb{P}(y_i = 1|z_i)$ depends only on $x_i = x(z_i)$, we could just as well condition on x_i instead, giving (23). Therefore, if the log-linear IPP model is correct, it implies the individual $y_i|x_i$ follow a logistic regression with the same slope parameters β.[6]

[6] The y_i are technically not conditionally independent (if we knew the other $n_1 + n_0 - 1$ labels, we would know the last as well). This is always true in case-control studies, but it is typically ignored since the dependence is weak for large samples.

Thus, given any finite sample of presence and background points, if we believe in the IPP model, then we could either maximize the numerical IPP likelihood or the logistic regression likelihood, and in either case we would be estimating the same population parameter β. This does not guarantee we will obtain the same estimates $\hat{\beta}$ in any given finite sample, but if the model is correct, then either method gives a consistent estimator of β.

Note that if we change the marginal class ratio n_1/n_0 by some factor e^c, the only effect will be to multiply the odds of $y_i = 1$ given x_i by the same factor, that is, add c to η and leave β unchanged. Hence, under correct specification, $\hat{\beta} \to \beta$ regardless of the limiting ratio n_1/n_0.

4.2. *Case-control sampling under misspecification.* Now, suppose that $\lambda(z)$ is not really log-linear in our features x. Then, the fitted slopes $\hat{\beta}$ for logistic regression and the numerical IPP will not converge to the same limiting β if n_1 and n_0 grow large together. In fact, the limiting logistic regression parameters depend on the limiting ratio of n_1/n_0 [Xie and Manski (1989)].

To gain some intuition for why this is so, suppose we have a single covariate x, with $\lambda(z) = e^{\alpha + x(z)^2}$. Then the derivation of (24)–(26) gives

$$\mathbb{P}(y_i = 1 | x_i) = \frac{e^{\eta + x_i^2}}{1 + e^{\eta + x_i^2}} \tag{27}$$

with η as before. In the large-sample limit, then, our estimation problem amounts to finding $\hat{\eta}, \hat{\beta}$ for which

$$\hat{\eta} + \hat{\beta} x \approx \eta + x^2 = \log \frac{n_1 |\mathcal{D}|}{n_0 \Lambda(\mathcal{D})} + x^2 \tag{28}$$

in the population from which we are sampling. Now, since changing n_1/n_0 only adds a vertical shift to the right-hand side of (28), it may seem rather counterintuitive that this should have any impact on the *slope* $\hat{\beta}$ of our approximation on the left-hand side.

To understand why, we must come to grips with the sense in which we make the approximation in (28). The logistic regression log-likelihood is

$$\ell_{\mathrm{LR}}(\eta, \beta) = \sum_i (\eta + \beta' x_i) y_i - \sum_i \log(1 + e^{\eta + \beta' x_i}). \tag{29}$$

Its first derivatives with respect to η and β can be written in terms of the fitted conditional probabilities $\hat{y}_i(\eta, \beta) = \mathbb{P}_{\eta,\beta}(y = 1 | x = x_i)$:

$$\frac{\partial \ell_{\mathrm{LR}}}{\partial \eta} = \sum_i \left(y_i - \frac{e^{\eta + \beta' x_i}}{1 + e^{\eta + \beta' x_i}} \right) = \sum_i (y_i - \hat{y}_i), \tag{30}$$

$$\frac{\partial \ell_{\mathrm{LR}}}{\partial \beta} = \sum_i x_i \left(y_i - \frac{e^{\eta + \beta' x_i}}{1 + e^{\eta + \beta' x_i}} \right) = \sum_i x_i (y_i - \hat{y}_i). \tag{31}$$

