

# Accounting for Imperfect Detection and Survey Bias in Statistical Analysis of Presence-only Data

Robert M. Dorazio, Southeast Ecological Science Center, U.S. Geological Survey, 7920  
NW 71st Street, Gainesville, Florida 32653, USA. E-mail: bdorazio@usgs.gov

## Abstract

**Aim** During the past decade ecologists have attempted to estimate the parameters of species distribution models by combining species presence locations observed in opportunistic surveys with spatially referenced covariates of occurrence. Several statistical models have been proposed for the analysis of presence-only data, but these models have largely ignored the effects of imperfect detection and survey bias. In this paper I describe a model-based approach for the analysis of presence-only data that accounts for errors in detection of individuals and for biased selection of survey locations.

**Innovation** I develop a hierarchical, statistical model that allows presence-only data to be analyzed in conjunction with data acquired independently in planned surveys. One component of the model specifies the spatial distribution of individuals within a bounded, geographic region as a realization of a spatial point process. A second component of the model specifies two kinds of observations, the detections of individuals encountered during opportunistic surveys and the detections of individuals encountered during planned surveys.

**Main conclusions** Using mathematical proof and simulation-based comparisons, I demonstrate that biases induced by errors in detection or biased selection of survey locations can be reduced or eliminated by using the hierarchical model to analyze presence-only data in conjunction with counts observed in planned surveys. I show that a relatively small amount of high-quality data (from planned surveys) can be used to leverage the information in presence-only observations, which usually have broad spatial coverage but may not be informative of both occurrence and detectability of individuals. Because a variety of sampling

1 protocols can be used in planned surveys, this approach to the analysis of presence-only data  
2 is widely applicable. In addition, since the point-process model is formulated at the level of  
3 an individual, it can be extended to account for biological interactions between individuals  
4 and temporal changes in their spatial distributions.

5 **Key-words:** ecological niche model, N-mixture model, predictive biogeography, site occu-  
6 pancy model, spatial point process, species distribution model

7 **Running title:** Statistical analysis of presence-only data

8 **Word count in abstract:** 302

9 **Word count in body:** 6178

10 **Number of references:** 53

## 11 **Introduction**

12 Species distribution models (SDMs) are used to predict the spatial distribution of a  
13 species, often as a function of spatially varying covariates. SDMs have a variety of uses  
14 (Elith and Leathwick, 2009), but predicting the occurrence of a species within its potential  
15 geographic range is perhaps the most common (Scott et al., 2002). These predictions are  
16 particularly relevant in conservation or management problems that require assessments of the  
17 effects of environmental modifications on a species' geographic distribution (Cabeza et al.,  
18 2004).

19 During the past decade ecologists have attempted to estimate the parameters of a SDM  
20 using only locations where a species is observed. These so-called *presence-only* observations  
21 are often made during opportunistic surveys and are recorded in museum collections or online  
22 databases. To fit a SDM, the presence-only observations are supplemented with measures  
23 of potential covariates of species occurrence stored in geographical information systems or  
24 other geographical databases. These covariate measurements are generally available at a  
25 grid of locations that spans the study area.

1       Analyses of presence-only data, which include the presence locations and the spatially  
2       referenced covariate measurements, are motivated, in part, by the difficulty and expense of  
3       conducting planned surveys of natural populations. As the region of interest grows in size  
4       and spatial complexity, so does the number of surveys required to predict species occurrence  
5       accurately. Because presence-only data usually have broad spatial coverage, they provide  
6       attractive sources of information for SDMs.

7       A variety of statistical models have been used to analyze presence-only data (Elith et al.,  
8       2006). Many of these models are strictly appropriate for the analysis of data observed  
9       in planned surveys, where presence or absence of individuals is observable at each survey  
10       location. The predictive accuracy of these models generally suffers when individuals are  
11       present at locations that have not been surveyed, as with presence-only data. In this case  
12       alternative models that allow for presence of individuals at unsurveyed locations are generally  
13       more accurate in predicting the spatial distribution of species occurrences. These models  
14       include Maxent (Phillips et al., 2006; Elith et al., 2010), models of case-augmented binary  
15       outcomes (Lee et al., 2006; Lele and Keim, 2006), and models of spatial point patterns  
16       (Warton and Shepherd, 2010).

17       In the latter class of models, presence-only locations are modeled as a realization of  
18       a spatial point process – specifically, a Poisson process wherein the first-order intensity  
19       function specifies a SDM. This approach offers several advantages to analyses of presence-  
20       only data. First, the parameters of a point-process model are invariant to spatial scale,  
21       so the model can be used to predict the abundance or occurrence of individuals for any  
22       subregion located within the region of interest. In contrast, the parameters of other models  
23       (Maxent and case-augmented binary regression) vary with the spatial resolution of the data.  
24       Technical equivalences have recently been established between the parameters of Poisson  
25       process models and those of Maxent models (Fithian and Hastie, 2013; Renner and Warton,  
26       2013) and case-augmented binary regression models (Dorazio, 2012). Specifically, as the  
27       spatial resolution of the data is increased, the scale-dependent parameters of the latter two

1 models converge to those of Poisson process models; therefore, it can be argued that spatial  
2 point-process models provide a conceptual unification for classes of models whose parameters  
3 are not invariant to spatial scale.

4 One limitation of the point-process model of Warton and Shepherd (2010) is that it  
5 fails to account for the effects of errors in detection of individuals. This limitation is im-  
6 portant because nearly all surveys of natural populations, including opportunistic surveys  
7 that produce presence-only observations, are prone to detection errors (Yoccoz et al., 2001;  
8 Chen et al., 2013). In addition, Dorazio (2012) proved that failure to account for imperfect  
9 detectability in models of presence-only data induces bias in estimates of SDMs when the  
10 covariates of abundance are not distinct and stochastically independent of the covariates of  
11 detectability (also see Lahoz-Monfort et al. (2014)). Bias in estimates of SDMs can also  
12 occur if the presence-only locations are unrepresentative of the region of interest (Phillips  
13 et al., 2009; Yackulic et al., 2013). For example, this source of bias might be produced if  
14 survey locations are selected based on their accessibility or convenience. To alleviate these  
15 biases, Chakraborty et al. (2011) and Fithian and Hastie (2013) proposed models based on a  
16 location-dependent thinning of a Poisson process; however, the parameters of these models  
17 are not identifiable unless the covariates of abundance are distinct and linearly independent  
18 of the covariates of detectability.

19 In this paper I propose a hierarchical, statistical model that allows presence-only data to  
20 be analyzed in conjunction with data acquired independently in planned surveys. Previously,  
21 such data have been used only to validate the predictions of SDMs fitted to presence-only  
22 data (Newbold et al., 2010; Peterman et al., 2011; Gormley et al., 2013). Here I show that  
23 jointly modeling both kinds of data allows the parameters of a SDM to be estimated while  
24 accounting for the effects of imperfect detection and survey bias. In the model’s hierarchy  
25 one component specifies the SDM as a spatial point process and a second component, which  
26 is conditional on the first, specifies two kinds of observations: presence-only locations in  
27 opportunistic surveys and location-specific counts in planned surveys. I establish a set of

1 restrictive conditions under which the parameters of a SDM can be estimated from presence-  
2 only data alone; however, I also show that these conditions may safely be ignored when the  
3 counts observed in planned surveys are analyzed in conjunction with the presence-only data.  
4 I show that this approach is widely applicable owing to the variety of sampling protocols –  
5 including site-occupancy sampling – that can be used to conduct planned surveys.

## 6 **Modeling the Spatial Distribution of a Species and its** 7 **Detection in Opportunistic and Planned Surveys**

8 In this section I describe a hierarchical model composed of two components. One compo-  
9 nent specifies the spatial distribution of individuals within a bounded, geographic region that  
10 is relevant in the context of some scientific or management-related problem. This component  
11 is the true, but unknown, SDM and specifies how the limiting, expected density of individ-  
12 uals – which will be given a mathematically precise definition – varies geographically as a  
13 function of one or more spatially varying covariates (such as measures of habitat quality).

14 The second component specifies models for two kinds of observations: (1) detections of  
15 individuals encountered during opportunistic surveys (i.e., presence-only observations) and  
16 (2) detections of individuals encountered during planned surveys of locations selected using  
17 a prescribed sampling design. These observations depend on several factors, such as the area  
18 of the surveyed locations, the number of individuals present at these locations, the effects  
19 of spatially varying covariates on an observer’s detection ability, and heterogeneity among  
20 observers in detection skills. In the following sections I describe both components of the  
21 hierarchical model.

### 22 **Spatial point-process model of individual locations**

23 I assume that the spatial distribution of individuals – or, more correctly, of the activity  
24 centers of mobile individuals – may be modeled using a Poisson point process. This model

1 has been used in the analysis of spatial capture-recapture data (Efford, 2004; Borchers and  
 2 Efford, 2008; Dorazio, 2013) and in the analysis of presence-only locations (Warton and  
 3 Shepherd, 2010). The latter analysis was intended as a SDM, but it did not consider the  
 4 ramifications of detection errors or sources of survey bias. The hierarchical model proposed  
 5 here is intended to correct this deficiency.

6 To formulate the SDM, I consider individuals that reside within a bounded, geographic  
 7 region  $B \subset \mathbb{R}^2$ , where  $\mathbb{R}^2$  denotes the real plane. The choice of  $B$  is often driven by issues  
 8 related to science, management, or conservation of a species. I assume that the activity  
 9 centers of these individuals are a realization of a Poisson point process parameterized by a  
 10 first-order intensity function  $\lambda(\mathbf{s})$ , where  $\mathbf{s}$  denotes a location (point) in  $B$ . In the context  
 11 of SDMs,  $\lambda(\mathbf{s})$  denotes the limiting, *expected* density of individuals (number of individuals  
 12 per unit area) at location  $\mathbf{s}$ . That is, for a small region  $d\mathbf{s}$  of area  $A(d\mathbf{s})$  centered at  $\mathbf{s}$ ,

$$\lambda(\mathbf{s}) = \lim_{A(d\mathbf{s}) \rightarrow 0} \text{E}\{N(d\mathbf{s})\}/A(d\mathbf{s})$$

13 for the Poisson point process  $N$  (Cressie and Wikle, 2011). The total number  $N(B)$  of indi-  
 14 viduals in the region is a Poisson random variable that depends on the mean intensity of the  
 15 process over  $B$ ,  $\mu(B) = \int_B \lambda(\mathbf{s}) d\mathbf{s}$ . In other words,  $N(B) \sim \text{Poisson}(\mu(B))$ . Furthermore, if  
 16  $B$  is partitioned into a set of disjoint (nonoverlapping) subregions, say  $C_1 \cup \dots \cup C_K = B$ , the  
 17 number of individuals in each subregion also is a Poisson random variable. In other words,  
 18  $N(C_k) \sim \text{Poisson}(\mu(C_k))$ , where  $\mu(C_k) = \int_{C_k} \lambda(\mathbf{s}) d\mathbf{s}$  (Møller and Waagepetersen, 2004; Il-  
 19 lian et al., 2008). The number of individuals present in one subregion is also independent of  
 20 the number in another subregion, a property that will be exploited in other components of  
 21 the hierarchical model.

