Species Occupancy Modeling for Detection Data Collected Along a Transect

Gurutzeta GUILLERA-ARROITA, Byron J.T. MORGAN,
Martin S. RIDOUT, and Matthew LINKIE

The proportion of sampling sites occupied by a species is a concept of interest in ecology and biodiversity conservation. Occupancy surveys based on collecting detection data along transects have become increasingly popular to monitor some species. To date, the analysis of such data has been carried out by discretizing the data, dividing the transects into discrete segments. Here we propose alternative occupancy models which describe the detection process as a continuous point process. These models provide a more natural description of the data and eliminate the need to divide transects into segments, which can be arbitrary and may lead to increased bias in the estimator of occupancy or increased chances of obtaining estimates on the boundary of the parameter space. We present a model that assumes independence between detections and an alternative model that describes the detection process as a Markov modulated Poisson process to account for potential clustering in the detections. The utility of these models is illustrated with the analysis of data from a recent survey of the Sumatran tiger *Panthera tigris sumatrae*. The models can also be applied to surveys that collect continuous data in time, such as those based on the use of camera-trap devices. Supplementary materials for this article are available online.

Key Words: Clustered data; Markov modulated Poisson process; Wildlife monitoring; Zero-inflated Poisson process.

1. INTRODUCTION

Occupancy, the probability of a site being occupied by a species, is a parameter of interest in ecology where it is frequently used for modeling habitat relationships, as a state

Gurutzeta Guillera-Arroita (✉) is a Ph.D. Student (E-mail: *gg63@kent.ac.uk*), Byron J.T. Morgan is Professor of Applied Statistics (E-mail: *b.j.t.morgan@kent.ac.uk*) and Martin S. Ridout is Professor of Applied Statistics (E-mail: *m.s.ridout@kent.ac.uk*), School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NZ, UK and are part of the National Centre for Statistical Ecology. Matthew Linkie (E-mail: *matthew.linkie@ffi.or.id*) is Programme Manager, Fauna & Flora International (FFI) in Aceh, Jln Tgk. Chik Dipasi No. 50, Desa Limpok, Darussalam, Aceh Besar, Nanggroe Aceh Darussalam 23373, Indonesia and Honorary Research Fellow at the Durrell Institute of Conservation and Ecology, University of Kent, Canterbury, CT2 7NR, UK.

variable for wildlife monitoring programmes and in metapopulation studies. To obtain an unbiased estimator of occupancy it is necessary to take into account the fact that species often remain undetected even when present at a site. The issue of imperfect detection in the context of occupancy modeling has received much attention in recent years. MacKenzie et al. (2002) proposed a model that allows the simultaneous estimation of occupancy and detection probability. This model assumes that a number of discrete independent replicate surveys are carried out in each sampling site and describes the detection/non-detection data as coming from independent Bernoulli trials (i.e. a discrete Bernoulli process). Since then a number of developments and refinements of this model have been proposed including extensions to multiple seasons (MacKenzie et al. 2003), multiple species (MacKenzie, Bailey, and Nichols 2004) and heterogeneous detection probability (Royle 2006).

The basic sampling situation described by MacKenzie et al. (2002) to account for imperfect detection requires replicate surveys at a number of sampling sites at distinct times. Other methods to achieve replication include the simultaneous use of independent observers or detection methods and surveying different sectors within each sampled site. This last method would require a sampling procedure that ensures independence of detections among the spatial replicates so as not to violate the assumptions of the basic occupancy model (MacKenzie et al. 2006). For instance, dependence may be induced by the way sectors are chosen spatially, if sites are only partially occupied by the species (or its tracks). The occupancy status of sectors in close proximity will tend to be correlated, and so will be the detections. However, for logistical convenience, sometimes replicates are assigned to consecutive transect segments. For instance, this type of sampling has been applied recently to various surveys designed to monitor large carnivores, such as tiger *Panthera tigris* and snow leopard *Panthera uncia*. To address the potential lack of independence among consecutive replicates when data are collected in this fashion, Hines et al. (2010) relax the independence assumption of the basic occupancy model by incorporating Markovian dependence among consecutive replicates. Detection/non-detection data continue to be modeled as arising from Bernoulli trials but these trials are no longer assumed to be independent.

Here we present a new framework for modeling occupancy when detection/non-detection data are collected along a continuous axis (time or space). Instead of modeling occurrences based on a discretization of the data, as done conventionally (e.g. Hines et al. 2010), we describe the species detections as a continuous point process (Cox and Isham 1980). This framework provides a more natural description for data collected along a transect and eliminates the need to divide the transect into discrete segments, which can be arbitrary and may lead to increased bias in the estimator of occupancy or increased chances of obtaining estimates on the boundary of the parameter space. In this paper we start by presenting in Section 2 an occupancy model that describes the detection process as a Poisson process. The Poisson process is the simplest point process and assumes that events occur totally at random along the axis. Although mathematically convenient for some scenarios, generalizations of the Poisson process that relax the requirement of independence are often a more appropriate representation of reality. In Section 3 we propose an occupancy model that accounts for clustering in the detections using one such generalization: the two-state Markov modulated Poisson process (2-MMPP). The 2-MMPP models

a system with two different event rates. Which of the two rates is effective at each point of time/space is governed by a two-state Markov process. This structure allows the modeling of clustering in the occurrence of events and thus may be of use for describing the detections of certain species along a line transect (e.g. minke whales, Skaug 2006). In our study we use maximum-likelihood estimation for model fitting and evaluate the performance of the described models via simulations in Section 4. Finally in Section 5 we apply the models to data from a survey of the Sumatran tiger *Panthera tigris sumatrae* carried out in Kerinci Seblat National Park, in which transects were walked within a number of sampling sites. Tigers are wide-ranging animals. They maintain large home ranges and walk long distances. Even resident breeders, which keep a relatively fixed home range, move over 10–15 km per day (Karanth and Nichols 2002). There are various mechanisms that can give rise to clustering in the detection of tiger footprints. Given the ruggedness of the landscape in the Kerinci area, the transects surveyed often followed ridges along which walking is relatively easier. It is believed that tigers frequently use ridges to move across their territory so it is reasonable to expect some clustering in the data. Clustering has been reported in transect survey data from tigers in India, as they tend to walk along forest roads and trails (Hines et al. 2010).