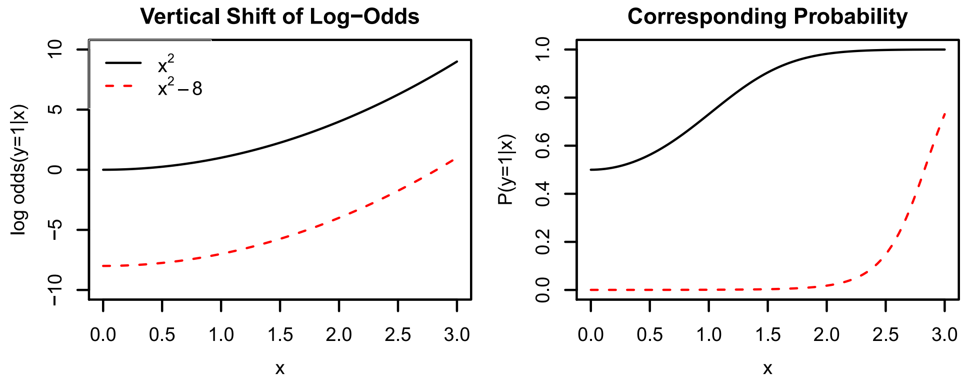

FIG. 3. *The dashed red curve in the left panel is a vertical shift of the solid black curve. However, vertically shifting the log-odds changes the conditional probability in a more complex way.*

If we define $r_i = y_i - \hat{y}_i$, then $\hat{\eta}, \hat{\beta}$ maximize the likelihood if and only if $\sum_i r_i = 0$ and $x \perp r$. The crucial point is that the residuals of our approximation, $y_i - \hat{y}_i$, are measured on the probability scale, and not the log-odds scale.

The black and red curves in the left panel of Figure 3 show the conditional log-odds $\log \frac{\mathbb{P}(y_i=1|x_i=x)}{\mathbb{P}(y_i=0|x_i=x)}$ for our misspecified model with two different values of η, 0 and -8, respectively. On the log-odds scale, one is no steeper than the other. But when we look at the same two curves on the conditional probability scale (right panel), now the red looks steeper than the black. This is due to a "ceiling" effect for the black curve: in the region where the log-odds x^2 is changing fast, the probability $\hat{y} = \frac{e^{x^2}}{1+e^{x^2}}$ has already saturated at 1. The actual estimates of $\hat{\eta}$ and $\hat{\beta}$ depend on the background density of x as well as n_1/n_0; see Section 4.5 for a full simulation.

As Warton and Shepherd (2010) prove, this ceiling effect vanishes in the limit where $n_1/n_0 \to 0$; in that case $\hat{\eta} \to -\infty$, $\hat{y}_i = \frac{e^{\hat{\eta}+\hat{\beta}}}{1+e^{\hat{\eta}+\hat{\beta}}} \approx e^{\hat{\eta}+\hat{\beta}}$, and the logistic regression and IPP estimates are identical. Hence, there is no difference when the background sample grows so large that it dwarfs the presence records in the population from which we are sampling. Dorazio (2012) considers a similar framework, called the case-augmented design, and proves a similar equivalency to the IPP as $n_0 \to \infty$.

4.3. *Infinitely weighted logistic regression.* If we modify the logistic regression procedure a bit, we can resolve the discrepancy in the previous section and recover the same $\hat{\beta}$ that we would estimate with an IPP using the same presence and background samples.

We can remove the ceiling effect of the previous section if we add case weights to the samples

$$w_i = \begin{cases} W, & y_i = 0, \\ 1, & \text{otherwise}, \end{cases} \tag{32}$$

for some large number W. We then obtain the weighted log-likelihood

$$\ell_{\mathrm{WLR}}(\eta, \beta) = \sum_i w_i[y_i(\eta + \beta' x_i) - \log(1 + e^{\eta+\beta' x_i})] \tag{33}$$

$$= \sum_{i:y_i=1} \eta + \beta' x_i - \sum_i W^{1-y_i} \log(1 + e^{\eta+\beta' x_i}). \tag{34}$$

