University of Alberta

Site Occupancy Models

by

Monica Rocio Moreno Prieto

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Mathematical and Statistical Sciences Department

Fall, 2011

Edmonton, Alberta.

to Juan and Gregorio

Abstract

The probability of a species to be present on a certain site is a quantity of interest for monitoring programs. Data for the occupancy of a species is recorded along with habitat covariates that are suspected to relate with its status (presence/absence). The objective of the analysis is to estimate the proportion of sites in which the species is present and the effects of the habitat covariates on the status of the species. Nevertheless, it is possible to have some errors in the observations. The most popular approach to estimate occupancy while accounting for the detection error is that of multiple surveys, for which every site is visited several times. The effectiveness of this approach relies on two main assumptions: 1) during the time of the study the population is closed and 2) the replicate visits at every site are independent from each other.

In my thesis I evaluated the multiple surveys approach under two perspectives: the statistical properties and the sensibility of its assumptions. For the former, I found that the estimates are unstable when the number of visited sites and the number of surveys are small. To overcome these flaws in the estimation procedure, I developed an alternative estimation method, based on penalized likelihood, that provides better estimates for small number of sites and surveys. The analysis of the sensibility of the assumptions revealed that the violation of the assumptions could lead to biased estimates. The single survey model does not require the closed population assumption, but the popular belief for the non-identifiability of the parameters sanctioned its use. I tested the reliability of the estimates for the probability of occupancy and detection from information collected from a single survey and found that, contrary to popular belief, the parameters are identifiable under certain conditions. Finally, I developed a model (the cluster sampling model) to include the dependence between sites. This model allows estimation of the site occupancy using information collected on a single survey from sites that are correlated.

Acknowledgements

First and foremost, I want to acknowledge Juan for he has been my support, my inspiration, the person I can turn to when in need of strength and who will encourage me during the most difficult times. He is a reason for me to smile.

I also want to acknowledge my sisters, my brother, my mother and specially my father for showing my what perseverance, hard work and dedication are and how far they can take me.

I also want to thank my supervisor, Dr. Subhash Lele. He has continuously given me advice for life as a researcher, he taught me the importance of building an opinion of my own and question everything before acknowledging it as truth. I am thankful for his support, guidance and advice in every academic aspects of my PhD as it will help me building a strong research career.

I want to thank Erin Bayne for the academic discussions we had, his encouragement to keep the models attached to the reality and for thoroughly proofreading the English style and grammar in the final version of my thesis. I also thank Dr. Prasad for being part of the supervisory committee, and Dr. Fangliang He and Dr. Carl Schwarz for their active participation on the evaluation of the thesis. I want also to thank Dr. Brian R Gray for providing the data for the cluster sampling model and the interpretation of the results.

I want also to recognize the support staff in the Mathematical and Statistical Sciences Department at the University of Alberta. In particular I want to thank Tara Schuetz and Patti Bobowski for the help they have provided me with in the moments when I needed it most.

Finally, I would like to acknowledge the Universidad de Los Andes (Bogota, Colombia), the Department of Mathematical and Statistical Sciences at the University of Alberta and the Alberta Biodiversity Monitoring Institute for supporting my research.

Contents

List of Tables

List of Figures

List of symbols

δ	**Probability of detection**
ψ	**Probability of occupancy**
ψ_{naive}	**Naive probablity of occupancy (ignoring detection error)**
AIC	**Akaike's Information Criterion**
AUC	**Area Under the ROC**
FIM	**Fisher Information Matrix**
MCMC	**Markov Chain Monte Carlo**
MLE	**Maximum Likelihood Estimator**
MPaLE	**Maximum Partial Likelihood Estimator**
MPLE	**Maximum Penalized Likelihood Estimator**

N	**Abundance**
n	**Number of surveyed sites**
ROC	**Receiver Operating Curve**
RSPF	**Resource Selection Probability Function**
SMLE	**Simulated Maximum Likelihood Estimator**
ZIB	**Zero Inflated Binomial**

Chapter 1

Introduction

The systems studied in Ecology can be described according to its level of complexity in a hierarchical structure. At the first level is the population of a particular species, the second level consists on a group of populations called a metapopulation, the last level is an assemblage of different species called community. Different metrics and models have been developed to study every level of this hierarchy. This thesis is focused on the study of one particular metric, the *site occupancy*. The site occupancy is defined as the proportion of patches in which the species is present at an specific time. This metric has been used at the metapopulation level to describe its status, how it changes over time, and to identify key habitat factors for its persistence. It has also been used in a community level to determine the number of species and the interaction between them. The statistical models developed to estimate this proportion are called *site occupancy models*.

In this thesis we illustrate how and why the site occupancy is an important metric in Ecology. We also evaluate the current statistical models used to estimate it; in particular, those models applied to sampling circumstances in which the presence/absence of the species at the study field cannot be determined without error. Finally, we proposed some alternative models that overcome its current flaws.

This document is organized as follows: chapter 2 contains a summary of the main theoretical models developed to study every level of the ecological hierarchy mentioned above, along with some definitions that will be used in the rest of the document. The goal of this summary is to provide a general understanding of how the site occupancy probability is linked to the study of metapopulations and communities. Chapter 3 contains an overview

of the current available statistical models to estimate the site occupancy probability of a metapopulation in a fixed period of time, with a particular emphasis in the Zero Inflated Binomial (ZIB) model, the most popular method to estimate the site occupancy probability when the probability of detection is less than one. It also contains a brief description on how the ZIB model has been extended to study the dynamics of a metapopulation and the structure and dynamics of a community. Chapter 4 contains an exhaustive evaluation of the ZIB model based on the statistical properties of its estimates, the feasibility of the assumptions and the robustness of the estimates against the violation of the assumptions. Some of the criticisms raised in chapter 4 lead to the development of the models presented in the rest of the document.

Chapter 5 introduces an alternative estimation method that, based on penalized likelihood, improves the estimates of the parameters of the ZIB. In chapter 6, the general belief that multiple surveys are required to correctly estimate the site occupancy probability is confronted with a model whose estimation is based on a single survey sampling scheme. In chapter 7, an extension of the single survey model where the assumption of independence between visited sites is relaxed is presented. This extended model incorporates into the estimation the correlation between adjacent sites, a common feature of site occupancy studies. The last two chapters show some of the projects I intend to pursue after the completion of my doctoral studies, together with a summary of the conclusions from this work.

Chapter 2

Ecological models for the study of populations and communities

The elements of ecology systems can be organized according to their level of complexity, usually in a hierarchical structure. At the first level in that hierarchy is a population of a particular species, at the second level is a metapopulation, and the third and more complex element is a community. Ecologists have been interested on studying every element in this hierarchy, more specifically, the structure in a fixed period of time and how that structure changes over time (dynamics of the system). Sections 2.1, 2.2, and 2.3 provide a definition for every level in the hierarchy, a brief description of the statistics typically used to describe its structure and some of the theoretical models that have been developed to study dynamics within that hierarchy.

2.1 Population

A population is defined as a group of individuals of the same species occupying a particular space at a particular time[55]. Spatial boundaries defining populations sometimes are easily identified (e.g., individuals inhabiting small islands or isolated habitat patches), but more frequently are difficult to determine. For that reason the spatial boundaries of a population are often arbitrarily defined by the investigator[116]. The structure of a population at time t is described in terms of the number of individuals N_t, also called *abundance*. Temporal changes in abundance can be expressed by the following difference equation:

$$N_{t+1} = N_t + B_t + I_t - D_t - E_t, \tag{2.1}$$

where N_{t+1} is the abundance at time $t+1$; B_t is the number of individuals that were born after the time t and before $t+1$; I_t is the number of new individuals recruited by immigration, D_t is the number of individuals that died during that interval of time, and E_t is the number of individuals lost by emigration. The variables B_t, I_t, D_t and E_t are associated with the primary processes that drive changes in population size: natality, mortality, immigration and emigration.

One of the main research interests in ecology is to estimate whether a population will persist, more specifically, to determine the probability that a population will go extinct within a given number of years. This is typically referred as population viability analysis. A time series of the abundance of a populations may be required for this kind of analysis.

2.2 Metapopulation

The concept of a metapopulation was first introduced by Richard Levins in 1969 [63]. Levins defined a metapopulation as a set of local populations inhabiting similar patches[1], where typically migration from one local population to at least some other patches is possible (Figure 2.1a). The assumptions of Levins' model were the size of the local populations occupying these patches is either 0 or K (carrying capacity[2]), the patches have equal area, the spatial arrangement of the populations has no effect on the dynamics of the system, and migration to occupied patches does not affect local population dynamics. The structure of a metapopulation, according to Levins' model, can be described as the proportion of occupied patches, denoted by ψ. Changes in the proportion of occupied sites (ψ) in continuous time are given by equation 2.2, where m and e are the colonization and extinction rates respectively.

$$\frac{d\psi}{dt} = m\psi\,(1-\psi) - e\psi \tag{2.2}$$

It can be shown that the equilibrium value of ψ for equation 2.2 is $\psi^* = 1 - {}^{e}/_{m}$, for which ψ^* is positive if $m > e$. This implies that for a metapopulation to persist, the colonization rate must exceeds a certain threshold value for any given extinction rate. Empirical studies

[1] A patch can be defined as a continuous area of space with all necessary resources for the persistence of a local population[41]

[2] Maximum number of individuals of a particular species that can be supported indefinitely in a specific environment.

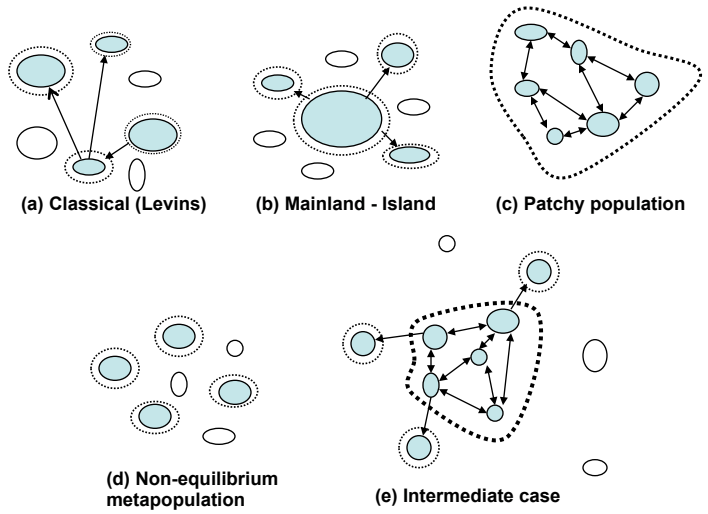


Figure 2.1: Metapopulation structures. Filled circles: occupied habitat patches; empty circles, vacant habitat patches; dotted lines, boundaries of local populations; arrows, dispersal. Adapted from Harrison and Taylor[43].

have shown that the extinction rate decreases in larger habitat patches [28, 81, 119], and that colonization rate decreases with increasing isolation [109, 110, 91, 36, 24, 88, 7, 114]. Consequently, according to Levins' model, a species may go extinct on systems with patches with small area and/or from systems with a large degree of isolation.

Many of Levin's assumptions are too simplistic to accurately model dynamics of biological populations. This has lead to the development of various models to incorporate information about the spatial location and differences in area of metapopulation patches. For instance, a metapopulation may have large variance among the size of the populations. In this type of system, the persistence of the metapopulation tends to be determined by the persistence of the largest single population (Figure 2.1b). Local extinctions affect the small populations, but the system can persist as long as the largest patch, also known as mainland, persists (for more information about this model see[67]). Another example is that of systems in which the patches are so close together that local extinctions are unlikely to occur (Figure 2.1c). Moreover, Levins' model assumed that every local extinction creates an empty habitat available for colonization. However, it may not be the case when the species is declining leading to a non-equilibrium system where colonization events are unlikely (Figure 2.1d). For a critical assessment of the metapopulation approach in field studies see Harrison and Taylor (1997)[43].

Jared Diamond in 1975 introduced the concept of *Incidence Functions* as a simple way

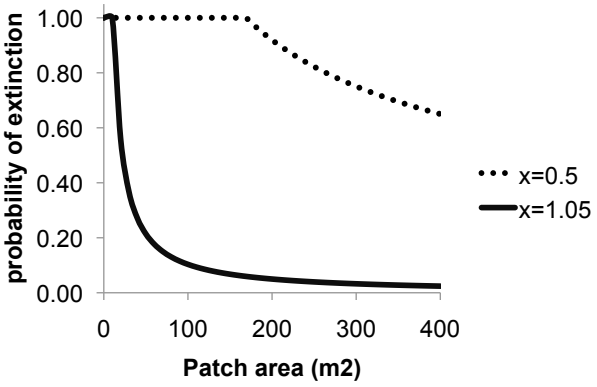


Figure 2.2: Probability of extinction vs patch size.

to depict the relationship between the probability of occurrence of a species on a island and the island's area[18]. He obtained the incidence function for different species of birds in islands around New Guinea. He found that the position and shape of the incidence function differed from one species to another. This observation lead him to conclude that incidence functions "can be interpreted in terms of the biology of a particular species[19]."

Ilkka Hanski in 1992 [37] proposed a simple model, based on the incidence functions defined by Diamond, to make inferences about the pattern of occupancy of species on islands. In this model the occupancy of a patch was described by a first order Markov Chain with two states: occupied or empty. The probability of a patch going from occupied to empty or extinction probability, was set to depend on the area A of the patch according to the following function $E(A) = \frac{e}{A^x}$. The larger the area, the smaller the probability of extinction. The parameter x described the strength of the relationship between the area and the probability of extinction; for $x > 1$ there is a threshold area that if exceeded, it is unlikely the species will go to extinction, on the other hand if $x < 1$ there is no such threshold and no matter how large the patch is, there will always be a substantial risk of extinction(Figure 2.2). On the other hand, the probability a patch goes from empty to occupied, colonization probability, was defined as a constant and denoted by C. The equilibrium fraction of islands that are occupied, or the incidence of the species was given by:

$$J(A) = \frac{1}{1 + \frac{e}{CA^x}} \tag{2.3}$$

Hanski in 1994 introduced a modified version of the Incidence Function model in which some of the assumptions were relaxed, and more information about the structure of the

metapopulation were considered. The colonization probability was modified to depend on the distribution of the species (presence/absence on the patches), isolation of the patches and the area. The model allowed the use of real data to estimate the parameters related to colonization and extinction rates for a specific species in a specific set of habitat patches. These estimates could then be used to simulate the dynamics of the system, and ultimately to predict the value of patch occupancy at equilibrium. For some studies using this model see [38, 42, 17, 79, 80, 40, 39]. The inferences made using Hanski's model are based on presence/absence data observed in the field, for which it is assumed that the status of the site (occupied/empty) is determined without error. Hanski's model established the idea that the metapopulation models were not only a theoretical approximation to understand the dynamics of a system, but could also be used to make inferences by using real data. Since then, there has been a huge amount of contributions in theory, models and field studies using this model.

Nowadays, there are several types of metapopulation models in the literature. Hanski's model constitutes what is now called *site occupancy models.* These models have the simplest structure since they describe each population as present or absent. More complex models describe each population in terms of their age structure[4]. These models incorporate spatial dynamics by modeling dispersal and temporal correlation among populations. There are also models that describe spatial structure within each individual population[56]. Some other models define the habitat as a regular grid where each cell of the grid can be modeled as a potential patch[92]. All these models have been developed with the goal of answering specific management questions (e.g., [12, 86, 85]). How to choose the correct model depends on the complexity of the problem, the assumptions, and the available data[3].

2.3 Community

A community is an assemblage of populations of living organisms in a prescribed area or habitat[55]. A community may include all the different plants and animal species represented in the space, or more commonly, may refer to a subset of species defined by taxonomy (e.g., the bird community), functional relationships (e.g., herbivore community), or other criteria relevant to the question of interest[116]. As in a population, ecologists are interested in studying two properties of communities: the structure in a specific time and the dynamics of the system.

Community structure

Community structure refers to static properties such as diversity and composition of the community. Some of the metrics used to define species diversity are species richness and species evenness. Species richness can be defined as the number of species in a community. Species evenness is a measure of how balanced the community is in terms of the abundance of every species in the community.

Studies related to community structure can be seen as a group of simultaneous population studies; if the purpose is to estimate the species evenness, the parameter of interest will be the abundance for every species. On the other hand, if the purpose is to determine the species richness; only the information about the presence/absence of every species is needed[116]. For a comprehensive review of methods for measuring diversity see Magurran (2004)[76].

Community dynamics

There are two ways of modeling community dynamics: one is by modeling *interspecific interactions,* the other is a more descriptive approach in which the focus is on modeling changes in the number of different species in the community. The later constitutes a field called *Island Biogeography.*

Interspecific interactions

The main interest for studies of interspecific interactions is to determine how the vital rates of one species change because of the interaction with another species. There are three major classes of interspecific interactions: competition, predation and mutualism[10]. Competition is a mutually detrimental interaction between individuals; organisms that share requirements of the same essential resources must compete with each other to gain access to those resources. Predation can be defined as any interaction between two species in which one benefits and the other suffers, more specifically, a predator species is one that has a negative effect on the immediate per capita population growth of the prey species[15]. Mutualism is an interaction in which both species benefit from the association. Parasitism is an interaction in which an species lives in or on another species (its host) and benefits by taking nutrients at the host's expense.

These interspecific interactions have been thought to be among the more important processes to determine the structure of communities[16] and for that reason they have been the subject of many contributions with theoretical models and field studies. An example of

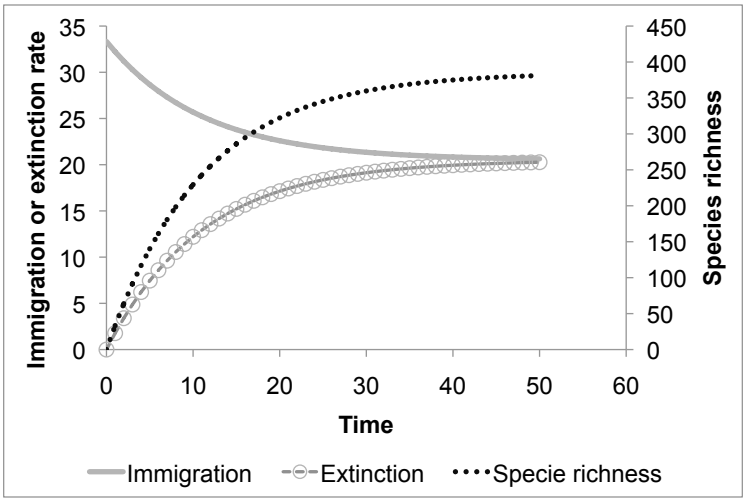


Figure 2.3: Example of the species accumulation dynamics according to MacArthur and Wilson[67] model. P=1000, A=200, D=300, c=0.1 and q=0.2.

models that have been developed to study interspecific interactions are the Lotka-Volterra Predator- Prey model[95]. According to this model, in a time continuum the change in the number of predators and prey can be described by the differential equations 2.4 and 2.5, where V is the number of prey, P is the number of predators, $\frac{dP}{dt}$ and $\frac{dV}{dt}$ represent the change in predator and prey population over time respectively, and α, β, γ, *and* δ are parameters representing the interaction between the two species.

$$\frac{dP}{dt} = \alpha VP - \beta P \tag{2.4}$$

$$\frac{dV}{dt} = \gamma V - \delta VP \tag{2.5}$$

Models like the Lotka-Volterra model have been useful in analyzing the consequences of interspecific interactions from a theoretical point of view. These theories have been supported with field studies (for some examples see [7, 49].)

Island biogeography

Island biogeography is a field of research that studies and explains the factors that affect species richness in communities. This field of research started when ecologists observed that larger islands and islands closer to a mainland support a greater number of species than smaller or more distant islands.

MacArthur and Wilson (1967) modeled species richness on an set of islands in which there is a big island (called, mainland) surrounded by a group of small islands (Figure

2.1b). The number of species on an island was considered to be the result of two processes: immigration and extinction. In their model, species from the mainland randomly migrate to an island. The rate at which new species arrive at the island was determined by three factors: the distance to the island from the mainland, the number of species remaining in the mainland that have not yet established themselves on the island, and the probability that a given species will disperse from the mainland to the island. Immigration rate was then defined as:

$$I = \frac{c\,(P - S)}{fD}, \tag{2.6}$$

c is the colonization probability, P is the number of species in the mainland pool, S is the species richness of the island, f is a scaling factor for distance, and D is the distance to the island from the mainland. Therefore, the farther an island lies from the mainland, the lower the rate of immigration will be. Since immigrants are drawn from a finite pool, as more species establish themselves on the island, fewer species that have not already established themselves on the island will remain in the pool. Similarly, the rate at which species on the island go extinct E (equation 2.7) was considered to be inversely related to the area A of the island, and directly proportional to the number of species present on the island.

$$E = \frac{qS}{A^m} \tag{2.7}$$

In the simplest version of the model, all species have equal probability q of reaching the island and of going extinct once there. The dynamics of the system can then be described by the discrete time model defined by:

$$\begin{aligned} S_{t+1} &= S_t + I_t - E_t \\ S_{t+1} &= S_t + \frac{c\,(P - S_t)}{fD} - \frac{qS_t}{A^m} \end{aligned} \tag{2.8}$$

Figure 2.3 shows an example of the dynamics of the system. It is observed that the *equilibrium species richness* is determined by a balance between immigration and extinction. This model is often used in conservation biology to predict the number of species that would be expected to persist or go extinct in nature reserves.

2.4 Summary

There are two parameters of interest when making inferences about the structure of an

ecological system in a fixed period of time: abundance N and the site occupancy probability ψ. Both parameters provide essential information to ecologists; having a reliable estimate of these two parameters can lead to make accurate inferences about the dynamics of the system. Similarly, if the estimates of the parameters are inaccurate, the inferences about the dynamic of the system are likely to be spurious.

Abundance is used to determine the status of a single population; series of estimates of abundance over the time can be used to make predictions about the persistence of the species and to determine which factors have an important impact on the probability of extinction for that species. Site occupancy probability can be used to describe the general status of all the populations contained within a metapopulation. It also provides information that is used to make inferences about the persistence of the metapopulation. In many situations it is preferred to estimate the probability of occupancy rather than to estimate the abundance because the former is more expensive and requires more survey effort [101, 72].

Chapter 3

Overview

Occurrence in a patch is a binary variable that describes the status of the species: present or absent. It is indexed by its location and the time. In this document, the occurrence of the species at the i^{th} patch at time t is denoted by z_{it} , where $z_{it} = 1$ indicates that the i^{th} patch is occupied by the target species at time t, and $z_{it} = 0$ indicates that the target species is absent from the i^{th} patch at time t.

From a statistical point of view, occurrence in a patch is a realization of a binary random variable, here denoted by Z_{it}, that follows a Bernoulli distribution with probability ψ_{it}. As mentioned in the previous chapter, this probability and how it relates to biotic and abiotic factors is of interest for ecological research and monitoring programs. This chapter contains an overview of the current available methods to estimate the probability ψ_{it}. Section 3.1 contains a description of the models that are applied to the study of the structure of a metapopulation in a fixed period of time, hereafter site occupancy models; sections 3.2 and 3.3 contain a brief description of how the site occupancy models have been extended to study the dynamics of a metapopulation and the structure of a community.

3.1 Site occupancy models

Field studies are conducted to estimate the probability of occurrence and its relationship with biotic and abiotic factors. Usually these studies are conducted by surveying a set of n patches[1]. The presence/absence of the target species and the values for some habitat covari-

[1] A patch is defined as a continuous area of space with all necessary resources for the persistence of a local population.

ates are recorded for every patch. If the status of the species at every patch is determined without error, an estimate of the probability of occupancy and the effects of the covariates can be estimated using *logistic regression[2].*

However, it is possible that a species that is occupying a site can go undetected by the observer during the time of the survey, in other words, it is possible to have false negatives within the vector of observations. It has been shown that if the detection error is ignored, it can lead to biased estimates of the probability of occupancy and of the effects of the covariates[34, 78]. The risk of such biases has led to the development of methods to account for imperfect detectability when estimating the site occupancy probability. Some of the most important contributions to this area of research are discussed.

Correcting the bias using the Horvitz-Thompson estimator

Paul Geissler and Mark Fuller in 1986[30] published what is considered to be the first approach to estimate the site occupancy probability when the probability of detection is less than one. In their model, the effects of biotic factors are ignored or assumed to be insignificant, thus, the probability of occupancy and detection were assumed to be constant. It is also assumed that it is not possible to misidentify the species (i.e., no false positive errors), and that the population is closed during the time of the surveys (i.e., the surveys are close enough in time that no individuals die, are born, move into the patch or move out of the patch between surveys).

The proposed method consisted of estimating the probability of detection by using replicate visits to the same patches. Once the probability of detection is estimated, an estimate of the probability of occupancy is obtained using the Horvitz-Thompson estimator[47], which is an unbiased estimator of a population total that is used in cases where the sampling probabilities are not the same for all individuals in the population. In the context of the Geissler and Fuller method, the sampling probabilities for the patches were considered to be the probability of detection at every patch. The sampling methodology and estimation procedure are as follows.

n patches are selected for the study. Each patch is visited k_i times. The observations at every site are recorded in a vector $\underline{y}_i = \{y_{i1}, ..., y_{ik_i}\}$, where $y_{ij} = 1$ if the species was detected at the i^{th} patch during the j^{th} survey, and 0 otherwise.

An estimate of the probability of detection, $\widehat{\delta}_i$, is then obtained by using the information

collected at those patches that are occupied with certainty (i.e., patches where the species was detected at least once). The estimated probability of detection is given by

$$\widehat{\delta}_i = \sum_{j=t_i+1}^{k_i} \frac{y_{ij}}{k_i - t_i}, \tag{3.1}$$

where t_i is the number of the survey in which the species was observed for the first time at the i^{th} patch. It is assumed that the probability of detecting the species in one survey is the same for all the patches. Hence, the probability that the species is detected at least once during the k_i visits to the patch is estimated by:

$$\widehat{p_i} = 1 - \left(1 - \overline{\widehat{\delta}}_{\cdot}\right)^{k_i}, \tag{3.2}$$

where $\overline{\widehat{\delta}}_{\cdot}$ is the average of the estimated detection probability obtained using equation 3.1. The Horvitz-Thompson estimator for the occupancy is then:

$$\widehat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \frac{w_i}{\widehat{p_i}}, \tag{3.3}$$

where $w_i = 1$ if the species was detected at least once over all the visits, and 0 otherwise.

Modeling the detection using a truncated geometric distribution

In 1990, David Azuma, James Baldwin and Barry Noon introduced a new model to estimate the site occupancy probability[5]. The model was developed to assess the distribution of the Spotted Owl over the Washington, Oregon and California region. The assumptions for this model are the same as to those of the Geissler and Fuller model: closed population, constant probability of occupancy, constant probability of detection, independence between surveys and no false positive errors. A patch was considered to be occupied if at least a pair of owls were at the patch.

The sampling procedure consisted of selecting n patches. These patches were visited until occupancy was established (i.e., detecting a pair of owls), or until s surveys were completed. The information of the surveys was recorded into the vector $\underline{v} = \{v_1, ..., v_{n_1}\}$ where v_i is the number of surveys that were conducted at the i^{th} patch until the first detection, and n_1 is the number of patches in which the species was detected.

Their approach to obtain an unbiased estimate of the probability of occupancy started by considering the probability of establishing that a site is occupied (i.e., detecting the

species at least one time after conducting s surveys). This probability is known as the naive probability of occupancy ψ_{naive}, and is equal to $\psi\left(1-(1-\delta)^{s}\right)$, where ψ is the probability of occupancy and δ is the probability of detecting the species in one survey. An estimate of the probability of occupancy was then given by:

$$\hat{\psi}=\frac{\hat{\psi}_{naive}}{1-\left(1-\hat{\delta}\right)^{s}} \tag{3.4}$$

The observed proportion of occupied sites $\hat{\psi}_{naive}={}^{n_1}/_{n}$ is an unbiased estimator of ψ_{naive}. An estimate of δ can be obtained from the information collected from the repeated surveys to the patches. Consider that the number of surveys required to detect the species is a random variable V with a *truncated geometric distribution.* The vector $\underline{v}$ contains n_1 realizations of this random variable. Using the method of moments, and knowing that the expected value of V is given by $E\left(V\right)=\frac{1}{\delta}-\frac{s(1-\delta)^{s}}{1-(1-\delta)^{s}}$, an estimate of δ can then be obtained from the following expression:

$$\overline{v}=\frac{1}{\hat{\delta}}-\frac{s\left(1-\hat{\delta}\right)^{s}}{1-\left(1-\hat{\delta}\right)^{s}}, \tag{3.5}$$

where $\overline{v}=n_1^{-1}\sum_{i=1}^{n_1}v_i$. The estimates of ψ_{naive} and δ can then be used in equation 3.4 to obtain an unbiased estimate of the probability of occupancy.

The zero inflated binomial model

In 2002, what is now the most recognized method used today by ecologists to account for imperfect detection was published, hereafter MacKenzie's model [73]. Similar to the previously discussed models, the estimation procedure for MacKezie's model requires repeated visits to the same site. The assumptions for this model are: closed population, independent surveys and no false positive errors.

The sampling protocol is as follows: n patches are selected at random, each patch is visited k_i times, and the absence/presence of the species is recorded for each site at each visit. The information from the surveys is then collected in the vectors $\underline{y}_i=\{y_{i1},...,y_{ik_i}\}$ for $i=1,...n$, where $y_{ij}=1$ if the species is detected at the i^{th} patch during the j^{th} survey, and 0 otherwise.

The model can be constructed hierarchically by defining a latent variable Z_i to indicate

the true status of the i^{th} patch. $Z_i = 1$ if the patch is occupied and 0 otherwise. This latent variable has a Bernoulli distribution with probability ψ .

$$Z_i \sim Bernoulli\,(\psi) \tag{3.6}$$

The random variable Y_i counts the number of surveys in which the species was detected at the i^{th} patch, $Y_i = \sum_{j=1}^{k_i} Y_{ij}$. The distribution of this random variable is conditioned on the true status of the patch. If the patch is occupied, Y_i follows a Binomial distribution with parameters k_i and δ , where δ is the probability to detect the species at the i^{th} patch. On the other hand, if the patch is empty, Y_i follows a degenerate Binomial distribution with probability 0. The marginal distribution of Y_i is the following zero-inflated binomial distribution [22] :

$$f\left(y_i/\psi,\delta\right) = \psi\binom{k_i}{y_i}\delta^{y_i}\left(1-\delta\right)^{k_i-y_i} + \left(1-\psi\right)I\left(y_i=0\right), \tag{3.7}$$

where $I\left(\bullet\right)$ is an indicator function that is equal to one if its argument is true and 0 otherwise. The values of the probability of occupancy and the probability of detection are estimated by maximizing the likelihood function:

$$L\left(\psi,\delta;y_1,...,y_n\right) = \prod_{i=1}^{n}\psi\binom{k_i}{y_i}\delta^{y_i}\left(1-\delta\right)^{k_i-y_i} + \left(1-\psi\right)I\left(y_i=0\right) \tag{3.8}$$

If patch specific covariates are available, they can be incorporated to the estimation of ψ by using the Logistic link, for instance let us denote by $\underline{x}_i = \{x_{i1},...,x_{ip}\}$ a vector of habitat covariates associated with the i^{th} patch. The probability of occupancy can then be defined as follows:

$$\psi_i = \frac{exp\left(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}\right)}{1+exp\left(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}\right)} \tag{3.9}$$

The parameters β_0, β_1,..., β_p quantify the effect of the habitat covariates over the probability of occupancy. If information of covariates related to the probability of detection is available, it can be incorporated into the estimation in the same manner.

For those patches for which its status were not established during the time of the study (i.e., the species was not detected at any of the surveys), the estimated parameters can be used to calculate what is the probability they were actually be occupied. This probability can be estimated as follows:

$$Pr\left(Z_i=1/y_i=0\right)=\frac{\psi\left(1-\delta\right)^{k_i}}{\left(\psi\left(1-\delta\right)^{k_i}\right)+\left(1-\psi\right)}, \tag{3.10}$$

which is the probability that a patch is occupied conditioned on that the species was not detected after k_i surveys.

For a complete guide in how to allocate survey effort (number of patches vs. number of surveys) see MacKenzie and Royle 2005. Table 3.1 presents the suggested number of surveys according to an approximated value of the probabilities of occupancy and detection. Notice that if a species is rare, it is suggested to survey less intensively than if the species is common. A procedure to assess the goodness of fit for this model can be found in MacKenzie and Bailey 2004[70].

Table 3.1: Optimum number of surveys to conduct at each site for a standard multiple survey site occupancy study, ψ: probability of occupancy, δ: probability of detection. Reproduced from MacKenzie and Royle 2005[74].

	ψ								
δ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	14	15	16	17	18	20	23	26	34
0.2	7	7	8	8	9	10	11	13	16
0.3	5	5	5	5	6	6	7	8	10
0.4	3	4	4	4	4	5	5	6	7
0.5	3	3	3	3	3	3	4	4	5
0.6	2	2	2	2	3	3	3	3	4
0.7	2	2	2	2	2	2	2	3	3
0.8	2	2	2	2	2	2	2	2	2
0.9	2	2	2	2	2	2	2	2	2

Incorporating heterogeneity on the probability of detection

In the models that have been discussed so far it is assumed that the probability of detection is either constant or varies according to some survey specific covariates. The model of Royle and Nichols (2003)[101] differs in that it assumes that the probability of detection depends on the abundance in every patch. Although the probability of occupancy is not explicitly

modeled, it can be derived from the model. The sampling protocol and the assumptions of the model are the same as that of MacKenzie at al (2002).

The statistical model is as follows: assume that the abundance at every patch, denoted by N_i, is a realization of a random variable N that follows a poisson distribution with parameter λ. Hence, the probability for a patch to inhabit N_k individuals is as follows:

$$Pr\left(N = N_k\right) = \frac{\lambda^{N_k} e^{-\lambda}}{N_k!} \tag{3.11}$$

Moreover, assume that the probability of detecting the species at the i^{th} patch depends on the abundance according to the following function:

$$\delta_i = 1 - \left(1 - r\right)^{N_i}, \tag{3.12}$$

where r is the probability for a particular individual to be detected, and δ_i is the probability of establishing that the i^{th} patch is occupied (i.e., detecting at least one individual). If the abundance at every patch were known the likelihood (equation 3.13) could be easily maximized to obtain the estimates of the parameters λ and r.

$$L\left(\lambda, r; y_1, ..., y_n, N_1, ..., N_n\right) = \prod_{i=1}^{n} \binom{k_i}{y_i} \delta_i^{y_i} \left(1 - \delta_i\right)^{k_i - y_i} \tag{3.13}$$

However, since that it is not the case, the variables N_i are treated as random variables; the estimates of the parameters of interest are obtained by maximizing the likelihood of the marginal distribution of the observations $y_1, ..., y_n$

$$L\left(\lambda, r; y_1, ..., y_n\right) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} \binom{k_i}{y_i} \delta_i^{y_i} \left(1 - \delta_i\right)^{k_i - y_i} Pr\left(N = j\right) \right) \tag{3.14}$$

The estimated probability of occupancy can then be calculated as follows:

$$\hat{\psi} = 1 - Pr\left(N = 0\right) = 1 - e^{-\hat{\lambda}} \tag{3.15}$$

Additional sources of heterogeneity in abundance can be added by allowing the Poisson mean to vary randomly among patches (modeling N as a negative binomial) or systematically as a function of patch-specific covariates[20]. For other forms of incorporating heterogeneity in the probability of detection see Royle 2006 [97].

Incorporating false positive errors

An extension of MacKenzie's' model was published in 2006 in a paper by Andrew Royle and William Link [100]. The extension consisted of eliminating from the model the assumption of no false positive errors, while the assumptions of closed population and independent surveys were kept. The sampling protocol is the same as that of MacKenzie et al 2002.