2. INDEPENDENT DETECTIONS: POISSON PROCESS OCCUPANCY MODEL

2.1. BASIC MODEL

Let us consider a study with S sampling sites where detection/non-detection surveys have been carried out along transects, providing the location of each detection. Let us assume that each site is closed to changes in occupancy, has a probability ψ of being occupied and that detections along each transect can be considered independent and so can be modeled as a Poisson process with intensity λ, where λ represents the average number of detections per unit length. Note that at this stage we assume that occupancy probability ψ and detection intensity λ are constant across sites. The likelihood function for this model is constructed as the likelihood function of a Poisson process but allowing for zero-inflation to account for unoccupied sites. For sites where the species was detected at least once the likelihood contribution is

$$\psi \prod_{j=1}^{R_i} \lambda \exp(-\lambda l_{ij1}) \cdots \lambda \exp(-\lambda l_{ijd_{ij}}) \exp(-\lambda l_{ijd_{ij}+1}) = \psi \lambda^{d_i} \exp(-\lambda L_i),$$

where R_i is the number of independent transects in cell i, d_{ij} is the number of detections in transect j at site i and $l_{ij1}, \ldots, l_{ijd_{ij}+1}$ are the inter-detection distances, with l_{ij1} defined as the distance to first detection from the beginning of the transect and $l_{ijd_{ij}+1}$ the distance from the last detection until the end of the transect; d_i and L_i represent the total number of detections and the total length surveyed at site i respectively, so $d_i = \sum_{j=1}^{R_i} d_{ij}$ and $L_i = \sum_{j=1}^{R_i} \sum_{k=1}^{d_{ij}+1} l_{ijk}$. The likelihood contribution for cells with no detections is

$$(1 - \psi) + \psi \exp(-\lambda L_i)$$

that is, either the species was not present at the site or it was present but it was not detected in a total surveyed length L_i. Assuming independence of the sites, the likelihood for the whole detection data set can be constructed as the product of site likelihoods and written as follows:

$$L(\psi, \lambda) = \prod_{i=1}^{S} \{\psi \lambda^{d_i} \exp(-\lambda L_i) + (1 - \psi)\mathrm{I}(d_i = 0)\},$$

where $\mathrm{I}(\cdot)$ represents the indicator function. Note that for this model the data can be summarized by the total number of detections at each site $\{d_i\}$, given design parameters S and $\{L_i\}$.

2.2. MLEs and Asymptotic Design Recommendations

Let us assume that the length surveyed within each site is constant L. Although this simplification may not always be realistic in practice, it allows us to provide general study design recommendations. The likelihood can now be written in a compact form as follows:

$$L(\psi, \lambda) = \psi^{S_d} \lambda^{d_T} \exp(-S_d \lambda L)\{(1 - \psi) + \psi \exp(-\lambda L)\}^{S - S_d}, \tag{2.1}$$

where d_T is the total number of detections in the survey and S_d is the number of sites where the species was detected at least once. In this case (S_d, d_T) is a sufficient statistic. The likelihood in (2.1) can be rewritten using a reparameterization of the type suggested by Morgan, Revell, and Freeman (2007) for likelihood functions of this form. Setting $\theta = \psi(1 - e^{-\lambda L})$, the probability of detecting the species at an occupied site leads to:

$$L(\theta, \lambda) = \left\{\theta^{S_d}(1 - \theta)^{S - S_d}\right\} \left\{\frac{\lambda^{d_T}}{(\exp(\lambda L) - 1)^{S_d}}\right\}.$$

From this factorization we can easily derive expressions for the maximum-likelihood parameter estimates (MLEs). The estimate $\hat{\theta}$ is the proportion of cells where the species was detected $\hat{\theta} = S_d/S$, and therefore the estimate of occupancy ψ satisfies

$$\hat{\psi} = \frac{S_d/S}{1 - \exp(-\hat{\lambda}L)}. \tag{2.2}$$

The MLE for the detection rate $\hat{\lambda}$ satisfies

$$\frac{\hat{\lambda}}{1 - \exp(-\hat{\lambda}L)} = \frac{d_T}{S_d L}. \tag{2.3}$$

An explicit expression for $\hat{\lambda}$ can be written using the Lambert function (Corless et al. 2005), which is defined as the inverse of xe^x, as follows:

$$\hat{\lambda} = \frac{1}{L} \left\{\frac{d_T}{S_d} + W_0\left(-\frac{d_T}{S_d}\exp\left(-\frac{d_T}{S_d}\right)\right)\right\}, \tag{2.4}$$

where $W_0(\cdot)$ is the principal branch of the Lambert function. It can be shown that these MLE expressions hold when the observed detection history fulfills the condition

$$\ln\left(\frac{S - S_d}{S}\right) > -\left\{\frac{d_T}{S_d} + W_0\left(-\frac{d_T}{S_d} - \exp\left(-\frac{d_T}{S_d}\right)\right)\right\},$$

and that otherwise, the MLEs are:

$$\hat{\psi} = 1, \qquad \hat{\lambda} = \frac{d_T}{SL}. \tag{2.5}$$

It can be seen from (2.2) that, as $1 - \exp(-\hat{\lambda}L)$ decreases, that is, as the estimated probability of detecting the species at an occupied site when a transect of length L is surveyed decreases, the estimate of occupancy ($\hat{\psi}$) increases relative to the naïve estimate obtained assuming that species is always detected at occupied sites ($\hat{\psi}_{\text{naive}} = \hat{\theta} = S_d/S$), until it reaches the boundary (2.5). For a comparison with the analytical expressions for the MLEs in the basic Bernoulli occupancy model, see Guillera-Arroita, Ridout, and Morgan (2010).

Survey design recommendations can be based on the asymptotic variance-covariance matrix as it allows us to explore analytically how the performance of the estimator changes for different design parameters, under the assumption of sufficiently large sample size. It can be shown that for the basic Poisson process occupancy model the elements of the asymptotic variance-covariance matrix are:

$$\text{var}(\hat{\psi}) = \frac{\psi}{S} \left\{ (1 - \psi) + \frac{1 - \lambda^*}{\lambda^* - (1 - \lambda^*)\lambda L} \right\}, \tag{2.6}$$

$$\text{var}(\hat{\lambda}) = \frac{\lambda}{\psi SL} \left\{ \frac{\lambda^*}{\lambda^* - (1 - \lambda^*)\lambda L} \right\}, \tag{2.7}$$

$$\text{cov}(\hat{\psi}, \hat{\lambda}) = \frac{-\lambda}{S} \left\{ \frac{1 - \lambda^*}{\lambda^* - (1 - \lambda^*)\lambda} \right\}, \tag{2.8}$$

where $\lambda^* = 1 - \exp(-\lambda L)$. As λ^* approaches unity (i.e. the probability of missing the species at occupied sites approaches zero), (1) the variance of the occupancy estimator (2.6) tends to the variance dictated by the binomial proportion $\psi(1 - \psi)/S$ and decreases as the number of sites increases; (2) the variance of the intensity parameter, in (2.7), tends to $\lambda/(\psi SL)$ and decreases as the total effort increases regardless of whether it is spent on surveying more sites or longer transects within each site; (3) the covariance (2.8) approaches zero.