Proposition 2. *Let $J(\beta)$ be any convex penalty, and suppose $\ell_{\mathrm{IPP}}(\alpha, \beta) - J(\beta)$ has a unique maximizer $(\hat\alpha_{\mathrm{IPP}}, \hat\beta_{\mathrm{IPP}})$. Then if $(\hat\eta_W, \hat\beta_W)$ maximize $\ell_{\mathrm{WLR}}(\eta, \beta) - J(\beta)$ for weight W,*

$$\lim_{W\to\infty} \hat\beta_W = \hat\beta_{\mathrm{IPP}}. \tag{35}$$

Proof. Reparameterizing (33) with $\alpha = \eta + \log(W n_0/|\mathcal{D}|)$ and ignoring constants, we obtain

$$\begin{aligned} \ell_{\mathrm{WLR}}(\alpha, \beta) = {} & \sum_{i:y_i=1} \alpha + \beta' x_i - \sum_{i:y_i=0} W \log\left(1 + \frac{|\mathcal{D}|}{W n_0} e^{\alpha+\beta' x_i}\right) \\ & - \sum_{i:y_i=1} \log\left(1 + \frac{|\mathcal{D}|}{W n_0} e^{\alpha+\beta' x_i}\right). \end{aligned} \tag{36}$$

Fixing (α, β) and taking $W \to \infty$, each term in the second sum converges to $\frac{|\mathcal{D}|}{n_0} e^{\alpha+\beta' x_i}$ while the third sum converges to 0. Hence, ignoring constants, (36) converges to the numerical IPP log-likelihood (9), and this convergence occurs uniformly on compact subsets of the parameter space.

Now, both $\ell_{\mathrm{WLR}}(\alpha, \beta) - J(\beta)$ and $\ell_{\mathrm{IPP}}(\alpha, \beta) - J(\beta)$ are concave, and the latter is strictly concave by assumption; hence, the maximizer of the first converges to the maximizer of the second. $\square$

From the above, we see that IWLR is not really a new statistical method, but rather a technical device for optimizing the IPP/Maxent log-likelihood using already-implemented GLM software.

Although technically $\hat\beta_W \neq \hat\beta_{\mathrm{IPP}}$ for any finite W (hence the name "infinitely weighted"), in practice, we only need W large enough that the approximation of $\ell_{\mathrm{WLR}}(\alpha, \beta)$ to $\ell_{\mathrm{IPP}}(\alpha, \beta)$ is good near $(\hat\alpha, \hat\beta)$.

Essentially, if $\frac{|\mathcal{D}|}{W n_0} e^{\alpha+\beta' x_i} \approx 0$ for each i (say, all are less than 0.001), then the Taylor approximation should be good. We can assess this easily if we observe that

$$\hat{y}_i = \frac{|\mathcal{D}| e^{\hat{\alpha}+\hat{\beta}' x_i}/(W n_0)}{1 + |\mathcal{D}| e^{\hat{\alpha}+\hat{\beta}' x_i}/(W n_0)} \approx \frac{|\mathcal{D}|}{W n_0} e^{\hat{\alpha}+\hat{\beta}' x_i}, \tag{37}$$

when all of the above are small. To rephrase, then, if $\max_i \hat{y}_i$ from the logistic regression is less than 0.001 or so, it seems to us that W should be sufficiently large. If not, we can set $W \leftarrow \frac{\max_i \hat{y}_i}{0.001} W$ and check that the fitted $\hat{y}_i$ are now small enough. If any uncertainty remains whether W is large enough, one can always increase it by (say) another factor of 100 and check that the estimates do not change appreciably.

4.4. *Logistic regression as density estimation.* One interpretation of the results we have just reviewed is that in the context of presence-only data, logistic regression solves the same parametric density estimation problem as Maxent and the IPP do. Moreover, our infinitely weighted logistic regression yields an identical estimate of the density.

Using logistic regression for density estimation has been proposed before. For example, Hastie, Tibshirani and Friedman (2009) discuss it as a means for turning an unsupervised density estimation problem into a supervised classification problem. Their proposal uses a different weighting scheme (assigning half the total weight to the presence samples) which, unlike infinitely weighted logistic regression, does not give exactly the IPP solution.