22 To specify a SDM,  $\lambda(\mathbf{s})$  is formulated as a log-linear function of unknown parameters  
 23  $\boldsymbol{\beta}$  and location-specific regressors  $\mathbf{x}(\mathbf{s})$  as follows:  $\log\{\lambda(\mathbf{s})\} = \boldsymbol{\beta}'\mathbf{x}(\mathbf{s})$ . (The prime symbol  
 24 indicates the transpose of a matrix or vector.) The regressors  $\mathbf{x}(\mathbf{s})$  may be computed from

1 covariates measured at a grid of locations that span  $B$  (e.g., measures of habitat quality).  
 2 Therefore, accurate estimates of  $\beta$  may be used to predict the abundance or occurrence of  
 3 individuals for any location or any subregion within  $B$ .

4 Let  $n$  denote the unknown total number of individuals in region  $B$ . I have established  
 5 that

$$\Pr(N(B) = n) = \exp\{-\mu(B)\}\{\mu(B)\}^n/n! \quad (1)$$

6 Let  $\mathbf{S}_i \in B$  denote a random variable for the activity center of the  $i$ th individual residing  
 7 in region  $B$  ( $i = 1, \dots, n$ ). Conditional on  $n$ , the joint probability density of the  $n$  activity  
 8 centers is

$$f(\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n | n) = \prod_{i=1}^n \lambda(\mathbf{s}_i) / \mu(B) \quad (2)$$

(Møller and Waagepetersen, 2004; Illian et al., 2008). If  $n$  and  $\{\mathbf{s}_i\}$  were observable, their  
 joint density,

$$g(\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n, n) = \frac{\exp\{-\mu(B)\}}{n!} \prod_{i=1}^n \lambda(\mathbf{s}_i),$$

9 (obtained by combining (1) and (2)) could be used as a likelihood function for estimating  $\beta$   
 10 (Warton and Shepherd, 2010). However, only some of the  $n$  individuals are actually observed  
 11 owing to errors in detection and survey biases. The second component of the hierarchical  
 12 model (described in the following two sections) allows  $\beta$  to be estimated while accounting  
 13 for unobserved individuals.

## 14 **Detections of individuals in opportunistic surveys**

15 To model the detections of individuals encountered during opportunistic surveys, I make  
 16 several assumptions. First, I assume that each individual is detected independently by  
 17 an observer with probability  $p(\mathbf{s})$  that depends only on the location  $\mathbf{s}$  of an individual.  
 18 Therefore,  $p(\mathbf{s})$  includes both an observer's detection ability and choice of survey location.  
 19 For example,  $p(\mathbf{s})$  will be zero if an observer does not survey individuals at location  $\mathbf{s}$

1 because the location is inaccessible. I assume also that the presence of other individuals  
 2 encountered during an opportunistic survey has no effect on an observer’s detection rate.  
 3 Instead, the probability of detecting an individual located at  $\mathbf{s}$  is assumed to be a logit-  
 4 linear function of unknown parameters  $\boldsymbol{\alpha}$  and location-specific regressors  $\mathbf{w}(\mathbf{s})$  as follows:  
 5  $\text{logit}\{p(\mathbf{s})\} = \boldsymbol{\alpha}'\mathbf{w}(\mathbf{s})$ . The regressors  $\mathbf{w}(\mathbf{s})$  are assumed to be computable at all locations in  
 6  $B$  and may include features thought to influence an observer’s detection ability or choice of  
 7 survey location. For example, distance to nearest road could be included in  $\mathbf{w}(\mathbf{s})$  if observers  
 8 were known to have chosen survey locations based on their proximity to roads.

9 Although the identity of each observer potentially could be used as a regressor of  $p(\mathbf{s})$ ,  
 10 especially in cases with few observers, most presence-only data include many observers and  
 11 their identities are not always available. Therefore, I make the simplifying assumption that  
 12  $\boldsymbol{\alpha}$  contains a single intercept parameter. Although differences in  $p(\mathbf{s})$  may exist due to  
 13 differing abilities of observers, I assume here that these differences are small in comparison  
 14 to the effects of spatially varying covariates on an observer’s detection ability and choice of  
 15 survey location.

16 A final assumption concerns the effects of movements of individuals around their activity  
 17 centers. Unlike spatial capture-recapture data, presence-only data contain only a single  
 18 observation (and location) for each individual. This leaves no information to estimate the  
 19 magnitude of individual movements; therefore, I assume that individuals are detected at  
 20 their activity centers, recognizing that this assumption is likely to be violated if individuals  
 21 are highly mobile or possess large territories (that is, large relative to the size of region  $B$ ).

22 Given these assumptions, the detections of individuals in opportunistic surveys may be  
 23 modeled as a location-dependent thinning of the Poisson point process described earlier. Let  
 24  $Y$  denote a binary random variable whose observed value indicates whether an individual  
 25 residing in region  $B$  is detected ( $Y = 1$ ) or not ( $Y = 0$ ) during incidental surveys of the  
 26 region. I assume  $Y_i|\mathbf{s}_i \sim \text{Bernoulli}(p(\mathbf{s}_i))$ , which specifies a marked point-process model of  
 27  $(y_i, \mathbf{s}_i)$  ( $i = 1, \dots, n$ ). However, the individuals detected in opportunistic surveys all have

1 the same mark (that is,  $Y = 1$  for these individuals). Therefore, let  $M(B)$  denote a random  
 2 variable for the number of individuals that are present and detected during opportunistic  
 3 surveys of region  $B$ . It is easily shown that

$$\Pr(M(B) = m) = \exp\{-\nu(B)\}\{\nu(B)\}^m/m! \quad (3)$$

4 where  $\nu(B) = \int_B \lambda(\mathbf{s})p(\mathbf{s}) d\mathbf{s}$  and that  $M$  is a Poisson process resulting from an independent  
 5 thinning of the original process  $N$  with probability  $1 - p(\mathbf{s})$  (Møller and Waagepetersen,  
 6 2004; Illian et al., 2008). It follows that  $\nu(B) = E(M(B))$  and that the joint density of  $m$   
 7 and the locations of individuals detected in the opportunistic surveys is

$$h(\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m, m) = \frac{\exp\{-\nu(B)\}}{m!} \prod_{i=1}^m \lambda(\mathbf{s}_i) p(\mathbf{s}_i) \quad (4)$$

8 where the first  $m$  of  $n$  locations are assumed to correspond to those of detected individuals<sup>1</sup>.  
 9 Note that the integral required to compute  $\nu(B)$  cannot be evaluated in closed form. In  
 10 practice  $\nu(B)$  is approximated as a Riemann sum by partitioning  $B$  into a sufficiently fine  
 11 grid.

12 The joint density of the observed data (i.e., (4)) may be used as a likelihood function  
 13  $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$  for estimating the parameters  $\boldsymbol{\beta}$  and  $\boldsymbol{\alpha}$ . In later sections I establish conditions  
 14 required for identifiability of parameters based on this likelihood function.

## 15 **Detections of individuals in planned surveys**

16 Several survey protocols allow the abundance of imperfectly detected individuals to be  
 17 estimated from data observed in spatial sample units. Examples of these protocols include  
 18 double-observer surveys, removal surveys, and replicated point-count surveys (Royle and  
 19 Dorazio, 2008, chapter 8). In practice the choice of protocol often depends on the species  
 20 and on the methods used to detect or capture individuals.

---

<sup>1</sup>The order of the  $n$  observations can be changed without loss of generality.

1        Suppose one of these protocols is used to supplement the information obtained through  
2 opportunistic surveys. The idea is to select a representative sample of region  $B$  using a planned  
3 design and to survey individuals at each sample location using a protocol that is informative  
4 of both abundance and detectability. The number of locations in this sample usually will  
5 be small in comparison to the number of presence-only locations observed in opportunistic  
6 surveys. To illustrate this approach, I describe here a model of replicated point counts at  
7 each location, though any of the survey protocols mentioned earlier can be applied without  
8 loss of generality.

9        For the purposes of sampling, assume that region  $B$  is partitioned into a finite number of  
10 disjoint sample units (e.g., quadrats, strip transects, etc.). Let  $C_1, C_2, \dots, C_K$  denote a repre-  
11 sentative (but not necessarily random) sample of these units, and let  $A(C_1), A(C_2), \dots, A(C_K)$   
12 denote their areas. Under the assumptions of the SDM, the number  $N(C_k)$  of individual ac-  
13 tivity centers in sample unit  $C_k$  is a Poisson random variable with mean  $\mu(C_k) = \int_{C_k} \lambda(\mathbf{s}) d\mathbf{s}$ .  
14 If the values of spatially varying regressors  $x(\mathbf{s})$  are constant within a sample unit, which is  
15 often true in practice, the Poisson mean may be expressed as a function of the area of the  
16 unit and the fixed value of the regressors  $\mathbf{x}(C_k)$  as follows:  $\mu(C_k) = A(C_k) \exp\{\boldsymbol{\beta}' \mathbf{x}(C_k)\}$ .  
17 Thus, the expected number of individual activity centers in each sample unit increases with  
18 its area.