Suppose that a study can employ a fixed amount of surveying effort ($E = SL$) and that we wish to maximize the precision of $\hat{\psi}$, the estimator of occupancy. Table 1 shows the optimal survey design based on (2.6). The optimal length to survey per site is determined by the parameter values (ψ and λ), irrespective of the total effort allocated to the survey. The probability of occupancy (ψ) determines the mean number of detections at occupied sites (λL) that maximizes the precision of $\hat{\psi}$, from which the optimal length can be derived. For rare species (i.e. low ψ) it is best to sample more sites (at the expense of increasing the probability of false absence), while for commoner species (i.e. higher ψ) it is best to allocate the effort so that fewer sites are more intensively surveyed. MacKenzie and Royle (2005) provide similar observations regarding the Bernoulli occupancy model. Note that the optimal length is the same if the study is designed to minimize the total surveying effort for a given precision of the occupancy estimator.

Table 1. Mean number of detections at occupied sites (λL) to minimize the variance of the occupancy estimator in the Poisson process occupancy model for different levels of occupancy (ψ). From this, the optimum survey length per site can be derived by dividing by λ. The corresponding probability of detecting the species at occupied sites (λ^*) is also shown.

ψ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
λL	1.44	1.52	1.62	1.74	1.88	2.08	2.34	2.75	3.53
λ^*	0.76	0.78	0.80	0.82	0.85	0.87	0.90	0.94	0.97

2.3. INTRODUCING COVARIATES

The model may readily be expanded to allow the probability of occupancy (ψ) and/or the detection intensity (λ) to depend upon site characteristics, such as habitat type or level of human disturbance, using a generalized linear model framework, with a logistic link function for the vector of site occupancies $\boldsymbol{\psi}$ and a log link function for the vector of detection intensities $\boldsymbol{\lambda}$:

$$\boldsymbol{\psi} = \frac{1}{1 + \exp(-\boldsymbol{C\alpha})}, \qquad \boldsymbol{\lambda} = \exp(\boldsymbol{D\beta}), \tag{2.9}$$

where $\boldsymbol{C}$ and $\boldsymbol{D}$ are matrices with site covariate information and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of model parameters. The model can also be extended to accommodate within-site variation in the detection rate, by describing the detection process as a non-homogeneous Poisson process (Cox and Isham 1980, p. 48). This results in the following construction for the likelihood function:

$$L(\psi, \lambda) = \prod_{i=1}^{S} \left\{ \psi \prod_{j=1}^{d_i} \lambda(t_{ij}) \exp\left(-\int_{T_{L_i}} \lambda(t)\, dt\right) + (1 - \psi)\mathrm{I}(d_i = 0) \right\}, \tag{2.10}$$

where t_{ij} are the locations of detections and T_{L_i} is the interval of length L_i surveyed. The detection rate can now be expressed as a function of covariates that vary along the transect via a log link function:

$$\lambda(t) = \exp(\beta_0 + \beta_1 D_1(t) + \beta_2 D_2(t) + \cdots).$$

Most commonly in practice there will not be an explicit function describing the variation of the covariates along the transect, but rather discrete covariate values corresponding to transect sections, so the integration in (2.10) reduces to a finite summation as follows:

$$L(\psi, \lambda) = \prod_{i=1}^{S} \left\{ \psi \prod_{j=1}^{d_i} \lambda(t_{ij}) \exp\left(-\sum_{n=1}^{N_i} \lambda_{in} L_{in}\right) + (1 - \psi)\mathrm{I}(d_i = 0) \right\}, \tag{2.11}$$

where N_i is the number of transect sections at site i with distinct values for the intensity parameter, λ_{in} refers to the intensity parameter at transect section n and L_{in} refers to the length of transect section n. Note that the model assumes that sites are closed in terms of occupancy, so its structure does not allow for changes of occupancy within sites.

2.4. RELATION WITH THE BERNOULLI PROCESS OCCUPANCY MODEL

The Bernoulli process occupancy model proposed by MacKenzie et al. (2002) models the detections at a site as coming from K independent Bernoulli trials each with probability of success p, the detection probability. The binomial distribution is a good approximation to the Poisson distribution when the number of trials is large and the probability of success at each trial is small. Therefore the Bernoulli process occupancy model will be a good approximation to the Poisson process occupancy model described here if the continuous detection data are discretized using sufficiently small intervals, that is, if the transect is cut into sufficiently small segments. The parameter p will approximately be $1 - \exp(-\lambda L/K)$, that is, the probability of having at least one detection in a segment of length L/K given a Poisson process with detection rate λ.

3. CLUSTERED DETECTIONS: 2-MMPP OCCUPANCY MODEL

3.1. BASIC MODEL

Suppose that species detections within sampling sites exhibit some degree of clustering and thus cannot be considered independent. We propose the use of a two-state Markov modulated Poisson process (2-MMPP) to model the clustering in the detections. The 2-MMPP is a particular type of doubly stochastic Poisson process or Cox process (Cox and Isham 1980, p. 70) where the intensity is governed by an unobserved two-state Markov process. According to a 2-MMPP, species detections occur at two different rates, λ_1 and λ_2, and the interval spent surveying in each of these two states is stochastic and defined by parameters μ_{12} (switching intensity from λ_1 to λ_2) and μ_{21} (switching intensity from λ_2 to λ_1). The particular case with one detection rate equal to zero is often referred to as an Interrupted Poisson Process (IPP). The 2-MMPP may be, for instance, an adequate representation for transect survey data from a species that only partially covers occupied sampling sites or when surveying along trails that a species may cross or intermittently follow. The likelihood for the occupancy model can now be written as