4.5. *Simulation study: Weighted vs unweighted logistic regression.* We have seen that both infinitely weighted logistic regression (a.k.a. numerical IPP) and unweighted logistic regression estimate the same β parameter of the same log-linear IPP model, and when the background sample is much larger than the presence sample the estimates $\hat{\beta}$ are close to each other.

However, the infinitely weighted logistic regression estimate can converge much faster to the large-background-sample limit if the linear model is misspecified, as we illustrate here with a simulation study.

Consider a geographic region with a single covariate x whose background density is $p_0(x) = N(0, 1)$. Now, suppose a species follows our log-linear IPP model with slope β, so that $\lambda(x(z)) \propto e^{\beta x}$. Then the density of presence samples in feature space is $p_1(x) = e^{\beta x} p_0(x)/(\int e^{\beta u} p_0(u)\, du) = N(\beta, 1)$.

Suppose our species is in fact a mixture of two subspecies, one of which comprises 95% of the population and prefers x large, while the remaining 5% prefer x small. If each subspecies follows our model with coefficients 1.5 and -2, respectively, then

$$\lambda(x) \propto 0.95 e^{1.5x} + 0.05 e^{-2x}, \tag{38}$$

which no longer follows the log-linear model. $p_0(x)$ and $p_1(x)$ are depicted in the upper panel of Figure 4 as the dashed and solid black lines. The

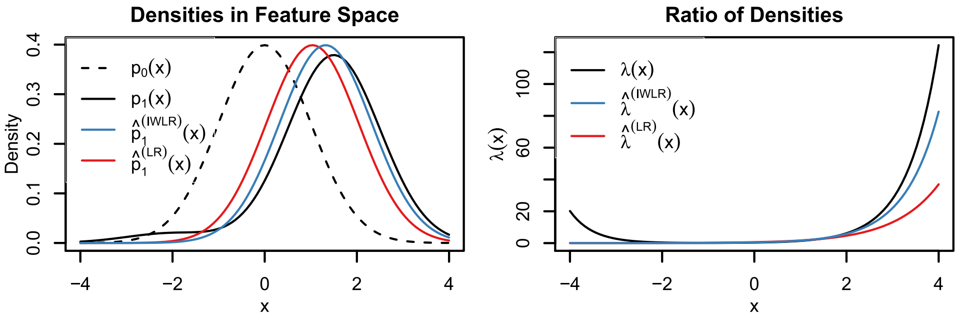

FIG. 4. *Large-sample estimates for the simulation study, misspecified case. The black curves represent the true presence density (left panel) and intensity (right panel). The blue and red curves show the fitted densities using IWLR and standard logistic regression with $n_0 = n_1$.*

black line in the left panel shows $\lambda(x) = p_1(x)/p_0(x)$, the relative intensity as a function of the covariate x. In the left panel all the curves have been normalized so that $\Lambda(\mathcal{D}) = \int \lambda(x) p_0(x)\,dx = 1$.

If we fit an infinitely-weighted logistic regression (or, equivalently, a log-linear IPP) to a large presence and background sample, our fitted $\hat{\beta}^{(\mathrm{IWLR})}$ will tend to $\mu_1 = \mathbb{E}_{p_1}(x) = 1.325$. We have plotted the corresponding large-sample estimates $\hat{\lambda}^{(\mathrm{IWLR})}(x)$ and $\hat{p}_1^{(\mathrm{IWLR})}(x)$ as blue lines in the respective panels of Figure 4.

If, alternatively, we fit an unweighted logistic regression to the same data set with large $n_0 = n_1$, the estimate $\hat{\beta}^{(\mathrm{LR})}$ will tend to roughly 1.04. The resulting large-sample estimates $\hat{p}_1^{(\mathrm{LR})}(x)$ and $\hat{\lambda}^{(\mathrm{LR})}(x)$ are plotted in red.