19        Suppose  $J_k$  independent point-count surveys of individuals in unit  $C_k$  are completed  
20 during a period where the number of individuals present in  $C_k$  remains constant. Inde-  
21 pendence can be achieved by using different observers or methods of detection, or by al-  
22 lowing sufficient time to elapse between successive surveys. These surveys yield a vector  
23  $\mathbf{y}_k = (y_{k1}, y_{k2}, \dots, y_{kJ_k})'$  of counts that contains the number of individuals detected during  
24 each survey of unit  $C_k$ . Generally speaking, the number of individuals available to be de-  
25 tected in a sample unit depends on the locations of individual activity centers and on the  
26 spatial extent of individual movements about their activity centers. If the survey protocol is  
27 repeated over time, the number of individuals that are present and available to be detected

1 in a sample unit can be estimated (Chandler et al., 2011). However, for many species tem-  
 2 poral replication of the survey protocol is often physically or logistically infeasible owing to  
 3 sampling constraints. In the absence of such replication I assume that only individuals whose  
 4 activity centers lie within a sample unit are available to be detected; thus,  $N(C_k)$  denotes a  
 5 random variable for the number of individuals present and available to be detected in unit  
 6  $C_k$ . This assumption may not be valid for highly mobile species (large birds or mammals),  
 7 but it will be satisfied for species whose movements are more limited (plants, clams, some  
 8 amphibians, small nesting birds, etc.).

To specify the effects of imperfect detection on the observed point counts, I use the  
 product-binomial model proposed by Royle (2004):

$$\Pr(\mathbf{Y}_k = \mathbf{y}_k | N(C_k) = n_k) = \prod_{j=1}^{J_k} \binom{n_k}{y_{kj}} p_{kj}^{y_{kj}} (1 - p_{kj})^{n_k - y_{kj}}$$

9 where  $p_{kj}$  is the conditional probability of detecting an individual given that it is present and  
 10 available to be observed during the  $j$ th survey of sample unit  $C_k$ . The detection probability  
 11  $p_{kj}$  may be specified as a function of unknown parameters and covariate measurements that  
 12 differ among sample units or surveys. These covariates may include those used to model  
 13 detections in opportunistic surveys, with the exception of the covariates used to specify  
 14 effects of survey bias. The latter covariates are obviously unnecessary since the locations of  
 15 sample units in planned surveys are prescribed by design. I assume here that  $p_{kj}$  is a logit-  
 16 linear function of unknown parameters  $\boldsymbol{\gamma}$  and regressors  $\mathbf{v}(C_k)$  whose values are constant  
 17 within unit  $C_k$ :  $\text{logit}(p_{kj}) = \boldsymbol{\gamma}'\mathbf{v}(C_k)$ . This function implies that the probability of detecting  
 18 an individual during planned surveys can differ among sample units (i.e., spatially) but not  
 19 among surveys within a unit.

The unconditional probability of the point counts observed in sample unit  $C_k$  is obtained

by marginalizing the joint density of  $n_k$  and  $\mathbf{y}_k$  over the admissible values of  $n_k$  as follows:

$$\Pr(\mathbf{Y}_k = \mathbf{y}_k) = \sum_{n_k = \max(\mathbf{y}_k)}^{\infty} \frac{\exp\{-\mu(C_k)\} \{\mu(C_k)\}^{n_k}}{n_k!} \prod_{j=1}^{J_k} \binom{n_k}{y_{kj}} p_{kj}^{y_{kj}} (1 - p_{kj})^{n_k - y_{kj}}$$

1 In practice, the infinite upper limit of summation is replaced with an integer that is suf-  
 2 ficiently large to ignore the contributions of additional terms. A likelihood function for  
 3 estimating the parameters  $\beta$  and  $\gamma$  from the point counts is

$$L(\beta, \gamma) = \prod_{k=1}^K \Pr(\mathbf{Y}_k = \mathbf{y}_k) \quad (5)$$

4 which stems from the independence of abundances and counts among sample units. If two  
 5 or more point counts are observed in each sample unit, the parameters of this model ( $\beta$   
 6 and  $\gamma$ ) are identifiable (Royle, 2004). If only one point count is observed in each sample  
 7 unit, the model's parameters may be identifiable if some of the covariate measurements used  
 8 as abundance regressors are distinct (or at least linearly independent) from the covariate  
 9 measurements used as regressors of detection probability (Sólymos et al., 2012).

## 10 **Estimating SDMs from Detections of Individuals in Op-** 11 **portunistic Surveys**

12 In this section I identify requirements for estimating SDMs using only presence-only  
 13 observations. First, if detection probability  $p$  is assumed to be constant at all locations (that  
 14 is, if  $\alpha = \alpha_0 = \text{logit}(p)$ ), then  $\beta_0$  and  $\alpha_0$  are not identifiable and the SDM is not estimable.

1 This result is easily deduced from the log-likelihood function of this model's parameters:

$$\begin{aligned}
\log\{L(\boldsymbol{\beta}, p)\} &= - \int_B \lambda(\mathbf{s}) p d\mathbf{s} + \sum_{i=1}^m \log\{\lambda(\mathbf{s}_i) p\} \\
&= -p \int_B \exp\{\beta_0 + \tilde{\boldsymbol{\beta}}' \tilde{\mathbf{x}}(\mathbf{s})\} d\mathbf{s} + m\{\beta_0 + \log(p)\} + \sum_{i=1}^m \tilde{\boldsymbol{\beta}}' \tilde{\mathbf{x}}(\mathbf{s}_i) \\
&= - \exp\{\beta_0 + \log(p)\} \int_B \exp\{\tilde{\boldsymbol{\beta}}' \tilde{\mathbf{x}}(\mathbf{s})\} d\mathbf{s} + m\{\beta_0 + \log(p)\} + \sum_{i=1}^m \tilde{\boldsymbol{\beta}}' \tilde{\mathbf{x}}(\mathbf{s}_i)
\end{aligned}$$

2 where the  $\tilde{\boldsymbol{\beta}}' \tilde{\mathbf{x}}(\mathbf{s})$  corresponds to terms in the log-linear predictor of  $\lambda(\mathbf{s})$  that do not include  
3  $\beta_0$ . In this log-likelihood function the parameters  $\beta_0$  and  $p$  appear only as a sum ( $\beta_0 + \log(p)$ );  
4 therefore, only the sum of these parameters is identified, and unique estimates of  $\beta_0$  and  $p$   
5 cannot be obtained.

6 More generally, consider models of presence-only data that assume spatial variation in  
7 both  $\lambda$  and  $p$ . For these models the log-likelihood function is

$$\log\{L(\boldsymbol{\beta}, \boldsymbol{\alpha})\} = - \int_B \lambda(\mathbf{s}) p(\mathbf{s}) d\mathbf{s} + \sum_{i=1}^m \log\{\lambda(\mathbf{s}_i) p(\mathbf{s}_i)\} \quad (6)$$

(= the logarithm of the right-hand side of (4), ignoring terms that don't include parameters).  
Mathematical proofs of parameter identifiability based on (6) are challenging because the  
integral cannot be evaluated in closed form and because neither the product of  $\lambda(\mathbf{s})$  and  
 $p(\mathbf{s})$ , which equals

$$\lambda(\mathbf{s}) p(\mathbf{s}) = \frac{\exp\{\boldsymbol{\beta}' \mathbf{x}(\mathbf{s}) + \boldsymbol{\alpha}' \mathbf{w}(\mathbf{s})\}}{1 + \exp\{\boldsymbol{\alpha}' \mathbf{w}(\mathbf{s})\}},$$

8 nor its logarithm can be expressed as a linear combination of  $\boldsymbol{\beta}$  and  $\boldsymbol{\alpha}$ .

9 One exception occurs when the probabilities of detecting individuals are relatively low  
10 (say,  $p(\mathbf{s}) < 0.2$ ) at all locations in  $B$ . In this case,  $\lambda(\mathbf{s}) p(\mathbf{s}) \doteq \exp\{\boldsymbol{\beta}' \mathbf{x}(\mathbf{s}) + \boldsymbol{\alpha}' \mathbf{w}(\mathbf{s})\}$ ,  
11 so the parameters  $\beta_0$  and  $\alpha_0$  occur only as a sum ( $\beta_0 + \alpha_0$ ) and again are not identifiable.  
12 Similarly, other elements of  $\boldsymbol{\beta}$  and  $\boldsymbol{\alpha}$  are not identified when the regressors  $\mathbf{x}$  and  $\mathbf{w}$  are  
13 linearly dependent (Fithian and Hastie, 2013). This implies that in circumstances where  $p$

1 is relatively low at all locations in  $B$ , the regressors of  $\lambda$  and  $p$  should include covariates that  
2 are distinct and not strongly correlated (either positively or negatively).

3 To assess the identifiability of parameters more generally for models in which  $\lambda$  and  $p$   
4 are assumed to vary among locations, I derived the Fisher information matrix  $\mathbf{I}(\boldsymbol{\theta})$  for log-  
5 likelihood function (6), where  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ . Recall that  $\mathbf{I}(\boldsymbol{\theta})$  equals the negative of the  
6 expected value of the Hessian matrix obtained by taking second partial derivatives of the  
7 log-likelihood function. If the Fisher information matrix has full rank (i.e., if  $\mathbf{I}(\boldsymbol{\theta})$  is positive  
8 definite and invertible), the parameters in  $\boldsymbol{\theta}$  are identifiable (Bowden, 1973). Therefore,  
9 parameter identifiability may be assessed by examining the condition number of  $\mathbf{I}(\boldsymbol{\theta})$ , which  
10 equals the ratio of largest to smallest eigenvalues of this matrix.

11 The elements of  $\mathbf{I}(\boldsymbol{\theta})$  cannot be expressed in closed form, but they can be evaluated  
12 for fixed values of the parameters ( $\boldsymbol{\beta}$  and  $\boldsymbol{\alpha}$ ) and the regressors ( $\mathbf{x}$  and  $\mathbf{w}$ ) using numerical  
13 integration (see Appendix S1). The eigenvalues of  $\mathbf{I}(\boldsymbol{\theta})$  can then be calculated from these in-  
14 tegral approximations. If  $\mathbf{I}(\boldsymbol{\theta})$  is positive definite, all of its eigenvalues are positive; however,  
15 if  $\mathbf{I}(\boldsymbol{\theta})$  is positive semi-definite (not full rank), the ratio of smallest to largest eigenvalues  
16 (reciprocal of condition number) will be zero, though numerical evaluations of this ratio  
17 may not equal zero exactly. Therefore, this ratio provides a useful diagnostic for assessing  
18 whether the parameters of a SDM are identifiable.

