POISSON POINT PROCESS MODELS SOLVE THE "PSEUDO-ABSENCE PROBLEM" FOR PRESENCE-ONLY DATA IN ECOLOGY[1]

BY DAVID I. WARTON AND LEAH C. SHEPHERD

University of New South Wales

Presence-only data, point locations where a species has been recorded as being present, are often used in modeling the distribution of a species as a function of a set of explanatory variables—whether to map species occurrence, to understand its association with the environment, or to predict its response to environmental change. Currently, ecologists most commonly analyze presence-only data by adding randomly chosen "pseudo-absences" to the data such that it can be analyzed using logistic regression, an approach which has weaknesses in model specification, in interpretation, and in implementation. To address these issues, we propose Poisson point process modeling of the intensity of presences. We also derive a link between the proposed approach and logistic regression—specifically, we show that as the number of pseudo-absences increases (in a regular or uniform random arrangement), logistic regression slope parameters and their standard errors converge to those of the corresponding Poisson point process model. We discuss the practical implications of these results. In particular, point process modeling offers a framework for choice of the number and location of pseudo-absences, both of which are currently chosen by ad hoc and sometimes ineffective methods in ecology, a point which we illustrate by example.

1. Background. Pearce and Boyce (2006) define presence-only data as "consisting only of observations of the organism but with no reliable data on where the species was not found. Sources for these data include atlases, museum and herbarium records, species lists, incidental observation databases and radio-tracking studies." Note that such data arise as *point locations* where the organism is observed, which we denote as **y** in this article. An example is given in Figure 1(a). This figure gives all locations where a particular tree species (*Angophora costata*) has been reported by park rangers since 1972, within 100 km of the Greater Blue Mountains World Heritage Area, near Sydney, Australia. Note that this does not consist of all locations where an *Angophora costata* tree is found—rather it is the locations where the species has been *reported* to be found. We would like to use these presence points, together with maps of explanatory variables describing the environment (often referred to in ecology as "environmental variables"), to predict

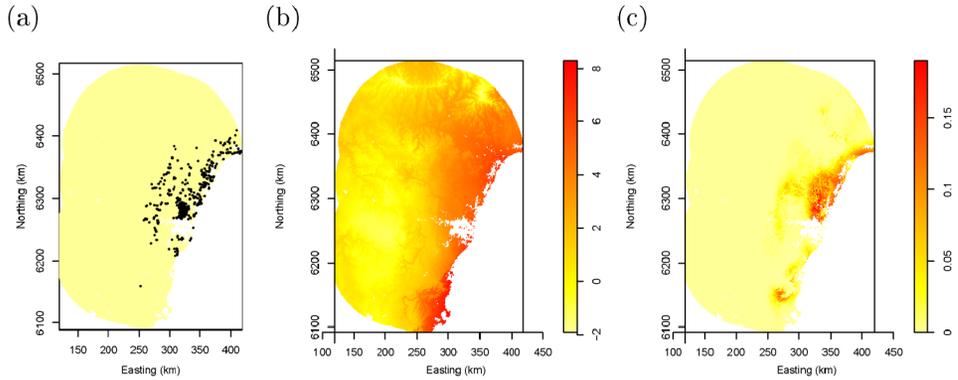(a)   (b)   (c)

FIG. 1.   (a) *Example presence-only data—atlas records of where the tree species* Angophora costata *has been reported to be present, west of Sydney, Australia. The study region is shaded.* (b) *A map of minimum temperature* (°C) *over the study region. Variables such as this are used to model how intensity of* A. costata *presence relates to the environment.* (c) *A species distribution model, modeling the association between* A. costata *and a suite of environmental variables. This is the fitted intensity function for* A. costata *records per km^2, modeled as a quadratic function of four environmental variables using a point process model as in Section* 4.

the location of *A. costata* and how it varies as a function of explanatory variables (Figure 1).

Presence-only data are used extensively in ecology to model species distributions—while the term "presence-only data" was rarely used before the 1990s, ISI Web of Science reports that it was used in 343 publications from 2005 to 2008. The use of presence-only data in modeling is a relatively recent development, presumably aided by the movement toward electronic record keeping and recent advances in Geographic Information Systems. One reason for the current widespread usage of presence-only data is that often this is the best available information concerning the distribution of a species, as there is often little or no information on species distribution being available from systematic surveys [Elith and Leathwick (2007)].

Species distribution models, sometimes referred to as habitat models or habitat classification models [Zarnetske, Edwards and Moisen (2007)], are regression models for the likelihood that a species is present at a given location, as a function of explanatory variables that are available over the whole study region. Such models are used to construct maps predicting the full spatial distribution of a species [given GIS maps of explanatory variables such as in Figure 1(b)]. When surveys have recorded the presence and absence of a species in a pre-defined study area ("presence/absence data"), logistic regression approaches and modern generalizations [Elith, Leathwick and Hastie (2008)] are typically used for species distribution modeling. If instead presence-only data are to be used in species distribution modeling, then a common approach to analysis is to first create "pseudo-absences," denoted as $\mathbf{y}_0$, usually achieved by randomly choosing point locations in the region of interest and treating them as absences. Then the presence/pseudo-absence data

set is analyzed using standard analysis methods for presence/absence data [Pearce and Boyce (2006); Elith and Leathwick (2007)], which have been used in species distribution modeling for a long time [Austin (1985)]. Ward et al. (2009) recently proposed a modification of the pseudo-absence logistic regression approach for the analysis of presence-only data, when the probability π that a randomly chosen pseudo-absence point of a presence is known. However, π is not known in practice.

We see three key weaknesses of the "pseudo-absence" approach so widely used in ecology for analyzing presence-only data, which we describe concisely as problems of model specification, interpretation, and implementation. A sounder *model specification* would involve constructing a model for the observed data **y** *only*, rather than requiring us to generate new data $\mathbf{y}_0$ prior to constructing a model. *Interpretation* of results is difficult, because some model parameters of interest (such as p_i of Section 3) are a function of the number of pseudo-absences and their location. For example, we explain in Section 3 that as the number of pseudo-absences approaches infinity, $p_i \to 0$, for a given presence-only data set **y**. *Implementation* of the approach is problematic because it is unclear how pseudo-absences should be chosen [Elith and Leathwick (2007); Guisan et al. (2007); Zarnetske, Edwards and Moisen (2007); Phillips et al. (2009)], and one can obtain qualitatively different results depending on the method of choice of pseudo-absences [Chefaoui and Lobo (2008)].

In this paper we make two key contributions. First, we propose point process models (Section 2) as an appropriate tool for species distribution modeling of presence-only data, given that presence-only data arise as a set of point events— a set of locations where a species has been reported to have been seen. A point process model specification addresses each of the three concerns raised above regarding pseudo-absence approaches. Our second key contribution is a proof that the pseudo-absence logistic regression approach, when applied with an increasing number of regularly spaced or randomly chosen pseudo-absences, yields estimates of slope parameters that converge to the point process slope estimates (Section 3). These two key results have important ramifications for species distribution modeling in ecology (Section 5), in particular, we provide a solution to the problem of how to select pseudo-absences. We illustrate our results for the *A. costata* data of Figure 1(a) (Section 4).