If we fit an unweighted logistic regression to a large sample with a different ratio n_1/n_0, we would get a different estimate, which would tend toward the IPP estimate of 1.325 if and only if this ratio tended to 0. By the same token, when n_1 and n_0 are fixed, the ratio between them can play a significant role in determining the estimated β. In contrast, the IWLR/IPP estimate tends to 1.325 in large samples no matter what the ratio n_1/n_0.

The left panel of Figure 5 illustrates this with a simulation study of the example just discussed. We first generate a single presence sample of size $n_1 = 3000$ from this species, then generate 20 sets of n_0 background samples from $p_0 = N(0,1)$ for each of a range of values n_0 ranging from 10^3 to 10^6.

For each background sample, we fit both an "infinitely" weighted ($W = 10^4$) and unweighted logistic regression to the combination of presence and background points. For relatively large sizes of background sample, there is very little sampling variability, but the logistic regression estimates carry a large bias that depends greatly on the size of the background sample. The limiting $\hat{\beta}$, to which both methods would converge given an infinite background sample, is depicted with a horizontal line.

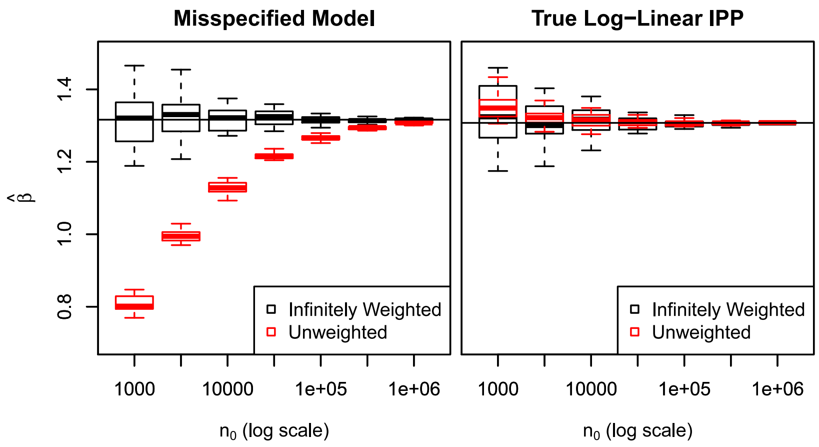

FIG. 5. *$\hat{\beta}$ estimates for simulation study with $n_1 = 3000$ and varying n_0. Unweighted logistic regression may require a very large background sample before convergence when the model is misspecified.*

In the right panel, we repeat this study with a presence sample from $N(\mu_1, 1)$, the correctly-specified model with the same mean as our misspecified model. Now the situation is very different; no matter what the mix of presence and background samples, the log-odds are truly linear with slope $\beta = \mu_1$. Consequently, $\hat{\beta}^{(\mathrm{LR})} \xrightarrow{p} \beta$ as $n_0 \to \infty$ and $n_1 \to \infty$, regardless of the limiting ratio n_1/n_0.

Since the choice of background sample size is primarily a matter of convenience, it is preferable to use an estimator that depends on it as little as possible. When the linear model is misspecified (which is nearly always the case), we recommend the infinitely weighted logistic regression over unweighted logistic regression for this reason.

We emphasize here that although IWLR resolves the issue of bias that we discussed in Section 4.2, using IWLR does *not* guarantee that we will obtain a good estimate for small n_0. The smaller n_0 is, the larger the variance of our estimate, so a larger background set is always better unless computational constraints apply.

What is more, the variability in our estimate due to the background sample is not reflected in the default standard error outputs from GLM software—only the variability due to the presence records is. Because $\ell_{\mathrm{IWLR}}(\alpha, \beta) \approx \ell_{\mathrm{IPP}}(\alpha, \beta)$ for large W, its Hessian will also converge to the Hessian of the IPP.

Even if our background sample was extremely large, the standard error estimates for any of the models we have discussed are based on asymptotic normal approximations that hold when the log-linear model is correctly specified. Resampling methods such as the bootstrap are more generally reliable, but even the bootstrap will depend crucially on the assumption that presence records (and in the case of logistic regression, background records) are *independent* observations. In terms of the IPP model, this assumption rules out spatial clustering of presence records. Renner and Warton (2013) provide evidence that this assumption may not hold for presence-only data. Therefore, model-based estimates of standard error should be viewed with suspicion no matter what method we choose.