19 To illustrate the utility of the Fisher information matrix, I compared two different  
20 presence-only models. The SDM of both models was identical – a single spatially varying  
21 covariate (Figure 1, upper panel) was used to predict  $\lambda(\mathbf{s})$  as follows:

$$\log\{\lambda(\mathbf{s})\} = \log(8000) + 0.5x(\mathbf{s}) \quad (7)$$

22 The covariate measurements were centered and scaled so that  $x$  had zero mean and unit  
23 variance. The two models of presence-only data differed in their specification of spatially  
24 varying detection probabilities. In one model a single covariate  $w$  (Figure 1, lower panel)

1 whose values were computed independently of  $x$  was used to predict  $p(\mathbf{s})$  as follows:

$$\text{logit}\{p(\mathbf{s})\} = \alpha_0 - 1.0 w(\mathbf{s}) \quad (8)$$

2 The covariate measurements were centered and scaled so that  $w$  had zero mean and unit  
3 variance. The parameter  $\alpha_0$  was assigned values ranging from -5 to 5 so that detection  
4 probabilities at the average value of the covariate ranged from very low values (near zero) to  
5 very high values (near one). In the second model the covariate  $x$ , which was used to predict  
6  $\lambda(\mathbf{s})$ , also was used to predict  $p(\mathbf{s})$ , that is

$$\text{logit}\{p(\mathbf{s})\} = \alpha_0 - 1.0 x(\mathbf{s}) \quad (9)$$

7 The values of  $\alpha_0$  and  $\alpha_1$  used in both models of detection probabilities were identical. The  
8 second model was intended to be representative of a worst-case scenario wherein the regres-  
9 sors of  $\lambda(\mathbf{s})$  and  $p(\mathbf{s})$  were perfectly correlated.

10 Whereas the expected number of individuals detected in region  $B$  ( $\nu(B)$ ) increased from  
11 242 to 35,470 over the assumed range of  $\alpha_0$  values, the difference in  $\nu(B)$  between models for  
12 any single value of  $\alpha_0$  was minor in comparison (Figure 2). In contrast, the ratio of smallest  
13 to largest eigenvalues of  $\mathbf{I}(\boldsymbol{\theta})$  signaled striking differences in the identifiability of these two  
14 models' parameters. For example, regardless of the value of  $\alpha_0$ , the ratio of smallest to largest  
15 eigenvalues was nearly constant (approximately zero) when identical regressors were assumed  
16 for  $\lambda(\mathbf{s})$  and  $p(\mathbf{s})$  (Figure 2). For the other model, which assumed independent regressors, the  
17 ratio of smallest to largest eigenvalues was highest at  $\alpha_0 = 0$  (corresponding to a detection  
18 probability of 0.5 at the average value of the covariate) and declined to approximately zero  
19 at extreme values of  $\alpha_0$  ( $< -4$  or  $> 4$ ). The reason for this pattern is clear. At extreme  
20 values of  $\alpha_0$ , detection probabilities are nearly constant over  $B$  (i.e.,  $p(\mathbf{s})$  is nearly zero (if  
21  $\alpha_0 < -4$ ) or one (if  $\alpha_0 > 4$ ) at every location in  $B$ ). At these extremes the parameters  $\beta_0$   
22 and  $\alpha_0$  are not identifiable and cannot be estimated uniquely, as established earlier.

1 This example illustrates that presence-only data may contain only limited information  
 2 about the parameters of a SDM. If at least some of the regressors used to specify spatial  
 3 differences in  $\lambda$  and  $p$  are not independent, SDMs will be unidentified. The parameters  
 4 of SDMs also can be difficult to estimate if detection probabilities are either very small or  
 5 very large at all locations. However, in these circumstances the ratio of smallest to largest  
 6 eigenvalues of  $\mathbf{I}(\boldsymbol{\theta})$  provides a useful diagnostic. This ratio is also invariant to the magnitude  
 7 of  $\beta_0$  (see Appendix S1) and is therefore relevant to the analysis of presence-only observations  
 8 from populations of all sizes.

## 9 **Estimating SDMs from Detections of Individuals in Op-** 10 **portunistic and Planned Surveys**

11 One way of overcoming the limited information in presence-only observations is to an-  
 12alyze these data in conjunction with the detections of individuals in planned surveys. A  
 13 joint analysis of these two kinds of data leverages the spatial coverage of presence-only ob-  
 14servations, which is usually large, with the strength of information about abundance and  
 15 detection in planned surveys.

16 As an example of this approach, consider a joint analysis of presence-only data and  
 17 replicated point-count survey data. The likelihood functions for these two data sets (i.e.,  
 18 (4) and (5)) both include the parameter  $\boldsymbol{\beta}$ , which determines the SDM. Observations in  
 19 opportunistic and planned surveys are obtained independently, so these two data sets can  
 20 be analyzed together using the following likelihood function:

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = L(\boldsymbol{\beta}, \boldsymbol{\alpha}) \times L(\boldsymbol{\beta}, \boldsymbol{\gamma}) \quad (10)$$

21 Estimates of  $\boldsymbol{\beta}$  obtained by maximizing (10) should have less bias and greater precision  
 22 than estimates obtained by maximizing either (4) or (5) separately. In this section I describe

1 simulation studies that illustrate the benefits of a joint analysis of data observed in oppor-  
2 tunistic and planned surveys. I designed the simulation studies to build upon the models  
3 of presence-only data described in the previous section and to demonstrate that addition  
4 of planned survey data (in this case, point counts) can reduce or eliminate the estimation  
5 problems induced by parameter unidentifiability.

## 6 **Design of simulation studies**

7 In the simulation studies presence-only data were simulated using the two models de-  
8 scribed earlier. That is, in both models  $\lambda(\mathbf{s})$  was specified as a function of  $x(\mathbf{s})$  using (7). In  
9 one model  $p(\mathbf{s})$  was specified as a function of  $w(\mathbf{s})$  (using (8)); in the other model  $p(\mathbf{s})$  was  
10 specified as a function of  $x(\mathbf{s})$  (using (9)). The parameter vector  $\boldsymbol{\alpha}$  was assigned the same  
11 value in both models:  $\boldsymbol{\alpha} = (-1, -1)'$ .

12 The locations of individuals were simulated within a square region  $B$  using the covariate  
13 measurements shown in Figure 1. The spatial distribution of  $\lambda$  and of the expected densities  
14 of detections (one for each model) are shown in Figure 3. Based on these distributions, the  
15 expected number of individuals present in region  $B$  was 35,857, and the expected number  
16 of individuals detected in region  $B$  equaled either 11,251 (for the model where detection  
17 depended on  $w$ ) or 7,922 (for the model where detection depended on  $x$ ). Each simulated  
18 presence-only data set (i.e., the number and locations of detected individuals) was computed  
19 as a realization of a thinned Poisson process, as described earlier.

20 To simulate observations from point-count surveys, region  $B$  was partitioned into a grid of  
21 10,000 square quadrats of equal size. Samples of  $K = 50, 100, 200, 400,$  or  $800$  quadrats were  
22 selected randomly (without replacement), and  $J = 4$  independent point-count surveys were  
23 conducted in each of these quadrats. The sample sizes were deliberately chosen to be small  
24 (0.5 to 8% of the total number of quadrats) relative to the expected number of presence-only  
25 locations to simulate the usual situation where presence-only observations greatly outnumber  
26 the locations visited in planned surveys. The covariates of detectability were identical to

1 those used in the models of presence-only data – that is,  $v(C) = \int_C w(\mathbf{s})d\mathbf{s}$  for one model, and  
2  $v(C) = \int_C x(\mathbf{s})d\mathbf{s}$  for the other model. The logit-scale detection parameters were identical  
3 in both point-count models:  $\boldsymbol{\gamma} = (0, -1)'$ . Note that the parameters  $\alpha_1$  and  $\gamma_1$  were assigned  
4 the same value to ensure that the effects of covariates on detection probability were the  
5 same regardless of whether an individual was encountered during an opportunistic survey or  
6 a planned survey.

7 The spatial distributions of the expected number of individuals per quadrat and of the  
8 expected number of individuals detected per quadrat during a single survey are shown in  
9 Figure 4. The similarity between these distributions and those shown in Figure 3 is not  
10 surprising given the similarity in the detection probability models assumed for opportunistic  
11 and planned surveys. Each simulated set of point counts was computed by aggregating the  
12 realized locations of individuals in  $B$  into quadrats, by selecting a random sample of these  
13 quadrats, and by taking  $J$  independent binomial draws from the individuals present in each  
14 sampled quadrat. This sequence of steps produced  $J$  point counts for each sampled quadrat.

15 A total of 1000 data sets containing both presence-only observations and point counts  
16 were simulated for each of the two models. Maximum-likelihood estimates of  $\boldsymbol{\beta}$  (parameters  
17 of the SDM) were computed for each simulated data set by fitting the model of presence-  
18 only data (using (6)), the model of point counts (using (5)), and the model of combined data  
19 (presence-only observations and point counts) (using (10)). Operating characteristics of the  
20 estimators (such as bias and variance) were estimated from these simulation results and  
21 compared to illustrate the inferential benefits of including both presence-only observations  
22 and point counts in the analysis. Appendix S2 contains the R code (R Core Team, 2014)  
23 that was used to generate and analyze the simulated data sets.

## 24 **Results of simulation studies**

25 For the model in which detections of individuals depended only on covariate  $w$ , presence-  
26 only based estimators of  $\boldsymbol{\beta}$  appeared to have negligibly small bias (Figure 5). When presence-

1 only observations and point counts were analyzed together, the uncertainty in estimates of  
2  $\beta_0$  was reduced considerably, while uncertainty in the estimates of  $\beta_1$  was unchanged.

3 Conspicuously different results were obtained for the model in which detections of indi-  
4 viduals depended on the same covariate used in the SDM (i.e., covariate  $x$ ). For this model  
5 presence-only based estimators of  $\beta_0$  and  $\beta_1$  were strongly biased and highly variable (Fig-  
6 ure 6). This result was expected because the model’s parameters are not well identified, as  
7 established earlier. More surprising were the inferential benefits obtained by including point  
8 counts in the analysis. Adding point counts from as few as 50 quadrats (0.5% of the sample  
9 frame) to the analysis dramatically reduced the bias and the uncertainty in the estimates of  
10  $\beta_0$  and  $\beta_1$ . The estimators appeared to exhibit virtually no bias in point-count samples of  
11 200 or more quadrats.