2. Poisson point process models for presence-only data. Presence-only data are a set $\mathbf{y} = \{y_1, \ldots, y_n\}$ of point locations in a two-dimensional region $\mathcal{A}$, where the locations where presences are recorded (the y_i) are out of the control of the researcher, as is the total number of presence points n. We also observe a "map" of values over the entire region $\mathcal{A}$ for each of k explanatory variables, and we denote the values of these variables at y_i as $(x_{i1}, \ldots, x_{ik})$.

We propose analyzing $\mathbf{y} = \{y_1, \ldots, y_n\}$ as a point process, hence, we jointly model number of presence points n and their location (y_i). This has not previously

been proposed for the analysis of presence-only data, despite the extensive literature on the analysis of presence-only data. We consider inhomogeneous Poisson point process models [Cressie (1993); Diggle (2003)], which make the following two assumptions:

1. The locations of the n point events $(y_1, \ldots, y_n)$ are independent.
2. The intensity at point y_i [$\lambda(y_i)$, denoted as λ_i for convenience], the limiting expected number of presences per unit area [Cressie (1993)], can be modeled as a function of the k explanatory variables. We assume a log-linear specification:

$$(2.1) \qquad \log(\lambda_i) = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j,$$

  although note that the linearity assumption can be relaxed in the usual way (e.g., using quadratic terms or splines). The parameters of the model for the λ_i are stored in the vector $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$.

Note that the process being modeled here is locations where an organism *has been reported* rather than locations where individuals of the organism occur. Hence, the independence assumption would only be violated by interactions between records of sightings rather than by interactions between individual organisms per se. The atlas data of Figure 1 consist of 721 *A. costata* records accumulated over a period of 35 years in a region of 86,000 km^2, so independence of records seems a reasonable assumption in this case, given the rarity of event reporting. Nevertheless, the methods we review here can be generalized to handle dependence between point events [Baddeley and Turner (2005)].

Cressie (1993) shows that the log-likelihood for **y** can be written as

$$(2.2) \qquad l(\beta; \mathbf{y}) = \sum_{i=1}^{n} \log(\lambda_i) - \int_{y \in \mathcal{A}} \lambda(y)\, dy - \log(n!).$$

Berman and Turner (1992) showed that if the integral is estimated via numerical quadrature as $\int_{y \in \mathcal{A}} \lambda(y)\, dy \approx \sum_{i=1}^{m} w_i \lambda_i$, then the log-likelihood is (approximately) proportional to a weighted Poisson likelihood:

$$(2.3) \qquad l_{\text{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{w}) = \sum_{i=1}^{m} w_i \big(z_i \log(\lambda_i) - \lambda_i \big),$$

where $z_i = \frac{I(i \in \{1, \ldots, n\})}{w_i}$, $\mathbf{y}_0 = \{y_{n+1}, \ldots, y_m\}$ are quadrature points, the vector $\mathbf{w} = (w_1, \ldots, w_m)$ stores all quadrature weights, and $I(\cdot)$ is the indicator function. Being able to write $l(\beta; \mathbf{y})$ as a weighted Poisson likelihood has important practical significance because it implies that generalized linear modeling (GLM) techniques can be used for estimation and inference about β. Further, adaptations of GLM techniques to other settings, such as generalized additive models [Hastie

and Tibshirani (1990)], can then be readily applied to Poisson point process models also.

Before implementing this approach, however, we need to make two key decisions—how to choose quadrature points $\mathbf{y}_0 = \{y_{n+1}, \ldots, y_m\}$ and how to calculate the quadrature weight w_i at each point y_i.

We propose choosing quadrature points in a regular rectangular grid, and considering grids of increasing spatial resolution until the estimate of the maximized log-likelihood $l_{\mathrm{ppm}}(\hat{\beta}; \mathbf{y}, \mathbf{y}_0, \mathbf{w})$ has converged. A rectangular grid provides reasonably efficient coverage of the region $\mathcal{A}$, and is an arrangement for which environmental data $x_{i1}, \ldots, x_{ik}$ can be easily obtained via GIS software. We illustrate this method in Section 4. Note a large data set may be required—in Section 4 convergence was achieved at a spatial scale that required inclusion of approximately 86,000 quadrature points.

Quadrature weights are calculated as the area of the neighborhood A_i around each point y_i, according to some definition of the A_i such that $y_i \in A_i$ for each i, $A_i \cap A_{i'} = \varnothing$ for each $i \neq i'$, and $\bigcup_i A_i = \mathcal{A}$. In Section 4 we calculated quadrature weights using the tiling method implemented in the R package spatstat [Baddeley and Turner (2005)]. This crude approach breaks the region $\mathcal{A}$ into rectangular tiles and calculates the weight of a point as the inverse of the number of points per unit area in its tile. We fixed tile size at the size of the regular grid used to sample quadrature points, such that all tiles contained exactly one quadrature point. Dirichlet tessellation [Baddeley and Turner (2005)] offers an alternative method of estimating weights, but this was not practical for our sample sizes.

3. Asymptotic equivalence of pseudo-absence logistic regression and Poisson point process models.
Ecologists typically analyze presence-only data points $\mathbf{y} = \{y_1, \ldots, y_n\}$ by generating a set of "pseudo-absence" points $\mathbf{y}_0 = \{y_{n+1}, \ldots, y_m\}$, then using logistic regression to model the "response variable" $I(i \in \{1, \ldots, n\})$ as a function of explanatory variables [Pearce and Boyce (2006)]. Note that $I(i \in \{1, \ldots, n\})$ is not actually a stochastic quantity, nevertheless, the use of logistic regression to model this quantity as a Bernoulli response variable can be motivated via a case-control argument along the lines of Diggle (2003), Section 9.3.

In this section we will show that the approach to analysis currently used in ecology, logistic regression using pseudo-absences, is closely related to the Poisson point process model introduced in Section 2. Specifically, if the pseudo-absences are either generated on a regular grid or completely at random over the region $\mathcal{A}$, then as the number of pseudo-absences increase, all parameter estimators except for the intercept in the logistic regression model converge to the maximum likelihood estimators of the Poisson process model of Section 2. This asymptotic relationship between logistic regression and Poisson point process models does not appear to have been recognized previously in the literature.