The model is hierarchically defined using a latent random variable Z_i that describes the true status of the patch. The probability distribution of the observations Y_i is conditioned on Z_i. If the patch is occupied by the target species (i.e., $Z_i = 1$), then Y_i follows a Binomial distribution with parameters δ_1 and k_i, where δ_1 is the probability of detecting the target species. On the other hand, if the patch is not occupied by the target species (i.e., $Z_i = 0$), Y_{ij} follows a Binomial distribution with parameters δ_0 and k_i, where δ_0 is the probability of misidentifying the target species. The likelihood for the data can then be written as follows:

$$L\left(\psi, \delta_1, \delta_0; y_1, ..., y_n\right) \propto \prod_{i=1}^{n} \left\{ \psi \left(\delta_1^{y_i} \left(1 - \delta_1\right)^{k_i - y_i} \right) + \left(1 - \psi\right) \delta_0^{y_i} \left(1 - \delta_0\right)^{k_i - y_i} \right\} \qquad (3.16)$$

It is easy to show that this likelihood provides equal support for multiple set of parameters values. For example, the likelihood for $\{\psi = 0.8, \delta_1 = 0.7, \delta_0 = 0.4\}$ is the same as the likelihood for $\{\psi = 0.2, \delta_1 = 0.4, \delta_0 = 0.7\}$[100]. Royle and Link proposed to solve this problem by imposing a restriction over the parameters. Specifically, they assumed that the detection rate at occupied sites is larger than the false detection rate (i.e., $\delta_1 > \delta_0$).

Defining multiple occupancy states

In 2007 another extension of MacKenzie's' model was published. Nichols et al (2007) [84] proposed a new model that allowed users to define multiple types for occupancy. The model was illustrated using data from the California Spotted Owls. Two occupancy states were defined: occupied with no production of young and occupied with successful reproduction.

The true status of a patch was modeled by a latent variable Z_i, where $Z_i = 0$ indicated that the i^{th} patch was unoccupied, $Z_i = 1$ indicated that the i^{th}patch was occupied but there was no production of young, and $Z_i = 2$ indicated that the i^{th} patch was occupied and there was successful reproduction.

The sampling protocol consisted of surveying n patches, every patch was visited k_i times.

The observations at every patch were recorded in a vector $y_i = \{y_{i1}, ..., y_{ik_i}\}$ where $y_{ij} = 0$ if the species was not detected at the i^{th} patch during the j^{th} visit; $y_{ij} = 1$ if the species was detected but there was not certainty about its reproduction state; and $y_{ij} = 2$ if the species was detected and evidence of successful reproduction was found.

It was assumed that if $max\{y_{ij}, ..., y_{ik_i}\} = 2$ there was certainty about the true state of the i^{th} patch (i.e., $Z_i = 2$). On the other hand, if $max\{y_{ij}, ..., y_{ik_i}\} = 1$ there were two possibilities for true state of the i^{th} patch: either $Z_i = 1$ or $Z_i = 2$. The later indicates that evidence of successful reproduction was missed by the observer. Similarly, if $max\{y_{ij}, ..., y_{ik_i}\} = 0$ the true state for the i^{th} patch was uncertain, it could be any of the three possibilities $Z_i = 0$, $Z_i = 1$, or $Z_i = 2$. The model was parameterized according to the following probabilities:

ψ_i^1	probability that the i^{th} patch is occupied regardless of reproductive state
ψ_i^2	probability that young occurred, given that the i^{th} patch is occupied
p_{ij}^1	probability that occupancy is detected for the i^{th} patch during the j^{th} survey given that $Z_i = 1$
p_{ij}^2	probability that occupancy is detected for the i^{th} patch during the j^{th} survey given that $Z_i = 2$
δ_{ij}	probability that evidence of successful reproduction is found, given that the species was detected at the i^{th} patch during the j^{th} survey and that $Z_i = 2$

The probability of the observations at every path can then be calculated using these parameters. For instance, consider a site that is visited three times and the vector of observations is $y_i = \{1, 0, 2\}$, the probability of y_i is :

$$Pr\left(y_i = \{1, 0, 2\}\right) = \left(\psi_i^1 \psi_i^2\right) \cdot \left(p_{i1}^2 \left(1 - \delta_{i1}\right)\right) \cdot \left(1 - p_{i2}^2\right) \cdot \left(p_{i3}^2 \delta_{i3}\right) \tag{3.17}$$

Given these observations, it is assumed that the site is occupied and the reproduction is successful (this assumption is based on the observation of the third visit).The first part of equation 3.17 corresponds to the probability that $Z_i = 2$. The second part accounts for what is observed in the first visit: the species is detected but evidence of reproduction is missed. The third part is the probability that the species was missed by the observer. Finally, the last part accounts for the probability of detecting the species and detecting evidence of reproduction. The probability for the observations at any patch can be calculated in a

similar manner. The likelihood for the data is proportional to the product of the probability of the observations at every patch.

$$L\left(\psi^1, \psi^2, p^1, p^2, \delta; y_1, ..., y_n\right) \propto \prod_{i=1}^{n} Pr\left(y_i\right) \tag{3.18}$$

The parameters can be estimated by maximizing the likelihood.

Multiple detection methods and multi-scale occupancy

Nichols et al 2008[83] developed an approach to estimate the probability of occupancy when multiple detection devices are used. They also consider the use of two parameters for the occupancy model. One parameter, denoted by ψ, accounts for the probability of a patch to be occupied. The other parameter, denoted by θ_j, accounts for the probability that a species is available for sampling during the j^{th} survey (i.e., the species is within the range of the detection devices), conditional on that the species is occupying the patch.

The model is parameterized as follows : ψ is the probability a patch is occupied, i.e., the large-scale probability of occupancy. The product $\psi\theta_j$ represents the probability of small-scale occupancy, i.e., the probability that the species is present at a patch and is exposed to the detection devices during the j^{th} survey. Similarly, the product $\psi\left(1-\theta_j\right)$ is the probability that the species is present at the large-scale but it is not available for sampling during the j^{th} survey. Furthermore, p_j^m is the probability for the species to be detected by the m^{th} detection device during the j^{th} survey given that the species was available to be sampled during the j^{th} survey.

The purpose of this parameterization is to discriminates between the device-specific probabilities of detection and the probability that the species is available for sampling. This allows users to make a fair comparison about the devices used to detect the presence of the species.

The sampling methodology is as follows: consider that n patches are surveyed by using l different detection devices, every patch is surveyed k occasions. The sampling observations at the i^{th} patch are arranged in a vector as follows $y_i = \{\{y_{i11}, ..., y_{i1l}\}, ..., \{y_{ik1}, ..., y_{ikl}\}\}$, where $y_{ijm} = 1$ if the species was detected by the m^{th} detection device at the i^{th} patch, during the j^{th} survey, and 0 otherwise.

As in the previously discussed model, the likelihood is proportional to the product of probabilities corresponding to the observations obtained at every patch. The parameters

are then estimated by maximizing the likelihood. An example on how to calculate the site-specific probabilities is presented as follows. Consider $y = \{\{0,1\},\{0,0\}\}$, these are the observations obtained from surveying a patch in two occasions using two different detection devices.

$$Pr(y) = \psi \cdot \left[\theta_1 \left(1 - p_1^1\right) p_1^2\right] \cdot \left[\theta_2 \left(1 - p_2^1\right) \left(1 - p_2^2\right) + (1 - \theta_2)\right]$$

First notice that the species was detected one time by one of detection devices in the first survey, hence it is assumed that the patch is occupied at the large scale. Moreover, notice that the species was detected at the first survey, therefore, there is certainty that the patch was occupied during the first survey. The first component of the expression above accounts for the probability the patch is occupied. The second component accounts for what is observed during the first survey, θ_1 indicates the species was available for sampling, $\left(1 - p_1^1\right)$ indicates that the species was not detected by the first device, and p_1^2 indicates that the species was detected by the second device. The last component accounts for what it is observed during the second survey. Since the species was not detected by any of the devices, there is uncertainty about the occupancy status at the small-scale during the second visit. This uncertainty is taken into account by considering the two possibilities: either the species was available but it was missed by the two detection devices $\theta_2 \left(1 - p_2^1\right) \left(1 - p_2^2\right)$, or the species was not available at all $(1 - \theta_2)$.

Non independent surveys

The most recent extension of MacKenzie's' model is that of Hines et al (2010)[45]. This model was developed to incorporate spatial dependency between surveys. Similar to Nichols et al (2008) occupancy was defined using two spatial scales. The model was inspired and illustrated using data from a tiger survey that was conducted in Karnataka, India.

Consider that n transects are selected from the study area. Every transect is divided into k segments. Every segment of every transect is surveyed to determine the presence/absence of the target species. The observations are collected in a vector as follows $y_i = \{y_{i1}, ..., y_{ik}\}$ where $y_{ij} = 1$ if the species was detected at the j^{th} segment of the i^{th} transect, and 0 otherwise.

The probability of occupancy is modeled using two scales: on the large scale is the probability for a transect to be occupied, this probability is denoted by ψ. On a smaller scale is the probability for a segment of a transect to be occupied. On this scale, it is assumed that

the occupancy for segments that were visited consecutively exhibit a Markovian dependence as follows: θ denotes the probability that the segment $t+1$ is occupied given that the segment t was empty. Similarly, θ' is the probability for the segment $t+1$ to be occupied given that the segment t was occupied as well. The order of the segments is given by the order in which the segments were surveyed. It is assume that there is a beginning for the transect and a consecutive order for the segments. So, the segment on position 1 is followed by segment in position 2, which is followed by the segment on position 3, etc. as depicted in figure 3.1. The probability of detection is defined as the probability of detecting the species given that the species is present on the transect and present on the segment.

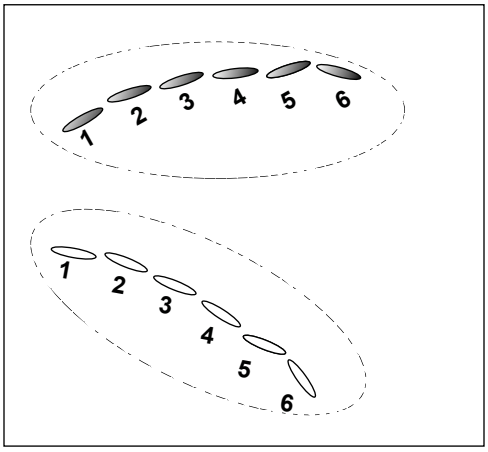


Figure 3.1: Sampling scheme considered for Hines et al (2010) model. Dotted lines defined the transects, circles defined the segments at every transect and numbers defined the segments of every transect. The sampling is assumed to be conducted in the order given by the numbers. Adapted from Hines et al (2010)[45].

The probability of the observations at every transect can be calculated using the parameters defined above, in a similar manner to the two previously discussed model (Nichols et al 2007 and Nichols et al 2008).

3.2 Dynamic occupancy models

One of the most important assumptions for the previously discussed models is that of closed population. In other words, the assumption that the probability of occupancy at an specific patch remains constant during the time of the study. However, if the study is conducted over a period of time in which changes to the population are likely to occur, it is sensible to waive this assumption. One way to do that is by modeling local extinction and colonization

events. MacKenzie et al (2003)[72] presents an extension of the zero inflated binomial model in which the closed population assumption is partially relaxed. This model allows estimation of not only the site occupancy probability, but also allows estimation of the colonization and extinction probabilities when the species is detected imperfectly.

The sampling scheme assumes that n patches are surveyed during T primary sampling periods, also called seasons. It is assumed that the population is closed during a season (i.e., no change in occupancy status), but it is open across seasons. Every patch is visited k times during every season. The observations at the i^{th} patch can be arranged into a vector as follows $y_i = \{\{y_{i11}, ..., y_{ik1}\}, ..., \{y_{i1T}, ..., y_{ikT}\}\}$ where $y_{ijt} = 1$ if the species was detected at the i^{th} patch, during the j^{th} surveyed during the season t.

The model is defined hierarchically by using the latent variable Z_{it} that describes the true status of the i^{th} patch at the season t; $Z_{it} = 1$ indicates that the species is present at the i^{th} patch during season t, $Z_{it} = 0$ indicates the opposite. ψ_{it} denotes the probability of occupancy at the i^{th} patch during season t, hence $Z_{it} \sim Bernoulli\,(\psi_{it})$. The number of detections at every patch during a specific season is denoted by y_{it}, for which its probability distribution is conditioned on the status of the patch. If the patch is occupied, $y_{it} \sim Binomial\,(\delta_{it}, k)$ where δ_{it} is the probability to detect the species at the i^{th} patch during the season t. If the patch is unoccupied, y_{it} is a degenerate *Binomial* distribution with zero probability. The changes in occupancy from one season to another are modeled by the parameters: ε_{it} and γ_{it}. ε_{it} is the the probability that an occupied patch at season t become unoccupied at season $t+1$, i.e., extinction probability. γ_{it} is the probability that an unoccupied patch at season t becomes occupied at season $t+1$ (i.e., colonization probability). Figure 3.2 displays the possible transitions from one season to another and their corresponding probabilities.

MacKenzie et al (2003) presents a likelihood approach to estimate these parameters. For a Bayesian approach see Royle and Kery (2007)[99].

3.3 Community models

Community models in general can be classified into two categories: models to estimate the species richness and models to study interspecific interaction between species. Similarly, site occupancy models have been extended to study these two aspects of a community. A general description on how these extensions are modeled is presented as follows.

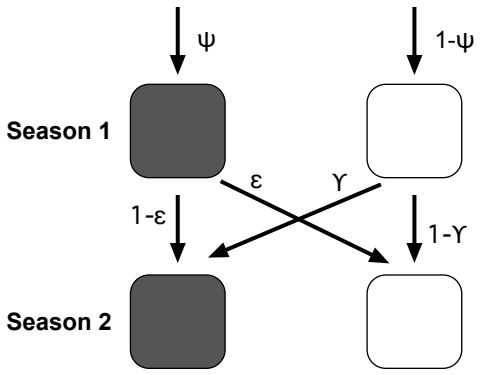


Figure 3.2: Dynamic site occupancy model. Black squares represent occupied patches, white squares represent empty patches.The arrows show the possible transitions and their corresponding probabilities. Reproduced from Royle and Dorazio (2008)[98].

Species richness models

The models to estimate species richness in a community can be seen as a multi-species occupancy model, for which the total number of species is the quantity of interest.

Dorazio and Royle (2005)[21] proposed a model to estimate species richness when the species are imperfectly detected. In their model, it is assumed that the community is closed during the time of the study (i.e., local extinctions or colonization by new species are unlikely). The sampling protocol is similar to that of the standard site occupancy model for a single species (zero inflated binomial)*;* n patches are selected from the study area, every patch is visited k times. The observer must record which species were detected at every patch in every survey. The observations can be arranged in a matrix as follows:

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{s1} & y_{s2} & & y_{sn} \\ \vdots & \vdots & & \vdots \\ y_{S1} & y_{S2} & & y_{Sn} \end{bmatrix}$$

where y_{ij} is the total number of times that the i^{th} species was detected at the j^{th} site; s is the number of species that were detected at least once, and S is a latent variable that denotes the true species richness.

Similar to the single species occupancy model, the true status of the i^{th} species at the j^{th} patch is denoted by $Z_{ij} = 0$ and modeled as a Bernoulli distribution with probability ψ_i. The number of detections is modeled as follows: if $Z_{ij} = 1$, then $Y_{ij} \sim Binomial\,(\delta_i, k)$;

on the other hand if $Z_{ij} = 0$, then $Y_{ij} = 0$ with probability one.

If it is assumed that the number of detected species equals the species richness of the community (i.e., s=S), the goal of the analysis is to estimate the parameters ψ_i and δ_i for $i = 1, ..., S$. The estimates for these parameters can be obtained by either doing a separate analysis for every species or by doing a joint analysis of the data. In the first case, the number of estimated parameters will increase as the number of observations increases, which can ultimately lead to inconsistent estimators. In the second case, it is possible to impose some restrictions on the parameters, and in doing so, to reduce its number. Dorazio and Royle (2005) assumed that the parameters ψ_i and δ_i were realizations of a bivariate normal distribution as follows:

$$\begin{pmatrix} logit\left(\psi_i\right) \\ logit\left(\delta_i\right) \end{pmatrix} \sim Normal\left(\begin{pmatrix} \beta \\ \alpha \end{pmatrix}, \Sigma\right), \tag{3.19}$$

where β and α are the logit-scale parameters for the mean probability of occupancy and the mean probability of detection respectively, and Σ is a measure of the heterogeneity in occurrence and detection probabilities among species.

However, it is unusual that all the species present in a community are detected during the time of the study ($s \leq S$), moreover, estimating the species richness is precisely the main goal of this kind of study. One approach to estimate S is to estimate the parameters α, β and Σ by maximizing the marginal likelihood of the observations. The estimated species richness is then given by:

$$\hat{S} = \frac{s}{1 - Pr\left({}^{y=0}/_{\hat{\alpha},\hat{\beta},\hat{\Sigma}}\right)}, \tag{3.20}$$

where $Pr\left({}^{y=0}/_{\hat{\alpha},\hat{\beta},\hat{\Sigma}}\right)$ is the probability of not detecting one species at any of the k surveys in any of the n patches. This is the approach used by Dorazio and Royle (2005). Another approach is that of the data augmentation proposed by Dorazio et al (2006)[23]. This is a Bayesian approach were information from unobserved species is included in the analysis, the advantage of this approach is that the model can be easily implemented using software such as WinBUGS[66].

Interspecific interaction models

The first extension of the site occupancy models to study interspecific interactions is that of MacKenzie et al (2004)[71]. The purpose of this model was to estimate the probability of

co-occurrence of two or more species, and doing so, to make inferences about the interspecific interactions between them. The assumptions and the sampling protocol for this model are the same as those for the single species model.

Assume for instance that there are 2 species of interest; n patches are surveyed at k occasions to determine the presence/absence of every species at every site. The observations at every patch during every survey are collected in a vector as follows: $y_{ij} = \{y_{ij1}, y_{ij2}\}$, where $y_{ijs} = 1$ if the s^{th} species was detected at the i^{th} patch during the j^{th} survey, $y_{ijs} = 0$ otherwise.

The true status of the i^{th} patch is described by the vector $Z_i = \{Z_{i1}, Z_{i2}\}$, where $Z_{ij} = 1$ if the j^{th} species is present at the i^{th} patch, $Z_{ij} = 0$ if the j^{th} species is absent from the i^{th} patch. The model is parameterized as follows:

$\psi_i^{11} = Pr\left(Z_i = (1,1)\right)$	Probability that the i^{th} patch is occupied by the two species
$\psi_i^{10} = Pr\left(Z_i = (1,0)\right)$	Probability that the i^{th} patch is occupied only by species 1
$\psi_i^{01} = Pr\left(Z_i = (0,1)\right)$	Probability that the i^{th} patch is occupied only by species 2
$\psi_i^{00} = Pr\left(Z_i = (0,0)\right)$	Probability that both species are absent from the i^{th} patch
$\delta_{ij}^{10} = Pr\left({}^{y_i=(1,0)}/_{z_i=(1,0)}\right)$	Probability of detecting species 1 in the i^{th} site during j^{th} survey given only species 1 is present
$\delta_{ij}^{01} = Pr\left({}^{y_i=(0,1)}/_{z_i=(0,1)}\right)$	Probability of detecting species 2 in the i^{th} site during j^{th} survey given only species 2 is present
$r_{ij}^{11} = Pr\left({}^{y_i=(1,1)}/_{z_i=(1,1)}\right)$	Probability of detecting both species at the i^{th} site during j^{th} survey given that both species are present
$r_{ij}^{10} = Pr\left({}^{y_i=(1,0)}/_{z_i=(1,1)}\right)$	Probability of detecting species 1 at the i^{th} site during j^{th} survey given that both species are present
$r_{ij}^{01} = Pr\left({}^{y_i=(0,1)}/_{z_i=(1,1)}\right)$	Probability of detecting species 2 at the i^{th} site during j^{th} survey given that both species are present

The probability of the observations at every patch in every survey can be calculated by using the parameters above. For instance, consider the following observations obtained from visiting a patch on three occasions: $y_{i1} = \{1,0\}, y_{i2} = \{0,0\}, y_{i3} = \{1,0\}$, its probability

can be written as:

$$Pr\left(y_i\right) = \left[\left(\psi_i^{10} - \psi_i^{11}\right)\left(\delta_{i1}^{10}\left(1-\delta_{i2}^{10}\right)\delta_{i3}^{10}\right)\right] + \left[\psi_i^{11} r_{i1}^{10} r_{i2}^{00} r_{i2}^{10}\right] \tag{3.21}$$

The first component of the expression above is the probability of the observations conditioned on that the species 1 is the only one present. The second component accounts for the probability that both species are present, but the second species was missed by the observer. The likelihood is then calculated by assuming that the detection histories collected at the n locations are independent. This model can be modified to make inferences for larger number of species with the caution that the number of parameters in the model increases exponentially with the number of species (MacKenzie et all 2004).

Waddle et al (2010)[115] developed a new parameterization for estimating co-occurrence of interacting species. In their model the occurrence of one species was assumed to depend on the occurrence of another species, but the occurrence of the latter species was assumed to be independent of the presence of the first species. The authors illustrated this assumption by using the predator-prey interaction, for which the occurrence of the prey is affected by the occurrence of the predator, but the occurrence of the predator is unaffected by the presence/absence of the prey. In their model it is assumed that each patch is visited multiple times and that the population is closed.

Denote by P the predator species and denoted by V the prey species (victim species). The occurrence of the predator and prey species is modeled as follows:

$$\begin{aligned} Z_i^P &\sim Bernoulli\left(\psi^P\right) \\ Z_i^V &\sim Bernoulli\left(Z_i^P \psi^{V/P} + \left(1 - Z_i^P\right)\psi^{V/\overline{P}}\right) \end{aligned}$$

where Z_i^P and Z_i^V denote the status of the predator and prey species at the i^{th} patch respectively, ψ^P is the probability of occurrence of the predator species, $\psi^{V/P}$ is the probability of occurrence of the prey species given that the predator is present, and $\psi^{V/\overline{P}}$ is the probability of occurrence of the prey species given that the predator is absent. The number of detections for the prey and predator species are modeled by a Binomial distribution with probability δ^P and $Z_i^P \delta^{V/P} + \left(1 - Z_i^P\right)\delta^{V/\overline{P}}$ respectively. The probability of the observations and the likelihood can be obtained by using the parameters above and applying similar arguments

to those of the previously discussed model.

3.4 Summary

The models previously discussed have two common characteristics: first, the assumption of closed population, and second, the use of replicate surveys. It is important to recall that the probability of occupancy is a parameter that depicts the status of a metapopulation in a snapshot, consequently, the assumption that the population is closed, at least during the time of the study, is essential for the inferences to be valid. Naturally, the closure assumption is strongly affected by the length of the time of the study. The longer the time, the more unlikely for the population to be closed. The second common characteristic is that of replicate surveys. The idea of using repeated visits to the same site is to be able to quantify the probability of detection, and doing that, to discriminate between a false negative and a true absence. If the population is closed, the larger the number of repeated visits to the same site, the better the estimate of the probability of detection will be.

The predicament is then whether a researcher should conduct a large number of repeated visits to obtain a good estimate of the probability of detection, even though the closure assumption may be violated; or to conduct a few number of surveys to obtain an imprecise estimate of the detection error while assuring that the population is closed. These considerations will be further discussed in the next chapter.

Chapter 4

Critical view of the multiple survey approach

As mentioned in the previous chapter, the zero inflated binomial (ZIB) model is the most popular approach to estimate site occupancy probability while accounting for detection error. This model requires that every patch in the study is visited on multiple occasions. These surveys need to be conducted in a sufficiently short period of time so that the population is closed, but also, the time between one survey and the next should be long enough to assume that every survey is independent from each other. One question that arises when considering these requirements is the number of required surveys. MacKenzie and Royle (2005)[74] proposed some guidelines for the design of a study. For instance, they suggested that if the species is common but difficult to detect, then every site should be visited 34 times (Table 3.1). On the other hand, if the species is rare and difficult to detect, then the number of suggested surveys is 16. Assuming there are no budget restrictions and the observer can make as many surveys as needed, then the question is to determine the time of the study so that the population can be assumed to be closed and the surveys can be assumed to be independent. However, if the assumption of closed population is not met by the data, what inferences can be made about the population? In practice, there are strong restrictions in sampling budget, hence only a few number of surveys can be conducted. Being this the case, is the information collected from a few surveys any good?

This chapter attempts to answer the previous questions by evaluating the zero inflated binomial model. The statistical properties of the ZIB are discussed in section 4.1. Sections

4.2 and 4.3 assess the robustness of the model to the violation of its assumptions.

4.1 Statistical properties

This section presents an assessment of the statistical properties of the ZIB model in circumstances where time, effort and cost limitations makes the number of visited sites small and the minimum number of repeated surveys are conducted ($k = 2$ visits per location). This assessment was carried out by using simulated data under two settings: one for which it was assumed that probabilities of occupancy and detection were constant; and another in which it was assumed that the probability of occupancy and detection depended on some habitat and other exogenous covariates.

For the first setting, all the possible combinations of three different values for the probability of occupancy and two values for the probability of detection were simulated ($\psi = \{0.80, 0.50, 0.30\}$, $\delta = \{0.30, 0.10\}$). On every combination the number of sites was set to be 30. For the second setting, the probability of occupancy and detection were assumed to depend on the covariates x_1, x_2, w_1 and w_2 according to the Logistic link defined as:

$$\psi_i = \frac{exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)}{1 + exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)}$$

and

$$\delta_{ij} = \frac{exp\left(\theta_0 + \theta_1 w_{1i} + \theta_2 w_{2i}\right)}{1 + exp\left(\theta_0 + \theta_1 w_{1i} + \theta_2 w_{2i}\right)}$$

where $x_{1i} \sim Normal\left(2, 1\right)$, $x_{2i} \sim Bernoulli\left(0.55\right)$, $w_{1ij} \sim Normal\left(1, 1.5\right)$ and $w_{2ij} \sim Bernoulli\left(0.65\right)$. The values of the parameters $\underline{\beta} = \{\beta_0, \beta_1, \beta_2\}$ and $\underline{\theta} = \{\theta_0, \theta_1, \theta_2\}$ were selected so that the mean probability of occupancy was 0.34 and the mean probability of detection was 0.15. The number of visited sites for this regression setting was 100. In both settings, 100 data sets were generated for each combination of parameters . The parameters for every data set were estimated using the Maximum Likelihood Estimator and Bayesian approaches. The specifics and results from each approach are described below.

4.1.1 Maximum Likelihood Estimator

The Maximum Likelihood Estimates (MLE) of the parameters were obtained assuming two different models: the ZIB model (equation 4.1) and the Naive model. The later model is

based on the assumption that the probability of detection is one, i.e., ignoring detection error. Equation 4.2 presents the likelihood for the Naive model for which the vector of observations at every site are transformed into a binary variable denoted by y_i^*, for which $y_i^* = 1$ if the species was detected at least once during the k surveys, and $y_i^* = 0$ if the species was not detected in any of the surveys.

$$L(\psi, \delta; y_1, ..., y_n) = \prod_{i=1}^{n} \psi \binom{k_i}{y_i} \delta^{y_i} (1-\delta)^{k_i - y_i} + (1-\psi) I(y_i = 0) \tag{4.1}$$

$$L(\psi, \delta; y_1^*, ..., y_n^*) = \prod_{i=1}^{n} \psi^{y_i^*} (1-\psi)^{1-y_i^*} \tag{4.2}$$

The likelihood in both cases was maximized using a quasi-Newton method algorithm implemented in R[113]. This algorithm not only provides the values of the parameters for which the likelihood is maximized, but also provides an estimate of the Fisher information that was subsequently used to estimate the standard error of the parameters. In addition, 100 bootstrap samples were generated to estimate the standard error and a confidence interval for the parameters of interest on each data set (See Appendix A for detailed algorithm).

Table 4.1 presents a summary of the simulation results for the first setting. In general the MLE of the ZIB tend to overestimate the probability of occupancy. In some cases, mean estimated values are more than twice the true occupancy. However, when the probability of occupancy is large (e.g. 0.80), the mean estimate of the ZIB is close to the true value even if the probability of detection is small. Notice that when the probability of detection is small (e.g. 0.10), no matter how large the probability of occupancy is, at least 50% of the time the estimated probability of occupancy is 1. On the other hand, the Naive model underestimates the true probability of occupancy in all cases: the larger the probability of detection is, the smaller the bias of the Naive estimates will be. It is also found that even though the estimates obtained by the Naive model are highly biased, they are more stable: the median and mean of the Naive estimates are close to each other and its standard errors are smaller than those of the ZIB.

Table 4.2 presents the true standard error for the parameters along with the mean estimated standard error obtained from the bootstrap sampling and the inverse of the Fisher Information Matrix (FIM). It is found that the estimated standard error obtained from the FIM is approximately 80% smaller than the true value for both models (ZIB and Naive). Moreover, when the detection is low, many of the FIM were singular, thus they could not be

Table 4.1: Summary of the maximum likelihood estimates for the site occupancy (ψ) for 100 simulated data sets with 30 sites, two surveys and constant probability of occupancy across the sites and the probability of detection is the same across the sites and surveys.

true values		ZIB				Naive			
δ	ψ	mean	median	se	mse	mean	median	se	mse
0.10	0.29	0.77	1.00	0.42	0.39	0.06	0.07	0.04	0.06
0.10	0.50	0.84	1.00	0.34	0.23	0.10	0.10	0.05	0.16
0.10	0.80	0.88	1.00	0.27	0.08	0.15	0.13	0.07	0.42
0.30	0.30	0.60	0.53	0.38	0.23	0.16	0.17	0.06	0.02
0.30	0.49	0.70	0.67	0.28	0.11	0.26	0.27	0.07	0.06
0.30	0.79	0.80	0.94	0.22	0.05	0.41	0.40	0.10	0.16

Table 4.2: Comparison of the true standard error and estimated standard error for ZIB and Naive models. True standard error (se), mean estimated standard error using bootstrap $\left(\overline{\widehat{se}}_B\right)$, mean estimated standard error using fisher information $\left(\overline{\widehat{se}}_F\right)$ and proportion of cases for which the estimated Fisher information could be used to estimate the standard error .

		ZIB				Naive			
δ	ψ	se	$\overline{\widehat{se}}_B$	$\overline{\widehat{se}}_F$	%	se	$\overline{\widehat{se}}_B$	$\overline{\widehat{se}}_F$	%
0.10	0.29	0.416	0.289	0.071	55%	0.039	0.037	0.002	100%
0.10	0.50	0.337	0.245	0.049	85%	0.052	0.051	0.003	100%
0.10	0.80	0.272	0.169	0.038	96%	0.065	0.062	0.004	100%
0.30	0.30	0.376	0.222	0.040	97%	0.064	0.064	0.004	100%
0.30	0.49	0.278	0.352	0.086	100%	0.069	0.094	0.006	100%
0.30	0.79	0.219	0.313	0.070	100%	0.100	0.123	0.008	100%

used to estimate the standard error. This indicates that the standard errors and confidence intervals based on the inverse of the FIM are inappropriate in small data sets. On the other hand, although the bootstrap samples also provide biased estimates of the standard errors, their bias is smaller than those obtained from FIM. For that reason, it is recommended to use bootstrap samples to estimate the confidence intervals and the standard errors of the parameters.

Table 4.3 presents the coverage of the bootstrap confidence intervals and some statistics related to their length. It is observed that the coverage of the confidence intervals for both the ZIB and the Naive models are below the nominal coverage of 90%. Notice that, for at least 50% of the cases, when the probability of detection and the probability of occupancy are both low, the confidence intervals for the ZIB cover the whole the range (0, 1). The confidence intervals for the Naive model are shorter than the confidence intervals for the

Table 4.3: Comparison of the coverage, mean and median length of the 90% confidence intervals for the constant probability of occupancy and constant probability of detection case.

		ZIB			Naive		
δ	ψ	%	mean length	median length	%	mean length	median length
0.10	0.29	65%	0.65	1.00	0%	0.11	0.13
0.10	0.50	55%	0.55	0.95	0%	0.16	0.17
0.10	0.80	34%	0.32	0.00	0%	0.20	0.20
0.30	0.30	62%	0.55	0.79	33%	0.20	0.20
0.30	0.49	97%	0.90	0.93	3%	0.29	0.30
0.30	0.79	92%	0.82	0.87	0%	0.39	0.37

ZIB and they have a poor coverage; in 4 out of the 6 simulated cases, none of the confidence intervals for the Naive model contained the true value of the parameter.

Table 4.4a presents the results of the simulation study under the covariates setting for the ZIB model. It is found that the median of the MLE is relatively close to the true value of the parameters, while the mean is highly biased. This indicates that the MLE for the ZIB are unstable. The FIM was singular for 92% of the cases and the estimated standard error obtained from the 8% left was extremely biased. Although the bootstrap sampling also provides biased estimates of the standard errors, it does a better job than the FIM. The bootstrap confidence intervals for the ZIB are very large, but coverage is relatively close to the nominal coverage. The mean estimate of the probability of occupancy is close to the true value (true:0.34, mean estimate: 0.37).

The Naive estimates for the regression setting are summarized in Table 4.4b. It is found that the median of the estimates for Naive model is close to the true value in all parameters but the intercept (β_0). The difference between the mean and the median of the estimates is not as large as the one observed for the ZIB model, which indicates that the estimates of the Naive model are more stable than those of the ZIB. Similarly, the standard errors of the Naive model are at least 12 times smaller than those for the ZIB model. Once more, the standard error of the parameters is better estimated by the bootstrap samples. The confidence intervals for the Naive model are at least 10 times shorter than those for the ZIB model, but their coverage is smaller than the nominal coverage, specially for the intercept (β_0). The Naive estimated mean occupancy is approximately 76% below the true value (true: 0.34, mean estimate: 0.08).

Table 4.4: Summary of the results of the estimated parameters for the occupancy for 100 simulated data sets, with n=100, two surveys and two covariates for occupancy.

(a) ZIB

							Confidence Intervals		
	true	mean	median	se	$\widehat{se}_B$	$\widehat{se}_F$ (n=8)	%	mean length	median length
β_0	0.500	46.93	1.039	108.3	413.65	47063491	94%	232.7	199.2
β_1	-1.00	-32.3	-1.46	64.89	391.06	48060	87%	164.6	122.3
β_2	1.200	29.34	2.004	65.35	203.49	46927638	85%	214.0	163.6

(b) Naive

							Confidence Intervals		
	true	mean	median	se	$\widehat{se}_B$	$\widehat{se}_F$ (n=85)	%	mean length	median length
β_0	0.50	-3.57	-1.83	5.45	9.28	2074191	11%	11.75	17.90
β_1	-1.00	-0.66	-0.68	0.39	24.62	0.19	78%	1.59	1.38
β_2	1.20	2.50	0.92	5.34	10.62	2074191	79%	12.26	18.12

In conclusion, the MLE of the ZIB provides unstable estimates of the parameters for small data sets. The standard error is better estimated by the bootstrap samples than by the FIM. The MLE of the Naive model are biased but stable.

4.1.2 Bayesian approach

The parameters were estimated under the Bayesian approach using non-informative priors. For the first case (constant probability of occupancy and detection) three sets of priors were considered: Uniform priors for the probabilities of occupancy and detection, Normal priors for the the log odds of the probability of occupancy and detection, and Uniform priors for the odds of the probability of occupancy and detection. For the second case (i.e. covariates setting), the same prior distribution was used for all the parameters (i.e., $\beta_i \sim Normal\,(0, 100)$ for i=0,1 and 2; and $\theta_i \sim Normal\,(0, 100)$ for i=0,1, and 2). The mean, standard deviation and 90% credible interval of the posterior distribution were estimated for every parameter using WinBUGS[66]. Figures 4.1, 4.2 and 4.3 present an example of the prior and posterior distributions that are obtained when using different sets of priors and various number of visited sites.