5. Discussion.

We have discussed several closely related models for a single presence-only sample. In this section we collect them all in one place and review their relationships:

Inhomogeneous Poisson process. The "mother" model, from which the others can be derived, is the inhomogeneous Poisson process (IPP), whose log-likelihood is

$$\sum_{i:y_i=1} (\alpha + \beta' x_i) - \int_{\mathcal{D}} e^{\alpha + \beta' x(z)}\, dz. \tag{39}$$

In practice, (39) is approximated numerically via

$$\sum_{i:y_i=1} (\alpha + \beta' x_i) - \frac{|\mathcal{D}|}{n_0} \sum_{i:y_i=0} e^{\alpha + \beta' x_i}. \tag{40}$$

Fitting this model amounts to solving for the density $p_\lambda(z) \propto e^{\beta' x(z)}$ for which the expected features $\mathbb{E}_{p_\lambda} x(z)$ match the empirical mean $\frac{1}{n_1}\sum_{i:y_i=1} x_i$, then multiplying that density by n_1.

Maxent. Conditioning on n_1, we obtain the exponential family density model $p(z) \propto e^{\beta' x(z)}$, resulting in the log-likelihood

$$\sum_{i:y_i=1} \beta' x_i - n_1 \log\left(\int_{\mathcal{D}} e^{\beta' x(z)}\, dz\right) \tag{41}$$

or its numerical counterpart. This is the log-likelihood maximized by Maxent, and it corresponds exactly to the log-likelihood (39) partially maximized with respect to α. Hence, both procedures give exactly the same estimates of β and p.

Logistic regression. The logistic regression log-likelihood is

$$\sum_i y_i(\eta + \beta' x_i) - \log(1 + e^{\eta + \beta' x_i}). \tag{42}$$

When the log-linear IPP model is correctly specified, this model is as well (aside from the fact that the $y_i|x_i$ are only approximately independent), with the same true β as in the IPP model. However, in finite samples the estimates for β given by maximizing (42) instead of (40) may be substantially different.

Infinitely weighted logistic regression. We can resolve this difference by upweighting all the background points by $W \gg 1$, obtaining weighted log-likelihood

$$\sum_{i:y_i=1} (\eta + \beta' x_i) - \sum_i W^{1-y_i} \log(1 + e^{\eta + \beta' x_i}). \tag{43}$$

In the limit where $W \to \infty$, we recover exactly the same $\hat{\beta}$ as we would by maximizing (40).

Discretized Poisson LLM. Another means for approximating the IPP log-likelihood with a GLM log-likelihood is the Berman and Turner method, which simply discretizes geographic space into pixels and assigns each presence point to a bin belonging to its nearest background point:

$$\sum_{i:y_i=0} N(A_i)(\alpha + \beta' x_i) - \frac{1}{n_0} \sum_{i:y_i=0} e^{\alpha + \beta' x_i}. \tag{44}$$

This discretization of presence features is unnecessary given that we can exactly fit the IPP likelihood using the infinitely weighted approach of (43).

5.1. *Extending the IPP model.* Logistic regression is one of the most widely applied methods in statistics. For decades, applied statisticians have been developing, studying and using variations on logistic regression to solve classification problems in statistics. R packages exist for fitting generalized additive models (GAMs), boosted regression trees, MARS and every manner of tailored regularization schemes [see, e.g., Hastie, Tibshirani and Friedman (2009)].

All of these methods are well understood within the context of logistic regression. We believe that the most important practical implication of the finite-sample equivalence between the IPP model and infinitely weighted logistic regression is that all of these methods can now be equally well understood and easily applied within the context of the IPP model.