12 Additional simulation-based comparisons (not shown) using different values of  $\beta_0$  and  
13  $J$  produced results that were qualitatively similar to those illustrated in Figures 5 and 6  
14 – that is, bias and uncertainty of presence-only based estimators of SDMs can be reduced  
15 considerably by analyzing presence-only observations and point counts together in a single  
16 model. An important caveat, however, is that  $J = 4$  replicate point-count surveys in each  
17 quadrat were required to achieve these inferential improvements. Bias and uncertainty of the  
18 presence-only based estimators of SDMs were not reduced if only  $J = 1$  point-count survey  
19 was conducted per quadrat. The reason, of course, is that the parameters of the point-count  
20 model cannot be identified when  $J = 1$  unless some of the covariate measurements used  
21 as abundance regressors are distinct (or at least linearly independent) from the covariate  
22 measurements used as regressors of detection probability (Sólymos et al., 2012).

## 23 Discussion

24 Several statistical models have been proposed for the analysis of presence-only data,  
25 but they have largely ignored the effects of imperfect detectability and survey bias (Dorazio,

1 2012; Lahoz-Monfort et al., 2014). In this paper I showed that the bias in estimates of SDMs  
2 induced by detection errors or survey bias can be reduced or eliminated by modeling presence-  
3 only data in conjunction with counts observed in planned surveys. In this modeling approach  
4 the SDM is specified using a spatial point process and the observable data (presence-only  
5 locations and counts) are specified conditional on a realization of this process.

6 If this approach is adopted but only the presence-only data are analyzed, the parameters  
7 of a SDM may not be identifiable. For example, if the probability of detecting an individual  
8 is identical at all locations, the parameters of a SDM are not identifiable. On the other hand,  
9 if detection probability differs among locations, the parameters of a SDM can be estimated  
10 if some of the regressors used to specify spatial differences in individual density are linearly  
11 independent of the regressors of detection probability. This restriction is similar to that  
12 identified by Fithian and Hastie (2013), who proposed a different thinned Poisson process  
13 model for the analysis of presence-only data. In practice, I showed that the condition number  
14 of the Fisher information matrix (Appendix S1) can be used to assess the identifiability of a  
15 model's parameters.

16 The simulation study I conducted illustrates how including counts from planned surveys  
17 in the analysis of presence-only data generally improves inferences about the parameters of  
18 a SDM. In this way a relatively small amount of high-quality data – which are informative  
19 of both detection and abundance of individuals – can be used to leverage the information in  
20 presence-only observations. This approach follows a recommendation of Phillips and Elith  
21 (2013) that additional data be collected if estimates of absolute species occurrences are  
22 desired.

23 Several benefits, both conceptual and practical, stem from using a point-process model  
24 as a SDM. As shown earlier, the parameters of a point-process model are defined on an areal  
25 basis and are invariant to spatial scale; thus, predictions may be computed for any summary  
26 of the spatial distribution of individuals within the region of interest. For example, under  
27 the assumptions of the Poisson process model that I described, the expected abundance

1  $E(N(C))$  and occurrence  $\Pr(N(C) > 0)$  of individuals in any subregion  $C \subset B$  are defined  
2 implicitly as follows:

$$\begin{aligned} E(N(C)) &= \mu(C) \\ \Pr(N(C) > 0) &= 1 - \exp\{-\mu(C)\} \end{aligned}$$

3 where  $\mu(C) = \int_C \lambda(\mathbf{s}) d\mathbf{s}$  is a function of the model's parameters  $\boldsymbol{\beta}$  and the values of the  
4 spatially varying regressors  $\boldsymbol{x}$  in  $C$ . Phillips and Elith (2013, page 1411) assert that the  
5 Poisson process model cannot be used to estimate the probability of presence, which equals  
6  $\Pr(N(C) > 0)$ . That claim is obviously incorrect. Furthermore, by defining occurrence prob-  
7 ability as a function of  $\mu(C)$ , any prediction of occurrence probability increases automatically  
8 with an increase in the area of subregion  $C$ . This property is not shared by models of occur-  
9 rence, such as Maxent, where occurrence probability is defined per grid cell and depends on  
10 the spatial resolution used in the analysis (Renner and Warton, 2013). For the same reason  
11 conventional site-occupancy models (MacKenzie et al., 2002; Tyre et al., 2003) are limited  
12 for use as SDMs – that is, the interpretation of occurrence probability in site-occupancy  
13 models depends on the spatial resolution used in the analysis. Because of this limitation,  
14 Efford and Dawson (2012) suggested that site-occupancy models be modified to account for  
15 spatial resolution and for the movements of individuals within their home ranges.

16 One practical benefit of using a point-process model as a SDM is that it clarifies how  
17 the location-specific measurements and interpolations of the covariates should be used in the  
18 analysis. In the SDM literature the number and spatial resolution of covariate measurements  
19 and interpolations needed in statistical analyses of presence-only data have been subjects of  
20 considerable debate (Pearce and Boyce, 2006; Guisan et al., 2007; Elith and Leathwick, 2009;  
21 Barbet-Massin et al., 2012). However, as Warton and Shepherd (2010) and I have shown,  
22 the role of the covariate values is clear when using a point-process model as a SDM. Both  
23 measured and interpolated values of covariates over the entire region of interest  $B$  are used

1 when  $B$  is partitioned to approximate the integral in the likelihood function as a Riemann  
 2 sum. The spatial resolution required for accurate approximation is easily determined by  
 3 increasing the resolution until the maximized value of the likelihood function stabilizes.

In this paper I used a model of point counts to illustrate the benefits of adding information from planned surveys to the analysis of presence-only data. Similarly, information from alternative sampling protocols, such as double-observer sampling, removal sampling, or even capture-recapture sampling, can be used to improve the analysis of presence-only data. These counts are then modeled conditional on the latent (unobserved) abundance of individuals within a unit (say,  $C_k$ ). This idea is an extension of Royle's (2004)  $N$ -mixture model of point counts and has been used to analyze many kinds of spatially-referenced counts (Royle and Dorazio, 2008, chapter 8). The key is to exploit the variation in abundance among sample units and to specify the effects of each unit's size, location, and habitat characteristics in the model of abundance. If abundance within sample units is not too high, this approach also can be applied using a site-occupancy survey protocol. For example, suppose  $J_k$  ( $> 1$ ) independent presence-absence surveys are conducted in unit  $C_k$ . Let  $Z_{kj}$  denote a binary random variable whose observed value indicates whether one or more individuals were detected ( $Z_{kj} = 1$ ) or not ( $Z_{kj} = 0$ ) during the  $j$ th survey of sample unit  $C_k$ . As with conventional site-occupancy models,  $Z_{kj}$  is modeled conditional on the latent presence of individuals in unit  $C_k$  as follows:

$$Z_{kj}|N(C_k) = n_k \sim \text{Bernoulli}(q_{jk} I(n_k > 0))$$

where  $q_{kj}$  is the conditional probability of detection during the  $j$ th survey given that one or more individuals are present in unit  $k$ . ( $I(e)$  denotes the indicator function, which equals one if Boolean argument  $e$  is true and zero if  $e$  is false.) Unlike conventional site-occupancy models, the probability of occurrence in unit  $C_k$  (say,  $\psi(C_k)$ ) is automatically scaled for the size of the sample unit because  $\psi(C_k) = \Pr(N(C_k) > 0) = 1 - \exp\{-\mu(C_k)\}$  and

$\mu(C_k) = \int_{C_k} \lambda(\mathbf{s})d\mathbf{s}$  increases with the area of  $C_k$ . The parameters of this site-occupancy model are therefore invariant to spatial scale. An alternative approach (proposed by Royle and Nichols (2003)) is to specify a model wherein detection probability  $q_{jk}$  is assumed to increase with abundance  $n_k$  as follows:  $q_{jk} = 1 - (1 - p_{kj})^{n_k}$ . The estimable parameter  $p_{jk}$  is the probability of detection per individual during the  $j$ th survey. In this approach  $Z_{kj}$  is modeled conditional on the latent abundance of individuals in unit  $C_k$  as follows:

$$Z_{kj}|N(C_k) = n_k \sim \text{Bernoulli}(1 - (1 - p_{kj})^{n_k})$$

1 This site-occupancy model is closely related to Royle’s (2004) model of point counts. In fact,  
 2 the two models provide mathematically equivalent estimators of site occupancy (Dorazio,  
 3 2007).

#### 4 **Conclusions and recommendations**

5 Estimates of SDMs obtained in analyses of presence-only data are vulnerable to biases  
 6 induced by detection errors and survey bias. These biases can be reduced or eliminated  
 7 by (1) including covariates of detection and survey bias in the presence-only model and  
 8 (2) analyzing counts observed in planned surveys in conjunction with the presence-only  
 9 data. These planned surveys must include a sampling protocol that is informative of both  
 10 abundance and detectability. For some protocols (double-observer, removal, or capture-  
 11 recapture sampling) this requirement is implied; in others (independent point counts or  
 12 presence-absence samples) two or more replicates are needed at each survey location to  
 13 obtain multiple observations of the individuals present at each location.

14 The benefits of of planned surveys in species distribution modeling may depend on the  
 15 species. For example, planned surveys of highly mobile species (large birds and mammals)  
 16 can be problematic if movements of individuals make them available to detection in more  
 17 than one sample unit. Combining presence-only data with observations in planned surveys

1 is more likely to benefit species whose movements are more limited relative to the size of a  
2 sample unit.