Figure 4.1 displays the prior probability distribution for the probability of occupancy (ψ), the log-odds $\left(log\left(\frac{\psi}{1-\psi}\right)\right)$, and the odds $\left(\frac{\psi}{1-\psi}\right)$ for the first set of priors ($\psi \sim Uniform\,(0, 1)$). Notice that when the ψ is uniformly distributed from 0 to 1, the cor-

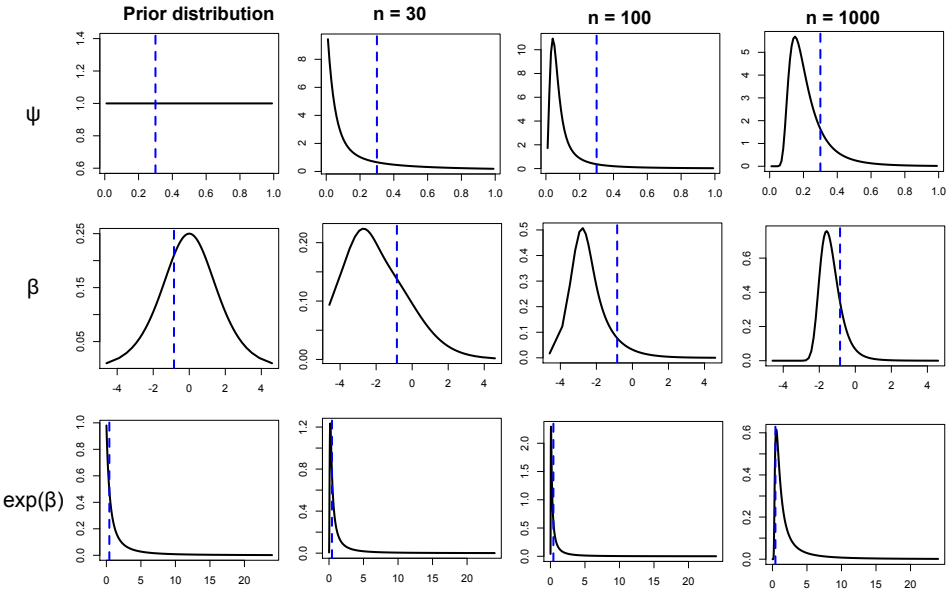


Figure 4.1: Uniform prior distribution for the probabilities. Probability distributions for the probability of occupancy ψ, the log odds $\beta = log\left(\frac{\psi}{1-\psi}\right)$ and the odds $exp(\beta) = \frac{\psi}{1-\psi}$. True value of the parameter is represented by the vertical dotted line.

responding prior distribution for the log-odds is unimodal and symmetric while the prior distribution of the odds is skewed to the right. It is observed that as the number of visited sites increases, the center of the posterior probability distribution gets closer to the true value of the parameter. However, even with 1000 visited sites, the densities of the posterior distribution for the probability of occupancy and the log odds are not yet centered at the true value.

Figure 4.2 presents the prior and posterior distributions for the second set of priors (log -odds$\sim Normal(0, 100)$). In this case, the prior distribution of the probability of occupancy resembles the shape of a Beta probability distribution with shape parameters smaller than 1. The probability distribution for the odds is once again skewed to the right. Notice that when the number of visited sites is 30 or 100 the densities of the posterior distribution of the parameters (i.e, probability of occupancy, log-odd and odds) are concentrated around values that are far from the true values. When the number of visited sites is 1000, only the posterior distribution of the log-odds is centered around the true value while the largest concentration of the densities for the other two parameters is still far from the true values. Finally, figure 4.3 displays the prior and posterior distributions for the third set of priors (odds $\sim Uniform(0, 1000)$). In this case, the prior probability distribution for the

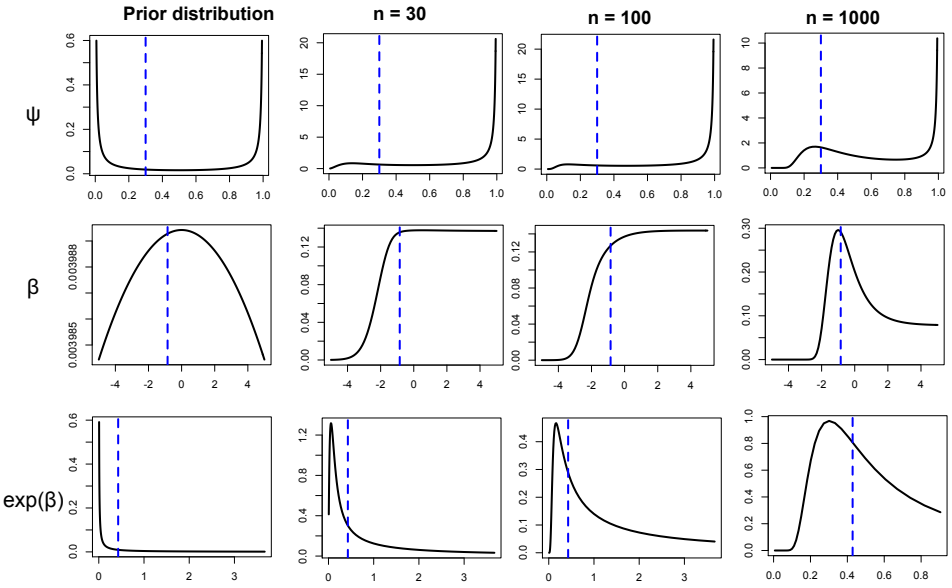


Figure 4.2: Normal prior distribution for the log-odds. Probability distributions for the probability of occupancy ψ, the log odds $\beta = log\left(\frac{\psi}{1-\psi}\right)$ and the odds $exp(\beta) = \frac{\psi}{1-\psi}$. True value of the parameter is represented by the vertical dotted line.

probability of occupancy is skewed to the left and concentrated at 1, and the prior probability distribution for the log-odds is also skewed to the left but concentrated around 4. The posterior distribution of the probability of occupancy preserves the same shape of the prior distribution in all the cases (i.e., when the number of visited sites is 30, 100 or 1000). This indicates that the information provided by the data does not have a big effect on the estimation procedure. A similar behavior is observed in the other two parameters (log-odds and odds): there are small differences between the posterior and the prior probability distributions, especially when the number of visited sites is 30 or 100.

Table 4.5 presents the summary of the simulation results for the constant probability case. It is found that the mean and median estimates obtained for the first set of priors are always unbiased, except when the probability of occupancy is large and the probability of detection is low (Table 4.5a). For the second set of priors, the estimates are highly biased in all the cases and the standard deviation is larger than that obtained when using Uniform priors for the probability (Table 4.5b). Lastly, for the third set of priors, a posterior mean of 0.99 is obtained for all the data sets in all the cases. This again is an indication of the poor effect of the data over the posterior distribution (Table 4.5c).

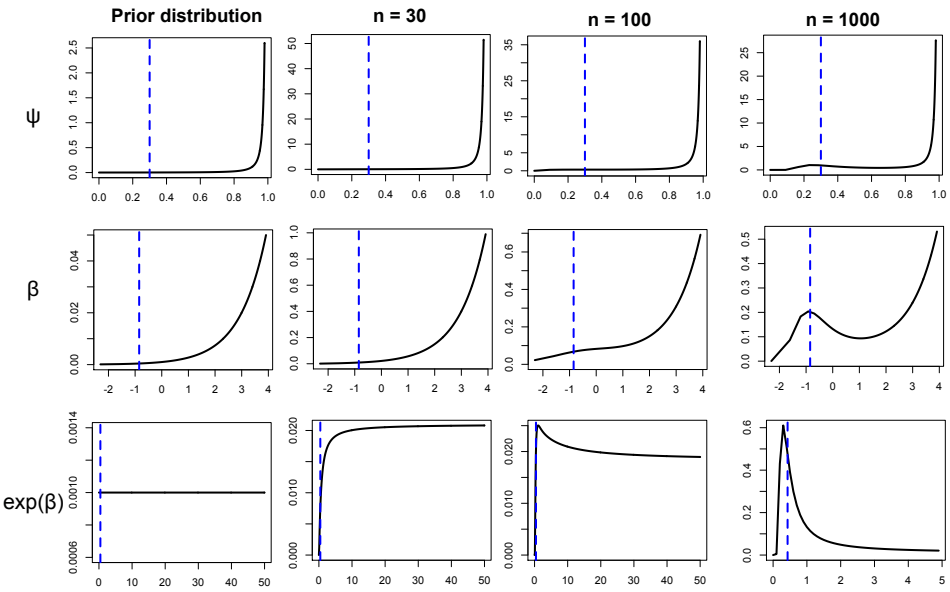


Figure 4.3: Uniform priors for the odds. Probability distributions for the probability of occupancy ψ, the log odds $\beta = log\left(\frac{\psi}{1-\psi}\right)$ and the odds $exp(\beta) = \frac{\psi}{1-\psi}$. True value of the parameter is represented by the vertical dotted line.

Table 4.6 presents a summary of the simulation results for the regression setting. It is observed that the parameters are overestimated, especially the intercept (β_0) for which the mean estimate is 20 times larger than the true value. The coverage of the credible intervals is smaller than the nominal coverage of 90%, and the estimated mean occupancy is larger than the true mean occupancy (true: 0.34, mean estimate: 0.84).

In summary, it is found that the accuracy and precision of the estimates of the probability of occupancy are strongly correlated to the prior distribution. For the constant probability setting, it was found that the estimates obtained from a single data set drastically differ depending on what prior is being used, even though all the priors are assumed to be non-informative. For the regression setting, it is found that the estimates are biased and the true probability of occupancy is largely overestimated.

4.1.3 Summary

The previous analysis showed that neither the Maximum Likelihood approach nor the Bayesian estimates provide reliable estimates of the probability of occupancy when detection or occupancy probability is small or when the number of sites and number of visits

Table 4.5: Bayesian estimates. Summary for 100 simulated data sets with 30 sites, two surveys, constant probability of occupancy and constant probability of detection. Mean, median and s.d. are summary statistics of the means of the posterior distribution.

(a) Uniform priors for the probabilities.

					Credible Intervals		
δ	ψ	mean	median	s.d.	%	mean length	median length
0.10	0.30	0.32	0.30	0.09	100%	0.79	0.79
0.10	0.50	0.40	0.37	0.11	100%	0.79	0.79
0.10	0.80	0.49	0.48	0.11	91%	0.74	0.74
0.30	0.30	0.39	0.43	0.12	100%	0.67	0.67
0.30	0.50	0.56	0.58	0.12	100%	0.63	0.63
0.30	0.80	0.72	0.73	0.08	100%	0.51	0.51

(b) Normal priors for the log-odds of the probabilities.

					Credible Intervals		
δ	ψ	mean	median	s.d.	%	mean length	median length
0.10	0.30	0.66	0.69	0.20	100%	0.93	0.93
0.10	0.50	0.77	0.79	0.13	100%	0.84	0.84
0.10	0.80	0.82	0.85	0.14	100%	0.72	0.72
0.30	0.30	0.67	0.74	0.25	74%	0.78	0.78
0.30	0.50	0.83	0.88	0.11	79%	0.61	0.61
0.30	0.80	0.93	0.94	0.04	91%	0.36	0.36

(c) Uniform priors for the odds of the probabilities.

					Credible Intervals		
δ	ψ	mean	median	s.d.	%	mean length	median length
0.10	0.30	0.99	0.99	0.00	0%	0.02	0.02
0.10	0.50	0.99	0.99	0.00	0%	0.02	0.02
0.10	0.80	0.99	0.99	0.00	0%	0.02	0.02
0.30	0.30	0.99	0.99	0.00	0%	0.02	0.02
0.30	0.50	0.99	0.99	0.00	0%	0.02	0.02
0.30	0.80	0.99	0.99	0.00	0%	0.02	0.02

Table 4.6: Bayesian estimates under the regression setting. Summary for 100 simulated data sets with 100 sites, two surveys and two covariates for the probability of occupancy. The prior distribution for all the parameters is Normal(0,100).

					Credible Intervals		
	true	mean	median	s.d.	%	mean	median
β_0	0.50	5.69	5.96	4.16	78%	20.77	21.31
β_1	-1.00	0.82	0.45	4.26	85%	18.27	21.02
β_2	1.20	5.01	4.63	3.57	80%	21.89	23.09

per site is small. The MLE of the ZIB have large biases, are numerically unstable and the corresponding confidence intervals have smaller than nominal coverage, while the Bayesian estimates in some cases are extremely biased and their credible intervals provide very poor coverage. Chapter 5 presents an alternative method of estimation, based on penalized likelihood. This method is numerically stable, the estimators have smaller mean square error than the MLE and associated confidence intervals have close to nominal coverage.

4.2 Closure assumption

The closure assumption is essential for site occupancy studies since it guarantees that the status of the metapopulation at the time of the study is accurately described by the collected data. There are many circumstances in which the assumption of a closed population is not met by the data. Let's consider, for instance, the case in which the target species randomly moved in and out of a sampling unit. In that case, according to MacKenzie (2005)[68], the estimate obtained from the ZIB model is an unbiased estimate of the proportion of used sites rather than the proportion of occupied sites.

Another circumstance is that in which some of the locations within the study area undergo extinction and colonization events during the time of the study. Rota et al. (2009)[96] proposed a modeling procedure that allowed testing for violations of closure under this circumstances. The procedure consisted of using a Likelihood Ratio Test to evaluate the relative support between a closed model, in which it is assumed that the population is closed, and an open model, in which the extinction and colonization probabilities are estimated by using the model introduced by MacKenzie et al. (2003)[72] (section 3.2). The procedure was applied by the authors to two avian point-count data sets collected in Montana and New Hampshire (USA). The first data set contained the information collected at 165 sites that were visited twice with on average two weeks between visits. These two visits served as primary sampling periods for the open model. The secondary sampling periods were defined by dividing each 10 minutes survey into four 2.5 minutes sampling intervals. The second data set contained information for 184 sites that were visited three times (6-8 days between surveys) using 10 minute surveys. Similar to the first data set, the days were used as primary sampling periods, and the 10 minute surveys were divided into three equally long intervals of time that served as secondary sampling periods. The hypothesis that colonization and extinction events occurred during the time of the study was better supported

for 71% and 100% of the species on each data set respectively. It was also found that for those species for which the open model was better supported, the estimated probability of occupancy obtained for the closed model was larger than that obtained for the open model. This may indicate that when the assumption of closed population is not met by the data the standard occupancy model tends to overestimate the probability of occupancy.

A simulation study conducted by Bayne et al.(2010)[6] corroborated this tendency. They used a spatial simulation to generate data that would occur in a multiple visit survey if birds move within sites between repeated visits. The data was then analyzed to estimate the density of birds for different levels of bird density, territory size and number of surveys. It was found that occupancy estimates of density overestimated the size of the population for large (5 ha) and intermediate territories (3 ha). On the contrary, it underestimated the population size when birds were abundant but had territories smaller that the point count area of 3 ha. The bias was highest when birds had territories larger that the point count area.

In summary, the study conducted by Rota et al. (2010) demonstrated that the assumption of closed population is not tenable for a large proportion of species even for short intervals of time between surveys. Bayne et al. (2010) showed that bias due to violations of the closure assumptions can be substantial and that its strength and direction largely depends upon the size of the target species' territory, which is very difficult to determine with certainty. In conclusion, both studies showed the need to develop models that, while accounting for the detection error, allow estimation of site occupancy without relying on the assumption of a closed population.

Chapter 6 presents an approach for estimating site occupancy probability in the presence of detection error that requires only a single survey and does not require the assumption of population closure. Therefore, this single survey approach facilitates analysis of data sets for which the assumption of closure is not met.

4.3 Independent surveys

It is still unclear what are the consequences of the violation of the assumption that the repeated visits are independent from each other. However, it is suspected that this assumption may be problematic particularly if the surveys are conducted over short periods of time[96]. Using sample data for three species of birds, Riddle et al. (2010)[93] found strong evidence

against independence between repeated surveys (in their study the repeated surveys were the result of dividing a 10-minutes surveys into 4 sampling intervals of 2.5 minutes), and more importantly, that ignoring the dependency between subsequent surveys could lead to bias in occupancy.

There are many scenarios in which independence between the observations from a site cannot be assumed. Sites may be correlated as a result of mechanisms such as dispersal or the influence of unobserved environmental variables [58, 117]. However, to date, the only modeling approach in which the assumption of the independence is relaxed is that of Hines et al. (2010) (section 3.1). For that reason, I decided to use the Hines model to illustrate the biases that can be obtained when the dependence between surveys is ignored.

According to Hines model, observations for the site occupancy study are collected from independent transects, where each transect is constituted by k segments. The probability of occupancy at every segment is conditioned on whether the species is present or absent at the corresponding transect, and on the presence/absence of the species at the precedent segment (Figure4.4). Hines et al. (2010) conducted a simulation study to evaluate the performance of their model. They considered the case of 200 transects, each transect consisted on 10 segments, and the probability of occupancy at the transect level was set to be 0.75. The probability for a segment to be occupied was set to 0.10 if the precedent segment was empty and 0.50 if it was occupied. The probability of detecting the species at every occupied segment was set to 0.80. The estimates obtained for a 1000 data sets, generated under the Markovian model, revealed that the estimated site occupancy obtained from the standard model (assuming independence) was approximately 30% smaller than the true value, while the estimates obtained from the true model (Markovian dependence) were unbiased although unstable (for 15% of the data sets the optimization algorithm was unable to maximize the likelihood).

Unfortunately, the simulation conducted by Hines et al.(2010) was restricted to a single sample size and using relatively large values for the probability of occupancy and detection. With the purpose of getting a more comprehensive evaluation of the model, I conducted a simulation study using different number of transects (10, 100, 150 and 200), two values for the number of segments on each transect (k=5 and k=10), three different levels of probability of occupancy (0.30, 0.50, 0.75) and a lower level of the probability of detection (0.30). The probabilities of occupancy at every segment were the same used by Hines et al.(2010) (i.e., $\theta = 0.10$ and $\theta' = 0.50$). For every combination of parameters, 100 data

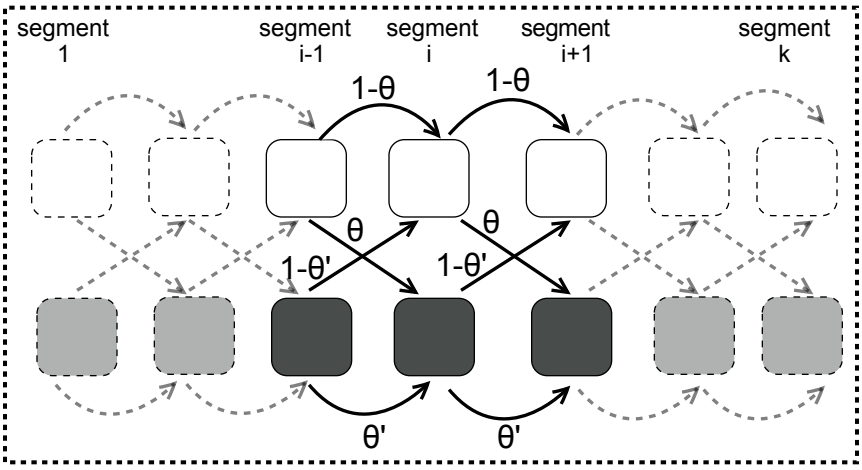


Figure 4.4: Markovian dependency on the observations from an occupied transect. White squares represent an empty segment, black squares represent an occupied segment. The probability for the i^{th} segment to be occupied depends on whether the previous segment was occupied or not.

sets were generated. The parameters were estimated using the likelihood for the true model (Markovian dependency) and the standard occupancy model (ZIB model). Figure 4.5 shows the mean percentage bias for the probability of occupancy. The results obtained from these simulations resemble the results obtained by Hines et al. (2010); for a large number of transects, a large number of segments and a large probability of occupancy, the Markovian estimates are unbiased while the estimates of the ZIB are in average 31% smaller than the true value. It was also found that if the probability of occupancy is high (0.75) or medium (0.50), the ZIB estimates are negatively biased.. In addition, the largest positive bias for the ZIB model was obtained for a low probability of occupancy and a small number of segments. For a large probability of occupancy, the estimates for the Markovian model are unbiased in all the cases. If the probability of occupancy is medium or low, then the Markovian estimates are biased for small number of segments.

In summary, the estimates obtained for the ZIB when the surveys were not independent were biased. The strength and direction of the bias depends on the level of the probability of occupancy and the sampling effort (number of transect/sites). However, as it was mentioned before, the model from Hines et al. (2010) explores only one of the several circumstances in which the independence assumption is not met. Another type of dependence is that of cluster sampling in which neighboring sites are correlated. Chapter 7 introduces a model that was developed to model the correlation between sites on that situation. This model allows estimation of site occupancy at the site level using information collected in a single

survey.

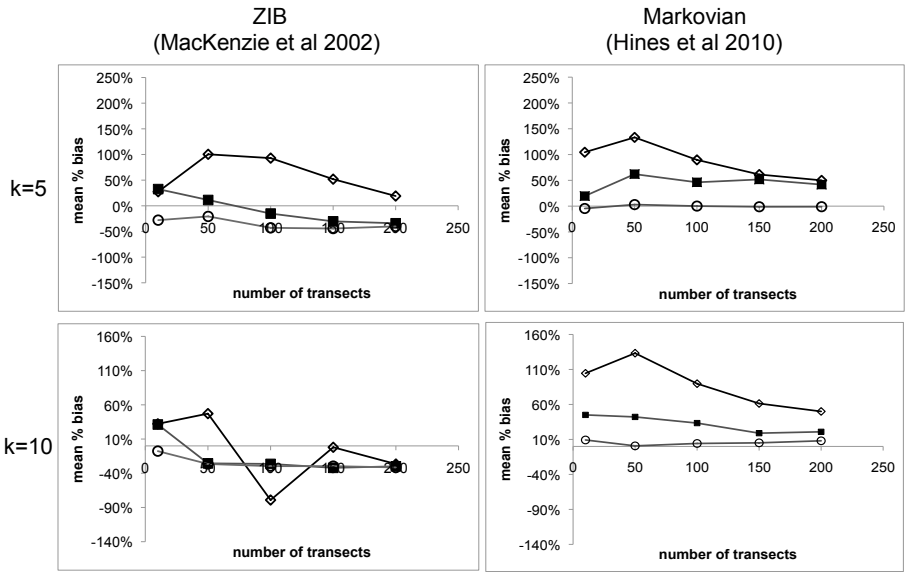


Figure 4.5: Mean percentage bias for the ZIB and the Markovian model. Diamonds show bias for low probability of occupancy (0.30), filled squares show bias for medium probability of occupancy (0.50) and circles shows bias for the low probability of occupancy (0.75).

Chapter 5

Penalized Likelihood: a way to improve MLE[1]

The simulation study discussed in section 4.1 revealed that the estimates of the site occupancy obtained from the ZIB model were unstable, in particular when the number of sites and surveys is small. It was also found that although the Naive estimates are biased, their likelihood is well behaved and provides stable estimates of the parameters. This chapter presents an alternative method of estimation that combines the correctness of the MLE with the stability of the Naive by means of Penalized likelihood. The estimation procedure is described in section 5.1. Section 5.2 presents the results of the simulation study conducted to evaluate the performance of the proposed method. The application of the method is then illustrated in section 5.3 for two site occupancy studies: one for the Blue Ridge Two Lined Salamander and the other for the Black-capped Chickadee.

5.1 Statistical model and estimation procedure

Consider a standard site occupancy study in which n sites are visited k times to determine the presence/absence of the target species. The surveys are assumed to be independent of each other and the sites are assumed to be independent of each other. ψ_i denotes the probability of occupancy for the i^{th} site and δ_{ij} denotes the probability of detecting the target species at the site during the j^{th} survey, given that the species is present. The

[1] A version of this chapter has been published. Moreno M and Lele S R. Ecology 2010, 91: 341-346.

observations at every site during every visit are denoted by y_{ij} for which $y_{ij} = 1$ indicates that the species was detected at the i^{th} site during the j^{th} survey and $y_{ij} = 0$ indicates that the species was not detected. The likelihood function for the ZIB model in the general case is:

$$L(\psi_i, \delta_{ij}) = \prod_{i=1}^{n} \left(\psi_i \left(\prod_{j=1}^{k} (\delta_{ij})^{y_{ij}} (1 - \delta_{ij})^{1-y_{ij}} \right) + (1 - \psi_i) I(y_{i.} = 0) \right) \tag{5.1}$$

where $y_{i.} = \sum_{j=1}^{k} y_{ij}$ and $I(\bullet)$ is an indicator function that is equal to 1 if its argument is true and 0 otherwise. If habitat or sampling covariates are available, these can be incorporated into the likelihood by using the logistic link. For instance, assuming that $\underline{x}_i = \{x_{i1}, ..., x_{ip}\}$ and $\underline{w}_{ij} = \{w_{ij1}, ..., w_{ijm}\}$ are the values of the covariates at the i^{th} site during the j^{th} survey, the probability of occupancy and detection can then be written as follows:

$$\psi_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}, \tag{5.2}$$

$$\delta_{ij} = \frac{exp(\theta_0 + \theta_1 w_{ij1} + ... + \theta_m w_{ijm})}{1 + exp(\theta_0 + \theta_1 w_{ij1} + ... + \theta_m w_{ijm})}, \tag{5.3}$$

if the probability of occupancy and probability of detection are constant, that is $\psi_i = \psi$ and $\delta_{ij} = \delta$. The likelihood function is then reduced to:

$$L(\psi, \delta) = \prod_{i=1}^{n} \left(\psi \left(\binom{k}{y_{i.}} \delta^{y_{i.}} (1 - \delta)^{k - y_{i.}} \right) + (1 - \psi) I(y_{i.} = 0) \right) \tag{5.4}$$

where $\psi = \frac{exp(\beta_0)}{1+exp(\beta_0)}$ and $\delta = \frac{exp(\theta_0)}{1+exp(\theta_0)}$. The MLE of the parameters $\underline{\beta}$ and $\underline{\theta}$ is obtained by maximizing the likelihood function (equation 5.1). The results in tables 5.1 and 5.2 illustrate that the estimators based on maximizing this likelihood function can be quite unstable, with large biases, large standard errors and incorrect coverage for the confidence intervals.

An alternative estimation method for the parameters related to the occupancy model $(\underline{\beta})$ is to ignore the detection error, hence to obtain the Naive estimate. The likelihood function in this case is simply $\prod_{i=1}^{n} (\psi_i)^{y_i^*} (1 - \psi_i)^{1-y_i^*}$ where $y_i^* = \max_j (y_{ij})$. It has been shown that this estimator, for small number of visits can have large negative bias, but is extremely stable with small standard errors (section 4.1). Nonetheless, if the number of surveys and/or the probability of detection is large, the Naive estimator can be unbiased (Figure 5.1).

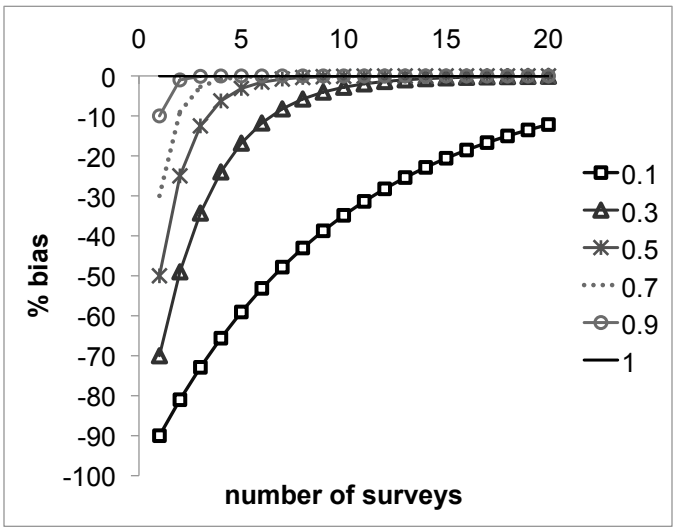


Figure 5.1: Percentage bias of Naive estimate for different values of the probability of detection (0.1, 0.3, 0.5, 0.7, 0.9, 1). Notice that as the probability of detection and the number of surveys increases the bias decreases.

In an attempt to combine the theoretical correctness of the MLE with the stability of the Naive estimator we proposed a Penalized likelihood approach. In general, the penalized likelihood is used in situations where the likelihood function is relatively flat or is too bumpy. Such a likelihood function makes the task of numerical maximization unstable. By imposing restrictions through the penalty function, the likelihood function is sharpened and the optimization problem is stabilized. For example, in non-parametric density estimation, imposing restrictions on the smoothness of the density estimate stabilizes the numerical optimization problem [33]. A review of the literature suggests that there is no unique specification of the penalty term in all instances nor is there a theoretical basis to choose one. The penalty term is usually chosen based on heuristic arguments. The main restriction on the penalty term is that as the sample size increases, it should converge to zero so that the asymptotic properties of the MLE are maintained. Consequently, we proposed the following penalized log-likelihood:

$$
\begin{aligned}
l_p(\psi,\delta) &= \left[\sum_{i=1}^{n} ln\left(\psi\binom{k}{y_{i.}}\delta^{y_{i.}}(1-\delta)^{k-y_{i.}} + (1-\psi)\,I(y_{i.}=0)\right)\right] \\
&\quad - \left[\lambda(k,\delta_0,n)\cdot f\left(\psi,\hat{\psi}_{naive}\right)\right]
\end{aligned}
\tag{5.5}
$$

where $\lambda\left(k, \overline{\hat{\delta}}_0, n\right)$ and $f\left(\psi, \hat{\psi}_{naive}\right)$ are

$$\lambda\left(k, \overline{\hat{\delta}}_0, n\right) = \sqrt{\sum_{i=0}^{m} \hat{var}\left(\hat{\delta}_i\right)} \cdot \left(1 - \left(1 - \overline{\hat{\delta}}_0\right)^k\right) \cdot \left(1 - \overline{\hat{\psi}}_{naive}\right) \tag{5.6}$$

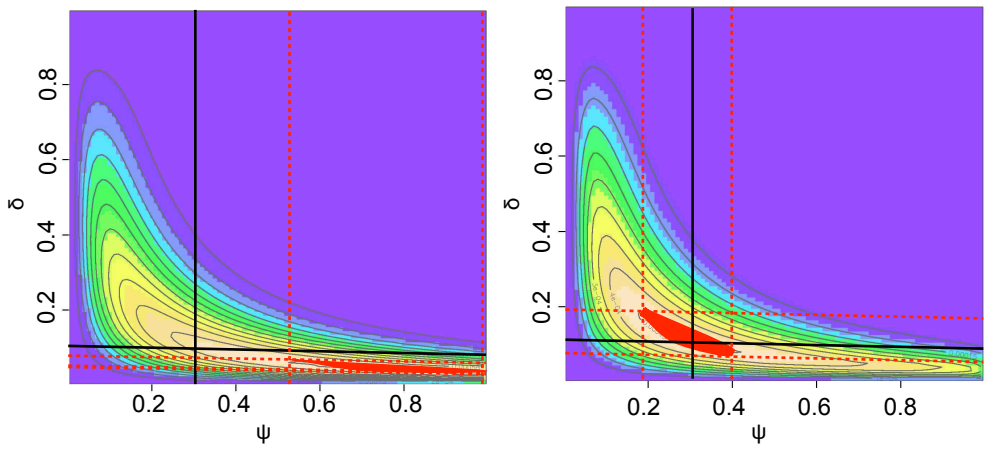
$$f\left(\psi, \widehat{\psi}_{naive}\right) = \sum_{i=0}^{p} \mid \beta_i - \tilde{\beta}_i \mid \tag{5.7}$$

The first term in equation 5.5 corresponds to the likelihood of the ZIB and the second term corresponds to the *penalty function*. The penalty function, in density estimation, penalizes or down-weights those values that are too far from the presumed properties of the density. For occupancy studies, the Naive estimator provides a ballpark estimate of where the true parameter might be. Thus, in our case, we penalize or down-weight those values of the likelihood that are "too far" from the Naive estimator. The distance from the Naive estimator is accounted by the term: $\sum_{i=0}^{p} \mid \beta_i - \hat{\beta}_{i,naive} \mid$. It is also obvious that if the detection probability is low, we should not rely too much on the Naive estimator unless the number of surveys is large. To reflect this, we use the initial estimate of the mean detection probability obtained using the MLE and weight this distance by: $\left(1 - \left(1 - \hat{\delta}_M\right)^k\right)$. Further, we know that if the occupancy is high, we should not penalize the MLE too much because in this case the likelihood function is well behaved. Thus, we multiply by the term $\left(1 - \overline{\hat{\psi}}_{naive}\right)$, which gives us a rough idea of how large the average probability of occupancy is. We should also take into account the fact that if the mean detection parameters are well estimated, the likelihood function is well behaved and hence we should not penalize it too much. To reflect this, we multiply by $\sqrt{\sum_{i=0}^{m} \hat{var}\left(\hat{\theta}_i\right)}$. This term also has an added benefit; as k or n increase, it converges to zero and the penalized likelihood function approaches the likelihood function (as it should).

One can also justify the penalty term from a Bayesian perspective. In this case, the penalty function can be seen as an approach to incorporate prior information (e.g. the density function is twice differentiable or has single mode) into the likelihood [32]. In our case, the prior information will come from the Naive estimate. Let's suppose we put independent, double exponential priors (with common scale parameter) on the occupancy parameters:

$$\pi\left(\beta_i\right) = \frac{\lambda\left(\cdot\right)}{2} exp\left(-\lambda\left(\cdot\right) \mid \beta_i - \tilde{\beta}_i \mid\right) \; for \; i = 0, ..., p$$

Where $\tilde{\beta}_i$ and $\lambda\left(\cdot\right)$ are the location and the scale parameter respectively. Now, let's



(a) Simple likelihood (b) Penalized likelihood

Figure 5.2: Comparison of the likelihood of the ZIB and the Penalized likelihood for one set of data with n=30 and k=2. Vertical black line indicates the true value of the probability of occupancy, horizontal black line indicates the true value of the probability of detection. Notice that the range of values for which the penalized likelihood is maximized is small and contains the true value of the parameters, unlike the ZIB likelihood.

suppose we take the location parameters equal to the naïve estimates and the common scale parameter equal to:

$$\lambda\left(k,\delta_0,n\right)=\sqrt{\sum_{i=0}^{m}\hat{var}\left(\hat{\delta}_i\right)}\left(1-\left(1-\overline{\hat{\delta}}_0\right)^k\right)\left(1-\overline{\hat{\psi}}_{naive}\right)$$

It is obvious that the mode of the posterior distribution will be obtained by maximizing the log-posterior density:

$$\begin{aligned} l_p &= log\left(\prod_{i=1}^{n}\psi_i\left(\prod_{j=1}^{k}\delta_{ij}^{y_{ij}}\left(1-\delta_{ij}\right)^{1-y_{ij}}\right)+\left(1-\psi_i\right)I\left(y_{i\cdot}=0\right)\right) \\ &- \left(\lambda\left(\bullet\right)\left(\sum_{i=0}^{p}\mid\beta_i-\tilde{\beta}_i\mid\right)\right) \end{aligned}$$

Notice that this is identical to the penalized likelihood function (equation5.5). It is expected that the effect of the prior distribution is reduced as the number of data points increases. Similarly, the effect of the penalty function is reduced as the number of data points increases.

In summary, by using the penalized likelihood we shrink the MLE towards the Naive estimator (as depicted in Figure 5.2). The shrinkage factor, $\lambda\left(k,\delta_0,n\right)$, is determined by the

Algorithm 5.1 Maximum Penalized likelihood Estimation Procedure

1. Obtain the MLE for the detection parameters: $\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_m$, their variances, $\hat{var}\left(\hat{\theta}_0\right), \hat{var}\left(\hat{\theta}_1\right), ..., \hat{var}\left(\hat{\theta}_m\right)$, and the mean probability of detection $\overline{\hat{\delta}}_0 = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\hat{\delta}_{ij}$.

2. Obtain the Naive estimator of the occupancy parameters $\hat{\tilde{\beta}}_0, \hat{\tilde{\beta}}_1, ..., \hat{\tilde{\beta}}_p$ and the mean estimated occupancy $\overline{\hat{\psi}}_{naive} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i$ where $\hat{\psi}_i = \frac{exp\left(\hat{\tilde{\beta}}_0+\hat{\tilde{\beta}}_1 x_{1i}+...+\hat{\tilde{\beta}}_p x_{pi}\right)}{1+exp\left(\hat{\tilde{\beta}}_0+\hat{\tilde{\beta}}_1 x_{1i}+...+\hat{\tilde{\beta}}_p x_{pi}\right)}$

3. Maximize the penalized likelihood function using Eq. 5.5 where $\lambda\left(k, \overline{\hat{\delta}}_0, n\right)$ is 5.8 and $f\left(\psi, \hat{\psi}_{naive}\right)$ is 5.9.

$$\lambda\left(k, \overline{\hat{\delta}}_0, n\right) = \sqrt{\sum_{i=0}^{m} \hat{var}\left(\hat{\delta}_i\right)} \cdot \left(1 - \left(1 - \overline{\hat{\delta}}_0\right)^k\right) \cdot \left(1 - \overline{\hat{\psi}}_{naive}\right) \tag{5.8}$$

$$f\left(\psi, \widehat{\psi}_{naive}\right) = \sum_{i=0}^{p} \mid \beta_i - \tilde{\beta}_i \mid \tag{5.9}$$

number of visits, the number of sites and the initial estimates of the average detection and occupancy probabilities. As the number of sites or number of visits increase, the likelihood function is well behaved and hence the penalty function is forced to converge to zero. If the detection probability is large, the naïve estimator is a good estimator, hence we can rely on it and the shrinkage factor can be large. On the other hand, when the occupancy probability is large, the MLE usually is stable and the shrinkage factor small.