First, we will specify a probability model for I $(i \in \{1, \ldots, n\})$ that permits a logistic regression model, and the study of its properties as $m \to \infty$. This can be achieved by considering a point chosen at random from $\{y_1, \ldots, y_m\}$ and defining U as the event that the randomly chosen point y_i is a presence. We are interested in modeling U conditionally on the explanatory variables observed at the randomly chosen point, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$. In this setting, U is a Bernoulli variable with conditional mean p_i, and we assume that

$$(3.1) \qquad \log\left(\frac{p_i}{1 - p_i}\right) = \gamma_0 - \log(m - n) + \sum_{j=1}^{k} x_{ij}\gamma_j.$$

The intercept term is written as $\gamma_0 - \log(m - n)$ because

$$(3.2) \qquad \frac{p_i}{1 - p_i} = \frac{f_1(\mathbf{x}_i | U = 1)}{f_0(\mathbf{x}_i | U = 0)} \cdot \frac{P(U = 1)}{P(U = 0)} = \frac{f_1(\mathbf{x}_i | U = 1)}{f_0(\mathbf{x}_i | U = 0)} \frac{n}{m - n},$$

where $f_1(\cdot)$ and $f_0(\cdot)$ are the densities of $\mathbf{x}_i$ conditional on $U = 1$ and $U = 0$ respectively. Provided that $f_0(\mathbf{x}_i | U = 0)$ is not a function of m (which is ensured, e.g., by using an identical process to select all pseudo-absence points), then the odds of a presence point $\frac{p_i}{1-p_i}$ is a function of m only via the multiplier $(m - n)^{-1}$.

It can be seen from equation (3.2) that if $m \to \infty$ in such a way that $f_0(\mathbf{x}_i | U = 0)$ is not a function of i, then $p_i \to 0$ at an asymptotic rate that is proportional to m^{-1}, and the intercept term in the logistic regression model approaches $-\infty$ at the rate $\log(m)$. This in turn means that the logistic regression log-likelihood, defined below, will also diverge as $m \to \infty$:

$$(3.3) \qquad l_{\text{bin}}(\gamma; \mathbf{y}, \mathbf{y}_0) = \sum_{i=1}^{n} \log(p_i) + \sum_{i=n+1}^{m} \log(1 - p_i).$$

Clearly, as $p_i \to 0$, $\log(p_i) \to -\infty$ and, hence, $l_{\text{bin}}(\gamma; \mathbf{y}, \mathbf{y}_0) \to -\infty$. Such divergence is a symptom that the original model has been incorrectly specified. The use of the more appropriate spatial point process model of Section 2 will not encounter such problems. However, it is shown in the following theorems that despite the problems inherent in the logistic regression model specification, and despite divergence of the intercept term, the remaining parameters converge to the corresponding parameters from the Poisson process model of equation (2.3). Further, pseudo-absences play the same role in the logistic regression model that quadrature points played in Section 2.

For notational convenience, we will define J_m to be the single-entry matrix whose first element is $\log m$:

$$(3.4) \qquad J_m = (\log m, 0, \ldots, 0).$$

This definition will be used in each of the theorems that follow. The J_m notation is immediately useful in writing out the parameters of the model for p_i in equation (3.1) as $\gamma - J_{m-n}$ where $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_k)$.

THEOREM 3.1.   *Consider a fixed set of n observations from a point process* $\mathbf{y} = \{y_1, \ldots, y_n\}$, *and a set of pseudo-absences* $\mathbf{y}_0 = \{y_{n+1}, \ldots, y_m\}$ *of variable size that is chosen via some identical process on* $\mathcal{A}$ *for* $i \in \{n+1, \ldots, m\}$. *We model U, whether or not a randomly chosen point is a presence point, via logistic regression as in equation* (3.1).

As $m \to \infty$, *the logistic regression log-likelihood of equation* (3.3) *approaches the Poisson point process log-likelihood* [*equation* (2.3)] *but with all quadrature weights set to one:*

$$l_{\mathrm{bin}}(\gamma; \mathbf{y}, \mathbf{y}_0) = l_{\mathrm{ppm}}(\gamma - J_m; \mathbf{y}, \mathbf{y}_0, \mathbf{1}) + O(m^{-1}),$$

where $\mathbf{1}$ *is a m-vector of ones, and* J_m *is defined in equation* (3.4).

The proofs to Theorem 3.1 and all other theorems are given in Appendix A.

Theorem 3.1 has two interesting practical implications.

First, it implies that the pseudo-absence points of presence-only logistic regression play the same role as quadrature points of a point process model, and so established guidelines on how to choose quadrature points (such as those of Section 2) can inform choice of pseudo-absences. Previously pseudo-absences have been generated according to ad hoc recommendations [Pearce and Boyce (2006); Zarnetske, Edwards and Moisen (2007)], given the lack of a theoretical framework for their selection. In contrast, quadrature points are generated in order to estimate the log-likelihood to a pre-determined level of accuracy, a criterion which guides the choice of locations and numbers of quadrature points $m - n$, as explained in Section 2 and as illustrated later in Section 4 [Figure 2(a)]. Interestingly, current methods of selecting pseudo-absences in ecology [Pearce and Boyce (2006); Zarnetske, Edwards and Moisen (2007)] do not appear to be consistent with the best practice in low-dimensional numerical quadrature—points are usually selected at random rather than on a regular grid, and the number of pseudo-absences $(m - n)$ is more commonly chosen relative to the magnitude of the number of presences (n) rather than based on some convergence criterion as in Figure 2(a).

Second, Theorem 3.1 implies that despite the apparent ad hoc nature of the pseudo-absence approach, *some form of point process model is being estimated.* However, logistic regression is only equivalent to a Poisson point process when $\mathbf{w} = \mathbf{1}$, that is, all quadrature weights are ignored. The implications of ignoring weights is considered in Theorems 3.2 and 3.3 below.

It should also be noted that Theorem 3.1 is closely related to results due to Owen (2007) and Ward (2007), although Theorem 3.1 differs from these results by relating pseudo-absence logistic regression specifically to point process modeling.

Owen (2007) also considered the logistic regression setting where the number of presence points is fixed, and the number of pseudo-absences increases to infinity, and referred to this as "infinitely unbalanced logistic regression." Owen (2007) derived conditions under which convergence of model parameters could be achieved as the number of pseudo-absences increased. The key condition is that the centroid
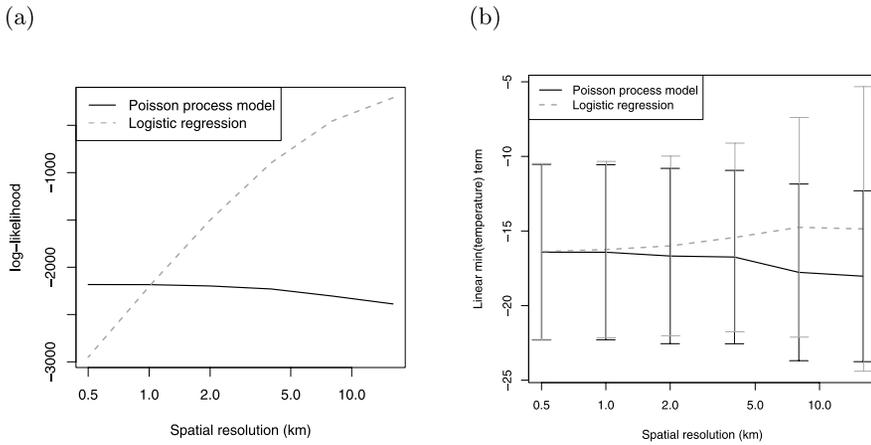
(a) (b)

FIG. 2. *Asymptotic behavior of Poisson point process and pseudo-absence logistic regression models when the number of quadrature points becomes large (via sampling in a regular grid with increasing spatial resolution). (a) The maximized log-likelihood converges for a Poisson point process, but not for pseudo-absence logistic regression. (b) The parameters and their standard errors converge for Poisson point process and logistic regression models, for sufficiently high spatial resolution. Linear coefficient of "minimum temperature" is given here (corresponding to the second entry in Table 2).*

of the points $\{y_1, \ldots, y_n\}$ in the design space is "surrounded"—see Definition 3 of Owen (2007) for details.