### 3 **Extensions of SDMs**

4 The hierarchical model developed in this paper can be extended to address a variety of  
5 inference problems in species distribution modeling. For example, several authors (Guisan  
6 and Thuiller, 2005; Goodsoe and Harmon, 2012; Higgins et al., 2012; Kissling et al., 2012;  
7 Wisz et al., 2013) have noted that biological interactions between individuals (e.g., com-  
8 petitive or predator-prey interactions) should be included in SDMs along with the effects  
9 of spatially varying habitat characteristics. Spatial point-process models have been used  
10 to infer the effects of competitive interactions within and among species of plants or ants  
11 (Högmander and Särkkä, 1999; Wiegand et al., 2007a,b; Grabarnik and Särkkä, 2004), but to  
12 my knowledge point-process models have not been formulated to specify the effects of these  
13 interactions mechanistically. Doing so, while accounting for species- and location-specific  
14 differences in detectability and survey bias, will no doubt present challenges for the analyst.  
15 Another important class of inference problems requires the construction of dynamic (i.e.,  
16 space-time) point-process models. Much of the SDM literature is limited to the construction  
17 of static (purely spatial) models, but the need for dynamic, process-based models is clearly  
18 evident and growing (Guisan and Thuiller, 2005; Elith and Leathwick, 2009; Dormann et al.,  
19 2012). Models are needed to predict changes in the spatial distribution of species due to  
20 changes in climate, habitat, levels of disturbance, and abundance of non-indigenous species.  
21 Point-process models are needed also to predict changes in the spatial distribution of exotic  
22 species following their initial introduction to a region of interest. Considerable data may  
23 be required to fit these dynamic models, but recent progress with invasive plant species  
24 (Balderama et al., 2012) illustrates the feasibility of this approach. I anticipate that the  
25 construction of point-process models driven by scientific needs and biological mechanisms  
26 will greatly enhance the inferences and predictions of future SDMs.

# Acknowledgements

Editorial suggestions from the Editor, C. Yackulic, and an anonymous referee greatly improved an earlier draft of this article. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

# References

- Balderama, E., Schoenberg, F. P., Murray, E., and Rundel, P. W. (2012). Application of branching models in the study of invasive species. *Journal of the American Statistical Association*, 107:467–476.
- Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 3:327–338.
- Borchers, D. L. and Efford, M. G. (2008). Spatially explicit maximum likelihood methods for capture-recapture studies. *Biometrics*, 64:377–385.
- Bowden, R. (1973). The theory of parametric identification. *Econometrica*, 41:1069–1074.
- Cabeza, M., Araújo, M. B., Wilson, R. J., Thomas, C. D., Cowley, M. J. R., and Moilanen, A. (2004). Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology*, 41:252–262.
- 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. *Applied Statistics*, 60:757–776.
- Chandler, R. B., Royle, J. A., and King, D. I. (2011). Inference about density and temporary emigration in unmarked populations. *Ecology*, 92:1429–1435.

- 1 Chen, G., Kéry, M., Plattner, M., Ma, K., and Gardner, B. (2013). Imperfect detection  
2 is the rule rather than the exception in plant distribution studies. *Journal of Ecology*,  
3 101:183–191.
- 4 Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons,  
5 Hoboken, New Jersey.
- 6 Dorazio, R. (2012). Predicting the geographic distribution of a species from presence-only  
7 data subject to detection errors. *Biometrics*, 68:1303–1312.
- 8 Dorazio, R. M. (2007). On the choice of statistical models for estimating occurrence and  
9 extinction from animal surveys. *Ecology*, 88:2773–2782.
- 10 Dorazio, R. M. (2013). Bayes and empirical Bayes estimators of abundance and density from  
11 spatial capture-recapture data. *PLoS ONE*, 8:e84017.
- 12 Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney,  
13 M., Morin, X., Römermann, C., Schröder, B., and Singer, A. (2012). Correlation and  
14 process in species distribution models: bridging a dichotomy. *Journal of Biogeography*,  
15 39:2119–2131.
- 16 Efford, M. (2004). Density estimation in live-trapping studies. *Oikos*, 106:598–610.
- 17 Efford, M. G. and Dawson, D. K. (2012). Occupancy in continuous habitat. *Ecosphere*,  
18 3:<http://dx.doi.org/10.1890/ES11-00308.1>.
- 19 Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans,  
20 R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle,  
21 B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson,  
22 A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón,  
23 J., Williams, S., Wisz, M. S., and Zimmerman, N. E. (2006). Novel methods improve  
24 prediction of species’ distributions from occurrence data. *Ecography*, 29:129–151.

- 1 Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and  
2 prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*,  
3 40:677–697.
- 4 Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E., and Yates, C. J. (2010). A  
5 statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17:43–57.
- 6 Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for  
7 presence-only data. *Annals of Applied Statistics*, 7:1917–1939.
- 8 Goodsoe, W. and Harmon, L. J. (2012). How do species interactions affect species distribu-  
9 tion models? *Ecography*, 35:811–820.
- 10 Gormley, A. M., Forsyth, D. M., Griffioen, P., Lindeman, M., Ramsey, D. S. L., Scroggie,  
11 M. P., and Woodford, L. (2013). Using presence-only and presence-absence data to es-  
12 timate the current and potential distributions of established invasive species. *Journal of*  
13 *Applied Ecology*, 48:25–34.
- 14 Grabarnik, P. and Särkkä, A. (2004). Modelling the spatial structure of forest stands by  
15 multivariate point processes with hierarchical interactions. *Ecological Modelling*, 220:1232–  
16 1240.
- 17 Guisan, A. et al. (2007). Sensitivity of predictive species distribution models to change in  
18 grain size. *Diversity and Distributions*, 13:332–340.
- 19 Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than  
20 simple habitat models. *Ecology Letters*, 8:993–1009.
- 21 Higgins, S. I., O’Hara, R. B., and Römermann, C. (2012). A niche for biology in species  
22 distribution models. *Journal of Biogeography*, 39:2091–2095.
- 23 Högmander, H. and Särkkä, A. (1999). Multitype spatial point patterns with hierarchical  
24 interactions. *Biometrics*, 55:1051–1058.

- 1 Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling*  
2 *of spatial point patterns*. John Wiley & Sons, West Sussex, England.
- 3 Kissling, W. D., Dormann, C. F., Groeneveld, J., Hickler, T., Kühn, I., McNerny, G. J.,  
4 Montoya, J. M., Römermann, C., Schiffers, K., Schurr, F. M., Singer, A., Svenning, J. C.,  
5 Zimmermann, N. E., and O'Hara, R. B. (2012). Towards novel approaches to modelling  
6 biotic interactions in multispecies assemblages at large spatial extents. *Journal of Bio-*  
7 *geography*, 39:2163–2178.
- 8 Lahoz-Monfort, J. J., Guillera-Arroita, G., and Wintle, B. A. (2014). Imperfect detection  
9 impacts the performance of species distribution models. *Global Ecology and Biogeography*,  
10 23:504–515.
- 11 Lee, A. J., Scott, A. J., and Wild, C. J. (2006). Fitting binary regression models with  
12 case-augmented samples. *Biometrika*, 93:385–397.
- 13 Lele, S. R. and Keim, J. L. (2006). Weighted distributions and estimation of resource  
14 selection probability functions. *Ecology*, 87:3021–3028.
- 15 MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm,  
16 C. A. (2002). Estimating site occupancy rates when detection probabilities are less than  
17 one. *Ecology*, 83:2248–2255.
- 18 Møller, J. and Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial*  
19 *point processes*. Chapman & Hall, Boca Raton.
- 20 Newbold, T., Reader, T., El-Gabbas, A., Berg, W., Shohdi, W. M., Zalat, S., El Din, S. B.,  
21 and Gilbert, F. (2010). Testing the accuracy of species distribution models using species  
22 records from a new field survey. *Oikos*, 119:1326–1334.
- 23 Pearce, J. L. and Boyce, M. S. (2006). Modelling distribution and abundance with presence-  
24 only data. *Journal of Applied Ecology*, 43:405–412.

- 1 Peterman, W. E., Crawford, J. A., and Kuhns, A. R. (2011). Using species distribution and  
2 occupancy modeling to guide survey efforts and assess species status. *Journal for Nature*  
3 *Conservation*, 21:114–121.
- 4 Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of  
5 species geographic distributions. *Ecological Modelling*, 190:231–259.
- 6 Phillips, S. J., Dudik, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier,  
7 S. (2009). Sample selection bias and presence-only distribution models: implications for  
8 background and pseudo-absence data. *Ecological Applications*, 19:181–197.
- 9 Phillips, S. J. and Elith, J. (2013). On estimating probability of presence from use-availability  
10 or presence-background data. *Ecology*, 94:1409–1419.
- 11 R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foun-  
12 dation for Statistical Computing, Vienna, Austria.
- 13 Renner, I. W. and Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process  
14 models for species distribution modeling in ecology. *Biometrics*, 69:274–281.
- 15 Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated  
16 counts. *Biometrics*, 60:108–115.
- 17 Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology*.  
18 Academic Press, Amsterdam.
- 19 Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence-absence  
20 data or point counts. *Ecology*, 84:777–790.
- 21 Scott, J. M., Heglund, P. J., Morrison, M. L., Haufler, J. B., Raphael, M. G., Wall, W. A.,  
22 and Samson, F. B. (2002). *Predicting species occurrences: issues of accuracy and scale*.  
23 Island Press, Washington.

- 1 Sóllymos, P., Lele, S., and Bayne, E. (2012). Conditional likelihood approach for analyzing  
2 single visit abundance survey data in the presence of zero inflation and detection error.  
3 *Environmetrics*, 23:197–205.
- 4 Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., and Possingham, H. P.  
5 (2003). Improving precision and reducing bias in biological surveys: estimating false-  
6 negative error rates. *Ecological Applications*, 13:1790–1801.
- 7 Warton, D. I. and Shepherd, L. C. (2010). Poisson point process models solve the “pseudo-  
8 absence problem” for presence-only data in ecology. *Annals of Applied Statistics*, 4:1383–  
9 1402.
- 10 Wiegand, T., Gunatilleke, S., and Gunatilleke, N. (2007a). Species associations in a het-  
11 erogenous Sri Lankan Dipterocarp forest. *American Naturalist*, 170:E77–E95.
- 12 Wiegand, T., Gunatilleke, S., Gunatilleke, N., and Okuda, T. (2007b). Analyzing the spatial  
13 structure of a Sri Lankan tree species with multiple scales of clustering. *Ecology*, 88:3088–  
14 3102.
- 15 Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann,  
16 C. F., Forchhammer, M. C., Grytnes, J., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn,  
17 I., Luoto, M., Maiorano, L., Nilsson, M., Normand, S., Öckinger, E., Schmidt, N. M.,  
18 Termansen, M., Timmermann, A., Wardle, D. A., Aastrup, P., and Svenning, J. (2013).  
19 The role of biotic interactions in shaping distributions and realised assemblages of species:  
20 implications for species distribution modelling. *Biological Reviews*, 88:15–30.
- 21 Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Grant, E. H. C.,  
22 and Veran, S. (2013). Presence-only modelling using MAXENT: when can we trust the  
23 inferences? *Methods in Ecology and Evolution*, 4:236–243.
- 24 Yoccoz, N. G., Nichols, J. D., and Boulinier, T. (2001). Monitoring of biological diversity in  
25 space and time. *Trends in Ecology and Evolution*, 16:446–453.