The algorithm to estimate the Maximum Penalized likelihood Estimator (MPLE) is summarized in the Algorithm 5.1.

5.2 Simulation analysis

A simulation study was conducted to compare the performance of the Maximum likelihood Estimator (MLE) and Maximum Penalized likelihood (MPLE) using the same cases considered in section 4.1. Some of the simulations results of section 4.1 are reproduced in here for easy of illustration.

Figure 5.3 displays the mean and median bias obtained for different values of probability of occupancy for the MLE, MPLE and Naive estimates. Notice that the bias obtained for the MLE is larger than that obtained for the MPLE, in particular for moderate values of

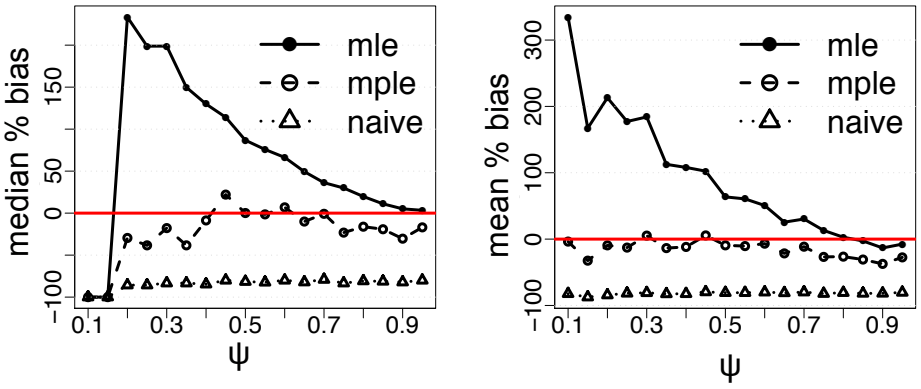


Figure 5.3: Median (left) and mean (right) percent bias for the MLE, MPLE and Naive for 100 data sets with 30 sites and 2 surveys per site. The probability of detection is fixed on 0.10. Notice that the MLE overestimates the probability of occupancy, while the MPLE tends to underestimate it, but it is closer to the true value than the MLE and Naive estimates.

the probability of occupancy $\psi \leq 0.70$. On the contrary, when the probability of occupancy is large $\psi > 0.80$, the MPLE tends to be more biased than the MLE. However it remains still close to the true value.

Table 5.1 presents the summary of the simulation results for the first case (constant probability of occupancy and detection). The mean and median of the MPLE are closer to the true values than MLE in every case except when the probability of occupancy is large. The standard errors of the MPLE are smaller than for the MLE in every case. The coverage of the bootstrap confidence intervals, based on MPLE, is close to the nominal coverage. The estimated standard error obtained from the bootstrap samples is more accurate for the MPLE than for the MLE, although it is still biased.

Table 5.2 presents the results of the simulation study under the regression setting where the probabilities depend on the covariates. In this case, there are two quantities of interest: the regression coefficients, which measure covariate effects; and the average occupancy. For the latter, the simulation results show that the MLE and the MPLE perform equally well, with the MLE slightly overestimating average occupancy and the MPLE slightly underestimating it. For the regression coefficients, the mean and the median of the MPLE are close to each other, indicating numerical stability. They are also closer to the true values of the parameters than the ML estimates. The standard errors for the MPLE are also substantially smaller than those of MLE. Finally, the lengths of the confidence intervals obtained

Table 5.1: Summary of simulation results for 100 simulated data sets with 30 sites, two surveys and constant probability of occupancy across the sites and the probability of detection is the same across the sites and surveys. Observe that MLE overestimates the occupancy whereas MPLE is nearly unbiased with smaller standard errors.

(a) MPLE

							Confidence Intervals		
δ	ψ	mean	median	se	$\overline{\widehat{se}}_B$	mse	%	mean length	median length
0.10	0.29	0.30	0.16	0.26	0.21	0.06	72%	0.50	0.61
0.10	0.50	0.46	0.46	0.26	0.20	0.07	86%	0.63	0.69
0.10	0.80	0.63	0.73	0.26	0.19	0.10	74%	0.56	0.60
0.30	0.30	0.45	0.41	0.27	0.19	0.10	75%	0.62	0.69
0.30	0.49	0.52	0.45	0.23	0.27	0.06	84%	0.63	0.70
0.30	0.79	0.71	0.73	0.20	0.27	0.04	87%	0.54	0.18

(b) MLE

							Confidence Intervals		
δ	ψ	mean	median	se	$\overline{\widehat{se}}_B$	mse	%	mean length	median length
0.10	0.29	0.77	1.00	0.42	0.289	0.39	65%	0.65	1.00
0.10	0.50	0.84	1.00	0.34	0.245	0.23	55%	0.55	0.95
0.10	0.80	0.88	1.00	0.27	0.169	0.08	34%	0.32	0.00
0.30	0.30	0.60	0.53	0.38	0.222	0.23	62%	0.55	0.79
0.30	0.49	0.70	0.67	0.28	0.352	0.11	97%	0.90	0.93
0.30	0.79	0.80	0.94	0.22	0.313	0.05	92%	0.82	0.87

Table 5.2: Summary of the results of the estimated parameters for the occupancy for 100 simulated data sets, with n=100, two surveys and two covariates for occupancy. Observe that MLE are unbiased and have large standard errors. Moreover, the MPLE based confidence intervals for the parameters are ten times shorter without sacrifice in coverage.

(a) MPLE

						Confidence Intervals		
	true	mean	median	se	$\widehat{se}_B$	%	mean length	median length
β_0	0.500	-0.37	0.148	7.748	6.77	0.770	17.42	19.53
β_1	-1.00	-1.56	-0.96	1.893	2.20	0.930	5.266	4.069
β_2	1.200	3.187	1.417	5.552	5.19	0.910	15.86	18.54
$\overline{\psi}$	0.340	0.314	0.277	0.172	0.201	0.900	0.453	0.445

(b) MLE

						Confidence Intervals		
	true	mean	median	se	$\widehat{se}_B$	%	mean length	median length
β_0	0.500	46.93	1.039	108.3	413.65	94%	232.7	199.2
β_1	-1.00	-32.3	-1.46	64.89	391.06	87%	164.6	122.3
β_2	1.200	29.34	2.004	65.35	203.49	85%	214.0	163.6
$\overline{\psi}$	0.340	0.373	0.334	0.169	0.210	96%	0.474	0.476

by the MPLE are at least 10 times shorter than the ones obtained by the ML method. More importantly, except for the intercept parameter, the actual coverage of the MPLE based confidence intervals is closer to the nominal coverage than for MLE based confidence intervals. This means that the effect of the covariates is better estimated by the MPLE than by the MLE. In summary, the simulation results show that: for those cases where MLE fails, the MPLE works extremely well; and at the same time when MLE does work well MPLE works equally well.

5.3 Example data analysis

We illustrate the MPLE method using two occupancy studies: one for Blue Ridge Two Lined Salamander and the other for Black-capped Chickadee.

The Blue-Ridge Two Lined Salamanders study was conducted in the Great Smoky Mountains National Park (USA) during 2001. The data are available as part of the software Presence[1]. The data consist of 39 sites visited once every two weeks for a total of five surveys. There are no covariates available for the occupancy and detection probability models. We used the simple model of constant probability of occupancy and constant probability

Table 5.3: Summary of the estimated probability of occupancy and its standard error for the Blue-Ridge two lined salamander data. MPLE have smaller standard errors and shorter confidence intervals as compared to the MLE. Also note that different combinations of two surveys leads to quite different point estimates making the closed population assumption suspect.

	MLE				MPLE			
surveys	$\hat{\psi}$	$\hat{se}$	90% CI		$\hat{\psi}$	$\hat{se}$	90% CI	
1,2,3,4,5	0.59	0.15	0.43	1.00	0.59	0.15	0.43	0.94
1,2	0.31	0.35	0.12	1.00	0.24	0.27	0.10	0.88
1,3	0.54	0.24	0.32	1.00	0.48	0.22	0.28	0.95
1,4	0.31	0.19	0.16	0.78	0.29	0.14	0.12	0.57
1,5	0.64	0.28	0.26	1.00	0.48	0.25	0.23	0.95
2,3	0.92	0.23	0.32	1.00	0.65	0.23	0.30	0.97
2,4	0.78	0.25	0.31	1.00	0.57	0.24	0.28	0.95
2,5	1.00	0.00	1.00	1.00	0.92	0.07	0.79	0.96
3,4	0.46	0.16	0.27	0.82	0.44	0.15	0.25	0.71
3,5	0.72	0.23	0.37	1.00	0.62	0.22	0.35	0.98
4,5	0.31	0.15	0.18	0.56	0.30	0.11	0.18	0.49

of detection. The goal of the analysis was to compare the estimated occupancy obtained by the MLE and MPLE and their confidence intervals and standard errors under various scenarios.

We first present the analysis using MLE and MPLE using all five visits. The results presented in table 4 show that the MLE and MPLE are quite similar, although standard errors and confidence intervals based on MPLE are somewhat shorter than for MLE. This is to be expected because, when the number of visits is large, the penalty function is small and MPLE and MLE become similar. Next we considered the possibility of only two visits. There are 10 such combinations possible. In table 5.3, we present the estimated occupancy obtained by using the ML and MPL estimators for every possible pair of visits. Notice that in all cases the standard errors of the MPL estimator are smaller than the ones obtained by the ML estimator, except for the combination of the second and fifth surveys for which the ML estimated occupancy is one for all bootstrap samples. The bootstrap confidence intervals based on the MPLE are shorter than those based on the MLE estimators. MPLE, thus provides a more precise representation of the occupancy than the MLE. It is also interesting to note that the inferences from different pairs of surveys vary substantially from each other with occupancy estimates ranging from 0.24 to 0.92. This suggests that perhaps the validity assumption of a closed population during the time of the study is questionable.

The second example corresponds to an occupancy study that was conducted on lands

Table 5.4: Estimated parameters, 90% confidence intervals and standard errors for the occupancy and detection model of the Blackcapped Chickadee. The standard errors and confidence intervals were estimated using 200 bootstrap samples. Notice that the confidence intervals for the occupancy model parameters obtained by the MLE are about 60% shorter than the confidence intervals obtained by the MLE.

	MLE		MPLE	
	estimate (90% CI)	se	estimate (90% CI)	se
Intercept	8.78 (2.52, 9.68)	2.12	3.22 (1.65, 4.24)	0.789
Year 2	-9.88 (-10.7, -3.62)	2.15	-4.32 (-5.58, -2.66)	0.858
Year 3	-10.31 (-11.5, 4.27)	2.11	-4.74 (-5.99, -3.24)	1.397
Intercept	-1.50 (-1.76, -1.22)	0.16	-1.47 (-1.72, -1.19)	0.164
$\overline{\psi}$	0.47 (0.44, 0.52)	0.02	0.46 (0.41, 0.51)	0.028
$\overline{\delta}$	0.18 (0.14, 0.22)	0.02	0.18 (0.15, 0.23)	0.025

managed by Millar Western Forest Products in western Alberta (Canada) from 2000 to 2002 (E. Bayne, unpublished manuscript). For illustrative purposes, only the data for the Black-capped Chickadee (BCHH) are used. The data were collected over a period of three years. Each year, 40 sites were visited once every week starting on May 15 and ending on July 28. Two different observers, randomly assigned to the sites, were used. The purpose of the analysis was to determine whether or not there was a trend for the occupancy of the BCCH over the three years of the study. The covariates for the occupancy model correspond to the year of the survey, the reference year (year 1) being year 2000. For the detection probability model covariates such as the observer, the Julian date and the time of the survey were evaluated. However, because they turned out to be non significant, a constant detection model was then fitted.

Table 5.4 presents the MLE and MPLE of the parameters for this model. The standard errors, as well as the 90% confidence intervals, were calculated using 200 bootstrap samples. It was found that the standard errors for the occupancy model provided by the MLE were substantially larger than the ones obtained by the MPLE, and that the MPLE's confidence intervals were shorter than the ones obtained by the MLE. On the other hand, the standard errors and confidence intervals for the detection model were almost the same for both the MLE and the MPLE. Using the MPLE estimates it can be concluded that there was a

decreasing trend for the occupancy of the BCCH. During the first year the estimated mean occupancy was about 0.9616, dropping for the second year to 0.2495 and decreasing further for the last year to 0.1790. These data are part of a large ecological study of how forest density affects occupancy. Nearly 50% of the trees were removed from the area between year 1 and year 2. The drop in occupancy is the likely to be an outcome of such a change in the forest density.

5.4 Summary

The penalized likelihood estimators have better statistical properties with smaller mean squared error. They also have a bootstrap confidence interval coverage closer to the nominal coverage than the ML estimator. Furthermore, the estimates for the occupancy model obtained by the MPLE are somewhat conservative while the estimates obtained by the MLE are optimistic (Figure 5.2). From the perspective of a monitoring program, it is prefered to have a conservative estimate rather than an optimistic one, the latter can prevent managers to take action to protect a species that can be at risk of extinction.

From a practical perspective, the use of penalized likelihood estimators can lead to a substantial reduction in the required number of surveys and sites. This ultimately can lead to a substantial reduction in the cost of implementing such surveys.

Chapter 6

The single survey approach[1]

A common requirement for current methods to estimate occupancy probability when detection probability is less than 1, is that sites must be sampled repeatedly. The premise behind this requirement is that only with repeated visits the discrimination between a true absence and a false negative will be possible. For example, Bolker (2008, page 333)[9] claims: "there is no way to identify catchability—the probability that you will observe an individual—from a single observational sample; you simply don't have the information to estimate how many animals or plants you failed to count[9]." This has also been claimed by Gu and Swihart (2004)[34], MacKenzie et al. (2003)[72], Dorazio et al. (2006)[23] and Dorazio and Royle (2005)[21]. The model introduced in this chapter disputes that claim by showing that multiple surveys are not always essential for estimating site occupancy parameters in the presence of detection error. In general, site occupancy and detection probability parameters can be estimated using a single survey provided two conditions: the probability of occupancy and probability of detection depend on covariates, and the set of covariates that affect occupancy and the set of covariates that affect detection differ by at least one variable.

The statistical model is described in section 6.1. The results of a simulation study conducted to evaluate the single survey approach are discussed in section 6.2. The use of the model is illustrated using data from the Breeding Bird Survey in section 6.3.

[1] A version of this chapter has been accepted for publication in the Journal of Plant Ecology.

6.1 Statistical model and estimation procedure

Consider a site occupancy study in which n sites are surveyed in the study area. Let's denote by Z_i the binary variable that describes the true status of the i^{th} site, $Z_i = 1$ if the i^{th} site is occupied and $Z_i = 0$ if the i^{th} site is unoccupied. These true states are unobserved. Let $Y_i = 1$ if the i^{th} site is "observed to be occupied" and $Y_i = 0$ if the i^{th} site is "observed to be unoccupied." The probability of occupancy is denoted by $Pr(Z_i = 1) = \psi_i$ and the probability of detection by $Pr\left({}^{Y_i=1}/_{Z_i=1}\right) = \delta_i$. It is assumed that if the species is not present, it will not be misidentified and hence $Pr\left({}^{Y_i=0}/_{Z_i=0}\right) = 1$. Simple probability calculations show that $Pr(Y_i = 0) = 1 - \delta_i\psi_i$ and $Pr(Y_i = 1) = Pr\left({}^{Y_i=1}/_{Z_i=1}\right) Pr(Z_i = 1) = \delta_i\psi_i$. These probabilities can depend on the habitat and other covariates. Let $\underline{x}$ denote the set of covariates that affect occupancy, and $\underline{w}$ the set of covariates that affect detection. Some covariates may affect only detection, some covariates may affect only occupancy and some covariates may affect both detection and occupancy. For example, type of forest cover may affect both occupancy and detection, whereas time of the day or weather conditions may affect only detection. Thus, some of the covariates in the sets $\underline{x}$ and $\underline{w}$ might be the same. With the notation, $\psi_i = \psi\left(\underline{x}_i, \underline{\beta}\right)$ and $\delta_i = \delta\left(\underline{w}_i, \underline{\theta}\right)$, the functions should be such that $0 \leq \psi\left(\underline{x}_i, \underline{\beta}\right) \leq 1$ and $0 \leq \delta\left(\underline{w}_i, \underline{\theta}\right) \leq 1$. The necessary conditions under which the parameters$\left(\underline{\beta}, \underline{\theta}\right)$ are identifiable using single survey data are:

1. There should exist at least one covariate that affects either the probability of detection or probability of occupancy.

2. Suppose A and B denote the covariate sets for detection and occupancy respectively. A and B should be such that the sets A-B and B-A are not empty.

The goal of the statistical analysis is to estimate $\left(\underline{\beta}, \underline{\theta}\right)$ given the observations $\underline{y} = (y_1, ..., y_n)$. The likelihood function for these data is:

$$L\left(\underline{\beta}, \underline{\theta}; \underline{y}\right) = \prod_{i=1}^{n} \left(\psi\left(\underline{x}_i, \underline{\beta}\right) \delta\left(\underline{w}_i, \underline{\theta}\right)\right)^{y_i} \left(1 - \left(\psi\left(\underline{x}_i, \underline{\beta}\right) \delta\left(\underline{w}_i, \underline{\theta}\right)\right)\right)^{1-y_i} \tag{6.1}$$

Maximum likelihood estimators are obtained by maximizing this function with respect to $\left(\underline{\beta}, \underline{\theta}\right)$. If the sample size is large, one can use any numerical optimization technique to obtain the MLE. However, this likelihood function, similar to that of the ZIB, is not well-behaved for small samples. The results from the previous chapter showed how the MLE can be improved by using a Penalized Likelihood approach. Similarly, for the single survey

Algorithm 6.1 MPL estimation procedure for the single survey data

1. Obtain the MLE for $\left(\underline{\beta}, \underline{\theta}\right)$by maximizing the likelihood function in equation 6.1. Let us denote these by $\hat{\beta}_M, \hat{\theta}_M$.

2. Obtain the naïve estimator of $\underline{\beta}$ by maximizing 6.2

$$L\left(\underline{\beta}; \underline{y}\right) = \prod_{i=1}^{n} \psi\left(\underline{x}_i, \beta\right)^{y_i} \left(1 - \psi\left(\underline{x}_i, \beta\right)\right)^{1-y_i} \tag{6.2}$$

   This estimator, denoted by $\hat{\beta}_{naive}$, is based on the assumption that there is no detection error. This is stable but biased with the magnitude of bias depending on how large the detection error is.

3. Obtain the naïve estimator of $\underline{\theta}$ by maximizing

$$L\left(\underline{\theta}; \underline{Y}\right) = \prod_{i=1}^{n} \delta\left(\underline{Z}_i, \underline{\theta}\right)^{Y_i} \left(1 - \delta\left(\underline{Z}_i, \underline{\theta}\right)\right)^{1-Y_i}$$

   This estimator, denoted by $\hat{\theta}_{naive}$, is based on the assumption that all sites are occupied. This estimator is stable but biased.

4. Maximize the penalized likelihood function with respect to $\left(\underline{\beta}, \underline{\theta}\right)$

$$log\left(PL\left(\underline{\beta}, \underline{\theta}; \underline{y}\right)\right) = log\left(L\left(\underline{\beta}, \underline{\theta}; \underline{y}\right)\right) - \lambda_1 \mid \underline{\beta} - \hat{\beta}_{naive} \mid - \lambda_2 \mid \underline{\theta} - \hat{\theta}_{naive} \mid$$

   where$\lambda_1 = \left(1 - \hat{\psi}_{naive}\right) \hat{\delta}_M \sqrt{tr\left(var\left(\hat{\theta}_M\right)\right)}$ and $\lambda_2 = \left(1 - \hat{\delta}_{naive}\right) \hat{\psi}_M \sqrt{tr\left(var\left(\hat{\beta}_M\right)\right)}$ and $\left(\hat{\delta}_{naive}, \hat{\psi}_{naive}\right)$ and $\left(\hat{\delta}_M, \hat{\psi}_M\right)$ and denote the average occupancy and detection probabilities under the naïve method of estimation and MLE respectively.

approach, a penalized likelihood can be used to stabilize the estimation procedure. The penalized likelihood estimators for the single visit case are obtained using algorithm 6.1.

The justification for the penalty function showed in the algorithm 6.1 is along the same lines as described in the previous chapter. Because $tr\left(var\left(\hat{\theta}_M\right)\right) \to 0$ and $tr\left(var\left(\hat{\beta}_M\right)\right) \to 0$ as the sample size increases, the penalized likelihood function approaches the likelihood function if the number of sites is large. Penalization simply stabilizes the likelihood function for small sample sizes. If the MLE of average detection probability is high, naïve estimates of the occupancy are reasonable. In this case, the first component of the penalty function is large, thus the occupancy parameters are shrunk towards their naïve estimates. Similarly, if the MLE of the average occupancy is high, the naïve estimates of the detection parameters will also be reasonable. In this case, the second component of the penalty function is large, therefore the detection parameters will be shrink towards their naïve estimates.

In Step 1 of the penalized likelihood estimation algorithm 6.1 we need to compute the MLE and its variance. If the number of sites is smaller than 100, using a gradient-based optimization technique to find the location of the maximum tends to be tricky as it is prone to lead to nearly singular Hessian matrices [82]. Because of this we cannot use the inverse of the Hessian matrix to approximate the variance of the MLE. Instead of using a local gradient-based technique to find the MLE and its variance in Step 1, we use a global stochastic search method-a variant of the well known simulated annealing method- called data cloning [60, 62]. In data cloning, as in simulated annealing, the MLE is obtained as the mean of the posterior distribution. This avoids the task of numerically differentiating a non-smooth function. To obtain the variance of the MLE, one can either use bootstrap samples [27] or it can also be approximated by the variance of the posterior distribution [60, 62]. This eliminates the need to invert a nearly singular Hessian matrix to approximate the variance of the MLE. The penalized likelihood function (Step 4 in algorithm 6.1) is maximized using the standard numerical optimization technique. The confidence intervals for MPLE can be based on the bootstrap technique and are shown to have good coverage (section 4.1). A computer program written in R to estimate the parameters using information from a single surveys is available in Solymos and Moreno (2010)[111].

6.2 Simulation study

The simulation study presented in this section has two goals. The first goal is to support the claim of estimability of the parameters using a single survey; if the parameters are consistently estimable, then, as we increase the sample size, the estimates should converge to the true values. The second goal is to show that these estimators give reasonable inferences in practical situations.

To achieve this goal, and for the purpose of considering a variety of scenarios commonly found in this type of analysis, a total of 54 cases were simulated. These cases were defined by considering different levels for factors such as the sample size, the type of link function, the probability of occupancy and detection, and the configuration of the covariates (i.e. whether or not there was a common covariate for both occupancy and detection).
In each case 100 data sets were generated using two covariates for occupancy and two covariates for detection. For the case with no common covariates the probability of occupancy

was calculated using the Logistic link

$$\psi_i = \frac{exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)}{1 + exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)} \tag{6.3}$$

where covariate values were generated using $x_{1i} \sim Normal\,(0, 1)$ and $x_{2i} \sim Bernoulli\,(0.55)$. Similarly, the probability of detection was calculated using either the Logistic link

$$\delta_i = \frac{exp\left(\theta_0 + \theta_1 w_{1i} + \theta_2 w_{2i}\right)}{1 + exp\left(\theta_0 + \theta_1 w_{1i} + \theta_2 w_{2i}\right)} \tag{6.4}$$

or the Log-Log link

$$\delta_i = exp\left(-exp\left(\theta_0 + \theta_1 w_{1i} + \theta_2 w_{2i}\right)\right) \tag{6.5}$$

where covariate values were generated using and $w_{i1} \sim Normal\,(0, 1)$ and $w_{2i} \sim Bernoulli\,(0.65)$.

For the common covariate cases, the covariate for the occupancy model was taken as the common one for both. For instance, if the common covariate is a continuous one and the link for both occupancy and detection is the Logistic link, the probability of occupancy for the i^{th} site is

$$\psi_i = \frac{exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)}{1 + exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\right)} \tag{6.6}$$

and the probability of detection is

$$\delta_i = \frac{exp\left(\theta_0 + \theta_1 x_{1i} + \theta_2 w_{2i}\right)}{1 + exp\left(\theta_0 + \theta_1 x_{1i} + \theta_2 w_{2i}\right)} \tag{6.7}$$

The set of parameters was selected to obtain the desired level of occupancy and detection required according to the case, and the estimates were obtained by using the Maximum Penalized Likelihood estimator described in the algorithm 6.1. Figures 6.1 and 6.2 present the results from two representative cases obtained from the simulations. A complete summary of the results obtained for the 54 cases is available in the appendix B .

Figure 6.1 shows the box plots of the estimated parameters when the mean probability of occupancy is 0.27, the mean probability of detection 0.27, and the covariates for occupancy and detection are separable. Clearly as the sample size increases, the distributions of the parameters become more symmetric and their centers get closer to the true value. It is also observed that as the sample size increases the spread of the distributions decreases. Figure 6.2 presents the results obtained for the case in which there is a discrete covariate that is common to both occupancy and detection. Again, in this case, the mean probabilities

of occupancy and detection are low (0.27 and 0.31 respectively). Similar to the separable covariates case, as the sample size increases the centers of the density functions get closer to the true value, and, the variance decreases as the sample size increases.

For most of the situations considered in our simulations (see appendix B), the mean occupancy and mean detection probabilities can be estimated reasonably well at sample sizes of 100 and 200 whereas a good estimation of regression coefficients occurs at sample sizes of 300 or larger. See Figure 6.3 for an example. If the main goal of an analysis is the estimation of mean occupancy rate, one does not have to worry as much about sample size as when accurate estimation of the regression coefficient per se is the objective.

These results along with the results in the appendix show that the occupancy and detection parameters are identifiable using single survey data. This holds even when the set of covariates for occupancy and detection have some overlap.

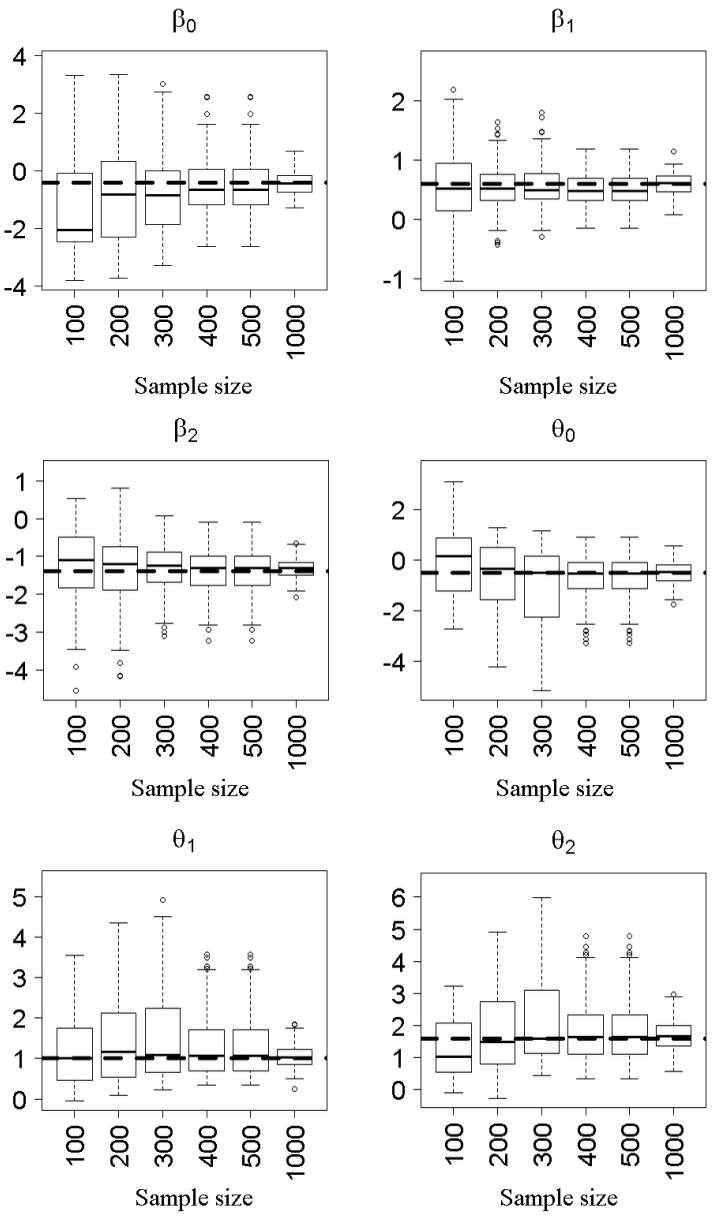


Figure 6.1: Simulations showing estimability of the parameters using a single survey when the covariates that affect occupancy and detection are separable. The parameters β_0, β_1 and β_2 correspond to the occupancy model; the parameters θ_0, θ_1 and θ_2 correspond to the detection model. As the sample size increases, the estimates converge to the true value.

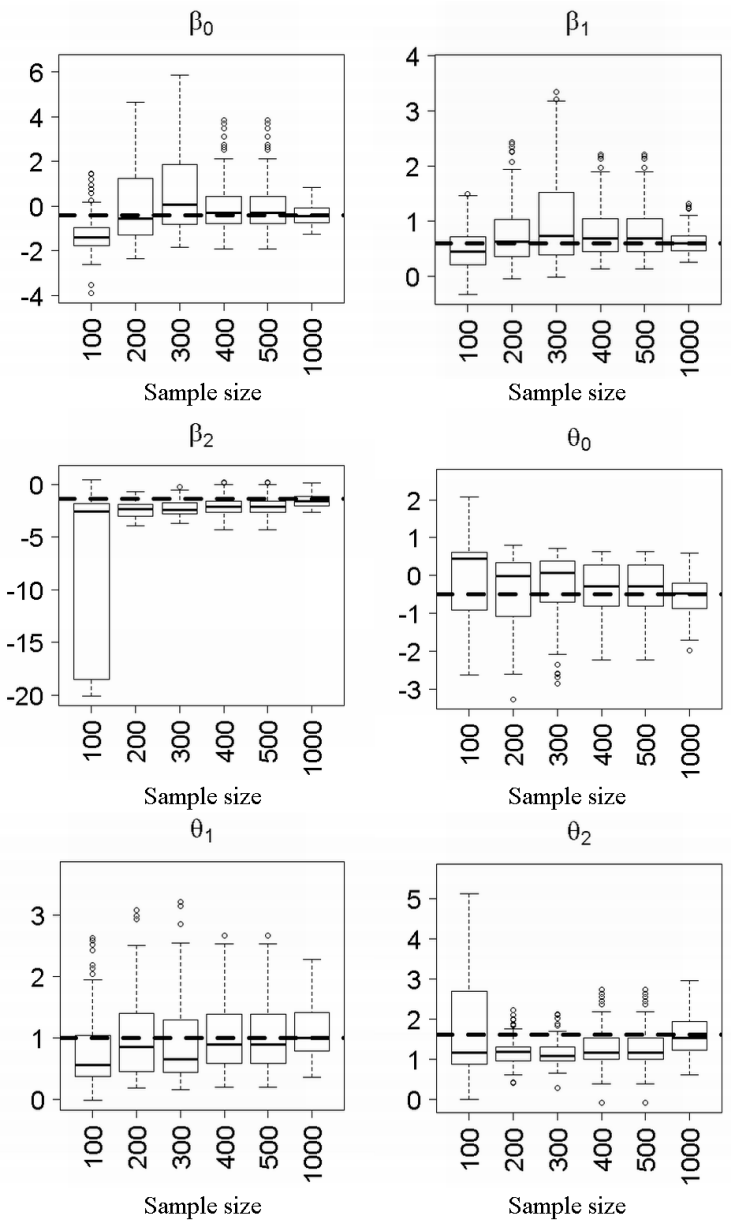


Figure 6.2: Simulations showing estimability of the parameters using a single survey when there is a categorical common covariate that affect occupancy and detection. The parameters β_0, β_1 and β_2 correspond to the occupancy model; the parameters θ_0, θ_1 and $\theta_2$correspond to the detection model. As the sample size increases, the estimates converge to the true value.

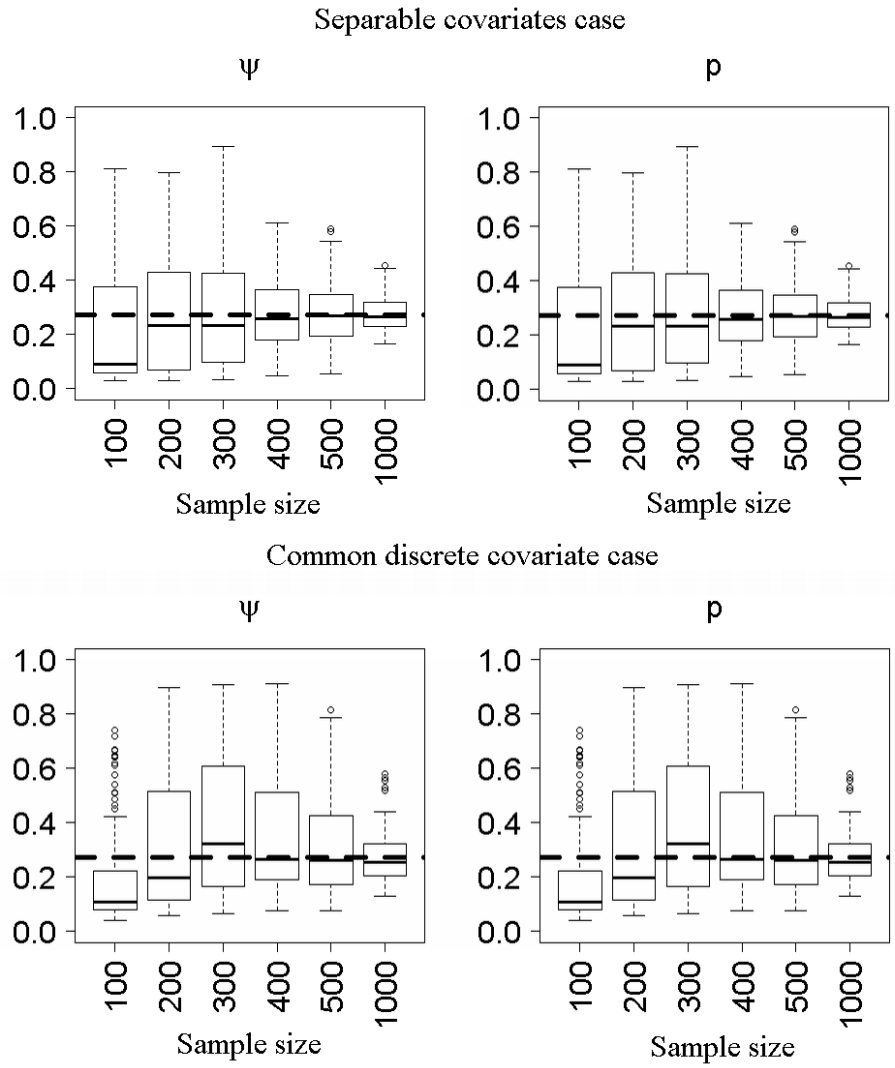
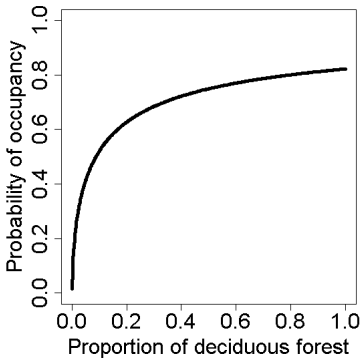


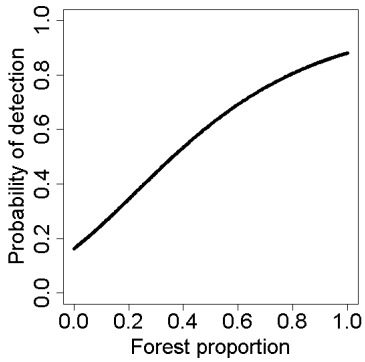
Figure 6.3: Simulations showing estimability of the mean occupancy and detection for both cases using separable covariates and using a categorical common covariate that affects occupancy and detection.