$$L(\psi, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \prod_{i=1}^{S} \left\{ \psi \prod_{j=1}^{R_i} M_{ij} + (1 - \psi) \mathrm{I}(d_i = 0) \right\},$$

where M_{ij} is the expression for the contribution to the likelihood of data from transect j at site i described as a $2-$MMPP. To construct M_{ij} we use a number of well-known results about the MMPP (see Fischer and Meier-Hellstern 1992 and Rydén 1994 for details). Let $\boldsymbol{\pi}$ be the initial distribution of the underlying 2-state Markov process, that is the probability of being in each of the two states (λ_1 or λ_2) at the beginning of the transect. If the start of the transect is chosen randomly, an appropriate specification for $\boldsymbol{\pi}$ is the equilibrium distribution of the underlying Markov process given by

$$\boldsymbol{\pi}' = [\pi_1' \quad \pi_2'] = \left[\frac{\mu_{21}}{\mu_{12} + \mu_{21}} \quad \frac{\mu_{12}}{\mu_{12} + \mu_{21}} \right]. \tag{3.1}$$

If the transect is started at a point of detection, an appropriate specification for the initial distribution $\boldsymbol{\pi}$ is

$$\boldsymbol{\pi}^* = \left[\frac{\lambda_1 \pi_1'}{\lambda_1 \pi_1' + \lambda_2 \pi_2'} \quad \frac{\lambda_2 \pi_2'}{\lambda_1 \pi_1' + \lambda_2 \pi_2'} \right], \tag{3.2}$$

where π_1' and π_2' are given by (3.1).

Let $\boldsymbol{Q}$ be the generator matrix of the underlying Markov process:

$$\boldsymbol{Q} = \left[\begin{matrix} -\mu_{12} & \mu_{12} \\ \mu_{21} & -\mu_{21} \end{matrix} \right].$$

Now M_{ij} can be written as

$$M_{ij} = \boldsymbol{\pi} \exp(\boldsymbol{C}l_{ij1})\boldsymbol{\Lambda} \cdots \exp(\boldsymbol{C}l_{ijd_{ij}})\boldsymbol{\Lambda} \exp(\boldsymbol{C}l_{ijd_{ij}+1})\boldsymbol{e}, \tag{3.3}$$

where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2\}$, $\boldsymbol{C} = \boldsymbol{Q} - \boldsymbol{\Lambda}$, $\boldsymbol{e}' = (1\ 1)$, $l_{ij1}, \ldots, l_{ijd_{ij}+1}$ are the inter-detection distances and 'exp' denotes the matrix exponential function. A closed form for the matrix exponential $\exp(\boldsymbol{C}l)$ is available for 2-MMPP, see Rydén (1994). For transects with no detections, (3.3) reduces to

$$M_{ij} = \boldsymbol{\pi} \exp(\boldsymbol{C}L_{ij})\boldsymbol{e}. \tag{3.4}$$

Note that (3.3) holds for higher-dimensional MMPP with the corresponding definitions for matrices $\boldsymbol{\pi}$, $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$.

3.2. IDENTIFIABILITY

In a 2-MMPP, clustering in the data arises as the two intensity parameters, λ_1 and λ_2, differ from each other. When the two intensities are the same, the model reduces to a homogeneous Poisson process and the parameters μ_{12} and μ_{21} are not identifiable. For any given value of $\psi = \tilde{\psi}$ the likelihood function takes the same value for all combinations of parameters satisfying $\lambda_1 = \lambda_2 = \tilde{\lambda}$, regardless of the values of μ_{12} and μ_{21}. An identical likelihood value is achieved for $\lambda_1 = \tilde{\lambda}$ when $\mu_{12} \ll \mu_{21}$, regardless of λ_2, as in effect this represents a case with only one state. Note that, if the data are such that there are no transects without detections at sites where the species was detected (e.g. if only one transect is surveyed per site), then the likelihood function will take the same value as above for $\psi = \tilde{\psi}(1 + \tilde{\mu}_{12}/\tilde{\mu}_{21})$, $\lambda_1 = \tilde{\lambda}$, $\lambda_2 = 0$, $\mu_{12} \to 0$, $\mu_{21} \to 0$. The model has two alternative explanations for cells without detections (unoccupied or occupied in state 2), so the parameter ψ and the ratio of the switching parameters μ_{12}/μ_{21} are not separately identifiable. This same lack of identifiability arises for species that occupy sampling sites partially (i.e. $\lambda_1 = \tilde{\lambda}$, $\lambda_2 = 0$), if transects are short and without replication within sites.

3.3. INTRODUCING COVARIATES

As in the PP occupancy model, information on site covariates can be incorporated easily in the 2-MMPP model using a logistic function for ψ and a log function for λ (2.9). Within-site detection covariates can also be incorporated in a similar way as in (2.11). Let us consider a case in which transect ij consists of two sections (A and B) with different

intensity parameter vectors, $\boldsymbol{\lambda}_A$ and $\boldsymbol{\lambda}_B$, and suppose that there were two detections in section A and none in section B. The expression for M_{ij} is:

$$M_{ij} = \boldsymbol{\pi}_A \exp(\boldsymbol{C}_A l_{ijA1}) \boldsymbol{\Lambda}_A \exp(\boldsymbol{C}_A l_{ijA2}) \boldsymbol{\Lambda}_A \exp(\boldsymbol{C}_A l_{ijA3}) \exp(\boldsymbol{C}_B l_{ijB1}) \boldsymbol{e}, \qquad (3.5)$$

where $\boldsymbol{\pi}_A$, $\boldsymbol{C}_A$ and $\boldsymbol{\Lambda}_A$ are calculated with the intensity parameters for section A and l_{ijAk} are the inter-detection distances in section A, with l_{ijA1} the distance to the first detection and l_{ijA3} the distance from the last detection to the end of the section; the same argument applies to section B.

3.4. RELATION WITH THE BERNOULLI PROCESS OCCUPANCY MODELS FOR CLUSTERING

The counterpart to the 2-MMPP occupancy model for discretized data would be a two-state Markov modulated Bernoulli process (2-MMBP) occupancy model. In such a model, detections at occupied sites are described as coming from Bernoulli trials with two possible probabilities of success (detection probabilities p_1 and p_2). Which of the two probabilities of success is effective at each replicate is governed by a Markov chain, with transition probabilities q_{12} (from state p_1 to p_2) and q_{21} (from state p_2 to p_1). We can expect the discrete 2-MMBP occupancy model to be a good approximation to the continuous 2-MMPP occupancy model if the continuous detection data are discretized using sufficiently small intervals, with $p_i \approx 1 - \exp(-\lambda_i L/K)$ and $q_{ij} \approx 1 - \exp(-\mu_{ij} L/K)$, where L/K is the length of transect segments.