Along the lines of Ward et al. (2009), Ward (2007) considered a pseudo-absence logistic regression formulation of the presence-only data problem, and defined the "population logistic model" as the model across "the full population" of locations in the region $\mathcal{A}$. The unconstrained log-likelihood of the population logistic model [Ward (2007), equation (7.6)] has a similar form to the point process log-likelihood $l(\beta; \mathbf{y})$ of Section 2. For presence-only logistic regression as in Ward et al. (2009), Ward (2007) shows that as the number of pseudo-absences approaches infinity, the log-likelihood converges to that of the population logistic model, a result that is analogous to Theorem 3.1.

Having shown that pseudo-absence logistic regression is equivalent to a point process model where weights are ignored, the implications of ignoring weights is now considered in Theorems 3.2 and 3.3.

THEOREM 3.2. *Consider a point process model with quadrature points y_{n+1}, $\ldots, y_m$ selected such that for all i $w_i = \frac{|\mathcal{A}|}{m}$, where $|\mathcal{A}|$ is the total area of the region $\mathcal{A}$. Assume also that the design matrix $\mathbf{X}$ has full rank.*

The maximum likelihood estimators of $l_{\mathrm{ppm}}(\gamma - J_m; \mathbf{y}, \mathbf{y}_0, \mathbf{1})$ and $l_{\mathrm{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \frac{|\mathcal{A}|}{m}\mathbf{1})$, $\hat{\gamma} - J_m$ and $\hat{\beta}$ respectively, satisfy

$$\hat{\gamma} = \hat{\beta} + J_{|\mathcal{A}|}.$$

Further, the Fisher information for $\hat{\gamma}$ and $\hat{\beta}$ is equal.

That is, provided that quadrature points have been selected such that quadrature weights are equal, ignoring quadrature weights in a Poisson point process model does not change slope parameters nor their standard errors, although the intercept term will differ by $\log(|\mathcal{A}|/m)$.

Theorem 3.2 refers to the special case where all points (including presence points) are sampled on a regular grid. This arises in the special case where the region $\mathcal{A}$ has been divided into grid cells of equal area, and each grid cell is assigned the value 1 only if it contains a presence point. This form of presence-only analysis is sometimes used in ecology [Phillips, Anderson and Schapire (2006), e.g.]. In addition, the setting of Theorem 3.2 provides a reasonable approximation to the approach to quadrature-point selection proposed in Section 2.

Note that together Theorems 3.1–3.2 suggest that when quadrature points (or, equivalently, pseudo-absences) are sampled in a regular grid at increasing resolution, the logistic regression parameter estimates and their standard errors will approach those of the point process model—with the exception of the intercept term, which diverges slowly to $-\infty$ as all $p_i \to 0$ at a rate inversely proportional to m. This nonconvergence of the intercept was also noticed by Owen (2007). Figure 2 illustrates these results for the *A. costata* data.

Theorem 3.3 below links the above results with the case where pseudo-absences are randomly sampled within the region $\mathcal{A}$, which is a more common approach in ecology than sampling on a regular grid [e.g., Elith and Leathwick (2007); Hernandez et al. (2008)].

THEOREM 3.3. *Consider again the conditions of Theorem 3.2, but now assume that the quadrature points* $\mathbf{y}_0$ *are selected uniformly at random within the region* $\mathcal{A}$. *As previously,* $\hat{\gamma}$ *is the maximum likelihood estimator of* $l_{\text{ppm}}(\gamma - J_m; \mathbf{y}, \mathbf{y}_0, \mathbf{1})$, *but now let* $\hat{\beta}$ *be the maximum likelihood estimator of* $l(\beta; \mathbf{y})$ *from equation* (2.2). *As* $m \to \infty$,

$$\hat{\gamma} \xrightarrow{\mathcal{P}} \hat{\beta} + J_{|\mathcal{A}|}.$$

That is, if quadrature points are randomly selected instead of being sampled on a regular grid, the result of Theorem 3.2 holds in probability rather than exactly.

Note that the stochastic convergence in Theorem 3.3 is with respect to m not n, that is, it is conditional on the observed point process.

Note also that one can think of randomly selecting pseudo-absence points as an implementation of "crude" Monte Carlo integration [Lepage (1978)] for estimating $\int_{y \in \mathcal{A}} \lambda(y) \, dy$ in the point process likelihood [equation (2.2)].

4. Modeling *Angophora costata* species distribution. As an illustration, we construct Poisson point process models for the intensity of *Angophora costata* records as a function of a set of explanatory variables. We consider modeling the log of intensity using linear and quadratic functions of the variables minimum and maximum temperature, mean annual rainfall, number of fires since 1943, and "wetness," a coefficient which can be considered as an indicator of local moisture. These five variables were recommended by local experts as likely to be important in determining *A. costata* distribution.

Our full model for intensity of *A. costata* records at the point y_i has the following form:

$$(4.1) \qquad \log(\lambda_i) = \beta_0 + \mathbf{x}_i^T \beta_1 + \mathbf{x}_i^T \mathbf{B} \mathbf{x}_i,$$

where β_1 is a vector of linear coefficients, $\mathbf{B}$ is a matrix of quadratic coefficients, and $\mathbf{x}_i$ is a vector containing measurements of the five environmental variables at point y_i. We consider a quadratic model for $\log(\lambda_i)$ because this enables fitting a nonlinear function and interaction between different environmental variables, both considered important in species distribution modeling [Elith, Leathwick and Hastie (2008)].

All analyses were carried out using purpose-written code on the R program [R Development Core Team (2009)].

We first considered the spatial resolution at which quadrature points needed to be sampled in order for the log-likelihood $l(\beta; \mathbf{y})$ to be suitably well approximated by $l_{\mathrm{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{w})$. We found [as in Figure 2(a)] that on increasing the number of quadrature points, the estimate of the maximized log-likelihood converged, and that there was minimal change in the solution beyond a resolution of one quadrature point every 1 km (the maximized log-likelihood changed by less than one when the number of quadrature points was increased 4-fold). Hence, the 1 km resolution was used in model-fitting, and these results are reported here. This involved a total of 86,227 quadrature points.

In order to study which environmental variables are associated with *A. costata* and how they are associated, we performed model selection where we considered different forms of models for log-intensity as a function of environmental variables, and we considered different subsets of the environmental variables via all-subsets selection. In both cases we used AIC as our model selection criterion, a simple and widely-used penalty-based model selection criterion [Burnham and Anderson (1998)].