6.3 Example data analysis

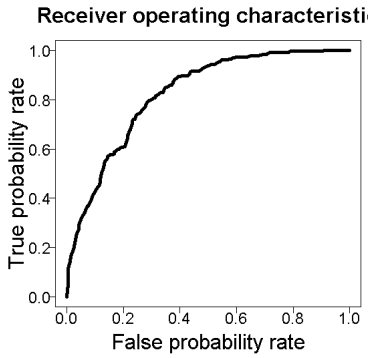
To illustrate the estimation of the parameters for an occupancy model using a single survey we consider detected and not detected data for Ovenbirds (*Seiurus aurocapilla*). Data were collected in 1999 using Breeding Bird Survey (BBS) Protocols [25] in the boreal plains eco-region of Saskatchewan, Canada. The goal of the study was to determine whether the occupancy of this species was influenced by the amount of forest around each survey point. Data were collected along 36 BBS routes each consisting of 50 survey locations with survey locations separated by 800 meters. To increase the independence of observations we used



(a) Probability of occupancy



(b) Probability of detection

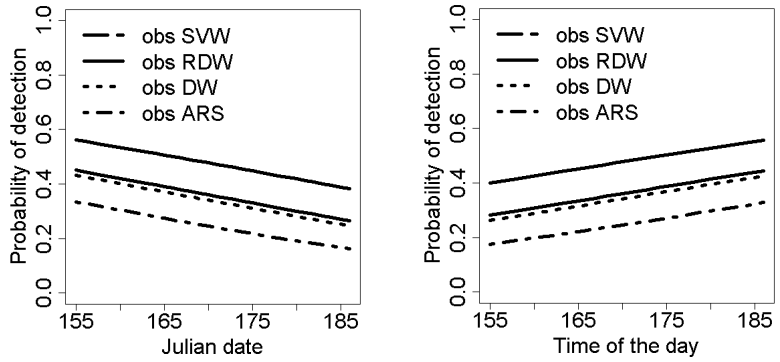


(c) Receiver Operating Curve

Figure 6.4: Ovenbird data analysis results

every second survey point along each route (thus each point was 1.6 km apart) in our analysis (n = 900 survey locations). Attributes of the forest type and amount of forest remaining within a 400 m radius were estimated from the Saskatchewan Digital Land Cover Project [75]. The habitat requirements of the Ovenbird are well understood [46] and it was expected that the probability a location is occupied by the Ovenbird would be positively influenced by the amount of deciduous forest remaining (i.e. the forest deciduous proportion). Longitude was also included as the study covered an east-west gradient over 1000 km in length.

The factors expected to influence detection probability were: observer, time of day, time of year, and amount of forest. Observers differ in their ability to hear birds in part because of their individual skill but also because of fundamental differences in the distance over which they hear birds. In general, male songbirds sing very regularly early in the breeding season making it easy to detect individuals that are present. As the breeding season progresses,



(a) Julian date (b) Time of the day

Figure 6.5: Estimates of the effects of the observers, Julian date and Time of the day over the probability of occupancy for the Ovenbird data.

however, the males spend less time singing as they focus on other activities. This often results in lower detectability later in the breeding season. Julian date was included as a variable influencing detection error. Male songbirds also have a tendency to sing earlier in the day, shortly after sunrise, and then later in the morning when they focus either on guarding the mate or foraging. To account for this, time of the day was included as a factor influencing detectability. Detectability can also be influenced by habitat attributes[103]. On the contrary, in areas with more forest, the chance of multiple males singing may be higher, hence increasing detection probability relative to areas with less forest where only one individual may be present. Several models were considered, the best model was selected using the Akaike's Information Criterion (AIC). Furthermore, the Receiver Operating Curve (ROC) and the Area under the ROC (AUC) were calculated to heuristically compare the fit of the full model, the detection and the occupancy altogether. Table 6.1 gives the details on the various models that were considered and the corresponding AIC and AUC values. The final model had an AUC of 0.82, indicating a fairly good predictive capacity for the full model, detection and occupancy. This model also had a smaller AIC value relative to other candidate models. Table 6.2 presents the estimated parameters, the 90% confidence intervals, and the estimated standard errors for occupancy and detectability. Figure 6.4 depicts graphically how the probability of occupancy and detection vary with the covariates. The confidence intervals and the standard errors were estimated using 200 bootstrap samples. As expected, the proportion of deciduous forest has a positive effect on the probability of occupancy. This relationship was best fit using a log-transformation of forest deciduous proportion. Longitude was not statistically significant but it suggested

Table 6.1: Models for the Ovenbird data sorted from smallest to largest Akaike's Information Criterion (AIC). We also provide the Schwarz Information Criterion (BIC) and the area under the Receiver Operating Characteristic curve (AUC). Smaller the AIC, BIC, better the model fit; larger the AUC, better is the fit.

model	occupancy model covariates	detection model covariates	AUC	AIC	BIC
1	log(proportion deciduous forest)	proportion of forest	0.823	825.5	844.6
2	log(proportion deciduous forest)	proportion of forest; julian date; time of day	0.826	826.5	855.2
3	log(proportion of deciduous forest);log(non deciduous forest)	proportion of forest	0.823	827.1	851.0
4	proportion of deciduous forest; longitude	proportion of forest; julian date; time of day; observer	0.828	827.9	875.7
5	log(proportion of deciduous forest); longitude	proportion of forest; julian date; time of day	0.826	828.4	862.0
6	log(proportion of deciduous forest); longitude	proportion of forest; julian date; time of day; observers	0.827	830.7	878.6
7	log(proportion of agricultural area); longitude	proportion of forest; julian date; time of day; observers	0.818	841.2	889.1
8	proportion of agricultural area; longitude	proportion of forest; julian date; time of day; observers	0.820	843.4	891.3

that Ovenbird occupancy rate increased as surveys were done further west.

The amount of forest cover was the strongest predictor of detection probability. Detection probability was highest in areas with higher forest cover. This suggests that larger number of birds in areas with more forest increase the probability of detecting the species. Conversely, in areas with low forest cover, the smaller number of birds means that the chances of detecting the species given they are present is lower. Although not strictly statistically relevant according to AIC (and, hence not included in the final model), detection probability did differ among observers and was affected by the time of the day and the Julian date in a

Table 6.2: Estimated parameters, confidence intervals and standard errors for the occupancy and detection model for the Ovenbird occupancy survey data.

model	covariates	point estimate (90% CI)
occupancy	intercept	-0.255 (-0.843, 0.566)
occupancy	log(proportion of deciduous forest)	1.546 (0.863, 3.561)
detection	intercept	0.476 (0.047, 1.127)
detection	proportion of forest area	0.951 (0.588, 1.763)
Average occupancy probability		0.496 (0.405, 0.643)
Average detection probability		0.489 (0.377, 0.608)
Naive estimate of average occupancy		0.297 (0.278, 0.318)

sensible fashion (Figure 6.5). For instance, one observer in particular (SVW) was much more likely to detect birds in areas with less forest than the others. Because previous experience from other projects has demonstrated that this individual is able to hear birds over far greater distances than other people, this result was not surprising. Detection probability was negatively related to Julian date indicating decreased singing activity later in the season was reducing observer ability to detect birds given they were present. Time of day had a positive relationship with detection probability, and this was somewhat unexpected. However, time of day had the least significant effect and surveys were done in a very narrow time window (4:00 to 9:00 local time). The estimated mean probability of occupancy for all the sites, based on the final model, was 0.496, with a mean probability of detection of 0.489. The mean probability of occurrence without correcting for detection error, on the other hand, was 0.297.

6.4 Summary

The simulation study demonstrated that the estimates of the site occupancy probability can be obtained using information from a single survey, provided the site occupancy probability and the detection probability significantly depend on habitat or other exogenous covariates, and that the set of covariates that affect occupancy and the set of covariates that affect detection differ by at least one covariate. From a survey of previous applications of site

occupancy models, it seems that most practical situations do satisfy these conditions. In fact, for 94 out of 100 cases the covariates that affect detection and the covariates that affect occupancy were disjoint. A limitation of our methodology is that the case of constant probability of occupancy and constant probability of detection cannot be estimated. However, it appears that this restriction is not important in practice as in most papers we have reviewed, the probability of occupancy and detection both were seldom constant. It is not possible to provide general results about identifiability conditions under every possible model. However, the data cloning method [60, 62] can be used for both estimation and detection of possible non-estimability. The parameters are estimable if and only if the posterior variance converges to zero as the number of clones increases[62]. This test is built into our software for the analysis of single survey site occupancy data [111].

When the crucial assumptions of population closure and independence of surveys are satisfied and costs are not a major issue, then multiple survey methods will generally provide statistically more efficient estimators than a single survey based approach. However, if the closed population assumption is not met, there is uncertanty of the meaning of the estimates obtained from the ZIB model. At this point, we want to emphasize that when using the single survey approach the estimated probability of occupancy can be defined as the instantaneous probability of occupancy. In conclusion, the development of the single survey approach provides an additional tool to ecologists that allows for correction of detection error, does not have the critical assumption of population closure, and has the logistical flexibility of conventional single-survey designs.

Chapter 7

Cluster sampling

With the purpose of keeping sampling effort and the use of resources to a minimum, it has been suggested that the requirement of repeated surveys can be accomplished by using strategies such as multiple observers surveying independently the same patch, or by surveying multiple locations within a patch[69]. Although these strategies can be effective on reducing sampling effort and the use of resources, they can also lead to violations of the assumption of independence between surveys. Hines et al. (2010) introduced a first approach to estimate site occupancy without the requirement of having independent surveys (section 4.3). They also showed that a violation of the requirement of independence of the surveys results in bias of the estimates obtained from the ZIB.

In this chapter we proposed an alternative model that allows estimation of site occupancy probability when the probability of detection is less than one and the information collected in a single survey from sites exhibit some level of correlation (i.e. violation of the independence requirement). For the model, we assumed a sampling scheme in which the surveyed sites are clustered as depicted in Figure 7.1. The locations within a cluster are correlated but the locations across clusters are independent from one another. Hereafter, a cluster will be referred as a sample unit. Each sample unit will be assumed to consist of k sampling locations. This chapter is organized in three sections: the statistical model is introduced in section 7.1; the results of a simulation study are shown in section 7.2, and the application of the model is illustrated in section 7.3.

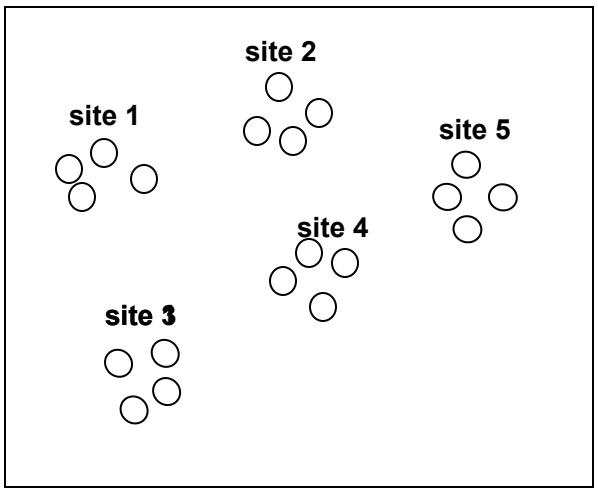


Figure 7.1: Example of the sampling scheme for the cluster sampling model. The sampling area consist on 5 sample units or clusters. Each sample unit is surveyed by sampling 4 locations within it. The samples units are assumed to be independent from each other. The locations within every sample unit are correlated.

7.1 Statistical model and estimation procedure

In this section we introduce the statistical model for the cluster sampling methodology. The model will be described in two parts. First, we will discuss the site occupancy probability component of the model, and then we will extend the model to incorporate the detection error in the observations.

Occupancy model

The site occupancy probability for a sampling scheme such as cluster sampling can be modeled by using the *auto-logistic model*. The auto-logistic model has been widely used to estimate the site occupancy probability for those cases in which the surveyed sites are expected to be correlated and an assumption of no error in the detection can be made [118, 48, 54, 102, 44].

Let's assume that n independent sample units are selected to monitor the target species and that the presence/absence of the species can be determined with no error (i.e., probability of detection is 1). In addition, let's assume that within every sample unit, k locations are surveyed once to determine the presence/absence of the species (figure 7.1). The true status of the target species (present/absent) on the j^{th} location at the i^{th} sample unit is therefore a binary variable denoted by z_{ij}, where $z_{ij} = 1$ indicates that the species is present at the j^{th}location of the i^{th} sample unit, and $z_{ij} = 0$ indicates that the species is absent

from the j^{th} location at the i^{th} sample unit.

Let's assume the conditional probabilities $Pr\left(z_{ij}/\{z_{im}:m\neq j\}\right)$ for $i=1,\ldots,n$ and $j=1,\ldots,k$ can be calculated. These conditional probabilities indicate that the presence/absence of the target species at the j^{th} location of the i^{th} sample unit is conditioned by its presence/absence in all the other locations within the i^{th} sample unit. The stochastic process defined by these conditional probabilities is called a Random Markov Field and it was first introduced by Besag (1974)[8]. The conditional probability mass function can be written as follows:

$$f\left(z_{ij}/z_{im}:m\neq j\right)=\frac{exp\left(z_{ij}\cdot A_{ij}\left(z_{im}:m\neq j,\underline{x}_{ij}\right)\right)}{1+exp\left(z_{ij}\cdot A_{ij}\left(z_{im}:m\neq j,\underline{x}_{ij}\right)\right)} \tag{7.1}$$

where $\left\{\underline{x}_{ij}:i=1,...,n;j=1,...,k\right\}$ are the habitat covariates associated with the j^{th} location at the i^{th} sample unit, and $A_{ij}\left(\cdot\right)$ is called the natural parameter function. Besag (1974) showed that for the binary case, $A_{ij}\left(\cdot\right)$ should be of the following form:

$$A_{ij}\left(z_{im}:m\neq j,\underline{x}_{ij}\right)=\underline{x}_{ij}^{T}\underline{\beta}+\gamma\sum_{m\neq j}z_{im} \tag{7.2}$$

where $\underline{\beta}$ is a vector of parameters that relates the habitat covariates with the probability of occupancy of the target species, and γ is the statistical dependence parameter that accounts for the dependence between locations within a sample unit. In this way, $\gamma=0$ indicates that the locations within the sample unit are independent. Notice that the conditional probability mass function in equation 7.1 resembles that of the Logistic regression model with an additional term that is a function of the binary response at the neighbor locations. For this reason this model is also known as the *Auto-logistic regression*.

It can be shown that under the parameterization given by equation 7.2 the odds for a location to be occupied relative to odds of the independence model ($\gamma=0$) increases for any nonzero neighbors, and can never decrease. Caragea and Kaiser (2009) proposed the following centered version of natural parameter function $A_{ij}\left(\cdot\right)$ [13]

$$A_{ij}\left(z_{im}:m\neq j,\underline{x}_{ij}\right)=\underline{x}_{ij}^{T}\underline{\beta}+\gamma\sum_{m\neq j}\left(z_{im}-\frac{exp\left(\underline{x}_{im}^{T}\underline{\beta}\right)}{1+exp\left(\underline{x}_{im}^{T}\underline{\beta}\right)}\right) \tag{7.3}$$

Unlike the traditional model (Equation 7.2), this parameterization allows for an interpretation of the parameters independent of the level of the statistical dependence. Using

the centered version of the natural parameter function, $\gamma > 0$ indicates that the odds of the location ij to be occupied ($z_{ij} = 1$) increase if the number of occupied neighbor locations is greater than the number of occupied neighbor locations under the independence model. Because we consider that this last parameterization provides a more natural interpretation of the parameters, we decided to incorporate it on our model.

Having the conditional probabilities given by equation 7.3, the next step is to obtain the joint probability distribution denoted by $Pr\left(z_{i1}, ..., z_{ik}\right)$. The Hammersley-Clifford theorem [35] establishes the form the joint distribution must take for it to become the joint probability measure of a Markov Random Field. The application of the Hammersley-Clifford theorem is illustrated for k=2 as follows:

$$\begin{aligned}
\frac{Pr\left(\underline{z}_i = (1,1)\right)}{Pr\left(\underline{z}_i = (0,0)\right)} &= \frac{Pr\left(z_{i1}=1/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)} \cdot \frac{Pr\left(z_{i1}=1/z_{i2}=1\right)}{Pr\left(z_{i1}=0/z_{i2}=1\right)} && (7.4)\\
&= \frac{\psi_{i11}}{\psi_{i00}} && (7.5)\\
&= exp\left(\underline{x}_{i1}^T\underline{\beta} + \underline{x}_{i2}^T\underline{\beta} + \gamma\left(1 - \mu_{i1} - \mu_{i2}\right)\right) && (7.6)
\end{aligned}$$

$$\begin{aligned}
\frac{Pr\left(\underline{z}_i = (0,1)\right)}{Pr\left(\underline{z}_i = (0,0)\right)} &= \frac{Pr\left(z_{i1}=0/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)} \cdot \frac{Pr\left(z_{i1}=1/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)} && (7.7)\\
&= \frac{\psi_{i01}}{\psi_{i00}} && (7.8)\\
&= exp\left(\underline{x}_{i2}^T\underline{\beta} + \gamma\left(-\mu_{i1}\right)\right) && (7.9)
\end{aligned}$$

$$\begin{aligned}
\frac{Pr\left(\underline{z}_i = (1,0)\right)}{Pr\left(\underline{z}_i = (0,0)\right)} &= \frac{Pr\left(z_{i1}=1/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)} \cdot \frac{Pr\left(z_{i1}=0/z_{i2}=1\right)}{Pr\left(z_{i1}=0/z_{i2}=1\right)} && (7.10)\\
&= \frac{\psi_{i10}}{\psi_{i00}} && (7.11)\\
&= exp\left(\underline{x}_{i1}^T\underline{\beta} + \gamma\left(-\mu_{i2}\right)\right) && (7.12)
\end{aligned}$$

It follows that

$$Pr\left(\underline{z}_i = (0,0)\right) = \psi_{i00} = \frac{1}{1 + \frac{Pr\left(\underline{z}_i=(1,1)\right)}{Pr\left(\underline{z}_i=(0,0)\right)} + \frac{Pr\left(\underline{z}_i=(0,1)\right)}{Pr\left(\underline{z}_i=(0,0)\right)} + \frac{Pr\left(\underline{z}_i=(1,0)\right)}{Pr\left(\underline{z}_i=(0,0)\right)}} \quad (7.13)$$

where $\mu_{ij} = \frac{exp(\underline{x}_{ij}^T \cdot \underline{\beta})}{1+exp(\underline{x}_{ij}^T \cdot \underline{\beta})}$ for $i = 1, ..., n$ and $j = 1, 2$. The joint distribution for a general number of locations k can be determined in a similar manner.

Detection model

Nevertheless, the true status of the species is not directly observable since it is likely for the species to be present but not detected during the survey. These observations are denoted by the binary variable y_{ij}, where $y_{ij} = 1$ if the target species is detected at the j^{th} location of the i^{th} sample unit (i.e., the species was present and detected during the survey), and $y_{ij} = 0$ if the species is not detected. This last case indicates that either the species was truly absent or that the species was present but failed to be detected during the survey.

Let δ_{ij} be the probability of detecting the species at the j^{th} location of the i^{th} site given that it is present. These probabilities can depend on covariates such as time of the day, weather conditions and some habitat characteristics. Let us denote by $\underline{w}_{ij}$ the set of covariates at the j^{th} location of the i^{th} sample unit that are associated to the probability of detection. These covariates can be incorporated into the probability of detection by using any link function such as the logit link, complemenary loglog link or probit link among others. For instance, using the complementary loglog link, the probability of detection is:

$$\delta_{ij} = Pr\left(y_{ij}=1/z_{ij}=1\right) = 1 - exp\left(-exp\left(\underline{w}_{ij}^T \underline{\theta}\right)\right) \; for \; i = 1, ..., n; j = 1, ..., k \tag{7.14}$$

where $\underline{\theta}$ is a vector of parameters that quantify the relationship between the covariates and the probability of detection.

The probability mass function for the observations is then found by combining the joint probability distribution of the Markov Random field and the probability of detection. For instance, if there are two locations within the i^{th} sample unit (i.e., k=2) and the species was not detected in the first location nor in the second location, then the observation vector for the sample unit is then $\underline{y}_i = (0, 0)$. The probability for this observation is:

$$Pr\left(\underline{y}_i = (0,0)\right) = \psi_{i00} + \left(\psi_{i10}\left(1 - \delta_{i1}\right)\right) + \left(\psi_{i01}\left(1 - \delta_{i2}\right)\right) + \left(\psi_{i11}\left(1 - \left(\delta_{i1}\delta_{i2}\right)\right)\right) \tag{7.15}$$

The first term corresponds to the probability that none of the two locations is occupied by the species, the second term corresponds to the probability that the species is present only in the first location but was not detected, the third term corresponds to the probability

that the species is present only in the second location but was not detected, and the last term corresponds to the probability that both locations are occupied but the species was not detected in neither. The likelihood is then calculated as the product of the observation probabilities collected at every sample unit.

Maximum likelihood estimators (MLE) are obtained by maximizing this function with respect to the parameters $\left(\underline{\beta}, \underline{\theta}, \gamma\right)$. If the number of sites is large, any optimization method can be used to obtain the MLE. However, if the number of visited sites is small, the likelihood function will tend to be flat, and it will be necessary the use of more sophisticated techniques to maximize it. The models introduced in chapters 5 and 6 showed how by penalizing the likelihood it is possible to obtain better estimates of the parameters in an occupancy model. This is also true for the cluster sampling model. The algorithm 7.1 describes how to obtain the Maximum Penalized Likelihood estimates (MPLE) for the cluster sampling model. It is important to mention that the purpose of the penalization is to stabilize the likelihood function when the number of visited sites is small. As the number of visited sites increases the penalty converges to zero, and the MPLE are the same as the MLE.

Notice that for the model described above it is assumed that there are covariates related to the probability of occupancy and the probability of detection. The reason for this is that there are some conditions in which the parameters for the model are not identifiable, one case arising when both the probability of occupancy and probability of detection are constant. In general, the identifiability of the parameters for any model can be tested using *Data Cloning* [60, 62]. In brief, *Data Cloning* is a method that using a large number of identical replicates of the data (clones) and the Markov chain Monte Carlo (MCMC) algorithms estimates a posterior distribution that is concentrated around the MLE. Lele et al (2011) demonstrated that if the parameters of an specific model are non estimable (non-identifiable), then as the number of clones increases the largest eigenvalue of the posterior variance matrix does not converge to zero[62]. This procedure is illustrated in section 7.3.

7.2 Simulation analysis

A simulation study was carried out to evaluate the estimability of the parameters and the statistical properties of the proposed estimation method. The number of surveyed locations (k=3, 6), the number of sample units (n=50,100, 200, 300, 400 and 500) and the level of dependence ($\gamma = 0.5$ and $\gamma = 1$) were the factors considered for the analysis.

Algorithm 7.1 Maximum Penalized Likelihood estimates for the Cluster Sampling model

1. Obtain the MLE for $\left(\underline{\beta}, \underline{\theta}, \gamma\right)$ by maximizing the likelihood function. Let us denote these by $\left(\hat{\underline{\beta}_M}, \hat{\underline{\theta}_M}, \hat{\gamma_M}\right)$.

2. Obtain the naïve estimator of $\left(\underline{\beta}, \gamma\right)$using the auto-logistic model. These estimators $\left(\hat{\underline{\beta}}_{naive}, \hat{\gamma}_{naive}\right)$are based on the assumption that there is no detection error.

3. Obtain the naïve estimator of $\underline{\theta}$ by maximizing

$$L\left(\underline{\theta}; \underline{y}\right) = \prod_{i=1}^{n} \prod_{j=1}^{k} \varphi\left(\underline{w}_{ij}, \underline{\theta}\right)^{y_{ij}} \left(1 - \varphi\left(\underline{w}_{ij}, \underline{\theta}\right)\right)^{1-y_{ij}}$$

   where $\varphi\left(\cdot\right)$ is the link function used for the probability of detection. This estimator $\left(\hat{\underline{\theta}}_{naive}\right)$ is based on the assumption that all sites are occupied.

4. Maximize the penalized likelihood function presented below with respect to $\left(\underline{\beta}, \underline{\theta}, \gamma\right)$

$$\begin{aligned} log\left(PL\left(\underline{\beta}, \underline{\theta}, \gamma\right)\right) &= log\left(L\left(\underline{\beta}, \underline{\theta}, \gamma\right)\right) - \lambda_1 \left[\sum_{l=1}^{s} \mid \beta_{l-} \hat{\beta}_{l,naive} \mid + \mid \gamma - \hat{\gamma}_{naive} \mid\right] \\ &\quad -\lambda_2 \left[\sum_{l=1}^{r} \mid \theta_{l-} \hat{\theta}_{l,naive} \mid\right] \end{aligned}$$

   where s and r are the number of parameters of the probability of occupancy and detection respectively. Where $\lambda_1 = \left(1 - \hat{\psi_{naive}}\right) \hat{\delta}_M \sqrt{tr\left(\hat{var}\left(\hat{\theta}_M\right)\right)}$ and $\lambda_2 = \left(1 - \hat{\delta_{naive}}\right) \hat{\psi}_M \sqrt{tr\left(\hat{var}\left(\hat{\beta}_M\right)\right)}$. $\left(\hat{\psi}_{naive}, \hat{\delta}_{naive}\right)$ and $\left(\hat{\psi}_M, \hat{\delta}_M\right)$ and denote the average probability of detection and occupancy under the naïve and MLE respectively.

Three configurations of the covariates for the probability of occupancy and detection model were considered. On the first configuration it was assumed that there are no common covariates for the detection and occupancy model (separable model). For the second configuration it was assumed that a binary covariate was common for both the occupancy and the detection models. The third configuration was similar to the second one differing only in that the common covariate was continuous.

For the separable model it was assumed that the probability of occupancy of the target species in every surveyed location depended on two covariates: a continuous covariate $x_{ij1} \sim Normal\,(0,1)$ and a binary covariate $x_{ij2} \sim Bernoulli\,(0.55)$. The probability of detection was calculated using the complementary loglog link function, and the assumption that it depends on a continuous covariate $w_{ij1} \sim Normal\,(0,1)$ and a binary covariate $w_{ij2} \sim Bernoulli\,(0.65)$. For the other two configurations (with common covariates) the probability of occupancy remained unchanged. However, the probability of detection was modified to include a covariate that was also included in the occupancy model. Consequently, for the common continuous covariate case, the probability of detection was set to depend on x_{ij1} and w_{ij2}. For the discrete common covariate case, the probability of detection was set to depend on w_{ij1} and x_{ij2}. The values of the parameters of the occupancy and detection ($\underline{\beta}$ and $\underline{\theta}$ respectively) models were selected to obtain a low probability of occupancy $\overline{\psi} \approx 0.30$ (under the independence model) and a low probability of detection $\overline{\delta} \approx 0.33$.

For every case in the simulation study, 200 data sets were generated using the following procedure. First, the values of the covariates at every location were randomly generated. Then, the presence/absence at every location was generated using a Gibbs sampling algorithm [14]. Subsequently, the probability of detection for the occupied locations was calculated. Finally, the detection/non detection of the species at the occupied locations was randomly generated using a Bernoulli distribution with probability of success equal to the probability of detection. The generated observations and the value of the covariates were used to obtain the MLE and MPLE for every data set. The MLE and MPLE were obtained using algorithms programmed in R [113] and WinBUGS[66]. A complete summary of the results of the simulations is presented in the Appendix C.

Figures 7.2, 7.3 and 7.4 show the estimated mean probability of occupancy for different sample sizes for the three considered cases (separate covariates, common discrete covariate and a common continuous covariate respectively). It was found that when the number of sample units is larger than 50 and the covariates are separable, the estimated mean

probability of occupancy is unbiased. In this case the estimates are more precise for a large cluster (k=6) and a large value of γ (Figure 7.2). Similar results were obtained for a common discrete covariate: the mean probability of occupancy is unbiased for sample sizes larger than 50, the smallest standard errors are present when a strong dependency between sites ($\gamma = 1$) exists and the cluster size is equal to 6 (figure 7.3). Figure 7.4 shows that for common continuous covariates and small dependency between sites ($\gamma = 0.50$), larger samples are required to obtain an unbiased estimate of the probability of occupancy. Nevertheless, when the dependency is strong ($\gamma = 1.00$) the results are similar to those obtained for the separable covariates and discrete common covariate cases.

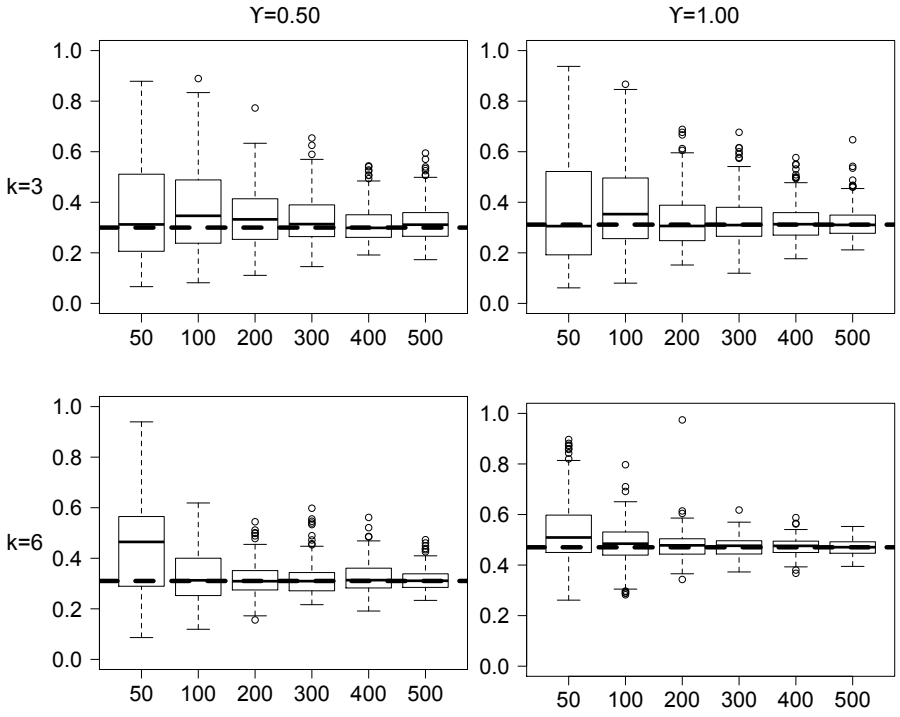


Figure 7.2: Estimated mean occupancy the separable covariates case. Dotted line indicates the true mean occupancy.

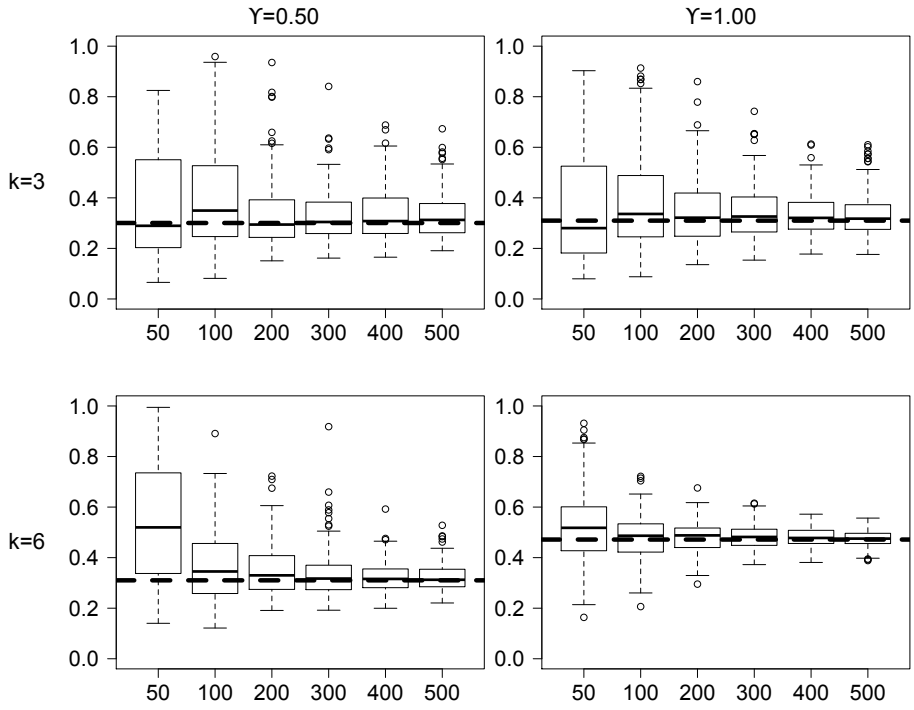


Figure 7.3: Estimated mean occupancy the discrete common covariate case. Dotted line indicates the true mean occupancy.

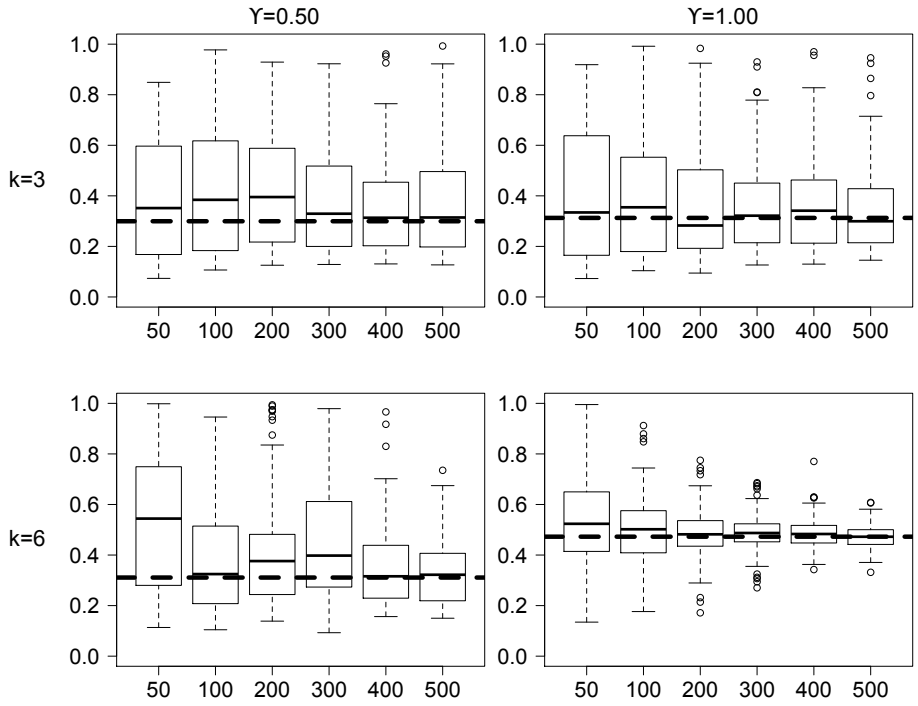


Figure 7.4: Estimated mean occupancy for the continuous common covariate case. Dotted line indicates the true mean occupancy.

Figures 7.5, 7.6 and 7.7 display the median of the percentage bias. The results depicted in these figures resemble those observed for the mean probability of occupancy. The bias of the estimated parameters gets reduced as the sampling effort (number of clusters and size of the cluster) increases. The best results are obtained when the dependency between sites is large and the cluster size is 6. It is also noticeable that the most difficult case for the estimation is when a continuous covariate is common, since it displays the largest bias for small number of sites ($n \leq 100$).

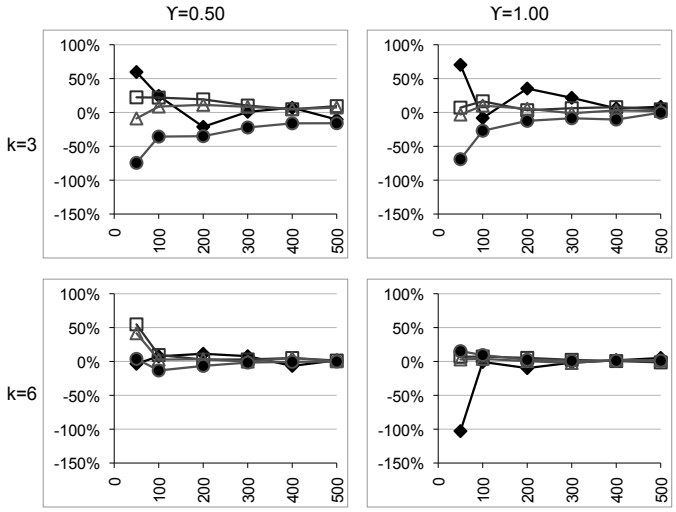


Figure 7.5: Median relative bias of the occupancy model parameters for the simulated cases with separable covariates. Filled diamonds represent the median % bias of the intercept (β_0), empty squares represent median % bias of the parameter associated with the continuous covariate(β_1), empty triangles represent the median % bias of the parameter associated with the binary covariate (β_2), and finally, the filled circles represent the median % bias of the dependence parameter (γ).