Hines et al. (2010) propose two models that refine the Bernoulli process occupancy model to account for dependence between consecutive spatial replicates. They refer to their models as a "Markov process for segment occupancy model" and a "trap response model". These two models are in effect particular cases of a 2-MMBP occupancy model where p_2 is set to zero, that is, detections are modeled as an interrupted Bernoulli process (IBP). In the "trap response model" p_1 is also fixed, set to 1 (see Supplementary Materials).

4. SIMULATION STUDY

4.1. INDEPENDENT DETECTIONS

We conducted a simulation study to evaluate the performance of the Poisson process occupancy model compared to the Bernoulli process occupancy model for detection data collected under continuous sampling. For each sampling site the occupancy status was determined as the outcome of a Bernoulli trial with probability ψ. Detection data for occupied sites were generated following a homogeneous Poisson process. We ran 10,000 simulations of a study design where the total survey effort available was 600 km and the length surveyed per site was 6 km, therefore resulting in 100 sampled sites. Occupancy was set to $\psi = 0.25, 0.5$ and 0.75. The detection rate was $\lambda = 0.1, 0.2$ or 0.3 detections/km, which give probabilities λ^* of detecting the species at an occupied site of 0.45, 0.70 and 0.83, respectively. For the Bernoulli process occupancy model the detection data were discretized based on three segment lengths (1, 2 and 3 km). The simulation results (Table 2a)

Table 2. Performance of the occupancy estimator in the Poisson process (PP) and Bernoulli process (BP) occupancy models when data are generated according to a Poisson process with detection rate λ, occupancy ψ, the total survey effort available 600 km and a sampling design based on either (a) $L = 6$ km or (b) the optimal L. Three segment lengths (1, 2 and 3 km) are tested for the BP model. Mean and mean square error (in square brackets) of the occupancy estimator $\hat{\psi}$ based on 10,000 simulations are shown.

	ψ	λ	L	$\hat{\psi}$ PP	BP 1 km	BP 2 km	BP 3 km
(a)	0.25	0.1	6.0	0.34 [0.055]	0.35 [0.069]	0.38 [0.090]	0.41 [0.123]
		0.2	6.0	0.26 [0.006]	0.27 [0.007]	0.27 [0.010]	0.28 [0.015]
		0.3	6.0	0.25 [0.003]	0.26 [0.003]	0.26 [0.003]	0.26 [0.004]
	0.50	0.1	6.0	0.55 [0.038]	0.56 [0.044]	0.57 [0.051]	0.58 [0.061]
		0.2	6.0	0.51 [0.008]	0.52 [0.009]	0.52 [0.011]	0.53 [0.015]
		0.3	6.0	0.50 [0.004]	0.51 [0.004]	0.51 [0.005]	0.51 [0.005]
	0.75	0.1	6.0	0.77 [0.026]	0.77 [0.029]	0.77 [0.031]	0.78 [0.035]
		0.2	6.0	0.76 [0.009]	0.76 [0.010]	0.77 [0.012]	0.77 [0.014]
		0.3	6.0	0.75 [0.004]	0.76 [0.005]	0.76 [0.005]	0.76 [0.006]
(b)	0.25	0.1	15.7	0.28 [0.016]	0.28 [0.018]	0.29 [0.022]	0.29 [0.024]
		0.2	7.9	0.26 [0.004]	0.26 [0.005]	0.27 [0.007]	0.27 [0.010]
		0.3	5.2	0.26 [0.003]	0.26 [0.003]	0.26 [0.005]	–
	0.50	0.1	18.8	0.51 [0.014]	0.52 [0.014]	0.52 [0.015]	0.52 [0.015]
		0.2	9.4	0.51 [0.006]	0.51 [0.006]	0.51 [0.007]	0.51 [0.007]
		0.3	6.3	0.50 [0.004]	0.51 [0.004]	0.51 [0.005]	0.51 [0.005]
	0.75	0.1	25.1	0.76 [0.012]	0.76 [0.013]	0.76 [0.013]	0.76 [0.013]
		0.2	12.5	0.75 [0.006]	0.75 [0.006]	0.75 [0.006]	0.76 [0.006]
		0.3	8.4	0.75 [0.004]	0.75 [0.004]	0.75 [0.004]	0.76 [0.004]

show that, at this sample size (600 km), the occupancy estimator has in general little bias. However, for rare and elusive species (low occupancy and low detection rate), the occupancy estimator is biased and its variance is larger than that predicted by the asymptotic approximation (e.g., for $\psi = 0.25$, $\lambda = 0.1$ and $L = 6$, the asymptotic MSE, which corresponds to the asymptotic variance given the estimator is asymptotically unbiased, is 0.013 while the actual MSE is 0.055). This reinforces the value of using simulations for study design. When the detection process follows a continuous process, the discretization of the data produces an occupancy estimator with larger bias and variance, especially when transects are divided into a few large segments. The estimator is more prone to estimates at the boundary of the parameter space (e.g., the proportion of estimates $\tilde{\psi} = 1$ is 0.052 in the Poisson process model, and 0.078, 0.118 and 0.188 in the Bernoulli process model with 1, 2 and 3 km segment lengths, when $\psi = 0.25$, $\lambda = 0.1$ and $L = 6$). Simulations run assuming a study design based on the optimal per-site survey length (Table 2b) illustrate how the estimator properties improve in this case.