Comparison of AIC values for linear and quadratic models suggest that a much better fit is achieved when using a quadratic model with interactions terms for all coefficients (Table 1). Hence, we have evidence that environmental variables interact in their effect on *A. costata*. Judging from the model coefficients and their relative size compared to standard errors, the interactions between maximum temperature, minimum temperature, and annual rainfall appear to be the major contributors.

TABLE 1
AIC values for linear and quadratic Poisson point process models for $\log(\lambda)$
of Angophora costata *presence. Model fitted at the* 1 *km by* 1 *km resolution*

Model	AIC
Linear terms only	5363.6
Quadratic (additive terms only)	4763.4
Quadratic (interactions included)	4400.6

All-subsets selection considered a total of 32 models, and found that the best-fitting model included four variables (Figure 3)—all except for "wetness." Parameter estimates and standard errors for this best-fitting model are given in Table 2(a).

An image of the fitted intensity surface from the best-fitting model is presented in Figure 1(c). The regions of highest predicted intensity are near the coast and just north of Sydney, which are indeed where the highest density of presence points appeared in Figure 1(a). We also compared intensity surfaces fitted at different spatial resolutions, and note that they appear identical when quadrature points are selected in a 500×500 m, 1×1 km, or 2×2 km grid, and that irrespective of spatial resolution, regions of higher intensity had 0.05–0.2 expected *A. costata* records per square kilometer, as in Figure 1(c). Note this is in contrast to logistic regression, where fitted probabilities in any given location are a function of number of pseudo-absences, and vary by a factor of 16 when moving from a 2×2 km to a 500×500 m grid.

To assist in interpreting parameters from the best-fitting model, we have constructed image plots of the fitted intensity in "environmental space" to elucidate
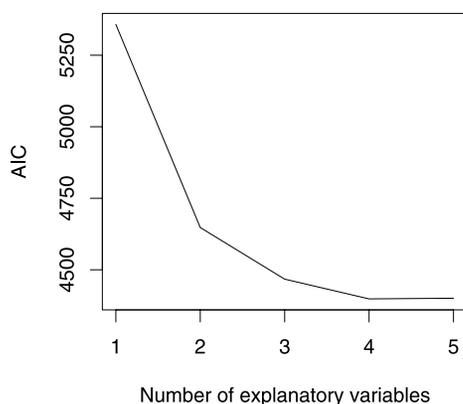

FIG. 3. *Results of all-subsets selection, expressed as AIC of the best-fitting model at each level of complexity. The respective best-fitting models included minimum temperature, then minimum and maximum temperature, then annual rainfall was added, then fire count, and finally wetness. The best-fitting model included four explanatory variables (all variables except wetness).*

TABLE 2

Parameter estimates and their standard errors for (a) *the Poisson point process model with a* 1×1 *km regular grid of quadrature points;* (b) *The logistic regression model with a* 1×1 *km regular grid of pseudo-absences;* (c) *The logistic model with* 1000 *randomly chosen pseudo-absence points. In each case we fitted a quadratic model of minimum temperature (MNT), maximum temperature (MXT), annual rainfall (RA), and fire count (FC). Notice that with few exceptions, terms are equivalent to 2–3 significant figures for the models fitted over a regular grid. But this is not the case for* (c), *and, in particular, standard errors are all 30–80% larger*

Term	(a)		(b)		(c)	
	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	$\hat{\beta}_j$	$se(\hat{\beta}_j)$
Intercept	-2130	169.4	-2119	171	-1999	227
MNT	-16.3	3.0	-16.2	3.0	-9.91	4.2
MNT^2	-0.21	0.027	-0.205	0.028	-0.185	0.050
MXT	128.7	10.1	128.1	10.1	120.2	13
$MNT * MXT$	0.539	0.090	0.535	0.091	0.377	0.13
MXT^2	-1.98	0.15	-1.97	0.15	-1.84	0.20
RA	0.759	0.065	0.755	0.066	0.714	0.089
$MNT * RA$	0.00345	0.00065	0.00339	0.00065	0.00147	0.00096
$MXT * RA$	-0.0218	0.0019	-0.0216	0.0019	-0.0203	0.0025
$RA^2/1000$	-0.0819	0.0072	-0.0815	0.0072	-0.0749	0.010
FC	6.24	3.37	5.98	3.42	4.08	4.9
$MNT * FC$	-0.101	0.040	-0.101	0.041	-0.207	0.070
$MXT * FC$	-0.123	0.10	-0.115	0.010	-0.0601	0.15
$RA * FC$	-0.00174	0.00066	-0.00171	0.00067	-0.000952	0.00095
FC^2	-0.127	0.024	-0.123	0.024	-0.107	0.041

the nature of the effect of each environmental variable on intensity of *A. costata* records (Figure 4). It can be seen in Figure 4 that there is a strong and negatively correlated response to maximum temperature and annual rainfall. Of the four environmental variables, the response to number of fires appears to be the weakest, with little apparent change in predicted intensity as number of fires increased, and no observable interaction with the three climatic variables.

For the purpose of comparison, parameter estimates and their standard errors are reported not just for the point-process model fit, but also for the analogous logistic regression model in Table 2(b), for a model fitted with quadrature points sampled in a 1×1 km regular grid. Note that most parameter estimates and standard errors differ by less than 1% between the logistic regression and point process models, as expected given Theorems 3.1–3.2.

We also report results when logistic regression is applied to $m - n = 1000$ pseudo-absences at randomly selected locations [Table 2(c)]. This is at the lower end of the range typically used [Pearce and Boyce (2006); Elith and Leathwick (2007); Hernandez et al. (2008)] for pseudo-absence selection in ecology. Note that the standard errors are substantially larger in this case, and no parameter esti-
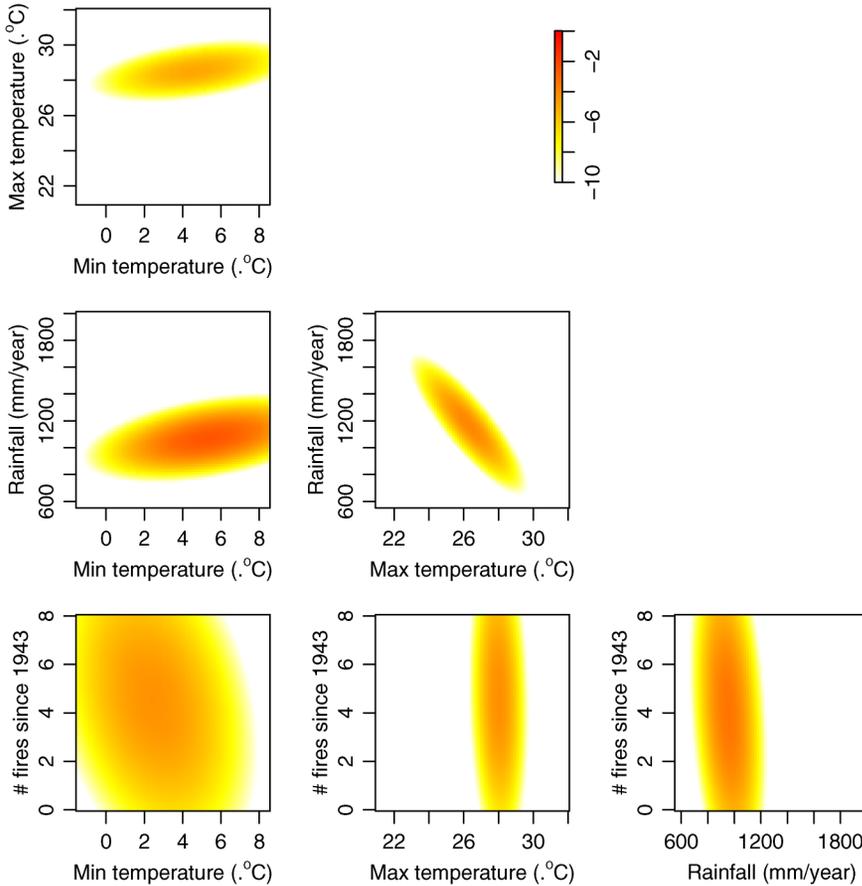
FIG. 4. *Image plots of the joint effects of environmental variables on predicted log* (*intensity*) *of* A. costata *records. Darker areas of the image correspond to higher predicted values of* $\log(\lambda_i)$. *Note the strong and highly correlated response of intensity to maximum temperature and annual rainfall, and the relatively weak response to # fires.*