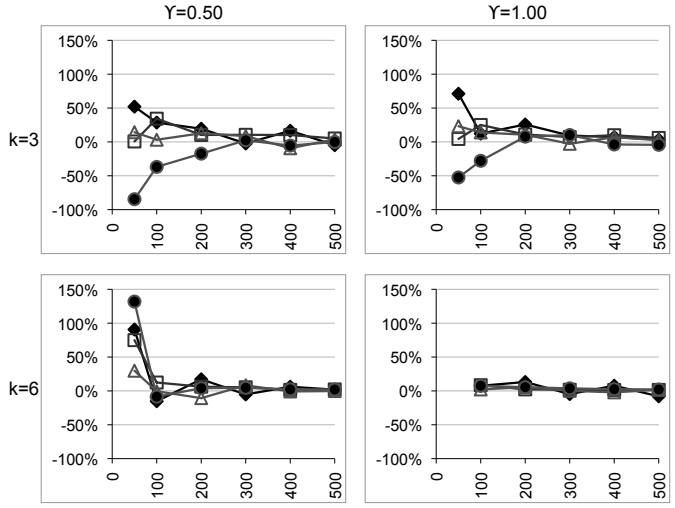


Figure 7.6: Median relative bias of the occupancy model parameters for the simulated cases with a common discrete covariate. Filled diamonds represent the median % bias of the intercept (β_0), empty squares represent median % bias of the parameter associated with the continuous covariate(β_1), empty triangles represent the median % bias of the parameter associated with the binary covariate (β_2), and finally, the filled circles represent the median % bias of the dependence parameter (γ).

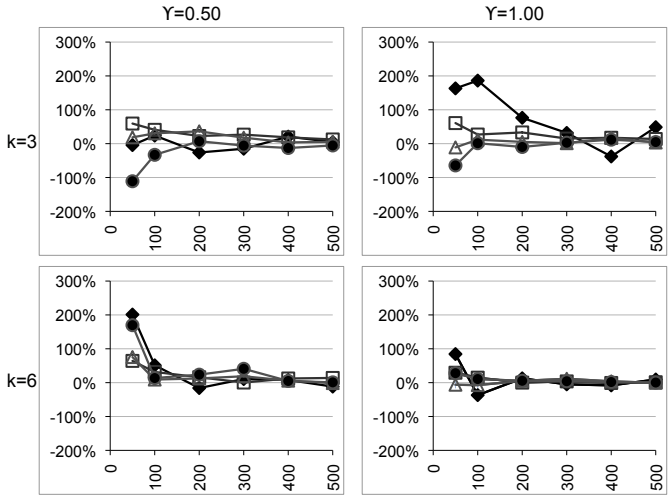
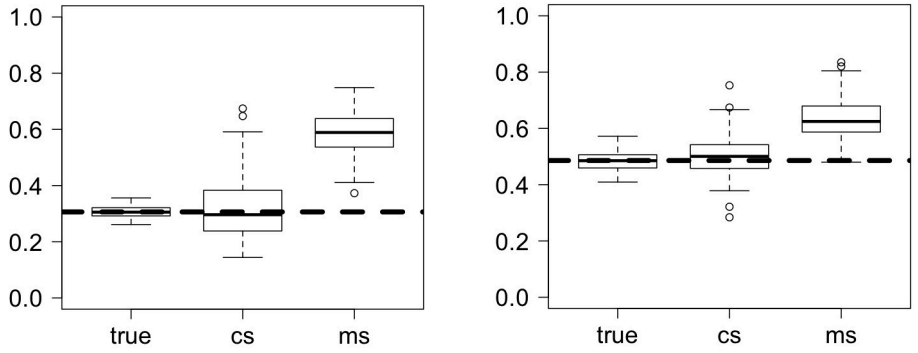


Figure 7.7: Median relative bias of the occupancy model parameters for the simulated cases with a common continuous covariate. Filled diamonds represent the median % bias of the intercept (β_0), empty squares represent median % bias of the parameter associated with the continuous covariate(β_1), empty triangles represent the median % bias of the parameter associated with the binary covariate (β_2), and finally, the filled circles represent the median % bias of the dependence parameter (γ).

We also wanted to compare how the estimates would change when the spatial correlation is ignored and the observations from every sample unit are assumed to be replicate observations from a single sample unit. For this we used a standard Zero Inflated Binomial model[73]. The table 7.1 shows the mean estimates and the standard errors for the parameters using both approaches: the cluster sampling model and the standard site occupancy model. It is found that ignoring the correlation between the surveyed locations leads to bias estimates of the parameters and large standard errors. It is also found that ignoring the spatial correlations leads to overestimation of the mean probability of occupancy (Figure 7.8).



(a) low spatial dependence ($\gamma = 0.5$) (b) Large spatial dependence ($\gamma = 1$)

Figure 7.8: Mean estimated occupancy for 100 simulated data sets with 100 sites and k=6. True: true distribution, CS: estimated values using the Cluster Sampling model, MS: estimated values using the Zero Inflated Binomial. When using the ZIB the mean probability of occupancy is overestimated.

Table 7.1: Summary statistics for 100 simulated data sets, k=6 and using Cluster Sampling model and the Zero Inflated Binomial (ZIB)

(a) Low spatial dependence

		Cluster Sampling		ZIB	
real		mean	se	mean	se
β_0	-0.40	-0.32	1.15	3.62	10.98
β_1	0.90	1.07	0.58	2.77	8.30
β_2	-1.20	-1.32	0.97	-2.94	6.14
θ_0	-0.50	-0.53	0.61	-1.50	0.22
θ_1	1.00	1.08	0.32	0.73	0.16
θ_2	-1.00	-0.97	0.49	-0.69	0.32
γ	0.50	0.42	0.46		

(b) Large spatial dependence

		Cluster Sampling		ZIB	
real		mean	se	mean	se
β_0	-0.40	-0.45	0.78	1.51	2.39
β_1	0.90	0.76	0.54	0.69	1.02
β_2	-1.20	-1.13	0.85	-0.92	2.24
θ_0	-0.50	-0.60	0.32	-0.98	0.21
θ_1	1.00	1.01	0.17	0.85	0.13
θ_2	-1.00	-0.93	0.30	-0.80	0.24
γ	1.00	1.11	0.33		

In conclusion, it is found that the parameters for the cluster sampling model are identifiable. The simulations show that the estimates are unbiased when the number of visited sites is larger than 50, especially if the size of the cluster is large (k=6).

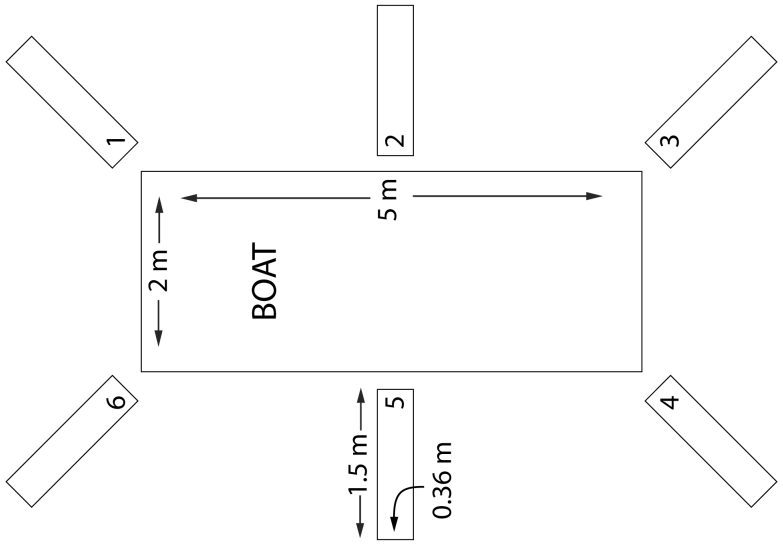


Figure 7.9: Illustration of the sampling methodology for the coontail species.

7.3 Example data analysis

The cluster sampling model was illustrated using the data from the submersed aquatic vegetation species *Ceratophyllum demersum L.* (coontail). This species is thought to be native in the U.S. and distributed in the Atlantic coastal plain, the Mississippi watershed and the Great Lakes region [31, 112].

The observations used in this analysis were collected during 2009 by the Upper Mississippi River's Long Term Resource Monitoring Program [50]. To sample the vegetation, a 3 meters long rake was dragged for 1.5 meters over riverine substrate and then removed. After removal, an observer noted if the species was present or not. Sampling was conducted at sites located no closer than 50 meters apart. At each sampling site, the boat was stopped and the rake method used at 6 locations within the sampling plot (see Figure 7.9). Further sampling information can be found in Yin et al. (2000)[120].

The following are the covariates that were considered for the analysis:

• Julian date: Sampling was conducted over an approximately 5 week period.

• Observer id: Identifier for the person who inspected the sample to determine the presence/absence of the species.

• Substrate: Ranges from 1 to 4: 1 indicates substrates that are predominately silt and/or clay, 2 indicates substrates that are mostly silt but have some sand, 3 indicates substrates that are mostly sand but have some silt or clay, and 4 indicates substrates that

Table 7.2: AIC of the 10 best candidate models for the coontail species. The models are ordered from the smallest AIC (best model) to the largest AIC, * indicates which covariates were included in the model. Model 1 is the selected model since it has the largest support.

		model									
		1	2	3	4	5	6	7	8	9	10
$\psi(\cdot)$	*intercept*	*	*	*	*	*	*	*	*	*	*
	julian	*		*	*	*	*	*	*	*	*
	$julian^2$	*				*	*	*	*	*	*
	depth	*	*	*	*	*	*	*	*	*	*
	$depth^2$	*	*	*	*	*	*	*	*		*
	east		*	*		*	*	*	*	*	*
	north		*					*	*	*	*
	substrate	*	*	*	*	*	*	*	*	*	*
	distance	*	*		*	*	*	*	*	*	*
$\delta(\cdot)$	*intercept*	*	*	*	*	*	*	*	*	*	*
	observer	*	*	*	*	*	*	*	*	*	*
	substrate	*	*	*	*	*	*	*	*	*	*
	julian	*	*	*	*	*	*	*	*	*	*
	$julian^2$	*					*	*	*		
	depth	*	*	*	*	*	*	*	*	*	*
	$depth^2$	*	*	*	*	*	*	*	*	*	*
	distance	*	*			*	*	*	*	*	*
	north										
	east										
	AIC	996	998	1000	1002	1003	1003	1006	1006	1007	1009

are gravel, rock or sand. This covariate was considered ordinal, assigning the larger the number to the coarser substrate.

- River mile: Distance from the Ohio river.
- Coordinates (north and east).
- Water Depth : Distance from the surface of the water to the settlement.

Different models were considered and the best model was selected using Akaike's Information Criterion(AIC)[11]. Table 7.2 shows the 10 best models and their corresponding AIC. It was found, according to the best model (model 1 in table 7.2), that the probability of occupancy is associated with the Julian date, the distance, the depth and the substrate. In addition, the probability of detection was found to be associated with observer, Julian date, distance, depth and substrate. The identifiability of the parameters for this model was verified using data cloning (see figure7.10).

Table 7.3 presents the estimated parameters and their corresponding confidence intervals. These confidence intervals were calculated using 200 bootstrap samples. It is found that all the parameters are significantly different from zero, except for the quadratic term associated

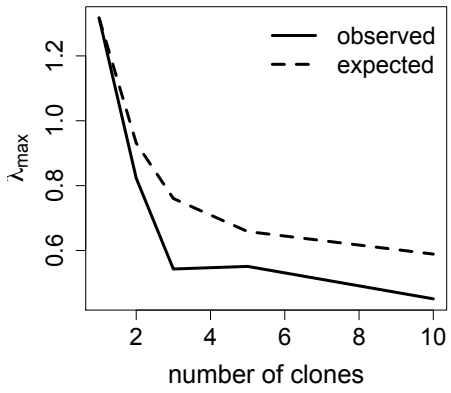


Figure 7.10: Maximum eigenvalue (λ_{max}) of the variance covariance matrix for different number of clones (1, 2, 3, 4, 5 and 10). . Solid line represents the obtained eigenvalues, dotted line represent the expected eigenvalues if the parameters are identifiable. It is found that λ_{max} decreases as the number of clones increases indicating that the parameters are identifiable.

with the Julian date for the probability of occupancy.

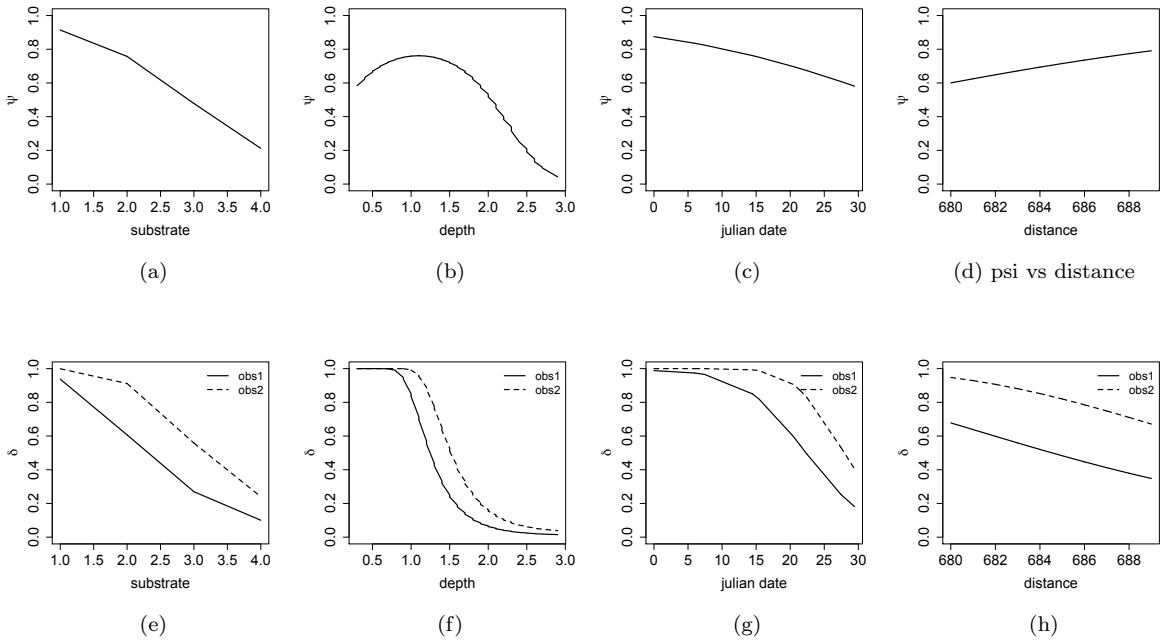
Figure 7.11 displays the relationship of the covariates with the probabilities of occupancy and detection. It appears, according to the best model (model 1; table 7.2), that the distance from the river has a weak positive relationship with the probability of occupancy (figure 7.11d). This relationship is expected since sites at smaller distances are farther downstream, therefore farther from the source. On the other hand, although it is expected for the probability of occupancy to increase with the Julian date, the estimated parameters tell a different story. According to figure 7.11c, the probability of occupancy is negatively associated with the Julian date. This result can be explained by the fact that the last visited sites (larger Julian dates) were more distant from the source, and so, although the prevalence of the coontail species is expected to increase as the time goes by, it may not be the case for sites located far from the source. The substrate was found to have a negative relationship with the probability of occupancy (figure 7.11a). To interpret this result it is important to recall that the larger the substrate value, the coarser the substrate. This relationship can be a reflection of lack of flow, since the probability of portions of coontail settling and staying in place will rise as flow decreases. The depth at the sample location was found to have a significant quadratic relationship with the probability of occupancy, for which the largest probability of occupancy was found at approximately 1.2 meters of depth. The probability of occupancy is nearly zero for values of depth larger than 2.5 meters. The

Table 7.3: Estimated parameters and confidence intervals on the logit scale for the coontail species using model 1 of table 7.2.

			90% CI	
		estimate	LL	UL
$\psi(\cdot)$	*intercept*	1.128	0.879	1.378
	julian	-0.530	-0.794	-0.257
	$julian^2$	-0.006	-0.202	0.195
	depth	-0.242	-0.435	-0.054
	$depth^2$	-0.285	-0.422	-0.160
	substrate	-1.174	-1.312	-1.023
	distance	0.259	0.081	0.432
	γ	0.894	0.861	0.927
$\delta(\cdot)$	*intercept*	-0.276	-0.501	-0.052
	observer	0.955	0.782	1.129
	substrate	-1.042	-1.145	-0.941
	julian	-1.320	-1.517	-1.119
	$julian^2$	-0.281	-0.435	-0.122
	depth	-1.720	-1.858	-1.570
	$depth^2$	0.194	0.088	0.308
	distance	-0.272	-0.402	-0.133

corrected mean probability of occupancy at every location was estimated to be 0.562, while the naive estimate was 0.336. The corrected mean probability of occupancy at the cluster level was estimated to be 0.723 while the naive estimated was 0.561.

Figures 7.11e, 7.11f, 7.11g, and 7.11h display the relationship of the substrate, depth, Julian date and the distance with the probability of detection for each observer. It was found that the observer coded with the number 2 performed consistently better than the observer coded with the number 1. It was also found that the relationship of the substrate, depth and Julian date with the probability of detection resembles that of the same covariates with the probability of occupancy. This leads us to conclude that most of the covariates support the hypothesis that an increase in the prevalence of the coontail species can lead to an increase in the chances for the species to be detected. However, it was also found that the association of distance with the probability of detection is the opposite of that of the distance with the probability of occupancy. Such a results are a likely an effect of the sampling schedule, for which sites at small distances (680 meters) were sampled last, when the coontail was bigger and more noticeable.



(a) (b) (c) (d) psi vs distance

(e) (f) (g) (h)

Figure 7.11: Covariates effects over the probability of occupancy and detection for the coontail species model.

7.4 Summary

In this chapter we introduced a model that allows estimation of the probability of occupancy using information from a single survey collected at sites that are expected to exhibit correlation. Our simulation study showed that for the cases under consideration the MLE of the parameters are consistent and identifiable.

We also discussed how the identifiability of any model can be tested by using readily available methods. We believe that checking the identifiability of the parameters of any model should be a regular practice that must be conducted before making any inference about the system under study.

The cluster sampling model provides another tool to correctly estimate the probability of occupancy when the surveys are not independent. The advantage of this model over the one introduced by Hines et al. (2010) is that our model provides estimates of the probability of occupancy at the sample unit level and at the location level, while Hines' model only allows estimation of the probability of occupancy at the sample unit level. On the other hand, the cluster sampling can be used to test the assumption of independece of surveys for

data collected under the multiple surveys sampling protocol. The test can be conducted by comparing the support of every model using a Likelihood Ratio Test.

Our model is constrained by the size of the cluster: in order to find the MLE it is necessary to obtain a closed form expression of the likelihood function, which results in an exponential increase of the complexity with the size of the cluster. Using our program, the largest sample unit size we were able to estimate parameters from was 10. Further research is required to determine the feasibility of using approximations to the likelihood to correctly estimate the parameters.

Chapter 8

Future research

In this chapter we introduce a couple of extensions to the models discussed in previous chapters, together with a brief description of how the MPLE can be incorporated to the estimation of abundance using capture-recapture models. In section 8.1 we introduced a model that using information from a single survey allows estimation of probability of occupancy for species that co-occupied the same territories. This section also contains some preliminary results from the simulation study conducted to evaluate this model performance under different scenarios. In section 8.2 we introduced the concept of Resource Selection Probability Function (RSPF) and a model that allows to estimate it while accounting for the detection error. Finally, section 8.3 contains a brief introduction to the capture-recapture models used to estimate abundance of a population and some ideas on how the MPLE can be used to improve its estimates.

8.1 Multiple species single survey model

An important area of research in ecology is the study of communities (see section 2.3), where one of its goals is to understand interspecific interactions between species. In section 3.3 we discussed the two main contributions for modeling site occupancy probability for the co-occurrence of multiple species (MacKenzie et al. (2004)[71] and Waddle et al. (2010)[115]). Both models require multiple and independent surveys of the same site.

Now we introduce a model to estimate the co-occurrence of multiple species using information from a single survey. The model we propose is to be applied to sampling schemes where n sites are surveyed one time to determine the presence/absence of at least two tar-

get species. The details of the statistical model are presented followed by a summary of the results of the simulation study that was conducted with the purpose of assessing the performance of the model.

8.1.1 Statistical model

The co-occurrence model

Let's assume n sites are surveyed once to monitor S species in the study area. The true state of the j^{th} species on the i^{th} site is a binary variable denoted by z_{ij} , where $z_{ij} = 1$ indicates the j^{th} species is present at the i^{th} site, and $z_{ij} = 0$ indicates the j^{th} species is absent from the i^{th} site. Let's also assume the conditional probabilities can be calculated:

$$Pr\left(z_{ij}/\{z_{im}:m\neq j\}\right) for\ i = 1, ...n\ and\ j = 1, ..., S \tag{8.1}$$

These probabilities indicate that the presence/absence of the j^{th} species on the i^{th} site is conditioned by the presence/absence of all other species on that same site. The stochastic process defined by these conditional probabilities was introduced by Besag (1974)[8] and is know as a Random Markov Field. The conditional probability mass function can be written as follows:

$$f\left(z_{ij}/\{z_{im}:m\neq j\}\right) = \frac{exp\left(z_{ij}A_{ij}\left(z_{im} : m \neq j, \underline{x}_i\right)\right)}{1 + exp\left(z_{ij}A_{ij}\left(z_{im} : m \neq j, \underline{x}_i\right)\right)} \tag{8.2}$$

where $\underline{x}_i$ are the set of habitat covariates associated with the i^{th} site and A_{ij} is called the natural parameter function. Besag (1974) showed that the natural parameter function for the binary case must be of the form

$$A_{ij}\left\{z_{im} : m \neq j, \underline{x}_i\right\} = \underline{x}_i^T \underline{\beta}_j + \sum_{m\neq j} \gamma_{mj} z_{im} \tag{8.3}$$

where $\underline{\beta}_j$ for $j = 1, \ldots, S$ is a vector of parameters that relate the habitat covariates with the probability of occupancy of the j^{th} species, and γ_{mj} (for $m = 1, \ldots, S$ and $j = 1, \ldots, S$) are the statistical dependence parameters that account for the interaction between species, for which $\gamma_{mj} = \gamma_{jm}$ for all $j = 1, \ldots, S$ and $m = 1, \ldots, S$.

For ease of explanation we present the model for two species. Nevertheless, the model presented below can be easily modified to include a larger number of species. For the two species case, the conditional probabilities can be written as

$$Pr\left(z_{i1}=z_{i1}/z_{i2}=z_{i2}\right)=\frac{exp\left(z_{i1}\left(\underline{x}_i^T\underline{\beta}_1+\gamma z_{i2}\right)\right)}{1+exp\left(\underline{x}_i^T\underline{\beta}_1+\gamma z_{i2}\right)} \tag{8.4}$$

$$Pr\left(z_{i2}=z_{i2}/z_{i1}=z_{i1}\right)=\frac{exp\left(z_{i2}\left(\underline{x}_i^T\underline{\beta}_2+\gamma z_{i1}\right)\right)}{1+exp\left(\underline{x}_i^T\underline{\beta}_2+\gamma z_{i1}\right)} \tag{8.5}$$

from where γ can be written as a function of the odds for the species 1 to be present

$$\gamma=log\left(\frac{Pr\left(z_{i1}=1/z_{i2}=1\right)}{Pr\left(z_{i1}=0/z_{i2}=1\right)}\right)-log\left(\frac{Pr\left(z_{i1}=1/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)}\right) \tag{8.6}$$

It is clear then that $\gamma>0$ indicates that the odds for the species 1 to be present at the i^{th} site are larger if the species 2 is present at the i^{th} site. On the other hand, if $\gamma<0$ then the odds for the species 1 to be present are larger if the species 2 is absent from the i^{th} site. Finally, $\gamma=0$ indicates that the occupancy probabilities for the two species are independent of each other.

Assuming that the presence/absence of the two species of interest on the i^{th} site are distributed according to the joint distribution $Pr\left(z_{i1},z_{i2}\right)$, the Hammersley-Clifford theorem (Hammersley and Clifford, 1971) establishes that the form that this joint distribution must take for it to be the joint probability measure of a Markov random field. Applying the Hammersley- Clifford theorem for the two species case the joint distribution is:

$$\frac{Pr\left(z_i=(1,1)\right)}{Pr\left(z_i=(0,0)\right)}=\frac{Pr\left(z_{i1}=1/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)}\cdot\frac{Pr\left(z_{i2}=1/z_{i1}=1\right)}{Pr\left(z_{i2}=0/z_{i1}=1\right)}=\frac{\psi_{i11}}{\psi_{i00}}=exp\left(\underline{x}_i^T\underline{\beta}_1+\underline{x}_i^T\underline{\beta}_2+\gamma\right) \tag{8.7}$$

$$\frac{Pr\left(z_i=(0,1)\right)}{Pr\left(z_i=(0,0)\right)}=\frac{Pr\left(z_{i1}=0/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)}\cdot\frac{Pr\left(z_{i2}=1/z_{i1}=0\right)}{Pr\left(z_{i2}=0/z_{i1}=1\right)}=\frac{\psi_{i01}}{\psi_{i00}}=exp\left(\underline{x}_i^T\underline{\beta}_2\right) \tag{8.8}$$

$$\frac{Pr\left(z_i=(1,0)\right)}{Pr\left(z_i=(0,0)\right)}=\frac{Pr\left(z_{i1}=1/z_{i2}=0\right)}{Pr\left(z_{i1}=0/z_{i2}=0\right)}\cdot\frac{Pr\left(z_{i2}=0/z_{i1}=1\right)}{Pr\left(z_{i2}=0/z_{i1}=1\right)}=\frac{\psi_{i10}}{\psi_{i00}}=exp\left(\underline{x}_i^T\underline{\beta}_1\right) \tag{8.9}$$

It follows that

$$Pr\left(\underline{z}_i=(0,0)\right)=\psi_{i00}=\frac{1}{1+exp\left(\underline{x}_i^T\underline{\beta}_1+\underline{x}_i^T\underline{\beta}_2+\gamma\right)+exp\left(\underline{x}_i^T\underline{\beta}_2\right)+exp\left(\underline{x}_i^T\underline{\beta}_1\right)} \tag{8.10}$$

and marginal probability of occupancy of the species 1 and 2 at the i^{th} site ($\psi_{i,s=1}$ and $\psi_{i,s=2}$ respectively) can be easily calculated from the joint distribution as

$$\psi_{i,s=1} = \psi_{i10} + \psi_{i11} \tag{8.11}$$

$$\psi_{i,s=2} = \psi_{i01} + \psi_{i11} \tag{8.12}$$

Detection error model

Nevertheless, it is likely for the species to be present but not detected during the survey, hence the true state of the species is not directly observable. The observations collected during the survey are denoted by the binary variable y_{ij} ,where y_{ij} equals 1 if the j^{th} species is detected at the i^{th} site, and 0 if the j^{th} species is not detected at the i^{th} site. Let $\delta_{ij} = Pr\left({}^{y_{ij}=1}/_{z_{ij}=1}\right)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, S$ be the probability of detecting the j^{th} species at the i^{th} site given that the j^{th} species is present. These probabilities can depend on covariates such as time of the day, weather conditions and the presence/absence of other species. Let us denote by $\underline{w}_i$ the set of covariates at the i^{th} site that are thought to be related to the probability of detection. These covariates can be incorporated into the probability of detection by using any link function (e.g. the Logit link, Complement Log-log link or the Probit link). For instance, using the Complement Log-log link, the probability of detection for the species 1 can be written as follows:

$$\delta_{i1/(z_{i1},z_{i2})} = Pr\left({}^{y_{i1}=1}/_{z_{i1}=1,z_{i2},\underline{w}_i}\right) = 1 - exp\left(-exp\left(\underline{w}_i^T\underline{\theta}_1 + \eta z_{i2}\right)\right) \tag{8.13}$$

where η is the parameter that accounts for the dependence of the probability of detection between species and $\underline{\theta}_1$ is a vector of parameters that determine the effect of the covariates $\underline{w}_i$ over the probability of detection. The interpretation of the parameter η will then be: $\eta > 0$ indicates that the presence of the species 2 increases the chances of detecting the species 1, $\eta < 0$ indicates that the presence of the species 2 decreases the chances of detecting the species 1, and $\eta = 0$ indicates that the probability of detecting the species 1 is independent of the presence/absence of the species 2. Similarly, the probability of detecting the species 2 is:

$$\delta_{i2/(z_{i1},z_{i2})} = Pr\left({}^{y_{i2}=1}/_{z_{i2}=1,z_{i1},\underline{w}_i}\right) = 1 - exp\left(-exp\left(\underline{w}_i^T\underline{\theta}_2 + \eta z_{i1}\right)\right) \tag{8.14}$$

The probability mass function for the observations $\underline{y}_i = (y_{i1}, y_{i2})$ for $i = 1, ..., S$ is as

follows:

$$
\begin{aligned}
Pr\left(\underline{y}_i = (0,0)\right) &= \psi_{i00} + \left[\psi_{i10}\left(1 - \delta_{i1/(1,0)}\right)\right] + \left[\psi_{i01}\left(1 - \delta_{i2/(0,1)}\right)\right] + && (8.15)\\
&\quad \left[\psi_{i11}\left(1 - \delta_{i1/(1,1)} - \delta_{i2/(1,1)} + \left(\delta_{i1/(1,1)} \cdot \delta_{i2/(1,1)}\right)\right)\right] && (8.16)\\
Pr\left(\underline{y}_i = (1,0)\right) &= \left[\psi_{i10} \cdot \delta_{i1/(1,0)}\right] + \left[\psi_{i11} \cdot \left(\delta_{i1/(1,1)} - \left(\delta_{i1/(1,1)} \cdot \delta_{i2/(1,1)}\right)\right)\right] && (8.17)\\
Pr\left(\underline{y}_i = (0,1)\right) &= \left[\psi_{i01} \cdot \delta_{i2/(0,1)}\right] + \left[\psi_{i11} \cdot \left(\delta_{i2/(1,1)} - \left(\delta_{i1/(1,1)} \cdot \delta_{i2/(1,1)}\right)\right)\right] && (8.18)\\
Pr\left(\underline{y}_i = (1,1)\right) &= \psi_{i11} \cdot \delta_{i1/(1,1)} \cdot \delta_{i2/(1,1)} && (8.19)
\end{aligned}
$$

Assuming that the observations between sites are independent, the likelihood is then calculated as the product of the probabilities of all the detection histories collected during the survey

$$
\begin{aligned}
L\left(\underline{\beta}_1, \underline{\beta}_2, \underline{\theta}_1, \underline{\theta}_2, \gamma, \eta; \underline{y}\right) &= \prod_{i=1}^{n} \left(\nu_{i00}\right)^{I\left(\underline{y}_i=(0,0)\right)} \cdot \left(\nu_{i01}\right)^{I\left(\underline{y}_i=(0,1)\right)} && (8.20)\\
&\quad \cdot \left(\nu_{i10}\right)^{I\left(\underline{y}_i=(1,0)\right)} \cdot \left(\nu_{i11}\right)^{I\left(\underline{y}_i=(1,1)\right)} && (8.21)
\end{aligned}
$$

where $I(\cdot)$ is the indicator function that is equal to one if its argument is true and zero otherwise, and, ν_{i00}, ν_{i10}, ν_{i01} and ν_{i11} are equal to $Pr\left(\underline{y}_i = (0,0)\right)$, $Pr\left(\underline{y}_i = (0,1)\right)$, $Pr\left(\underline{y}_i = (1,0)\right)$ and $Pr\left(\underline{y}_i = (1,1)\right)$ respectively. The MLE are obtained by maximizing $I(\cdot)$ with respect to the parameters $\left(\underline{\beta}_1, \underline{\beta}_2, \underline{\theta}_1, \underline{\theta}_2, \gamma, \eta\right)$.

If the number of sites is large, any optimization method can be used to obtain the MLE. Nevertheless if the number of sites is small, a penalty function similar to the one described in previous chapters can be effective on improving the estimated parameters for this model. The algorithm 8.1describes how to obtain the MPLE for the multiple species model.

8.1.2 Simulation study

The following simulation study was conducted to evaluate the performance of the co-occurrence model. Two cases were considered and analyzed. For the first case it was assumed that one species is detected independently of the presence/absence of the other species and vice versa, and for the second case it was assumed that the probability of detecting one of the species depended on the presence/absence of the other species and vice versa.

Algorithm 8.1 Maximum Penalized Likelihood Estimates (MPLE) for the multiple species single survey model

1. Obtain the MLE for $\left(\underline{\beta}_1, \underline{\beta}_2, \underline{\theta}_1, \underline{\theta}_2, \gamma, \eta\right)$ by maximizing the likelihood function. Let us denote these by $\left(\hat{\underline{\beta}}_1^M, \hat{\underline{\beta}}_2^M, \hat{\underline{\theta}}_1^M, \hat{\underline{\theta}}_2^M, \hat{\gamma}^M, \hat{\eta}^M\right)$.

2. Obtain the naïve estimator of $\hat{\underline{\beta}}_{\underline{j}}$ for j=1, 2 by maximizing

$$L\left(\underline{\beta}_j; \underline{y}_{\cdot j}\right) = \prod_{i=1}^{n} \left(\frac{exp\left(\underline{x}_i \underline{\beta}_j\right)}{1 + exp\left(\underline{x}_i \underline{\beta}_j\right)}\right)^{y_{ij}} \left(\frac{1}{1 + exp\left(\underline{x}_i \underline{\beta}_j\right)}\right)^{1-y_{ij}}$$

   This estimator is based on the assumption that there is no detection error and there is no interaction between species.

3. Obtain the naïve estimator of $\hat{\theta}_j^{naive}$ by maximizing

$$L\left(\underline{\theta}_j; \underline{y}_{\cdot j}\right) = \prod_{i=1}^{n} \varphi\left(\underline{w}_i, \underline{\theta}_j\right)^{y_{ij}} \left(1 - \varphi\left(\underline{w}_i, \underline{\theta}_j\right)\right)^{1-y_{ij}}$$

   where $\varphi(\cdot)$ is the link function used for the probability of detection. This estimator is based on the assumption that all sites are occupied and that the probability of detection of every species is independent of each other.

4. Maximize the penalized likelihood function with respect to $\left(\underline{\beta}_1, \underline{\beta}_2, \underline{\theta}_1, \underline{\theta}_2, \gamma, \eta\right)$

$$\begin{aligned} log\left(PL\left(\underline{\beta}_1, \underline{\beta}_2, \underline{\theta}_1, \underline{\theta}_2, \eta, \gamma\right)\right) &= log\left(L\left(\underline{\beta}_1, \underline{\beta}_2, \underline{\theta}_1, \underline{\theta}_2, \eta, \gamma\right)\right) \\ &- \sum_{j=1}^{2} \lambda_j \cdot \left|\underline{\beta}_j - \underline{\beta}_j^{naive}\right| - \sum_{j=1}^{2} \kappa_j \cdot \left|\underline{\theta}_j - \underline{\theta}_j^{naive}\right| \end{aligned}$$

   where

$$\lambda_j = \left(1 - \hat{\psi}_j^{naive}\right) \hat{\delta}_j^M \sqrt{tr\left(var\left(\hat{\theta}_j^M\right)\right)}$$

   and

$$\gamma_j = \left(1 - \hat{\delta}_j^{naive}\right) \hat{\psi}_j^{naive} \sqrt{tr\left(var\left(\hat{\underline{\beta}}_j^M\right)\right)}$$

and $\left(\hat{\underline{\delta}}_j^{naive}, \hat{\underline{\psi}}_j^{naive}\right)$ and $\left(\hat{\underline{\delta}}_j^M, \hat{\underline{\psi}}_j^M\right)$denote the average occupancy and detection probabilities for the j^{th} species under the naïve method of estimation and MLE respectively. It is important to mention that the purpose of the penalization is to stabilize the likelihood function when the number of visited sites is small. As the number of visited site increases the penalty converges to zero, and the MPLE are the same as the MLE.