4.2. CLUSTERED DETECTIONS

We used simulations to investigate the performance of the occupancy models presented for clustered detection data. We generated data based on a 2-MMPP with one state without detections and another state, in which less time is spent, with detections. The probability

Table 3. Performance of the occupancy estimator in the Poisson process (PP) and 2-state Markov modulated Poisson process (2-MMPP) occupancy models when data are generated according to a 2-MMPP with detection rates $\lambda = [0\ 5]$ and switching rates (a) $\mu = [1/15\ 1/10]$, (b) $\mu = [1/15\ 1/2]$, (c) $\mu = [1/15\ 1/1]$, (d) $\mu = [1/15\ 1/0.5]$ and (e) $\mu = [1/15\ 1/0.2]$. The simulated sampling design consists of surveying 100 sites, with one 20 km transect surveyed per site. Mean and mean square error (in square brackets) of the occupancy estimator $\hat{\psi}$ based on 500 simulations are shown. The probability of detecting the species at occupied sites is also shown.

| | | | $\hat{\psi}$ | |
ψ	λ, μ	p (detect)	PP	2-MMPP
0.25	(a)	0.83	0.21 [0.004]	0.26 [0.004]
	(b)	0.73	0.18 [0.006]	0.26 [0.005]
	(c)	0.68	0.17 [0.007]	0.27 [0.007]
	(d)	0.62	0.15 [0.010]	0.27 [0.007]
	(e)	0.49	0.13 [0.017]	0.27 [0.013]
0.50	(a)	0.83	0.41 [0.010]	0.51 [0.005]
	(b)	0.73	0.37 [0.020]	0.51 [0.007]
	(c)	0.68	0.34 [0.027]	0.51 [0.009]
	(d)	0.62	0.31 [0.038]	0.51 [0.011]
	(e)	0.49	0.26 [0.061]	0.51 [0.017]
0.75	(a)	0.83	0.62 [0.018]	0.76 [0.007]
	(b)	0.73	0.55 [0.042]	0.77 [0.010]
	(c)	0.68	0.51 [0.059]	0.75 [0.011]
	(d)	0.62	0.46 [0.084]	0.76 [0.014]
	(e)	0.49	0.38 [0.142]	0.77 [0.023]

of occupancy was set to $\psi = 0.25$, 0.5 and 0.75 and the sampling design had 100 sites with one transect of length $L = 20$ km surveyed per site. The 2-MMPP parameters used were detection rates $\lambda = [0\ 5]$ and switching rates $\mu = [1/15\ 1/10]$, $\mu = [1/15\ 1/2]$, $\mu = [1/15\ 1/1]$, $\mu = [1/15\ 1/0.5]$ and $\mu = [1/15\ 1/0.2]$. With these parameter values and sampling design the probabilities of detecting the species at an occupied site, $1 - \pi \exp(CL)e$, were 0.83, 0.73, 0.68, 0.62 and 0.49, respectively. The simulation results (Table 3) show that the occupancy estimator in the Poisson process occupancy model is negatively biased in the presence of clustering in the detection data.

5. APPLICATION TO SUMATRAN TIGER DETECTION DATA

5.1. DATA DESCRIPTION

We illustrate the use of the models proposed here with the analysis of detection/non-detection data from a survey of the Sumatran tiger carried out by Fauna and Flora International (FFI) / Durrell Institute of Conservation and Ecology (DICE) in the forested areas in and around Kerinci Seblat National Park, the largest national park in Sumatra (Indonesia), during 2007 and 2008. Surveys were carried out in 89 sampling sites (17×17 km squares) where survey teams walked transects in the forest noting down the location of any tiger footprints that were detected. A total distance of 2826.5 km was covered. The distance covered per site varied from 1.8 to 108.3 km, and was typically around 15–45 km (Figure 1). Within each cell the length surveyed was often made up of several transects which
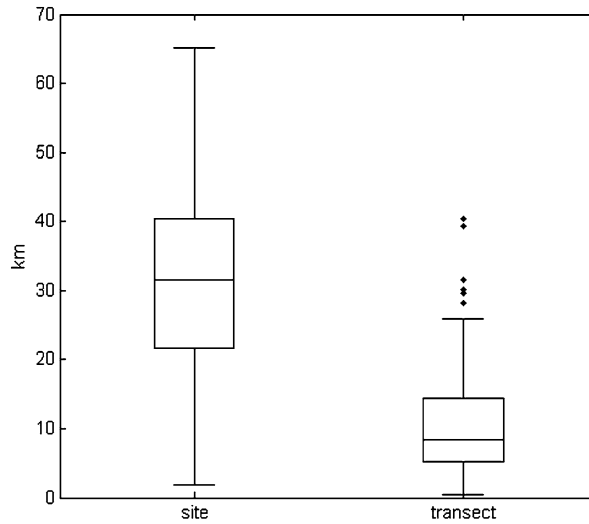
Figure 1.   Box plot of distances walked within each site and transect lengths.

Table 4.   Parameter estimates and AIC values for the occupancy models with a continuous detection process fitted to the Sumatran tiger data where PP: homogeneous Poisson process, 2-MMPP: 2-state Markov modulated Poisson process, IPP: Interrupted Poisson process and PP Mixture: mixture of two homogeneous Poisson processes. The unit of $\hat{\lambda}_i$ is km^{-1} and the unit of $\hat{\rho}_{ij} = 1/\hat{\mu}_{ij}$ is km. Standard errors (in brackets) were calculated using the delta method approximation from those obtained in the transformed domain (logit or log).

	PP	2-MMPP	IPP	PP Mixture
$\hat{\psi}$	0.82 (0.049)	0.96 (0.065)	0.97 (0.067)	0.96 (0.065)
$\hat{\lambda}_1$	0.11 (0.007)	0.23 (0.030)	0.19 (0.023)	0.22 (0.025)
$\hat{\lambda}_2$	–	0.03 (0.009)	Fixed to 0	0.03 (0.008)
$\hat{\rho}_{12}$	–	121 (216)	28 (18)	–
$\hat{\rho}_{21}$	–	243 (413)	28 (15)	–
$\hat{\pi}_1$	–	–	–	0.35 (0.064)
AIC	1722.1	1662.5	1674.4	1660.8
ΔAIC	61.3	1.7	13.6	0

varied in length from 0.5 to 40.1 km, with most in the range 3–21 km. For the analysis, these transects were assumed to be statistically independent, so that their contributions to the likelihood can be multiplied. Tiger footprints were detected in 66 of the cells, which gives a naïve occupancy estimate of 0.74.