# 1 Biosketch

2 Robert Dorazio is a Research Statistician at the U.S. Geological Survey’s Southeast Eco-  
3 logical Science Center. He also holds a Courtesy Associate Professorship in the Department  
4 of Statistics at the University of Florida. His research is motivated primarily by statistical  
5 inference problems that arise in the general areas of population dynamics, community ecol-  
6 ogy, and conservation biology. He is also interested in developing the theory and practice of  
7 adaptive decision making in problems of natural resource management.

## 8 List of Figures

|    |   |   |    |
|----|---|---|----|
| 9  | 1 | Spatial distributions of a covariate of individual density (upper panel) and a          |    |
| 10 |   | covariate of detection probability (lower panel). . . . .                               | 33 |
| 11 | 2 | Assessment of parameter identifiability for two point-process models of presence-       |    |
| 12 |   | only data. One model is based on independent regressors of abundance and                |    |
| 13 |   | detection (solid line); the other model is based on identical regressors (dashed        |    |
| 14 |   | line). Upper panel: Reciprocal of the condition number of the Fisher informa-           |    |
| 15 |   | tion matrix is plotted as a function of $\alpha_0$ (= logit-scale detection parameter). |    |
| 16 |   | Lower panel: Expected number of individuals detected in region $B$ is plotted           |    |
| 17 |   | as a function of $\alpha_0$ . . . . .   | 34 |
| 18 | 3 | Spatial distributions of expected density of individuals (upper panel) and              |    |
| 19 |   | of expected densities of detections in opportunistic surveys as a function of           |    |
| 20 |   | covariate $w$ (middle panel) or covariate $x$ (lower panel). Superimposed points        |    |
| 21 |   | indicate a single realization from each distribution. . . . .                           | 35 |
| 22 | 4 | Spatial distributions of expected number of individuals per sample unit (upper          |    |
| 23 |   | panel) and of expected number of individuals detected per survey as a function          |    |
| 24 |   | of covariate $w$ (middle panel) or covariate $x$ (lower panel). . . . .                 | 36 |

|    |   |   |    |
|----|---|---|----|
| 1  | 5 | Operating characteristics of maximum-likelihood estimators of the species distribution model parameters $\beta_0$ and $\beta_1$ . Detections of individuals were assumed to depend on covariate $w$ , which was independent of the covariate used in the SDM. Symbols indicate that estimates were obtained by fitting the model of presence-only data (triangles), the model of point counts (circles), or the model of combined data (squares). Error bars indicate 95% confidence intervals. . . . . | 37 |
| 2  |   |   |    |
| 3  |   |   |    |
| 4  |   |   |    |
| 5  |   |   |    |
| 6  |   |   |    |
| 7  |   |   |    |
| 8  | 6 | Operating characteristics of maximum-likelihood estimators of the species distribution model parameters $\beta_0$ and $\beta_1$ . Detections of individuals were assumed to depend on covariate $x$ , which was identical to the covariate used in the SDM. Symbols indicate that estimates were obtained by fitting the model of presence-only data (triangles), the model of point counts (circles), or the model of combined data (squares). Error bars indicate 95% confidence intervals. . . . .   | 38 |
| 9  |   |   |    |
| 10 |   |   |    |
| 11 |   |   |    |
| 12 |   |   |    |
| 13 |   |   |    |
| 14 |   |   |    |

## 15 **Supporting Information**

16 Additional supporting information may be found in the online version of this article at  
17 the publishers web-site.

18 **Appendix S1:** Fisher information matrix

19 **Appendix S2:** R code (R Core Team, 2014) used to simulate and analyze data in oppor-  
20 tunistic and planned surveys.

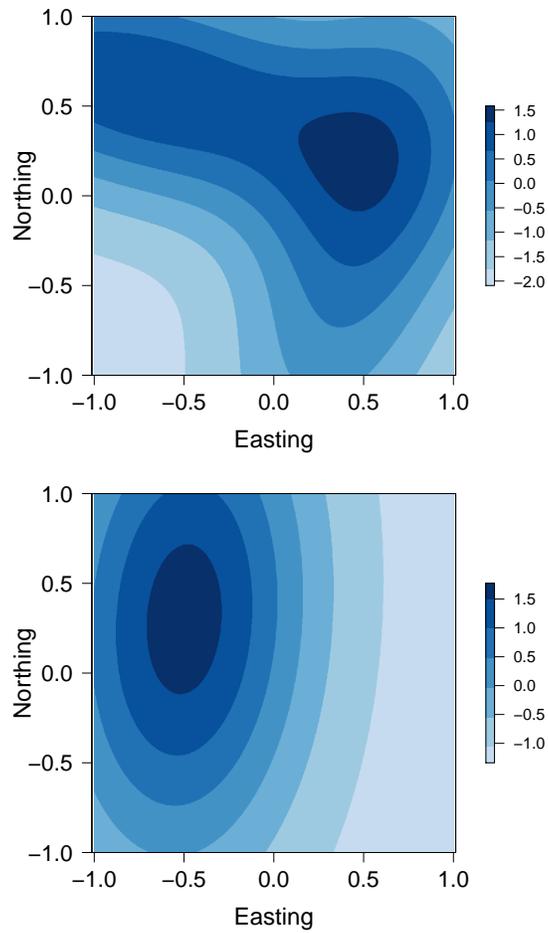


Figure 1: Spatial distributions of a covariate of individual density (upper panel) and a covariate of detection probability (lower panel).

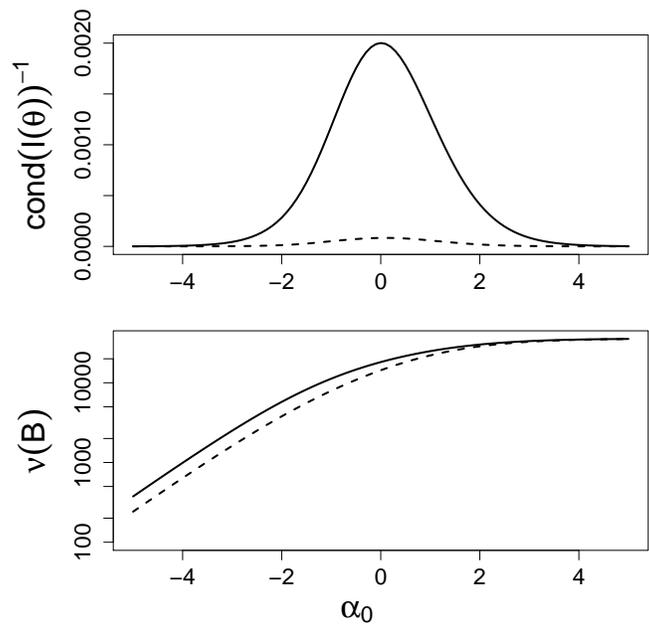


Figure 2: Assessment of parameter identifiability for two point-process models of presence-only data. One model is based on independent regressors of abundance and detection (solid line); the other model is based on identical regressors (dashed line). Upper panel: Reciprocal of the condition number of the Fisher information matrix is plotted as a function of  $\alpha_0$  (= logit-scale detection parameter). Lower panel: Expected number of individuals detected in region  $B$  is plotted as a function of  $\alpha_0$ .

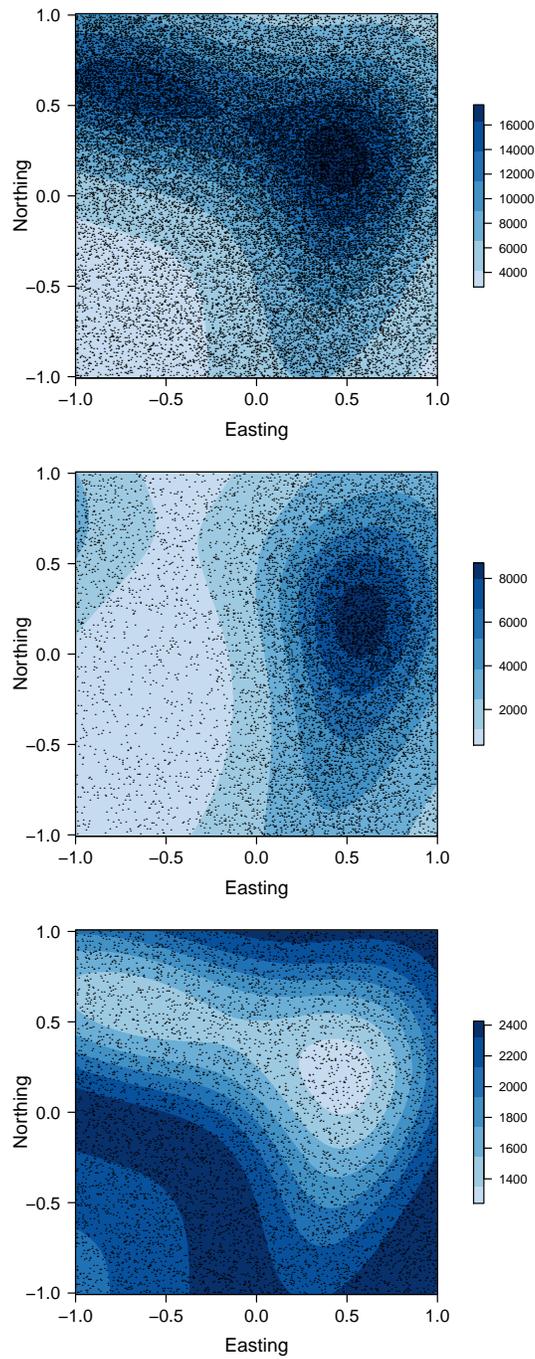


Figure 3: Spatial distributions of expected density of individuals (upper panel) and of expected densities of detections in opportunistic surveys as a function of covariate  $w$  (middle panel) or covariate  $x$  (lower panel). Superimposed points indicate a single realization from each distribution.