mates are correct past the first significant figure. This result exemplifies how current practice in ecology regarding the number of pseudo-absences ($m - n$) can lead to poor results. Instead, it is advisable to consider the sensitivity of results to different choices of $m - n$, along the lines of Figure 2.

To explore the goodness of fit of the best-fitting model, an inhomogeneous K-function [Baddeley, Moller and Waagepetersen (2000)] was plotted using the `kinhom` function from the `spatstat` package on R [Baddeley and Turner (2005)], and simulation envelopes around the fitted model were constructed. See Diggle (2003) for details concerning the use of K-functions to explore goodness of fit of point process models. The inhomogeneous K-function, as its name suggests, is a generalization of the K-function to the nonstationary case. It is defined over
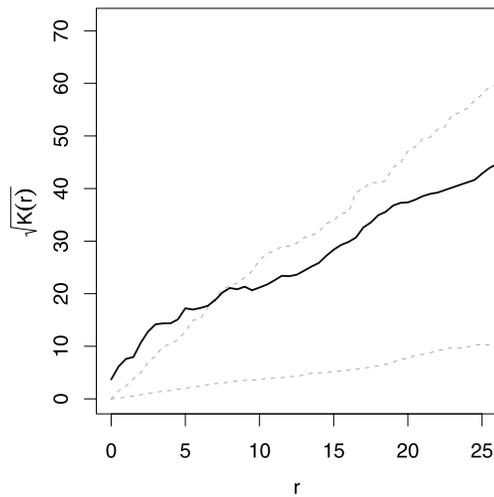
FIG. 5. *Goodness of fit plot of the quadratic Poisson process model—inhomogeneous K-function* (*solid line*) *with simulation envelope* (*broken lines*). *The envelope gives* 95% *confidence bands as estimated from* 500 *simulated data sets. Note that the K-function falls within those bounds over most of its range, although with a possible departure for r < 6 km.*

the region $\mathcal{A}$ as

$$K_{\text{inhom}}(r) = \frac{1}{|\mathcal{A}|} E \left\{ \sum_{y_i \in \mathbf{y}} \sum_{y_j \in \mathbf{y} \setminus \{y_i\}} \frac{I(\|y_i - y_j\| < r)}{\lambda_i \lambda_j} \right\}.$$

K_{inhom} reduces to the usual K function for a stationary process, and like the usual K-function, can be used to diagnose whether there are interactions in the point pattern $\mathbf{y}$ [Baddeley, Moller and Waagepetersen (2000)]. Results (Figure 5) suggest a reasonable fit of this model to the data, although with some lack of fit at small spatial scales ($r < 6$ km) suggestive of possible clustering in the data. One possible method of modeling this clustering is to fit an area-interaction process [Baddeley and van Lieshout (1995)], using the spatstat package. We have repeated analyses using such a model and found results to be generally consistent with those presented here, except of course that the equivalence with logistic regression no longer holds.

5. Discussion. In this paper we have proposed the use of Poisson point process models for the analysis of presence-only data in ecology, an important and widely-studied problem to which this methodology is well suited. We have also shown that this method is approximately equivalent to logistic regression, when a suitable number of regularly or randomly spaced pseudo-absences are used, hence, we provide a link between the proposed method and the approach most commonly used in ecology at the moment. But this raises the question: why use point process

models, if the method currently being used is (asymptotically) equivalent to logistic regression anyway? Several reasons are listed below.

Recall that in Section 1 we argued that the pseudo-absence approach has problems with model specification, interpretation, and implementation. We argue that each of these difficulties is resolved by using a point process modeling framework. *Model specification*—we believe that a point process model as in Section 2 is a plausible model for the data generation mechanism for presence-only data. In contrast, the logistic regression approach involves generating new data in order to fit a model originally designed for a different problem (analysis of binary data not analysis of point-events). Hence, the pseudo-absence approach as it is usually applied appears to involve coercing the data to fit the model rather than choosing a model that fits the original data. *Interpretation*—in the logistic regression approach we model p_i, the probability that a given point event is a presence not a pseudo-absence. This quantity has no physical meaning and clearly its interpretation is sensitive to our method of choice of pseudo-absences (and typically each $p_i \to 0$ as $m \to \infty$). In contrast, the intensity at a point λ_i has a natural interpretation as the (limiting) expected number of presences per unit area, and will *not* be sensitive to choice of quadrature points, provided that the number of quadrature points is sufficiently large. *Implementation*—in Section 2 we explain that point process models offer a framework for choosing quadrature points. Specifically, equation (2.3) is used to estimate the point process log-likelihood, and progressively more quadrature points are added until convergence of $l_{\mathrm{ppm}}(\hat{\beta}; \mathbf{y}, \mathbf{y}_0, \mathbf{w})$ is achieved as in Figure 2(a). No such framework for choice of pseudo-absences is offered by logistic regression, and instead choice of the location and number of pseudo-absences is ad hoc, with potentially poor results [Table 2(c)]. Ecologists are concerned about the issues of how many pseudo-absences to choose [Pearce and Boyce (2006)], where to put them [Elith and Leathwick (2007); Zarnetske, Edwards and Moisen (2007); Phillips et al. (2009)], and what spatial resolution to use in model-fitting [Guisan et al. (2007); Elith and Leathwick (2009)], all issues that have natural solutions given a point process model specification of the problem, as in Section 2.