5.2. MODEL FITTING

We implemented the models in MATLAB and obtained maximum-likelihood estimates by numerical maximization. The results from fitting the homogeneous Poisson process occupancy model and the 2-MMPP occupancy model (Table 4) indicate that the latter model fits the tiger data substantially better. Its Akaike Information Criterion (AIC) value was almost 60 units lower despite the penalty due to having three additional parameters. Fitting

an interrupted Poisson process for the detection process (i.e. λ_2 fixed to zero) was also better than the homogeneous Poisson process but considerably worse than the 2-MMPP. We note that the estimate of occupancy under the 2-MMPP model is higher than under the homogeneous Poisson process model, which confirms that disregarding the dependence between detections causes a negative bias in the occupancy estimator. Since for this parameter the symmetric 95% confidence interval derived from the point estimate extended beyond unity (0.835–1.090), we derived a new interval based on the profile log-likelihood obtaining (0.831–1.000). The estimates associated with the detection process in occupied cells suggest that the rate of encounter of tiger footprints is ten times higher in some areas compared to others and that the average distance traveled in areas with high encounter rate ($\rho_{12} = 1/\mu_{12}$) is half the average distance traveled in areas with low encounter rates ($\rho_{21} = 1/\mu_{21}$). We chose to use a parameterization for the embedded Markov process in terms of expected holding times (ρ) instead of switching rates (μ) as this provides a more straightforward interpretation of our results. The point estimates of the ρ parameters are larger than what we expected initially; however, their standard errors indicate poor precision. A reparameterization of the model using the ratio and sum of the ρ parameters ($R = \rho_{12}/\rho_{21}$ and $A = \rho_{12} + \rho_{21}$) gave estimates $\hat{R} = 0.5$ (0.15) and $\hat{A} = 365$ (628) showing that estimate of the ratio is relatively precise and that most of the uncertainty lies in their magnitude. This is interesting, as it suggests that the tiger data are informative in terms of the probability of being in each of the two states but carry little information on the actual rate at which transitions between states occur. The transects walked in this survey were short compared to the rate at which state transitions seem to take place, which prevents the accurate estimation of these parameters. Our results suggest that, due to their length, transitions are unlikely to occur within transects so we could expect that a mixture of two Poisson processes would describe our detection data well. Under this model the detection rate along a transect can take one of two values with a given probability, but does not change within the transect (see Supplementary Materials). Fitting this model provides a very similar likelihood value (826.4 vs. 826.2) and practically the same estimates of occupancy and detection rates as the 2-MMPP model. The estimate of the probability of being in the state with high detection rate $\hat{\pi}_1 = 0.35$ (0.064) also matches the corresponding estimate based on the 2-MMPP model $\hat{\pi}_1 = \hat{R}/(\hat{R} + 1) = 0.33$ (0.067).

5.3. MODEL DIAGNOSTICS

We explored two aspects of goodness-of-fit for the models fitted: how well the models describe (1) the inter-detection distances and (2) the distances from the beginning of each transect until the first sighting. The survivor function for inter-detection distances for the models fitted is given by

$$S_{\text{2-MMPP}}(l) = \boldsymbol{\pi}^* \exp(\boldsymbol{C}l)\boldsymbol{e}, \qquad S_{\text{PP}}(l) = \exp(-\lambda l)$$

and the survivor function for the distance until first sighting for the models fitted is given by

$$S_{\text{2-MMPP}}(l) = \psi \boldsymbol{\pi}' \exp(\boldsymbol{C}l)\boldsymbol{e} + (1 - \psi), \qquad S_{\text{PP}}(l) = \psi \exp(-\lambda l) + (1 - \psi).$$
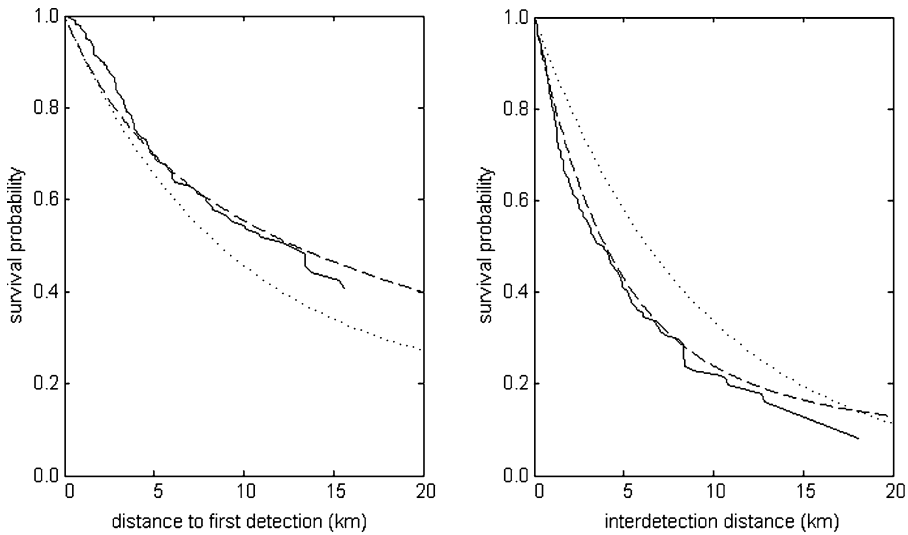
Figure 2. Survivor function for 'distance to first detection' (a) and 'inter-detection distances' (b) for: the Suma-tran tiger data (solid line), the 2-MMPP occupancy model (dashed line) and the Poisson process occupancy model (dotted line). The empirical survivor function is the Kaplan–Meier plot. The survivor functions for the models are derived using parameter estimates in Table 4.

A comparison of the fitted survivor functions (Figure 2) with the empirical survivor function (Kaplan–Meier curve) confirms that the 2-MMPP model has a better fit than the Poisson process model. We used the Kaplan–Meier curve (Kaplan and Meier 1958) as it accounts for right-censoring in the data, due to transects that ended before detecting the next (or any) tiger footprint.

As an additional assessment of fit we fitted a three-state MMPP occupancy model. The maximum likelihood estimates obtained were such that in effect the 3-MMPP model collapsed to the 2-MMPP model, therefore indicating that a model structure with two states provides good fit for the data.