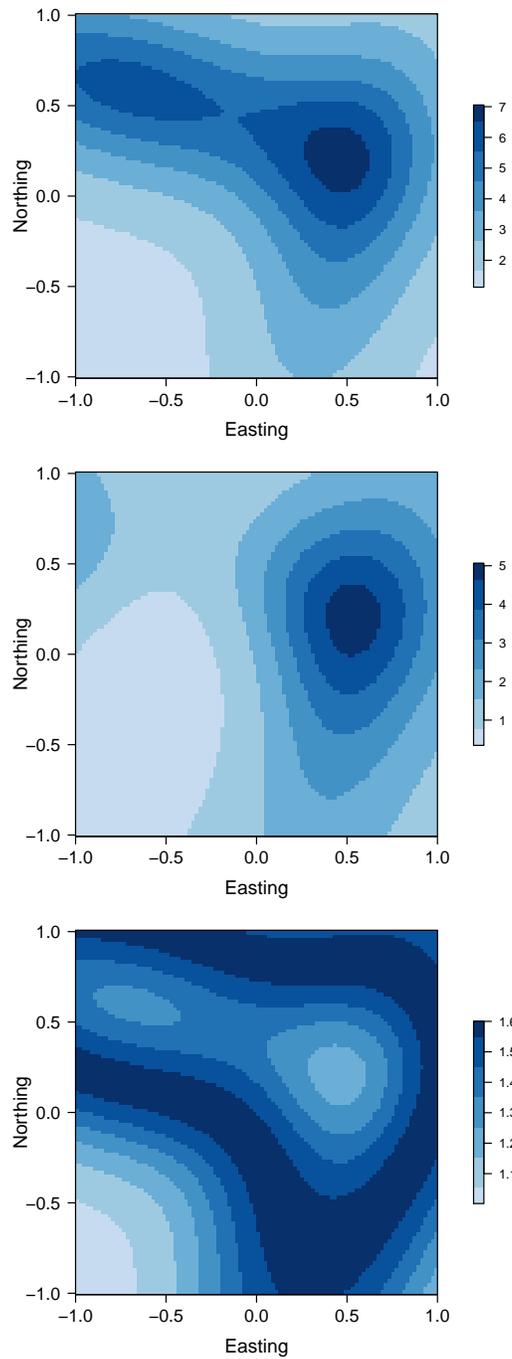


Figure 4: Spatial distributions of expected number of individuals per sample unit (upper panel) and of expected number of individuals detected per survey as a function of covariate  $w$  (middle panel) or covariate  $x$  (lower panel).

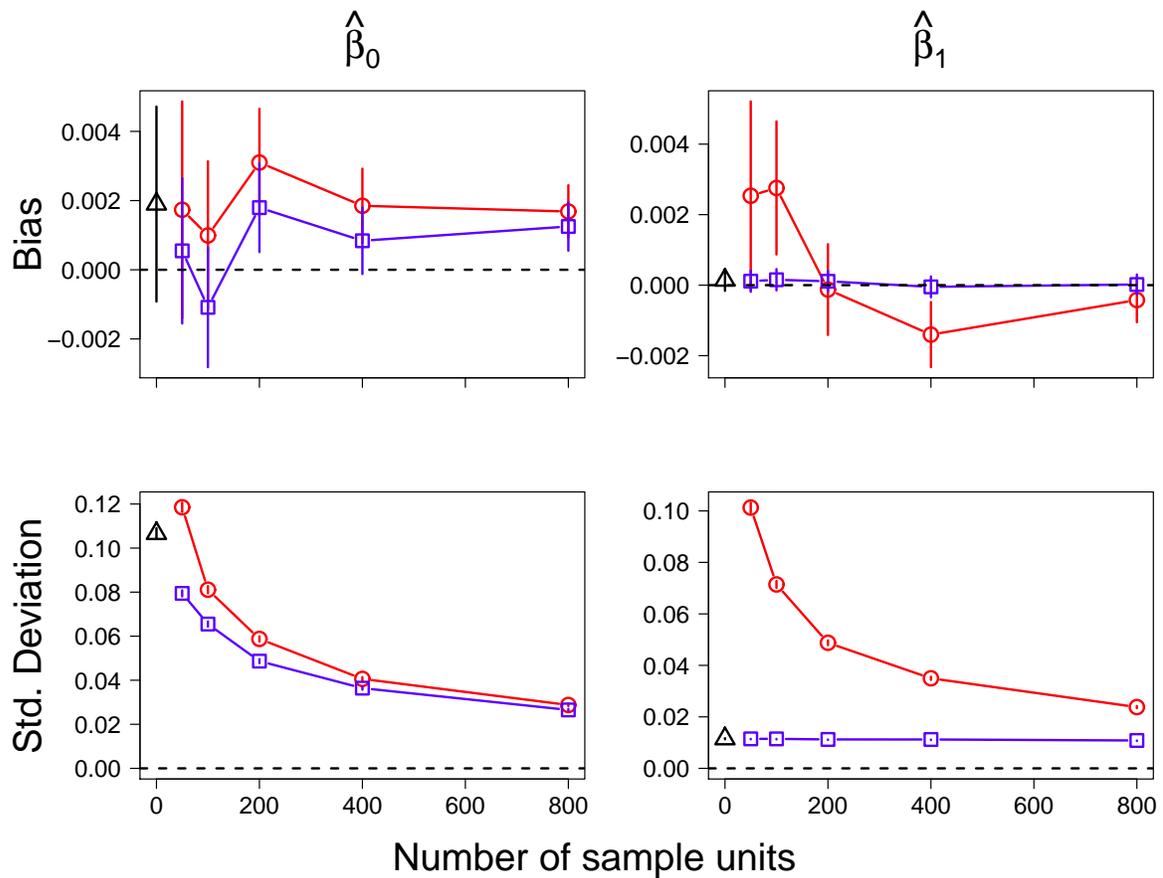


Figure 5: Operating characteristics of maximum-likelihood estimators of the species distribution model parameters  $\beta_0$  and  $\beta_1$ . Detections of individuals were assumed to depend on covariate  $w$ , which was independent of the covariate used in the SDM. Symbols indicate that estimates were obtained by fitting the model of presence-only data (triangles), the model of point counts (circles), or the model of combined data (squares). Error bars indicate 95% confidence intervals.

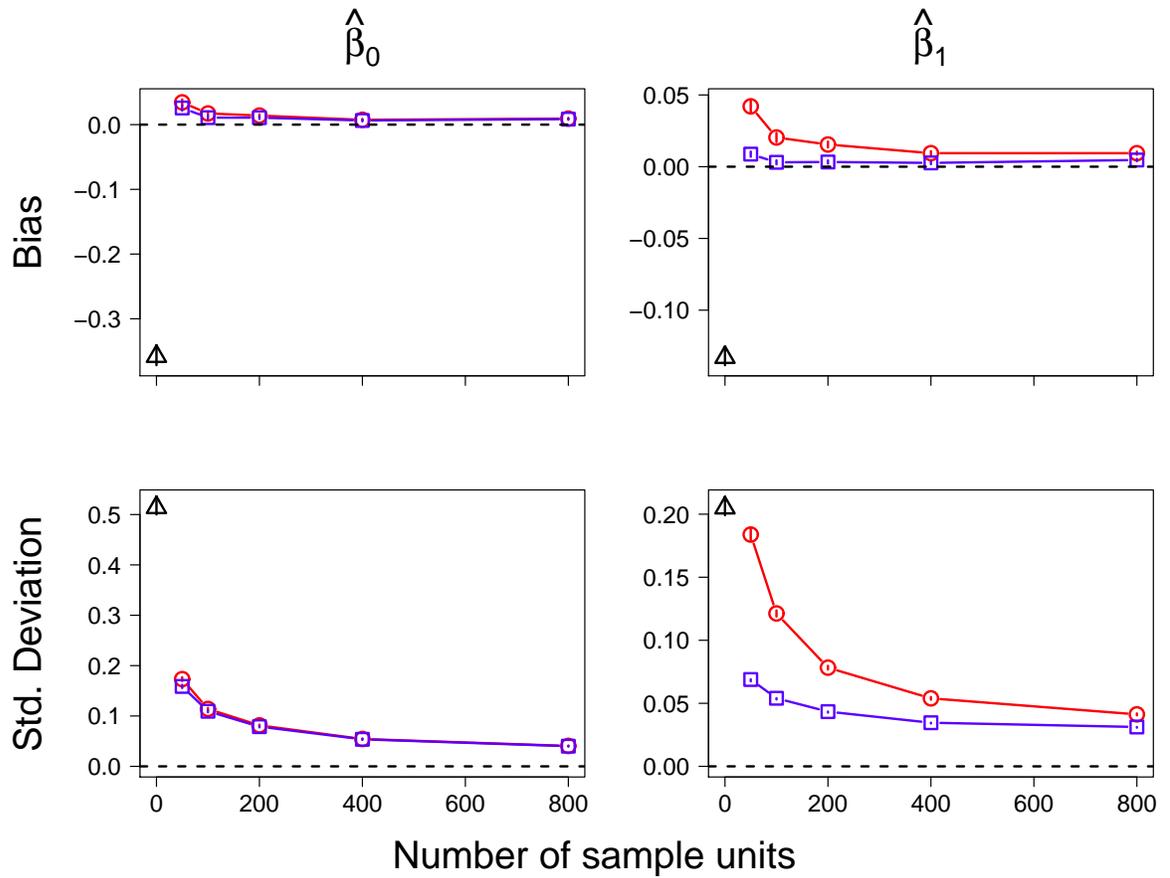


Figure 6: Operating characteristics of maximum-likelihood estimators of the species distribution model parameters  $\beta_0$  and  $\beta_1$ . Detections of individuals were assumed to depend on covariate  $x$ , which was identical to the covariate used in the SDM. Symbols indicate that estimates were obtained by fitting the model of presence-only data (triangles), the model of point counts (circles), or the model of combined data (squares). Error bars indicate 95% confidence intervals.