Case 1: independent probabilities of detection

Assuming only two species are of interest, the probability of occupancy for each species was set to depend both on a continuous covariate $x_i \sim Normal\,(0, 1.5)$ and the presence/absence of the other species:

$$Pr\left({}^{z_{i1}=z_{i1}}/_{z_{i2}=z_{i2}}\right) = \frac{exp\left(z_{i1}\left(\alpha_{11}+\alpha_{12}x_i+\beta z_{i2}\right)\right)}{1+exp\left(\alpha_{11}+\alpha_{12}x_i+\beta z_{i2}\right)} \tag{8.22}$$

$$Pr\left({}^{z_{i2}=z_{i2}}/_{z_{i1}=z_{i1}}\right) = \frac{exp\left(z_{i2}\left(\alpha_{21}+\alpha_{22}x_i+\beta z_{i1}\right)\right)}{1+exp\left(\alpha_{21}+\alpha_{22}x_i+\beta z_{i1}\right)} \tag{8.23}$$

In addition, the probability of detection for each species was assumed to depend on a continuous covariate $w_i \sim Normal\,(0, 1)$.

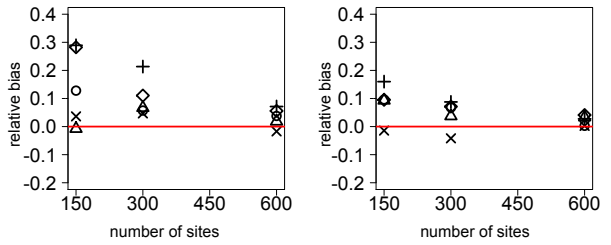
$$\delta_{i1} = Pr\left({}^{y_{i1}=1}/_{z_{i1}=1}\right) = 1 - exp\left(-exp\left(\theta_{11}+\theta_{12}w_i\right)\right) \tag{8.24}$$

$$\delta_{i2} = Pr\left({}^{y_{i2}=1}/_{z_{i2}=1}\right) = 1 - exp\left(-exp\left(\theta_{21}+\theta_{22}w_i\right)\right) \tag{8.25}$$

The values of the probability of occupancy and the number of visited sites were varied to obtain a total of 12 simulation sets (i.e. four different levels of probability of occupancy and three different sizes for number of visited sites). The levels for the probability of occupancy were selected considering both the marginal probability of occupancy for every species and the strength of the interaction between the two species. Every simulated level of the probability can be denoted by a vector $\left\{\overline{\psi}_{s=1}, \overline{\psi}_{s=2}, \overline{\psi}_{both}\right\}$ where $\overline{\psi}_{s=1}$ is the mean marginal probability of occupancy for the species 1, $\overline{\psi}_{s=2}$ is the mean marginal probability of occupancy for the species 2, and $\overline{\psi}_{both}$ is the mean probability of occupancy for both species. The four simulated levels were $\{0.40, 0.40, 0.10\}$, $\{0.39, 0.40, 0.22\}$, $\{0.63, 0.66, 0.43\}$, and $\{0.78, 0.78, 0.58\}$. The number of visited sites were set at 150, 300 and 600. The mean probability of detection was set to be 0.31 for the species 1, and 0.27 for the species 2. The parameters were selected to obtain the mean probability of occupancy and detection according to the level.

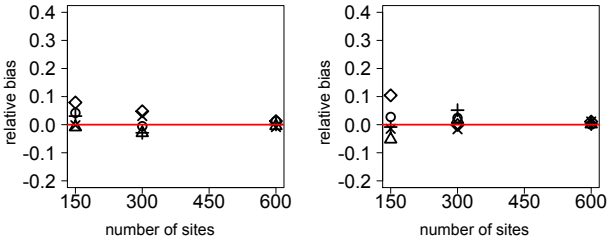
Figure 8.1 displays the median relative bias for the estimated probabilities. It is observed that the largest biases were obtained for the case in which the probability of occupancy for both species was lowest (figure 8.1a). However, for all the cases, as the number of sur-

veyed sites increased the bias decreased. Similar results were obtained for the parameters associated with the covariates. Figures 8.2 and 8.3 display the box plots of the estimated parameters for the low probability of occupancy case, for which the estimates of the parameters seem to be consistent. A complete summary of the results is presented on Appendix D.



(a) case1 :{0.40, 0.40, 0.10}

(b) case1 :{0.39, 0.40, 0.22}



(c) case1 :{0.63, 0.66, 0.43}

(d) case1 :{0.78, 0.78, 0.58}

Figure 8.1: Median relative bias for the estimated probabilities. The "o" represent $\psi_{s=1}$, the "△" represent $\psi_{s=2}$, the "+" represent ψ_{11} , "x" equises represent δ_1 and the "♢" represent δ_2.

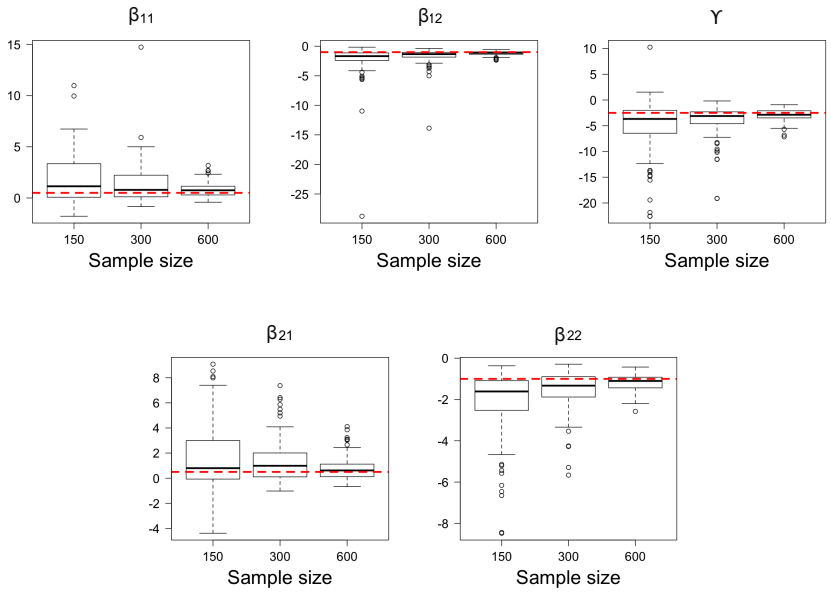


Figure 8.2: Box plots of estimated parameters for the probability of occupancy model. Summary of the 100 estimated parameters obtained when simulating the case $\{0.40, 0.40, 0.10\}$ for two species. For these simulations was assumed that the probability of detection of every species was independent of each other.

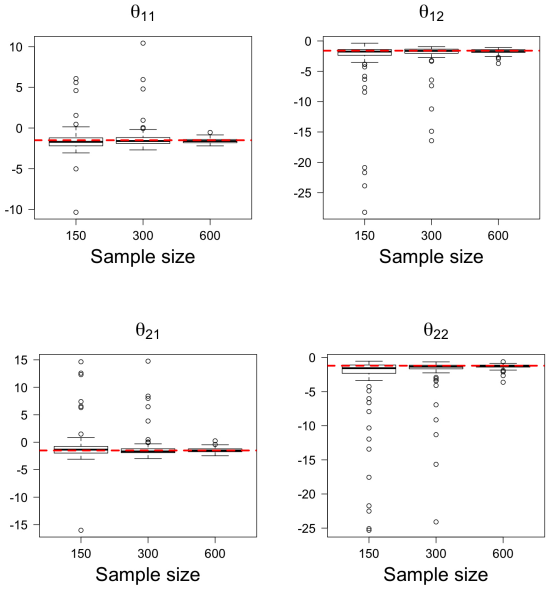


Figure 8.3: Box plots of estimated parameters for the probability of detection model. Summary of the 100 estimated parameters obtained when simulating the case $\{0.40, 0.40, 0.10\}$ for two species. For these simulations was assumed that the probability of detection of every species was independent of each other.

Case 2: dependent probabilities of detection

Similar to the first case, we assumed only two species were of interest. In this case, not only the probability of occupancy for a species depended on the presence/absence of the other species, but also the probability of detection of each species depended on the presence/absence of the other species.

$$\delta_{i1/zi2} = Pr\left(y_{i1}=1/z_{i1}=1\right) = 1 - exp\left(-exp\left(\theta_{11} + \theta_{12}w_i + \gamma_{12}z_{i2}\right)\right) \quad (8.26)$$

$$\delta_{i2/zi1} = Pr\left(y_{i2}=1/z_{i2}=1\right) = 1 - exp\left(-exp\left(\theta_{21} + \theta_{22}w_i + \gamma_{21}z_{i1}\right)\right) \quad (8.27)$$

Only one set of simulation were conducted. The parameters were selected in order to obtain small probabilities of occupancy but moderate levels for the probability of detection. The resulting probabilities were $\{\psi_{s=1} = 0.282, \psi_{s=2} = 0.280, \psi_{both} = 0.108\}$ and

$\left\{\overline{\delta}_{1/z_2=0} = 0.667, \overline{\delta}_{2/z_1=0} = 0.660, \overline{\delta}_{1/z_2=1} = 0.70, \overline{\delta}_{2/z_2=1} = 0.62\right\}$.

Unfortunately, the results of the simulation study were not satisfactory because for some of the generated data sets the parameters were not identifiable. Figure 8.4 displays the maximum eigenvalue of the variance-covariance matrix against the number of clones for four simulated data sets using 150 sites. It was found that while for two data sets the parameters were identifiable, for the other two data set they were not. Further research is required to determine whether a different parameterization of the model could lead toward better results.

8.2 RSPF with detection error

Biologists are often interested on identifying what resources are used by an animal and what the availability of those resources is. This information is crucial to develop plans for wildlife management and conservation. One of the tools commonly used to collect such information is the Resource Selection Probability Function (RSPF), a function that allows estimation of the probability for a specific resource to be used by an animal [77, page 27] (Manly et al: 2002:27).

Generally the sampling protocol used in these studies is the use-availability design (Sampling Protocol A, design I [77, page 15]). In this design it is assumed that a random sample of N locations is taken from the population of used sites, and that a random sample of M

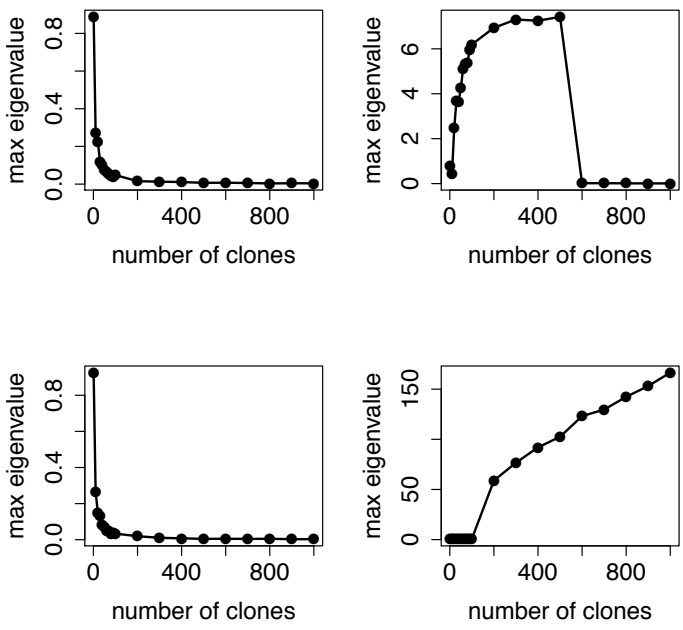


Figure 8.4: Maximum eigenvalue vs the number of clones for 4 simulated data sets for two species and dependent probabilities of detection.

locations is taken from the available sites. The latter sample will contain besides the unused sites other sites that might potentially have been already used. Every site is characterized by environmental factors that can be denoted by $\underline{x}_i = (x_{i1}, ..., x_{ip})$. The goal of the RSPF analysis is to study how the environmental factors $(\underline{x}_i)$ affect the probability of use, denoted by $\pi\left(\underline{x}_i, \underline{\beta}\right)$.

Let Z_i be a binary variable that equals 1 if the i^{th} site is used and 0 if the i^{th} site is not used. The probability of use (i.e. the RSPF) can be defined as $\pi\left(\underline{x}_i, \underline{\beta}\right) = Pr\left({}^{Z_i=1}/_{\underline{X}=\underline{x}_i}\right)$ where $0 \leq \pi\left(\underline{x}_i, \underline{\beta}\right) \leq 1$ for all possible values of $\underline{x}_i$ and $\underline{\beta}$.

Johnson et al. (2006)[51] showed that the use-availability study design is properly formulated in terms of weighted distributions [89], and that when using the exponential RSPF $(\pi\left(\underline{x}_i, \underline{\beta}\right) = exp\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\right)$ for $\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \leq 0$) all the parameters but the intercept (β_0) can be estimated using the standard logistic regression. Lele and Keim (2006)[61] extended the ideas from Johnson et al. (2006) and demonstrated that other parametric forms such as the logistic, probit and log-log link (Equations 5.1, 5.2 and 5.3 in Manly et al. 2002[77]) allow estimation of all the parameters, hence the estimation of the absolute probabilities.

One of the assumptions for the RSPF estimates to be valid under the use-availability study design is that the used locations are correctly classified.

> "If animals use an area and are unobserved or sign of their use is not observed, then this assumption will be violated"[77, Page 14].

The violation of this assumption is usually referred as the detection error. This error can be defined as the probability that a site that has been used by an animal is not classified as used because, either the animal was not detected during the survey, or that no evidence of its use was found. For instance, consider the case in which the used locations are determined by using GPS collars. It has been shown that GPS collars can be biased because vegetation or terrain interfere with the satellite signals necessary to acquire a location[29]. Under this circumstance the sample of used sites can be biased. A statistical model that incorporates the detection error into the RSPF estimation is discussed in the following section.

8.2.1 Statistical model and estimation procedure

Let's assume that every location at the study area is characterized by environmental variables denoted by $\underline{x}_i$. These locations are surveyed to determine whether the site has been or is still being used by the target species. As a result, there will be N locations that will be classified as "used" and M locations for which evidence of use was not found . It is assumed that the covariates vectors $\underline{x}_i$ are a random sample from a multivariate distribution $f^A(\underline{x})$, where $f^A(\underline{x})$ is the probability distribution of the available resources over the study area. Furthermore, since it is assumed that the resources are used disproportionally to their availability, the probability distribution of the resources used by the animal can be denoted by $f^U(\underline{x})$, thus calculated as 8.28.

$$f^U\left(\underline{X}=\underline{x}_i/Z_i=1;\underline{\beta}\right) = \frac{Pr\left(Z_i=1/\underline{X}=\underline{x}_i;\underline{\beta}\right) f^A(\underline{x})}{Pr(Z_i=1)} \tag{8.28}$$

where $\pi\left(\underline{x}_i,\underline{\beta}\right)$ is the resource selection probability function (RSPF) at a location characterized with the vector of covariates $\underline{x}_i$, and $Z_i = 1$ indicates that the i^{th} location has been used by the animal. The goal of the analysis is to estimate the parameters $\underline{\beta}$, hence to understand the effect of the covariates over the probability for that resource to be selected.

Incorporating the detection error into the RSPF

Classifying a used location as such is seldom perfect. Moreover, the classification of a location as "used" is the result of two processes: first, the location is selected by the animal; and second, evidence of its use is observed during the survey. These two processes are reflected on the following probability distribution:

$$\begin{aligned} f^U\left(\underline{X}=\underline{x}_i/_{Y_i=1;\underline{\beta},\underline{\theta}}\right) &= \frac{Pr\left(^{Y_i=1}/_{Z_i=1,\underline{W}=\underline{w}_i;\underline{\theta}}\right) Pr\left(^{Z_i=1}/_{\underline{X}=\underline{x}_i;\underline{\beta}}\right) f^A\left(\underline{x},\underline{w}\right)}{Pr\left(Y_i=1\right)} \\ &= \frac{\delta\left(\underline{w}_i,\underline{\theta}\right)\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\underline{w}\right)}{\int \delta\left(\underline{w}_i,\underline{\theta}\right)\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\underline{w}\right) dxdw} \\ &= \frac{\delta\left(\underline{w}_i,\underline{\theta}\right)\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\underline{w}\right)}{P\left(\underline{\beta},\underline{\theta}\right)} \end{aligned} \tag{8.29}$$

where Y_i is a binary variable that equals 1 if a used site is correctly classified (i.e. evidence that the site has been used by the target species was detected during the survey); $\delta\left(\underline{w}_i,\underline{\theta}\right)$ is the probability of correctly classifying a used site, and $\underline{w}_i$ are the set of covariates that affect the probability of detection.

It is easy to show that when the probability of detection is constant and equal for all the locations (i.e., $\delta\left(\underline{w}_i,\underline{\theta}\right)=\delta$ for all i), the RSPF can be estimated using the same weighted distribution presented in Johnson et al. (2006) and Lele and Keim (2006) (equation 8.30).

$$f^U\left(\underline{X}=\underline{x}_i/_{Y_i=1;\underline{\beta},\underline{\theta}}\right) = \frac{\delta\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\underline{w}\right)}{\int \delta\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\underline{w}\right) d\underline{x}} = \frac{\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\underline{w}\right)}{P\left(\underline{\beta}\right)} \tag{8.30}$$

On the other hand, the probability of detection can vary according to covariates related to the environment or the sampling methodology. If the covariates the RSPF depends on are separable from the covariates the probability of detection depends on (i.e., there are no common covariates), then the probability distribution for the used sites can be written as 8.31.

$$\begin{aligned} f^U\left(\underline{X}=\underline{x}_i/_{Y_i=1;\underline{\beta},\underline{\theta}}\right) &= \frac{\delta\left(\underline{w}_i,\underline{\theta}\right)\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\right) f^A\left(\underline{w}\right)}{\int \delta\left(\underline{w}_i,\underline{\theta}\right)\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\right) f^A\left(\underline{w}\right)} \\ &= \frac{\pi\left(\underline{x}_i,\underline{\beta}\right) f^A\left(\underline{x},\right)}{P\left(\underline{\beta}\right)} \cdot \frac{\delta\left(\underline{w}_i,\underline{\theta}\right) f^A\left(\underline{w}\right)}{P\left(\underline{\theta}\right)} \end{aligned} \tag{8.31}$$

It can be shown that the estimates of the parameters $\underline{\beta}$ obtained when maximizing the likelihood for the latter probability distribution are the same as the ones obtained when ignoring the detection error. However, there is one case when the detection error must be accounted for, this is when the covariates related to the RSPF are related to the probability of detection too. In this case, the parameters should be estimated by maximizing the following likelihood:

$$L\left(\underline{\beta},\underline{\theta}/\underline{x}_1,...,\underline{x}_N,\underline{w}_1,...,\underline{w}_N\right)=\prod_{i=1}^{N}\frac{\delta\left(\underline{w}_i,\underline{\theta}\right)\pi\left(\underline{x}_i,\underline{\beta}\right)f^A\left(\underline{x}_i,\underline{w}_i\right)}{P\left(\underline{\beta},\underline{\theta}\right)} \tag{8.32}$$

This likelihood can be maximized using the method of simulated maximum likelihood (Robert et al. (1999)[94], Lele and Keim, (2006) [61]). This estimation method consists on maximizing the log-likelihood using an estimate value of $P\left(\underline{\beta},\underline{\theta}\right)$. The log-likelihood can be written as follows:

$$\begin{aligned} log\left(L\left(\underline{\beta},\underline{\theta}/\underline{x}_1,...,\underline{x}_N,,w_1,...,\underline{w}_N\right)\right) &= \sum_{i=1}^{N} log\left(\delta\left(\underline{w}_i,\underline{\theta}\right)\right)+log\left(\pi\left(\underline{x}_i,\underline{\beta}\right)\right) \\ &- log\left(P\left(\underline{\beta},\underline{\theta}\right)\right)+log\left(f^A\left(\underline{x}_i,\underline{w}_i\right)\right) \end{aligned} \tag{8.33}$$

where the last term $log\left(f^A\left(\underline{x}_i,\underline{w}_i\right)\right)$ does not depend on the parameters $\left(\underline{\beta},\underline{\theta}\right)$ and can be disregarded when maximizing the log-likelihood. However, $P\left(\underline{\beta},\underline{\theta}\right)$ is not known analytically and depends on the parameters of interest. A way around for this problem is to use an estimate of $P\left(\underline{\beta},\underline{\theta}\right)$ that can be obtained from a random sample of the available sites (8.34).

$$\widehat{P\left(\underline{\beta},\underline{\theta}\right)}=\frac{\sum_{j=1}^{M}\pi\left(\underline{x}_j,\underline{\beta}\right)\delta\left(\underline{w}_j,\underline{\theta}\right)}{M} \tag{8.34}$$

So the log-likelihood can be written as

$$\begin{aligned} \hat{l}\left(\underline{\beta},\underline{\theta}/\underline{x}_1,...,\underline{x}_N,,w_1,...,\underline{w}_N\right) &= \sum_{i=1}^{N} log\left(\delta\left(\underline{w}_i,\underline{\theta}\right)\right)+log\left(\pi\left(\underline{x}_i,\underline{\beta}\right)\right) \\ &- log\left(\widehat{P\left(\underline{\beta},\underline{\theta}\right)}\right) \end{aligned} \tag{8.35}$$

The simulated maximum likelihood estimators (SMLE) are obtained by maximizing equation 8.35 with respect to $\left(\underline{\beta},\underline{\theta}\right)$. This estimation method was implemented on a program written in R [113]. Lele (2009) [59] proposed an alternative estimation method that consisted of maximizing the partial likelihood function by using a data cloning algorithm [60]. This

method provides estimators more stable than those obtained by maximizing the simulated likelihood. More importantly, by using the partial likelihood and the data cloning algorithm it is possible to determine the identifiability of the parameters[62]. The partial likelihood function for the complete model can be denoted by $PL\left(\underline{\beta},\underline{\theta},\alpha\right)$. The maximum partial likelihood estimators (MPaLE) are obtained by maximizing $PL\left(\underline{\beta},\underline{\theta},\alpha\right)$ under the restriction $0 \leq \alpha \leq 1$. The complete derivation of the partial likelihood function (equation8.36) is presented on the Appendix E.

$$\begin{aligned} PL\left(\underline{\beta},\underline{\theta},\alpha\right) &= \prod_{i=1}^{N} \frac{r\pi\left(\underline{x}_i^U,\underline{\beta}\right)\delta\left(\underline{w}_i^U,\underline{\theta}\right)}{r\pi\left(\underline{x}_i^U,\underline{\beta}\right)\delta\left(\underline{w}_i^U,\underline{\theta}\right)+(1-r)\,\alpha} \\ &\cdot \prod_{j=1}^{M} \frac{(1-r)\,\alpha}{r\pi\left(\underline{x}_j^A,\underline{\beta}\right)\delta\left(\underline{w}_j^A,\underline{\theta}\right)+(1-r)\,\alpha} \end{aligned} \tag{8.36}$$

8.2.2 Simulation analysis

A simulation study was conducted with two purposes: first, to determine the effect of the detection error over RSPF estimates; and second, to evaluate the statistical properties of the estimates under the full model (i.e., considering the detection error). Three factors were considered: the type of the common covariate, the link used for the detection model, and the number of used sites. These three factors were combined for a total of 18 simulated cases. For the first factor two cases were considered: the common covariate is continuous or the common covariate is a binary variable. Three different link functions were considered for the detection model: logistic, complement log-log and probit. Finally, three different sample sizes were used for the simulations: 500, 1000 and 2000. The number of available sites was set constant and equal to 1000.

The Logistic RSPF was considered and set to depend on two covariates: a continuous covariate $x_{i1} \sim Normal\,(0,1)$ and a binary covariate $x_{i2} \sim Bernoulli\,(0.55)$.

$$\pi\left(\underline{x}_i,\beta\right) = \frac{exp\left(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}\right)}{1+exp\left(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}\right)} \tag{8.37}$$

The probability of detection was set to depend on two covariates as well. For the simulated cases in which a continuous covariate was common, the probability of detection was set to depend on x_{i1} and $w_{i2} \sim Bernoulli\,(0.65)$. For the simulated cases with a common discrete covariate the probability of detection was set to depend on $w_{i1} \sim Normal\,(0,1)$ and x_{i2}. For every simulated case, 500 data sets were generated and the SMLE and the MPaLE were

calculated.

A complete summary of the simulation results is presented in appendix F. The results of the simulation corroborate the hypothesis: when no common covariates exist between the probability of detection and the RSPF, the estimates of the RSPF obtained for the full model are the same as those obtained for the naive model (i.e when the detection error is ignored). On the other hand, for those cases where common covariates exist, the RSPF estimates obtained from the naive model are biased. Figure 8.5 displays the box plot of the MPaLE and naive estimates for the common continuous covariate case. Notice that naive estimates of the parameter associated with the common covariate (β_1) are underestimated, while the MPaLE of the full model are unbiased. Similar results are obtained for the SMLE although their variance is larger than that obtained for the MPaLE .

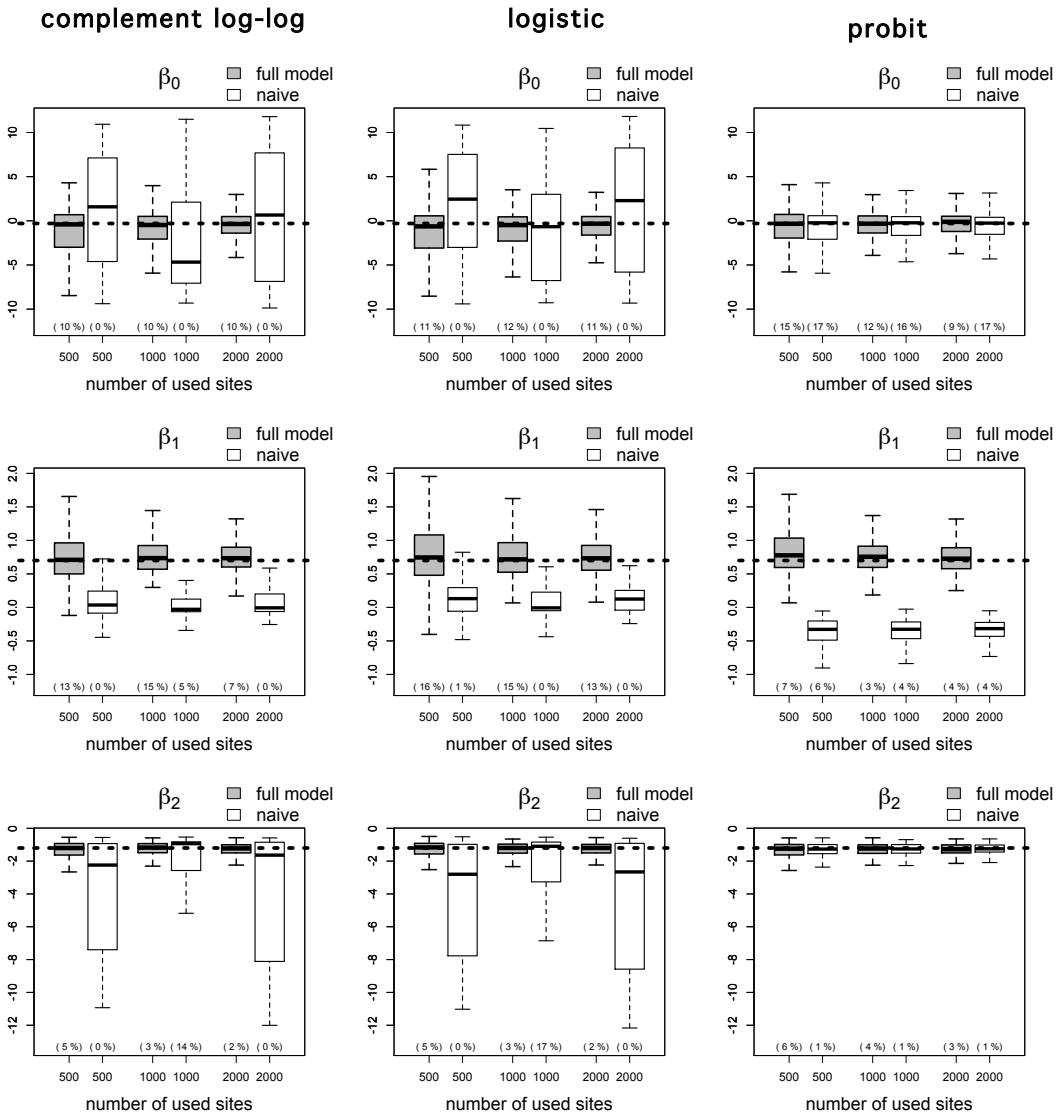


Figure 8.5: Box plots of the Maximum Partial Likelihood estimates obtained for 500 data sets using a continuous common covariate associated with the parameter β_1.

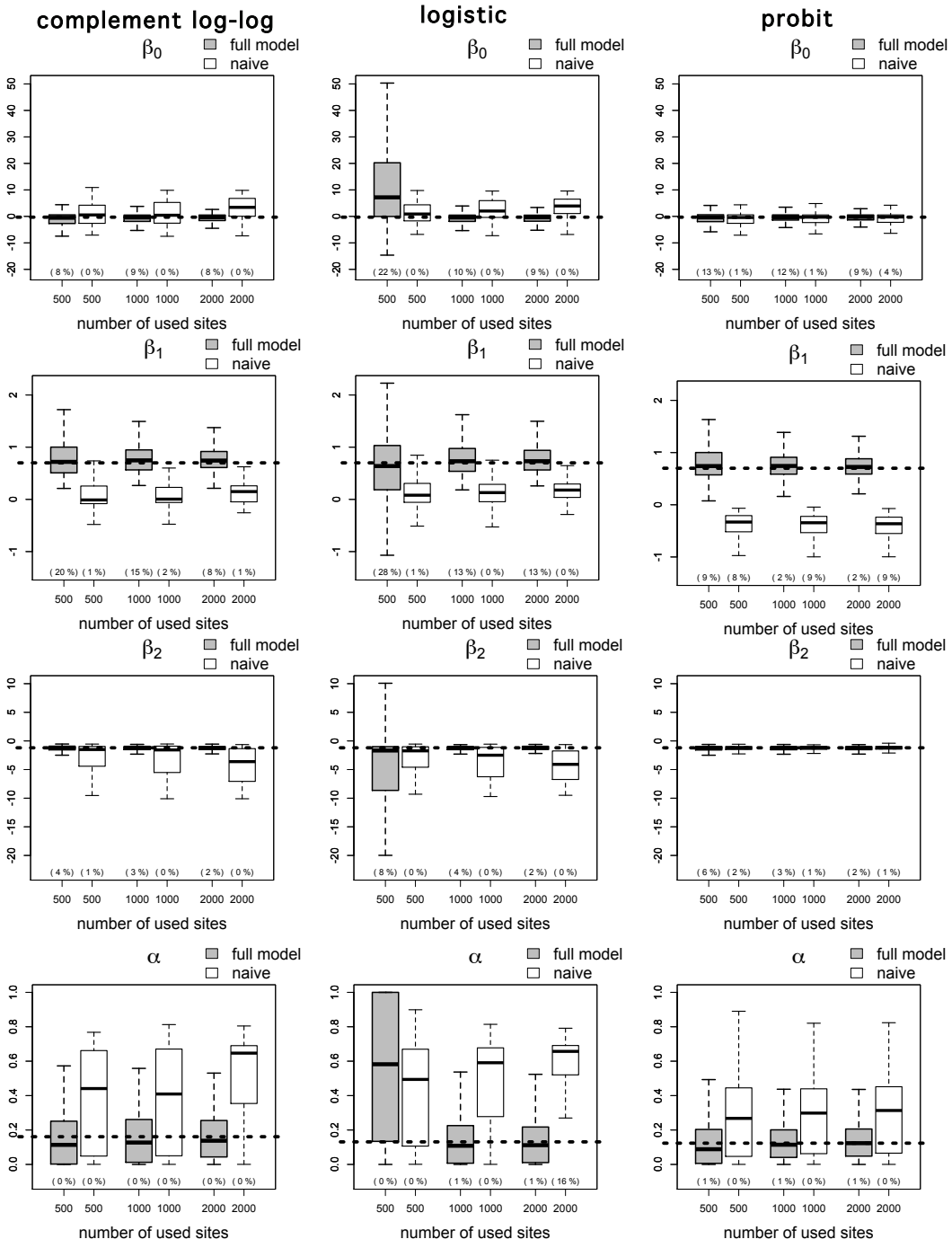


Figure 8.6: Box plots of the Simulated Maximum Likelihood estimates obtained for 500 data sets using a continuous common covariate associated with the parameter β_1.

8.3 Capture - Recapture models

The multiple surveys approach for site occupancy studies was inspired by the closed population mark-recapture model[73]. The mark-recapture model is used to estimate the abun-

dance of a population. Similar to the multiple surveys model for site occupancy, the mark-recapture model has been of great importance because it takes into account the fact that a researcher may fail to detect some individuals present within a population during the time of the survey. In general, the capture-recapture models to estimate abundance can be divided into two categories: closed population and open population models. The former refers to the models used to study a population that remains unchanged during the time of the study, while the latter refers to the models used for populations that can exhibit changes due to birth, death and migration processes[107].

The Lincoln-Petersen is considered to be the earliest contribution to the capture-recapture estimates for the abundance of a closed population. Its estimation procedure assumes that the study area is surveyed twice. The Lincoln-Petersen estimated population size is

$$\hat{N} = \frac{MC}{R} \tag{8.38}$$

where M is the total number of animals caught and marked on the first visit, C is the total number of animals captured on the second visit, and R is the number of recaptured animals on the second visit. The line of reasoning for this estimate is that if all individuals have the same probability of being captured on the second visit, the proportion of recaptured individuals $\left(\frac{R}{M}\right)$ must be the same as the proportion of marked animals in the whole population during the second visit $\left(\frac{C}{N}\right)$, thus, an estimate of the total population is given by equation 8.38. For applications of this method see [57, 64, 87].

On the other hand, the standard model for estimating abundance in open populations is the Jolly-Seber model[52, 108], in which the population is modeled as shown in Figure 8.7. This model allows to estimate both recruitment and survival while assuming homogeneity in capture and survival probabilities among animals in the population. It is known that violations of these assumptions can lead to biased estimates of the population abundance and other parameters of interest such as the effects of covariates[90, 65, 53]. Several models have been proposed to account for heterogeneity in the catchability of the animals[104, 26]. It has been found that for some of those models their likelihood is ill-behaved[105].

We hypothesize that the models developed in this thesis to improve estimation of the site occupancy, in particular the MPLE, can also be used to improve the estimates of the parameters of mark-recapture models for which the likelihood is ill-behaved. The general idea will be to define a Naive estimate, for which some of the components of the full model

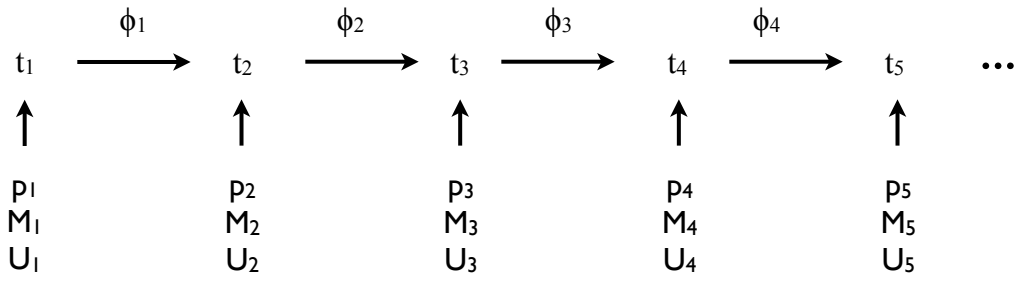


Figure 8.7: Original process model for Jolly-Seber experiments. p_i represents the probability of capture at occasion i; ϕ_i represents the probability of an animal surviving between occasions i and i+1; and Mi and Ui represent the number of marked animals and unmarked animals alive at occasion i. Reproduced from "Program MARK "A Gentle Introduction"[106].

are disregarded, and use that Naive estimate to stabilize the estimation of the parameters for the full model. Consider for instance a full model in which the population is open and the animals are classified on different strata (e.g., behavioral models). A Naive estimate can be obtained by assuming that the population is closed and homogeneous. Like for the site occupancy studies, the first step will be to determine the bias introduced on the estimated parameters when using the Naive model and subsequently to determine the appropriate penalty function for this case.

8.4 Summary

The two models introduced in this chapter are natural extensions of the work discussed in previous chapters. The model to estimate site occupancy for species that co-occur in the same area is an extension of the single survey model discussed in chapter 6. The simulation study for the co-occurrence model demonstrated that there are limitations for the current parameterization, namely, if both the probability of occupancy and the probability of detection for one species depend on the presence/absence of the other species the results are biased. Further research is required to determine whether other parameterization approaches would improve the identifiability of the parameters.

In this chapter we also introduced the concept of the Resource Selection Probability Function. We demonstrated that when the classification of a used site is affected by errors in the detection there are cases in which it is necessary to incorporate a model for the detection error into the RSPF estimation. The next step for the development of this project is to apply our method to real data.