It should be emphasized that we have demonstrated equivalence of point process modeling and pseudo-absence logistic regression only for large numbers of pseudo-absences and only for pseudo-absences that are either regularly spaced or located uniformly at random over $\mathcal{A}$. Current practice concerning selection of pseudo-absences in the ecology literature does not always involve sampling at random over $\mathcal{A}$ [e.g., Hernandez et al. (2008)] and does not involve sampling sufficiently many pseudo-absences for model convergence. Instead, choice of the number of pseudo-absences is ad hoc, and a total of 1000–10,000 pseudo-absences is usually used [Elith and Leathwick (2007); Hernandez et al. (2008)], although sometimes even fewer [Zarnetske, Edwards and Moisen (2007)]. On Figure 2, 1000–10,000 corresponds to a resolution of about 4–8 km, for which model convergence has not been achieved. When fitting a model using just 1000 pseudo absences, some parameter estimates are not equivalent to the high-resolution fits to

even one significant figure, and all standard error estimates were larger by 30–80%
(Table 2).

While only Poisson point processes were considered in this paper, the method-
ology implemented in Section 4 can be generalized to incorporate interactions be-
tween points in a straightforward fashion [Baddeley and Turner (2005)]. However,
the links between point process models and logistic regression identified in Sec-
tion 3 may be lost in this more general setting.

One issue not touched on in this paper is the problem of observer bias—that the
likelihood of a species being reported is a function of additional variables related
to properties of the observer and not of the target species, such as variation in
the level of accessibility of different parts of the region $\mathcal{A}$. For example, the high
number of *A. costata* records just north of Sydney may be due in part to proximity
to a large city, rather than simply being due to environmental conditions being
suitable for *A. costata*. This issue will be addressed in a related article.

APPENDIX A: PROOF OF THEOREMS

A.1. Proof of Theorem 3.1. The proof involves two steps. The first step in-
volves showing that $l_{\text{bin}}(\cdot)$, as a function of p_i, is asymptotically equivalent to
$l_{\text{ppm}}(\cdot)$ when written as a function of λ_i. The second step involves showing that
given the definitions of p_i and λ_i in equations (2.1) and (3.1), we can replace one
with the other without affecting the order of approximation.

Specifically, the log-likelihood function for U can be written as

$$l_{\text{bin}}(\gamma; \mathbf{y}, \mathbf{y}_0) = \sum_{i=1}^{n} \log(p_i) + \sum_{i=n+1}^{m} \log(1 - p_i)$$

and a Taylor expansion of $\log(1 - p_i)$ yields

$$= \sum_{i=1}^{n} \log(p_i) + \sum_{i=n+1}^{m} \{p_i + O(p_i^2)\},$$

but it can be seen from equation (3.2) that $p_i = O(m^{-1})$ and, hence, $\sum_{i=1}^{n} p_i = O(m^{-1})$ for fixed n, so

$$l_{\text{bin}}(\gamma; \mathbf{y}, \mathbf{y}_0) = \sum_{i=1}^{n} \log(p_i) + \sum_{i=n+1}^{m} p_i + O(m^{-1})$$

(A.1)
$$= \sum_{i=1}^{n} \log(p_i) + \sum_{i=1}^{m} p_i + O(m^{-1})$$

$$= \sum_{i=1}^{m} \{I(i \in \{1, \dots, n\}) \log(p_i) - p_i\} + O(m^{-1}).$$

Note that equation (A.1) has the form of the Poisson point process log-likelihood, but with all weights set to one and p_i being used in place of λ_i. We will now derive a relation between p_i and λ_i which motivates the replacement of p_i by λ_i.

First note that the Taylor expansion of $\log(1-x)$ implies both that $\log(1-p_i) = O(m^{-1})$, and that $\log(m-n) = \log(m) + \log(1-n/m) = \log(m) + O(m^{-1})$. So from equation (3.1),

$$\log p_i = \gamma_0 - \log(m-n) + \sum_{j=1}^{k} x_{ij}\gamma_j - \log(1-p_i)$$

$$= \gamma_0 - \log(m) + \sum_{j=1}^{k} x_{ij}\gamma_j + O(m^{-1}).$$

This has the form of equation (2.1), where $\beta = \gamma - J_m$. So when $\beta = \gamma - J_m$, $\log p_i = \log \lambda_i + O(m^{-1})$, and $\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \lambda_i\{1 + O(m^{-1})\} = \sum_{i=1}^{m} \lambda_i + O(m^{-1})$. Now plugging these results into equation (A.1) yields $l_{\mathrm{ppm}}(\gamma - J_m; \mathbf{y}, \mathbf{y}_0, \mathbf{1}) + O(m^{-1})$, completing the proof. $\square$

A.2. Proof of Theorem 3.2. The proof follows by inspection of the score equations. Specifically, let $s_j(\beta; \mathbf{w}) = \frac{\partial}{\partial \beta_j} l_{\mathrm{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{w})$. From equation (2.3), for $j \in \{1, \ldots, k\}$,

$$(A.2) \qquad s_j(\beta; \mathbf{w}) = \sum_{i=1}^{m} x_{ij}\lambda_i w_i \left(\frac{z_i - \lambda_i}{\lambda_i}\right) = \sum_{i=1}^{m} x_{ij} w_i (z_i - \lambda_i),$$

where $z_i = \frac{I(i \in 1, \ldots, n)}{w_i}$. If $j = 0$, equation (A.2) holds but with $x_{ij} = 1$ for each i.

Now $\hat{\beta}$ satisfies $s_j(\hat{\beta}; \frac{|\mathcal{A}|}{m}\mathbf{1}) = 0$ for each j, that is,

$$(A.3) \qquad 0 = \sum_{i=1}^{m} x_{ij}\left(I(i \in 1, \ldots, n) - \hat{\lambda}_i \frac{|\mathcal{A}|}{m}\right),$$

where from equation (2.1), $\log(\hat{\lambda}_i) = \hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\beta}_1$.

$\hat{\gamma}$ satisfies $s_j(\hat{\gamma} - J_m; \mathbf{1}) = 0$, for each j,

$$(A.4) \qquad 0 = \sum_{i=1}^{m} x_{ij}(I(i \in 1, \ldots, n) - \tilde{\lambda}_i),$$

where $\tilde{\lambda}_i$ is the maximum likelihood estimator of λ_i for $l_{\mathrm{ppm}}(\gamma - J_m; \mathbf{y}, \mathbf{y}_0, \mathbf{1})$, which satisfies $\log(\tilde{\lambda}_i) = \hat{\gamma}_0 - \log m + \sum_{j=1}^{k} x_{ij}\hat{\gamma}_1$.

The solutions to equations (A.3) and (A.4) are related by the identity $\tilde{\lambda}_i = \hat{\lambda}_i \frac{|\mathcal{A}|}{m}$ for each i, and if we take the logarithm of both sides,

$$\hat{\gamma}_0 - \log m + \sum_{j=1}^{k} x_{ij}\hat{\gamma}_j = \hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\beta}_j + \log|\mathcal{A}| - \log m.$$

Provided that the design matrix X has full rank, $\hat{\gamma} = \hat{\beta} + J_{|\mathcal{A}|}$.