6. DISCUSSION

Point process models provide a natural framework for the description of species detection data collected along transects. Although point processes have been used in the context of distance sampling studies (Hedley and Buckland 2004; Skaug 2006), their potential for occupancy modeling has not been explored to date. Transects have been used as the basis for collecting detection data in surveys targeted to study species occupancy but the analysis of such data has however been carried out by modeling the detection process as a discrete process after pooling detections from transect segments of a given length (e.g. Hines et al. 2010). Here we model the detection process in occupancy models based on transect surveys as a continuous process. In particular we develop occupancy models that use a Poisson process to describe detections that can be assumed independent and alternatively a 2-MMPP process for cases with clustered detections. We found that the 2-MMPP occupancy model was more appropriate than a model based on a homogeneous Poisson process

for the Sumatran tiger detection data. We also propose an occupancy model that describes species detections as a mixture of Poisson processes, which can be seen as a limiting case of the 2-MMPP occupancy model when $\mu_{12} + \mu_{21}$ approaches zero. This model is simpler to fit and provided an adequate description of our tiger data. However, in general, the 2-MMPP model will be a more appropriate choice for modeling data of this type as it accounts for potential state transitions within transects, which is relevant if transects are long compared to the scale of clustering in detections of the species.

The use of point processes for the description of the detection process in occupancy modeling opens the door to a number of different model developments, including the extension of the models described to account for abundance-induced heterogeneity, which we are currently investigating. Note that all the models and discussion addressed here are also applicable to surveys that collect detection data by monitoring each sampling site over a continuous interval of time. Camera-trap surveys provide an obvious example of such a design. To date these kind of data have been used to estimate occupancy by collapsing trapping times in intervals of a given duration (e.g. 2 weeks, Linkie et al. 2007). Another example is point count surveys, often targeted to avian species, which involve collecting detection data over an interval of time at a number of sampling stations. Note that, since the interval of time is usually relatively short in point counts (e.g. 10 minutes), independent repeated surveys at each sampling station may be needed to deal with the issue of closure. Otherwise the model would estimate the probability of the species being available for detection at a site during the time interval, rather than actual species occupancy. In the model structure presented in this paper these replicated surveys would be analogous to the independent transects surveyed within each sampling site.

When discussing the use of the discrete occupancy model, Hines et al. (2010) raise the question of whether there is an optimal transect segment length with respect to estimator properties (precision or MSE). Here we argue that it is best to model the detection process along transects as a continuous process, which in practice corresponds to dividing transects into infinitesimal segments. Pooling detections from larger transect segments may result in a poorer occupancy estimator, which is not surprising as data are lost in the discretization process. The use of relatively long transect segments can however help to mitigate the potential lack of dependence between adjacent segments, when not explicitly accounted for in the model.

Finally, it is interesting to note the parallelism of the occupancy models described here with models used in other applications. In survival analysis for instance the target is to model the time to an event, such as the onset of a medical condition or the failure of an electrical system (Cox and Oakes 1984). Survival data tend to be right-censored as experiments often end before all susceptible individuals develop the condition, in the same way as surveys may end before detecting the species of interest at occupied sites, causing the so-called "false absences". Furthermore, there are models that account for individuals that are "immune" to the condition of interest. These individuals, referred to as "long-term survivors", introduce zero-inflation in the same way that non-occupied sites do in species occupancy data. Survival analysis, counting processes and, in general, point process theory therefore provide an opportunity for bringing new ideas into species occupancy modeling, and vice versa.

SUPPLEMENTARY MATERIALS

"Markov process for segment occupancy model" and "trap response model" as 2-MMBP occupancy models: This supplementary material shows the occupancy models proposed by Hines et al. (2010) as models where the detection process is described as a 2-MMBP (SupMatA.pdf file).

"Poisson process mixture occupancy model": This supplementary material provides the likelihood function for the Poisson process mixture occupancy model (SupMatB.pdf file).

ACKNOWLEDGEMENTS

[Published Online January 2011.]

REFERENCES

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (2005), "On the Lambert W Function," *Advances in Computational Mathematics*, 5, 329–359.

Cox, D. R., and Isham, V. (1980), *Point Processes*, London: Chapman and Hall.

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman & Hall.

Fischer, W., and Meier-Hellstern, K. (1992), "The Markov-Modulated Poisson Process (MMPP) Cookbook," *Performance Evaluation*, 18, 149–171.

Guillera-Arroita, G., Ridout, M. S., and Morgan, B. J. T. (2010), "Design of Occupancy Studies with Imperfect Detection," *Methods in Ecology and Evolution*, 1, 131–139.

Hedley, S. L., and Buckland, S. T. (2004), "Spatial Models for Line Transect Sampling," *Journal of Agricultural, Biological, and Environmental Statistics*, 9, 181–199.

Hines, J. E., Nichols, J. D., Royle, J. A., MacKenzie, D. I, Gopalaswamy, A. M., Samba Kumar, N., and Karanth, K. U. (2010), "Tigers on Trails: Occupancy Modeling for Cluster Sampling," *Ecological Applications*, 20, 1456–1466.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Karanth, K. U., and Nichols, J. D. (2002), *Monitoring Tigers and Their Prey: A Manual for Researchers, Managers and Conservationists in Tropical Asia*, Bangalore: Centre for Wildlife Studies.

Linkie, M., Dinata, Y., Nugroho, A., and Haidir, I. A. (2007), "Estimating Occupancy of a Data Deficient Mammalian Species Living in Tropical Rainforests: Sun Bears in the Kerinci Seblat Region, Sumatra," *Biological Conservation*, 137, 20–27.

MacKenzie, D. I., and Royle, J. A. (2005), "Designing Occupancy Studies: General Advice and Allocating Survey Effort," *Journal of Applied Ecology*, 42, 1105–1114.

MacKenzie, D. I., Bailey, L. L., and Nichols, J. D. (2004), "Investigating Species Co-occurrence Patterns When Species are Detected Imperfectly," *Journal of Animal Ecology*, 73, 546–555.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A. (2002), "Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One," *Ecology*, 83, 2248–2255.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., and Franklin, A. B. (2003), "Estimating Site Occupancy, Colonization, and Local Extinction When a Species is Detected Imperfectly," *Ecology*, 84, 2200–2207.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., and Hines, J. E. (2006), *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*, New York: Academic Press.

Morgan, B. J. T., Revell, D. J., and Freeman, S. N. (2007), "A Note on Simplifying Likelihoods for Site Occupancy Models," *Biometrics*, 63, 618–621.

Royle, A. J. (2006), "Site Occupancy Models With Heterogeneous Detection Probabilities," *Biometrics*, 62, 97–102.

Rydén, T. (1994), "Parameter Estimation for Markov Modulated Poisson Processes," *Communications in Statistics—Stochastic Models*, 10, 795–829.

Skaug, H. J. (2006), "Markov Modulated Poisson Processes for Clustered Line Transect Data," *Environmental and Ecological Statistics*, 13, 199–211.