For instance, we can fit an IPP / Maxent version of boosted regression trees with the following single line of R:

```
boosted.ipp <- gbm(y~., family=‘‘bernoulli,’’
data=dat, weights=1E3^(1-y)).
```

For an IPP / Maxent version of LASSO, ridge, or the elastic net:[7]

```
lasso.ipp <- glmnet(dat.x, dat.y, family=‘‘binomial,’’
weights=1E3^(1-y)).
```

For an IPP GAM:

```
gam.ipp <- gam(y~s(x1)+x2, family=binomial, data=dat,
weights=1E3^(1-y)).
```

This added flexibility promises to provide a powerful tool to modelers of presence-only data.

[7] The user should be warned that `glmnet` automatically re-normalizes the weights so they sum to $n_0 + n_1$. To avoid issues, set `glmnet.control(pmin=1.0e-8, fdev=0)` in your R session, and keep in mind this renormalization when setting the Lagrange parameter λ.

5.2. *Model selection.* Regardless of which of the various related likelihoods we choose, there remains the issue of model selection. With the use of geographic information systems, ecologists often have access to a large number of predictor variables and may wish to winnow the field before modeling to avoid overfitting. Conversely, if some continuous variables are known to be important predictors, assuming a linear effect on the log-intensity may be too restrictive, and we may wish to expand the basis using splines, interactions, wavelets, etc. In either case, regularization may be called for.

Though it would be impossible to give a full treatment here of the many important considerations governing model selection, we note that these choices need not be governed by which likelihood we take as our starting point. In particular, the large set of derived features and ℓ_1 regularization used by Maxent software can just as well be applied to the IPP model or, for that matter, to logistic regression. Using the infinitely weighted logistic regression method, we can implement the exact loss function used by the Maxent with software for penalized GLMs.

Acknowledgments. The authors are grateful to Jane Elith for helpful discussions and suggestions. We were also very fortunate to have had very thorough and thoughtful reviewers; the current manuscript has benefited greatly from their attention.

REFERENCES

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

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Aust. N. Z. J. Stat.* **42** 283–322. MR1794056

Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., Shah, R. and Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. *Electron. J. Stat.* **4** 1151–1201. MR2735883

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with GLIM. *J. Appl. Stat.* **41** 31–38.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M. and Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 757–776. MR2844854

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. Revised reprint of the 1991 edition. MR1239641

Dorazio, R. M. (2012). Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics* **68** 1303–1312. MR3040037

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A. et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29** 129–151.

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

Gaetan, C. and Guyon, X. (2009). *Spatial Statistics and Modeling*. Springer, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

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

Lee, A. J., Scott, A. J. and Wild, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika* **93** 385–397. MR2278091

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

MacKenzie, D. I. (2006). *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, New York.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L. and Erickson, W. P. (2002). *Resource Selection by Animals: Statistical Analysis and Design for Field Studies*. Kluwer Academic, Dordrecht.

Margules, C. R., Austin, M. P., Mollison, D. and Smith, F. (1994). Biological models for monitoring species decline: The construction and use of data bases (with discussion). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **344** 69–75.

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. J., Dudík, M. and Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning* 83. ACM, New York.

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. and Elith, J. (2013). On estimating probability of presence from use-availability or presence-background data. *Ecology* **94** 1409–1419.

Renner, I. W. and Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69** 274–281.

Royle, J. A., Nichols, J. D. and Kéry, M. (2005). Modelling occurrence and abundance of species when detection is imperfect. *Oikos* **110** 353–359.

Ward, G., Hastie, T., Barry, S., Elith, J. and Leathwick, J. R. (2009). Presence-only data and the EM algorithm. *Biometrics* **65** 554–563. MR2751480

Warton, D. I. and Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Stat.* **4** 1383–1402. MR2758333

Xie, Y. and Manski, C. F. (1989). The logit model and response-based samples. *Sociol. Methods Res.* **17** 283–302.

Department of Statistics
Stanford University
390 Serra Mall
Stanford, California 94305-4065
USA
E-mail: wfithian@stanford.edu
hastie@stanford.edu