Also, note that the (j, j')th element of the Fisher information matrix of $l_{\text{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{w})$ is

$$(A.5) \qquad I_{jj'}(\beta; \mathbf{w}) = -E\left(\frac{\partial^2}{\partial \beta_j \beta_{j'}} l_{\text{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{w})\right) = \sum_{i=1}^{m} x_{ij} w_i x_{ij'} \lambda_i$$

and so $I_{jj'}(\hat{\beta}; \frac{|\mathcal{A}|}{m}\mathbf{1}) = \sum_{i=1}^{m} x_{ij} x_{ij'} \frac{|\mathcal{A}|}{m} \hat{\lambda}_i = \sum_{i=1}^{m} x_{ij} x_{ij'} \tilde{\lambda}_i = I_{jj'}(\hat{\gamma} - J_m; \mathbf{1})$ for each (j, j'). This completes the proof. $\square$

A.3. Proof of Theorem 3.3. Let $\delta = \hat{\gamma} - \hat{\beta} - J_{|\mathcal{A}|}$. We will prove the theorem by using a Taylor expansion of the score equations for $l_{\text{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{1})$ to show that for fixed n and $m \to \infty$, $\delta \xrightarrow{\mathcal{P}} 0$.

Let $\mathbf{S}(\beta; \mathbf{1})$ be the vector of score equations whose jth element is $s_j(\beta; \mathbf{1}) = \frac{\partial}{\partial \beta_j} l_{\text{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{1})$ and let $\mathbf{I}(\beta; \mathbf{1})$ be the corresponding Fisher information matrix. A Taylor expansion of $\mathbf{S}(\hat{\gamma} - J_m; \mathbf{1})$ about $\mathbf{S}(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1})$ yields

$$(A.6) \quad \mathbf{S}(\hat{\gamma} - J_m; \mathbf{1}) = \mathbf{S}(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1}) - \mathbf{I}(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1})\delta + O_p(\|\delta\|^2).$$

The left-hand side is zero, because it is evaluated at the maximizer of $l_{\text{ppm}}(\beta; \mathbf{y}, \mathbf{y}_0, \mathbf{1})$. Also, evaluating λ_i at $\hat{\beta} + J_{|\mathcal{A}|/m}$ gives $\hat{\lambda}_i \frac{|\mathcal{A}|}{m}$, and substituting this into equation (A.2) at $\mathbf{w} = \mathbf{1}$,

$$s_j(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1}) = \sum_{i=1}^{n} x_{ij} - \sum_{i=1}^{m} x_{ij} \hat{\lambda}_i \frac{|\mathcal{A}|}{m}$$

$$\xrightarrow{\mathcal{P}} \int_{y \in \mathcal{A}} x_j(y) \hat{\lambda}(y) \, dy$$

from the weak law of large numbers. But this is the derivative of $l(\beta; \mathbf{y})$ from equation (2.2), evaluated at the maximum likelihood estimate $\hat{\beta}$, and so it equals zero for each j and, hence, $\mathbf{S}(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1}) \xrightarrow{\mathcal{P}} \mathbf{0}$. Similarly, for each (j, j'), $I_{jj'}(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1}) \xrightarrow{\mathcal{P}} I_{jj'}(\hat{\beta})$, the (j, j')th element of the Fisher information matrix for $\hat{\beta}$ from $l(\beta; \mathbf{y})$. So returning to equation (A.6),

$$\delta = \mathbf{I}(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1})^{-1} \mathbf{S}(\hat{\beta} + J_{|\mathcal{A}|/m}; \mathbf{1}) + O_p(\|\delta\|^2) \xrightarrow{\mathcal{P}} 0,$$

completing the proof. $\square$

REFERENCES

AUSTIN, M. P. (1985). Continuum concept, ordination methods and niche theory. *Annual Review of Ecology, Evolution, and Systematics* **16** 39–61.

BADDELEY, A. and TURNER, R. (2005). Spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software* **12** 1–42.

BADDELEY, A. J. and VAN LIESHOUT, M. (1995). Area-interaction point processes. *Ann. Inst. Statist. Math.* **47** 601–619. MR1370279

BADDELEY, A. J., MOLLER, J. and WAAGEPETERSEN, R. (2000). Non- and semiparametric estimation of interaction in inhomogeneous point patterns. *Statist. Neerlandica* **54** 329–350. MR1804002

BERMAN, M. and TURNER, T. (1992). Approximating point process likelihoods with GLIM. *J. Roy. Statist. Soc. Ser. C* **41** 31–38.

BURNHAM, K. P. and ANDERSON, D. R. (1998). *Model Selection and Inference—A Practical Information-Theoretic Approach*. Springer, New York. MR1919620

CHEFAOUI, R. M. and LOBO, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling* **210** 478–486.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641

DIGGLE, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*, 2nd ed. Arnold, London. MR0743593

ELITH, J. and LEATHWICK, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13** 265–275.

ELITH, J. and LEATHWICK, J. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* **40** 677–697.

ELITH, J., LEATHWICK, J. R. and HASTIE, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology* **77** 802–813.

GUISAN, A., GRAHAM, C. H., ELITH, J. and HUETTMANN, F. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions* **13** 332–340.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman & Hall, Boca Raton, FL. MR1082147

HERNANDEZ, P. A., FRANKE, I., HERZOG, S. K., PACHECO, V., PANIAGUA, L., QUINTANA, H. L., SOTO, A., SWENSON, J. J., TOVAR, C., VALQUI, T. H., VARGAS, J. and YOUNG, B. E. (2008). Predicting species distributions in poorly-studied landscapes. *Biodiversity and Conservation* **17** 1353–1366.

LEPAGE, G. (1978). A new algorithm for adaptive multidimensional integration. *J. Comput. Phys.* **27** 192–203.

OWEN, A. B. (2007). Infinitely imbalanced logistic regression. *J. Mach. Learn. Res.* **8** 761–773. MR2320678

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

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

PHILLIPS, S. J., DUDÍK, M., ELITH, J., GRAHAM, C. H., 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** 181–197.

R DEVELOPMENT CORE TEAM (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

WARD, G. (2007). Statistics in ecological modelling; presence-only data and boosted MARS. PhD thesis, Dept. Statistics, Stanford Univ. Available at http://www-stat.stanford.edu/~hastie/THESES/Gill_Ward.pdf.

WARD, G., HASTIE, T., BARRY, S., ELITH, J. and LEATHWICK, J. R. (2009). Presence-only data and the EM algorithm. *Biometrics* **65** 554–563.

ZARNETSKE, P. L., EDWARDS, T. C., JR. and MOISEN, G. G. (2007). Habitat classification modeling with incomplete data: Pushing the habitat envelope. *Ecological Applications* **17** 1714–1726.

SCHOOL OF MATHEMATICS AND STATISTICS
 AND EVOLUTION & ECOLOGY RESEARCH CENTRE
UNIVERSITY OF NEW SOUTH WALES
SYDNEY, NSW 2052
AUSTRALIA
E-MAIL: David.Warton@unsw.edu.au

SCHOOL OF MATHEMATICS AND STATISTICS
UNIVERSITY OF NEW SOUTH WALES
SYDNEY, NSW 2052
AUSTRALIA
E-MAIL: leah.shepherd@iinet.net.au