The development of an estimation method to use the penalized likelihood for capture-

recapture models is currently at a very early stage. Further research is required to evaluate the statistical properties of the MLE for the Naive model, and doing so, to determine the appropriate penalty function for this case.

Chapter 9

Conclusions

In this thesis, I present a general description of ecological systems and the metrics that have been used to describe their status. This description allowed me to conclude that the abundance and site occupancy probability are two very important metrics used to describe the status of an ecological system in a fixed period of time. Afterwards, I presented an overview of the current available methods to estimate the site occupancy probability and a brief description of how these models have been extended to take into account detection error. This overview led me to recognize that current models to estimate site occupancy are based on two conflicting features: the assumption of closed population and the use of replicate and independent surveys. From a statistical point of view, the larger the number of replicated surveys, the better estimates of the probability of detection and consequently the estimates of the probability of occupancy are. However, if the time required to conduct these repeated surveys is long, it is unlikely for the population to remain closed, hence the inferences that can be made about the ecological system of interest will become inaccurate. The difficulties present to implement a site occupancy study revealed the need to develop models that can be freed from those assumptions to provide reliable estimates for the site occupancy probability.

The first step to improve the current methodology consisted of assessing its statistical properties. It was found that there are practical circumstances in which neither the Maximum Likelihood approach nor the Bayesian estimates are able to provide reliable estimates of the probability of occupancy nor of the probability of detection. As an alternative, an estimation method based on Penalized Likelihood was developed. This method provides es-

timates with better statistical properties than those obtained from the MLE or the Bayesian approaches, especially for those cases in which due to logistic or budget restrictions only a few number of sites can be visited and the number of replicated surveys has to be small. The model was illustrated using data from two site occupancy studies: one for the Blue Ridge Two Lined Salamander and the other for the Black-capped Chickadee. With the analysis for the Blue Ridge Two Lined Salamander I was able to illustrate how a site occupancy study with five surveys can lead to contrasting inferences when all the different pairs of surveys are analyzed separately. This study lead to the conjecture that in such study the closed population assumption might not hold. The data analysis for the Black-capper Chickadee demonstrated the accuracy on the estimates that can be gained by using the Penalized Likelihood approach.

The next step consisted of evaluating alternative models for which the requirement of repeated and independent surveys were relaxed. We demonstrated that the estimates of the site occupancy probability can be obtained using information from a single survey, provided the site occupancy probability and the detection probability significantly depend on habitat or other exogenous covariates, and that the sets of covariates that affect occupancy and detection differ each by at least one covariate. The simulation study demonstrated that the mean occupancy and mean detection probabilities can be estimated reasonably well with sample sizes of 100 and 200 whereas a good estimation of regression coefficients occurs at sample sizes of 300 or larger. It is worth mentioning that for the simulation study I decided to use only low levels of probability of detection ($\overline{\delta} < 0.30$). It is expected that for larger values of probability of detection, smaller sample sizes will be required.

The single survey model was further extended to incorporate the correlation between adjacent sites. The extended model, named "cluster sampling model", allows estimation of site occupancy probability and probability of detection both at a site level and at a cluster level. Once again, I decided for the simulation study for this model to focus on low levels of probability of detection. It was found that, for the simulated cases, the estimated parameters are consistent and identifiable. The best estimates were obtained for large cluster sizes and strong correlations between sites. The limitations for this model are similar to those of the single survey model, this is, it is required for the probability of occupancy and the probability of detection to depend on habitat or other covariates, and that the set of covariates that affect occupancy and detection differ by at least one covariate. We also discussed how the identifiability of the parameters for this model can be tested using the

data cloning algorithm [60, 62]. The application of the model and the identifiability test was illustrated using data from a submersed aquatic vegetation species, the *Ceratophyllum demersum L.* (coontail). The data analysis for the coontail species exemplifies a case in which, despite the fact that the only covariate that separates the occupancy model from the detection model is a binary variable associated with the observer, the parameters remain identifiable.

I also proposed an extension of the single survey model to the study of a community. This model, based on a Random Markov Field, would provide a tool to study the co-occurrence of multiple species. Unfortunately, the results of the simulation study for the model were not completely satisfactory. It was found that if both the probability of occupancy and the probability of detection depend on the other species of interest, the parameters might become unidentifiable.

Finally, I introduced the concept of the Resource Selection Probability Function, and the effect of the detection error on its estimates. My simulation study demonstrated that the incorrect classification of an used site may lead to biased estimates of the RSPF parameters.

In a more general context, the models developed in this thesis are important contributions to the estimation process for general linear models with a binary response in which the binary response is subject to error. From an ecological perspective, these models will allow users to analyze data for which repeated surveys are not available or the assumption of a closed population is not tenable. With the purpose of making these models available to ecologists, I am preparing a package in R called "detect". This package will contain all the models developed in this thesis, together with the identifiability test.

Appendix A

Algorithm to obtain the bootstrap confidence intervals

Let

n number of sites

k number of visits per site

b number of bootstrap samples

$\underline{y} = \{y_1, ..., y_n\}$ the observations for the n sites where $y_i = \{y_{i1}, ..., y_{ik}\}$

1. Select a random sample of the sites. This sample must be taken with replacement, and its size must be equal to n. The observations within a site are not to be resampled, but to remain the same for each selected site.

$$\underline{y}^* = \{y_1^*, ..., y_n^*\}$$

2. Using the bootstrap sample $\underline{y}^*$, apply the MPL to obtain the estimates of the parameters of interest. For instance:$\hat{\beta}_j = \left\{\hat{\beta}_{0j}, ..., \hat{\beta}_{gj}\right\}$, $\hat{\theta}_j = \left\{\hat{\theta}_{0j}, ..., \hat{\theta}_{mj}\right\}$ and $\hat{\psi}_j = \left\{\hat{\psi}_{1j}, ..., \hat{\psi}_{nj}\right\}$

3. Repeat 1, 2 b times.

4. At the end, a total of b estimates will be available for each parameter of the model. For instance, the estimates of the occupancy probability for the i^{th} site can be arranged

in a vector as follows:

$$\underline{\hat{\psi}}_{site=i} = \left\{\hat{\psi}_{i1}, \hat{\psi}_{i2}, ..., \hat{\psi}_{ib}\right\}$$

5. These estimates can be used to estimate the probability distribution for each parameter. The 90% confidence intervals will be found by selecting the 5th and 95th percentile of $\underline{\hat{\psi}}_{site=i}$

$$90\%\ CI = \left[\hat{\psi}_{i,0.05}; \hat{\psi}_{i,0.95}\right]$$

Appendix B

Single survey simulation results

Logistic link for occupancy, Log-log link for detection, the covariates are separable. Low probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	-0.40	4.65	-0.61	23.21	558.59	-1.38	-2.02	2.71	8.26
	β_1	0.60	9.28	0.65	29.11	914.26	0.66	0.56	1.31	1.69
	β_2	-1.40	-17.22	-1.87	36.59	1575.6	-3.23	-1.20	6.36	43.33
	θ_0	-0.50	-10.01	-1.45	21.43	545.32	-0.21	0.16	1.28	1.70
	θ_1	1.00	34.69	30.97	36.70	2467.9	1.26	1.01	1.06	1.17
	θ_2	1.60	34.94	18.87	39.43	2650.9	1.27	1.02	1.06	1.23
500	β_0	-0.40	1.30	-0.52	11.23	127.79	-0.23	-0.50	1.41	1.99
	β_1	0.60	2.01	0.67	9.40	89.52	0.78	0.64	0.83	0.72
	β_2	-1.40	-2.81	-1.44	8.63	75.78	-1.49	-1.27	0.87	0.76
	θ_0	-0.50	-2.49	-0.67	7.91	65.84	-0.78	-0.49	1.26	1.64
	θ_1	1.00	2.90	1.18	6.90	50.71	1.33	1.07	0.87	0.86
	θ_2	1.60	4.45	1.88	10.67	120.81	1.99	1.70	1.30	1.82

1000	β_0	-0.40	-0.38	-0.45	0.58	0.34	-0.36	-0.41	0.59	0.35
	β_1	0.60	0.64	0.62	0.27	0.07	0.62	0.61	0.25	0.06
	β_2	-1.40	-1.45	-1.39	0.47	0.23	-1.40	-1.33	0.46	0.21
	θ_0	-0.50	-0.95	-0.58	3.03	9.30	-0.61	-0.50	0.91	0.83
	θ_1	1.00	1.42	1.10	2.53	6.51	1.15	1.04	0.67	0.46
	θ_2	1.60	2.21	1.78	3.83	14.87	1.78	1.67	0.97	0.97

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.27	0.31	0.29	0.28
	$\overline{\delta}$	0.27	0.30	0.30	0.28
MPLE	$\overline{\psi}$	0.27	0.22	0.29	0.28
	$\overline{\delta}$	0.27	0.27	0.28	0.27

Logistic link for occupancy, Log-log link for detection, the covariates are separable. Medium probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	0.50	12.42	0.79	24.04	714.39	0.13	0.43	1.74	3.13
	β_1	0.80	14.66	1.42	30.58	1117.6	0.94	0.64	1.07	1.15
	β_2	-1.20	-7.79	-1.49	24.00	613.77	-0.99	-0.65	2.86	8.16
	θ_0	-0.50	-6.53	-0.70	17.30	332.76	-0.30	-0.02	0.97	0.97
	θ_1	1.00	15.53	1.52	29.15	1052.4	1.10	0.92	0.82	0.67
	θ_2	1.60	16.50	2.10	29.99	1112.2	1.34	1.10	0.82	0.74
500	β_0	0.50	1.11	0.49	2.05	4.54	0.90	0.55	1.19	1.57
	β_1	0.80	1.14	0.89	1.08	1.28	1.02	0.89	0.63	0.44
	β_2	-1.20	-1.64	-1.28	1.48	2.38	-1.37	-1.20	0.82	0.69
	θ_0	-0.50	-0.65	-0.50	0.71	0.52	-0.46	-0.39	0.52	0.27
	θ_1	1.00	1.21	1.05	0.68	0.51	1.06	0.97	0.42	0.17
	θ_2	1.60	1.88	1.63	0.97	1.02	1.62	1.47	0.63	0.39
1000	β_0	0.50	0.60	0.51	0.65	0.43	0.62	0.56	0.61	0.38
	β_1	0.80	0.89	0.84	0.30	0.10	0.89	0.84	0.29	0.09
	β_2	-1.20	-1.28	-1.21	0.43	0.19	-1.25	-1.18	0.40	0.16
	θ_0	-0.50	-0.59	-0.49	0.39	0.16	-0.51	-0.45	0.35	0.12
	θ_1	1.00	1.10	1.06	0.26	0.08	1.05	1.02	0.23	0.06
	θ_2	1.60	1.73	1.66	0.43	0.20	1.64	1.59	0.38	0.15

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.47	0.50	0.49	0.47
	$\overline{\delta}$	0.26	0.29	0.27	0.28
MPLE	$\overline{\psi}$	0.47	0.44	0.50	0.48
	$\overline{\delta}$	0.26	0.25	0.26	0.27

Logistic link for occupancy, Log-log link for detection, the covariates are separable. High probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	1.20	16.45	2.39	26.45	925.04	1.36	1.34	1.17	1.37
	β_1	1.60	21.58	2.43	34.59	1584.0	1.42	1.42	1.00	1.01
	β_2	1.40	11.32	1.84	25.91	762.99	0.99	0.72	1.21	1.61
	θ_0	-0.50	-2.11	-0.63	7.22	54.20	-0.41	-0.28	0.59	0.36
	θ_1	1.00	3.68	1.14	12.93	172.66	1.04	0.95	0.47	0.22
	θ_2	1.60	4.44	1.89	13.15	179.23	1.51	1.39	0.57	0.33
500	β_0	1.20	3.05	1.38	11.05	124.36	1.47	1.39	0.87	0.83
	β_1	1.60	3.05	1.91	8.05	66.32	1.80	1.81	0.67	0.48
	β_2	1.40	2.49	1.54	6.77	46.56	1.56	1.43	1.04	1.09
	θ_0	-0.50	-0.59	-0.52	0.36	0.14	-0.49	-0.46	0.29	0.08
	θ_1	1.00	1.11	1.04	0.37	0.15	1.03	0.99	0.25	0.06
	θ_2	1.60	1.75	1.66	0.49	0.26	1.60	1.54	0.36	0.13
1000	β_0	1.20	1.30	1.24	0.67	0.45	1.27	1.24	0.57	0.33
	β_1	1.60	1.76	1.66	0.61	0.39	1.69	1.63	0.51	0.26
	β_2	1.40	1.51	1.43	0.70	0.50	1.44	1.38	0.63	0.40
	θ_0	-0.50	-0.54	-0.52	0.19	0.04	-0.51	-0.50	0.17	0.03
	θ_1	1.00	1.04	1.01	0.14	0.02	1.02	0.99	0.13	0.02
	θ_2	1.60	1.66	1.62	0.23	0.06	1.62	1.59	0.21	0.04

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.79	0.75	0.78	0.78
	$\overline{\delta}$	0.27	0.30	0.28	0.28
MPLE	$\overline{\psi}$	0.79	0.75	0.79	0.78
	$\overline{\delta}$	0.27	0.26	0.27	0.27

Logistic link for occupancy, Log-log link for detection, a discrete covariates is common. Low probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	-0.40	13.74	-0.35	25.76	856.85	-0.68	-1.23	1.63	2.71
	β_1	0.60	13.02	0.90	31.76	1153.1	0.66	0.54	1.05	1.10
	β_2	-1.40	-14.39	-1.93	32.42	1209.7	-7.33	-2.60	7.89	96.73
	θ_0	-0.50	-10.14	-0.67	18.04	415.12	-0.06	0.46	1.03	1.25
	θ_1	1.00	25.58	2.80	33.74	1731.6	0.96	0.58	1.00	0.99
	θ_2	1.60	25.73	7.17	36.76	1920.0	1.71	1.17	1.19	1.40
500	β_0	-0.40	1.99	-0.50	12.06	149.64	0.20	-0.34	1.57	2.78
	β_1	0.60	2.36	0.66	8.94	82.22	0.94	0.64	0.99	1.10
	β_2	-1.40	-2.62	-1.38	11.21	125.87	-1.99	-2.02	1.06	1.45
	θ_0	-0.50	-2.48	-0.53	7.37	57.64	-0.64	-0.32	1.05	1.11
	θ_1	1.00	3.10	1.09	8.22	71.25	1.23	0.98	0.88	0.82
	θ_2	1.60	4.29	1.91	9.90	104.27	1.51	1.29	0.78	0.61
1000	β_0	-0.40	0.36	-0.44	6.45	41.73	-0.16	-0.40	0.92	0.89
	β_1	0.60	1.14	0.62	4.20	17.77	0.75	0.62	0.45	0.22
	β_2	-1.40	-1.51	-1.39	2.31	5.29	-1.51	-1.60	0.73	0.54
	θ_0	-0.50	-0.70	-0.61	0.70	0.53	-0.59	-0.55	0.62	0.39
	θ_1	1.00	1.21	1.05	0.61	0.41	1.13	1.01	0.52	0.29
	θ_2	1.60	1.90	1.77	0.77	0.67	1.68	1.53	0.63	0.40

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.27	0.43	0.35	0.32
	$\overline{\delta}$	0.31	0.34	0.33	0.32
MPLE	$\overline{\psi}$	0.27	0.23	0.34	0.31
	$\overline{\delta}$	0.31	0.25	0.33	0.32

Logistic link for occupancy, Log-log link for detection, a discrete covariates is common. Medium probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	0.50	12.35	0.37	23.09	668.07	0.66	0.57	1.55	2.41
	β_1	0.80	11.71	1.13	23.05	644.96	0.99	0.76	0.90	0.84
	β_2	-1.20	-9.23	-1.45	27.29	801.67	-2.82	-2.24	3.94	17.97
	θ_0	-0.50	-10.91	-0.85	18.24	437.56	-0.58	-0.22	0.88	0.78
	θ_1	1.00	18.01	1.65	31.42	1266.9	1.25	0.97	0.85	0.77
	θ_2	1.60	18.32	2.57	30.00	1170.6	1.43	1.33	0.77	0.62
500	β_0	0.50	2.46	0.49	12.03	147.22	1.03	0.62	1.28	1.91
	β_1	0.80	2.21	0.88	8.39	71.71	1.08	0.87	0.79	0.70
	β_2	-1.20	-1.37	-1.17	4.89	23.67	-1.34	-1.43	0.87	0.77
	θ_0	-0.50	-0.70	-0.52	0.75	0.59	-0.50	-0.39	0.58	0.33
	θ_1	1.00	1.23	1.09	0.60	0.41	1.09	1.01	0.45	0.21
	θ_2	1.60	1.93	1.70	0.81	0.76	1.63	1.45	0.55	0.30
1000	β_0	0.50	0.57	0.43	0.59	0.35	0.65	0.52	0.58	0.35
	β_1	0.80	0.88	0.80	0.34	0.12	0.89	0.81	0.32	0.11
	β_2	-1.20	-1.12	-1.12	0.55	0.31	-1.25	-1.24	0.50	0.25
	θ_0	-0.50	-0.60	-0.60	0.38	0.15	-0.53	-0.49	0.35	0.12
	θ_1	1.00	1.11	1.08	0.29	0.09	1.06	1.03	0.26	0.07
	θ_2	1.60	1.75	1.70	0.41	0.19	1.63	1.58	0.36	0.13

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.47	0.52	0.52	0.48
	$\overline{\delta}$	0.31	0.37	0.31	0.32
MPLE	$\overline{\psi}$	0.47	0.43	0.52	0.49
	$\overline{\delta}$	0.31	0.34	0.30	0.31

Logistic link for occupancy, Log-log link for detection, a discrete covariates is common. High probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	1.20	16.68	1.21	27.32	978.45	1.69	1.63	1.35	2.04
	β_1	1.60	15.61	1.77	29.83	1076.9	1.41	1.19	1.24	1.56
	β_2	1.40	3.05	0.88	23.92	569.22	-0.91	-0.98	2.28	10.49
	θ_0	-0.50	-3.96	-0.71	10.81	127.74	-0.27	-0.07	0.63	0.44
	θ_1	1.00	5.94	1.25	15.85	273.04	1.02	0.88	0.54	0.29
	θ_2	1.60	7.13	1.79	18.08	354.10	1.14	1.07	0.48	0.44
500	β_0	1.20	1.38	1.15	0.93	0.89	1.49	1.24	0.82	0.74
	β_1	1.60	1.74	1.53	0.82	0.68	1.68	1.56	0.63	0.40
	β_2	1.40	1.85	1.28	2.51	6.47	1.04	0.92	1.04	1.20
	θ_0	-0.50	-0.57	-0.55	0.25	0.07	-0.44	-0.44	0.23	0.06
	θ_1	1.00	1.11	1.09	0.21	0.06	1.03	1.01	0.19	0.04
	θ_2	1.60	1.69	1.68	0.35	0.13	1.50	1.47	0.31	0.10
1000	β_0	1.20	1.27	1.14	0.59	0.35	1.31	1.19	0.58	0.34
	β_1	1.60	1.66	1.51	0.50	0.26	1.64	1.52	0.47	0.22
	β_2	1.40	1.48	1.29	1.09	1.18	1.22	1.14	0.83	0.72
	θ_0	-0.50	-0.54	-0.54	0.19	0.04	-0.50	-0.51	0.18	0.03
	θ_1	1.00	1.06	1.04	0.14	0.02	1.04	1.02	0.13	0.02
	θ_2	1.60	1.66	1.62	0.25	0.06	1.59	1.55	0.23	0.05

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.79	0.72	0.77	0.78
	$\overline{\delta}$	0.31	0.36	0.32	0.31
MPLE	$\overline{\psi}$	0.79	0.65	0.77	0.77
	$\overline{\delta}$	0.31	0.30	0.30	0.31

Logistic link for occupancy, Log-log link for detection, a continuous covariates is common. Low probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	-0.40	16.31	0.69	31.99	1292.2	-5.79	-2.27	43.17	1874.2
	β_1	0.60	6.04	1.12	29.79	908.24	-2.29	-0.53	18.48	346.57
	β_2	-1.40	-21.72	-5.70	36.08	1701.4	-5.47	-1.59	11.22	141.22
	θ_0	-0.50	-9.98	-0.72	22.46	589.31	0.10	0.57	1.12	1.61
	θ_1	1.00	30.86	18.89	37.78	2304.6	0.38	0.31	0.31	0.48
	θ_2	1.60	34.75	13.35	40.33	2709.3	1.24	0.84	1.01	1.15
500	β_0	-0.40	0.03	-0.50	2.35	5.67	-0.26	-0.56	2.01	4.00
	β_1	0.60	0.81	0.57	1.00	1.03	0.36	0.26	1.08	1.21
	β_2	-1.40	-1.78	-1.54	1.30	1.83	-1.65	-1.46	0.90	0.87
	θ_0	-0.50	-5.94	-0.65	13.47	209.21	-1.29	-0.53	1.87	4.10
	θ_1	1.00	6.53	1.24	14.75	245.99	1.30	0.96	0.95	0.99
	θ_2	1.60	9.74	1.98	20.71	490.90	2.38	1.72	1.80	3.81
1000	β_0	-0.40	-0.26	-0.34	0.99	0.99	-0.24	-0.35	1.12	1.28
	β_1	0.60	0.68	0.65	0.42	0.18	0.62	0.62	0.46	0.21
	β_2	-1.40	-1.49	-1.45	0.48	0.23	-1.45	-1.42	0.49	0.24
	θ_0	-0.50	-1.26	-0.62	4.25	18.48	-0.80	-0.56	1.20	1.52
	θ_1	1.00	1.70	1.13	3.63	13.51	1.25	1.04	0.71	0.56
	θ_2	1.60	2.60	1.74	5.67	32.81	1.96	1.61	1.23	1.64

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.27	0.42	0.31	0.30
	$\overline{\delta}$	0.27	0.30	0.32	0.29
MPLE	$\overline{\psi}$	0.27	0.22	0.30	0.30
	$\overline{\delta}$	0.27	0.22	0.30	0.28

Logistic link for occupancy, Log-log link for detection, a continuous covariates is common. Medium probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	0.50	13.66	0.49	27.69	932.20	-0.96	-1.60	2.76	9.67
	β_1	0.80	0.20	0.72	39.58	1551.0	-0.37	-0.45	0.64	1.77
	β_2	-1.20	-6.75	-1.74	28.35	826.72	-1.07	-0.93	3.22	10.30
	θ_0	-0.50	-14.04	-1.88	20.69	607.36	-0.46	-0.16	1.08	1.15
	θ_1	1.00	17.37	2.26	27.09	994.38	0.43	0.32	0.38	0.47
	θ_2	1.60	25.69	7.57	33.67	1702.9	1.42	1.12	0.91	0.85
500	β_0	0.50	0.62	0.39	1.46	2.13	0.62	0.53	1.49	2.21
	β_1	0.80	0.92	0.84	0.75	0.57	0.80	0.74	0.76	0.57
	β_2	-1.20	-1.36	-1.26	0.67	0.47	-1.27	-1.15	0.62	0.38
	θ_0	-0.50	-2.92	-0.64	9.48	94.76	-0.91	-0.41	1.36	2.01
	θ_1	1.00	2.71	1.24	6.99	51.24	1.13	0.96	0.68	0.48
	θ_2	1.60	4.28	1.76	9.91	104.34	1.98	1.50	1.24	1.67
1000	β_0	0.50	0.57	0.52	0.78	0.61	0.65	0.59	0.81	0.66
	β_1	0.80	0.86	0.86	0.36	0.13	0.85	0.87	0.36	0.13
	β_2	-1.20	-1.26	-1.18	0.42	0.18	-1.24	-1.17	0.43	0.18
	θ_0	-0.50	-0.69	-0.54	0.57	0.35	-0.56	-0.44	0.52	0.27
	θ_1	1.00	1.12	1.08	0.36	0.14	1.02	0.99	0.31	0.10
	θ_2	1.60	1.79	1.63	0.55	0.34	1.65	1.52	0.48	0.23

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.46	0.47	0.45	0.47
	$\overline{\delta}$	0.27	0.33	0.31	0.29
MPLE	$\overline{\psi}$	0.46	0.30	0.46	0.48
	$\overline{\delta}$	0.27	0.29	0.29	0.27

Logistic link for occupancy, Log-log link for detection, a continuous covariates is common. High probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	1.20	12.38	0.97	22.69	634.84	0.01	0.10	1.62	4.01
	β_1	1.60	9.35	1.32	31.35	1033.0	0.00	-0.42	1.15	3.87
	β_2	1.40	14.27	1.76	29.53	1028.9	1.03	0.64	1.25	1.68
	θ_0	-0.50	-5.14	-0.82	10.90	139.08	-0.51	-0.20	0.80	0.64
	θ_1	1.00	5.75	1.39	11.33	149.77	0.59	0.52	0.34	0.29
	θ_2	1.60	8.74	2.07	16.94	335.16	1.39	1.23	0.63	0.44
500	β_0	1.20	2.78	1.12	8.70	77.40	1.33	1.14	1.24	1.53
	β_1	1.60	1.31	1.62	7.77	59.81	1.55	1.61	0.97	0.93
	β_2	1.40	1.96	1.44	4.27	18.36	1.45	1.36	0.77	0.59
	θ_0	-0.50	-0.78	-0.57	0.93	0.93	-0.56	-0.46	0.50	0.25
	θ_1	1.00	1.15	1.06	0.44	0.22	0.99	0.96	0.31	0.10
	θ_2	1.60	1.90	1.69	0.97	1.03	1.64	1.54	0.51	0.26
1000	β_0	1.20	1.17	1.10	0.76	0.57	1.21	1.15	0.72	0.51
	β_1	1.60	1.55	1.46	0.52	0.27	1.55	1.48	0.46	0.21
	β_2	1.40	1.41	1.37	0.54	0.29	1.40	1.38	0.53	0.27
	θ_0	-0.50	-0.60	-0.55	0.31	0.10	-0.54	-0.49	0.27	0.07
	θ_1	1.00	1.05	1.00	0.24	0.06	1.00	0.96	0.21	0.04
	θ_2	1.60	1.70	1.63	0.35	0.13	1.62	1.58	0.30	0.09

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.80	0.70	0.76	0.77
	$\overline{\delta}$	0.26	0.33	0.29	0.28
MPLE	$\overline{\psi}$	0.80	0.56	0.77	0.78
	$\overline{\delta}$	0.26	0.27	0.27	0.27

Logistic link for occupancy, Logistic link for detection, the covariates are separable. Low probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	-0.40	4.73	-0.17	12.54	181.89	-0.70	-0.88	2.01	4.11
	β_1	0.60	7.66	0.91	20.40	461.99	0.72	0.55	1.19	1.42
	β_2	-1.40	-13.07	-3.31	19.12	498.22	-3.06	-1.14	5.87	36.88
	θ_0	-0.60	0.52	-1.22	11.97	143.18	-1.77	-2.06	3.31	12.23
	θ_1	1.00	11.21	1.27	18.70	450.26	1.23	0.85	1.26	1.62
	θ_2	-1.60	-9.27	-2.10	18.52	398.28	-3.01	-1.22	6.92	49.33
500	β_0	-0.40	1.02	-0.40	6.61	45.26	0.13	-0.46	1.85	3.65
	β_1	0.60	1.66	0.69	6.08	37.77	0.87	0.63	0.99	1.05
	β_2	-1.40	-2.29	-1.60	4.07	17.19	-1.71	-1.42	1.18	1.47
	θ_0	-0.60	-0.21	-0.62	1.90	3.74	-0.50	-0.83	1.56	2.42
	θ_1	1.00	1.28	1.08	0.74	0.62	1.16	1.00	0.60	0.38
	θ_2	-1.60	-2.10	-1.73	1.64	2.91	-1.75	-1.47	1.09	1.19
1000	β_0	-0.40	0.42	-0.26	5.96	35.84	-0.06	-0.20	1.25	1.66
	β_1	0.60	1.09	0.67	4.00	16.11	0.73	0.68	0.57	0.34
	β_2	-1.40	-1.91	-1.52	3.04	9.38	-1.60	-1.44	0.69	0.51
	θ_0	-0.60	-0.48	-0.70	1.09	1.20	-0.61	-0.81	1.00	0.99
	θ_1	1.00	1.09	0.98	0.36	0.14	1.05	0.95	0.33	0.11
	θ_2	-1.60	-1.81	-1.69	0.82	0.71	-1.67	-1.58	0.69	0.48

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.27	0.34	0.32	0.31
	$\overline{\delta}$	0.22	0.25	0.25	0.23
MPLE	$\overline{\psi}$	0.27	0.31	0.33	0.31
	$\overline{\delta}$	0.22	0.22	0.24	0.22

Logistic link for occupancy, Logistic link for detection, the covariates are separable. Medium probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	0.50	6.89	0.18	12.96	207.05	-0.01	-0.09	1.73	3.22
	β_1	0.80	8.23	1.13	16.89	337.44	0.98	0.77	0.98	0.99
	β_2	-1.20	-3.42	-1.20	14.87	223.78	-0.91	-0.58	2.69	7.26
	θ_0	-0.60	4.43	-0.24	11.33	152.48	-0.44	-0.80	1.64	2.68
	θ_1	1.00	8.17	1.58	15.02	274.75	1.23	1.12	0.83	0.74
	θ_2	-1.60	-8.46	-2.15	15.99	300.18	-1.62	-1.33	2.07	4.24
500	β_0	0.50	2.49	0.38	9.21	87.91	0.88	0.51	1.65	2.83
	β_1	0.80	2.22	0.84	6.78	47.53	1.08	0.85	0.86	0.82
	β_2	-1.20	-2.20	-1.16	4.47	20.77	-1.25	-1.04	0.87	0.76
	θ_0	-0.60	-0.15	-0.56	2.38	5.79	-0.57	-0.75	0.98	0.94
	θ_1	1.00	1.50	1.08	2.86	8.33	1.16	1.01	0.72	0.54
	θ_2	-1.60	-1.87	-1.71	1.23	1.57	-1.58	-1.47	0.66	0.43
1000	β_0	0.50	0.86	0.67	1.32	1.85	0.84	0.69	1.10	1.32
	β_1	0.80	1.01	0.81	0.81	0.70	0.97	0.82	0.59	0.38
	β_2	-1.20	-1.33	-1.25	0.84	0.71	-1.29	-1.22	0.68	0.47
	θ_0	-0.60	-0.52	-0.66	0.60	0.37	-0.63	-0.74	0.57	0.32
	θ_1	1.00	1.09	1.05	0.29	0.09	1.04	1.02	0.26	0.07
	θ_2	-1.60	-1.74	-1.67	0.46	0.23	-1.62	-1.57	0.42	0.18

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.47	0.46	0.48	0.49
	$\overline{\delta}$	0.22	0.31	0.25	0.23
MPLE	$\overline{\psi}$	0.47	0.43	0.49	0.50
	$\overline{\delta}$	0.22	0.29	0.23	0.22

Logistic link for occupancy, Logistic link for detection, the covariates are separable. High probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	1.20	8.11	0.67	14.82	265.13	0.68	0.92	1.48	2.45
	β_1	1.60	10.04	1.13	21.01	508.01	1.00	0.89	1.21	1.81
	β_2	1.40	9.43	1.75	21.41	518.52	0.97	0.81	1.51	2.43
	θ_0	-0.60	1.77	-0.21	7.60	62.84	-0.42	-0.63	1.04	1.10
	θ_1	1.00	3.79	1.26	10.76	122.35	1.20	1.07	0.65	0.46
	θ_2	-1.60	-3.86	-2.17	6.66	49.06	-1.67	-1.71	0.64	0.41
500	β_0	1.20	5.22	1.38	11.70	151.64	1.58	1.35	1.31	1.84
	β_1	1.60	4.97	1.79	10.35	117.32	1.94	1.70	1.11	1.33
	β_2	1.40	3.68	1.53	7.40	59.37	1.59	1.47	1.24	1.55
	θ_0	-0.60	-0.51	-0.56	0.40	0.16	-0.63	-0.66	0.36	0.13
	θ_1	1.00	1.06	1.03	0.21	0.05	1.00	0.98	0.18	0.03
	θ_2	-1.60	-1.70	-1.65	0.36	0.14	-1.56	-1.53	0.31	0.10
1000	β_0	1.20	1.85	1.26	2.81	8.23	1.38	1.18	0.92	0.88
	β_1	1.60	2.35	1.76	2.71	7.84	1.83	1.69	0.76	0.63
	β_2	1.40	2.47	1.51	4.81	24.09	1.70	1.40	1.30	1.77
	θ_0	-0.60	-0.57	-0.61	0.25	0.07	-0.62	-0.64	0.25	0.06
	θ_1	1.00	1.00	0.98	0.13	0.02	0.97	0.96	0.12	0.02
	θ_2	-1.60	-1.62	-1.59	0.24	0.06	-1.55	-1.53	0.23	0.06

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.79	0.66	0.77	0.78
	$\overline{\delta}$	0.22	0.30	0.23	0.22
MPLE	$\overline{\psi}$	0.79	0.65	0.78	0.78
	$\overline{\delta}$	0.22	0.26	0.22	0.22

Logistic link for occupancy, Logistic link for detection, a discrete covariate is common. Low probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	-0.40	6.25	-0.41	13.27	218.52	-0.74	-1.40	1.75	3.15
	β_1	0.60	10.76	1.16	20.17	505.91	0.92	0.62	1.14	1.39
	β_2	-1.40	-9.44	-2.27	19.20	429.68	-6.16	-2.09	7.71	81.50
	θ_0	-0.60	3.45	-0.82	12.97	182.88	-0.85	-1.41	1.61	2.62
	θ_1	1.00	12.59	1.62	19.26	501.69	1.33	0.99	1.25	1.64
	θ_2	-1.60	-8.05	-2.36	18.83	392.59	-6.16	-2.24	7.72	79.72
500	β_0	-0.40	2.80	-0.49	11.28	136.18	0.20	-0.46	1.92	4.00
	β_1	0.60	2.96	0.72	7.87	66.95	1.07	0.66	1.06	1.33
	β_2	-1.40	-2.92	-1.78	9.72	95.92	-1.84	-1.81	2.05	4.36
	θ_0	-0.60	1.66	-0.55	8.80	81.78	-0.21	-0.64	1.77	3.27
	θ_1	1.00	3.73	1.13	9.62	99.10	1.61	1.07	1.57	2.82
	θ_2	-1.60	-2.77	-1.80	9.39	88.73	-1.75	-1.72	2.15	4.61
1000	β_0	-0.40	0.47	-0.46	6.07	37.23	0.04	-0.47	1.54	2.54
	β_1	0.60	1.15	0.70	2.67	7.38	0.92	0.66	0.69	0.57
	β_2	-1.40	-1.55	-1.50	4.52	20.29	-1.49	-1.58	1.09	1.18
	θ_0	-0.60	-0.28	-0.53	1.50	2.33	-0.40	-0.52	1.17	1.38
	θ_1	1.00	1.39	1.09	1.28	1.78	1.28	1.05	0.78	0.68
	θ_2	-1.60	-1.84	-1.76	2.09	4.36	-1.72	-1.94	1.11	1.22

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.27	0.38	0.36	0.36
	$\overline{\delta}$	0.24	0.36	0.34	0.30
MPLE	$\overline{\psi}$	0.27	0.25	0.36	0.35
	$\overline{\delta}$	0.24	0.25	0.31	0.29

Logistic link for occupancy, Logistic link for detection, a discrete covariate is common. Medium probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	0.50	10.13	0.41	15.58	332.96	0.54	0.13	1.65	2.70
	β_1	0.80	8.82	1.00	17.86	380.12	0.86	0.53	1.13	1.26
	β_2	-1.20	-9.77	-2.80	16.87	355.09	-3.51	-1.89	5.34	33.57
	θ_0	-0.60	2.61	-0.60	8.64	84.17	-0.37	-0.89	1.44	2.10
	θ_1	1.00	7.29	1.52	14.02	234.22	1.50	1.18	1.07	1.38
	θ_2	-1.60	-5.55	-1.80	15.80	262.80	-3.53	-1.91	5.41	32.67
500	β_0	0.50	2.95	0.63	10.49	114.85	0.99	0.85	1.51	2.50
	β_1	0.80	2.57	0.96	7.18	54.13	1.13	0.92	0.90	0.91
	β_2	-1.20	-1.80	-1.82	5.03	25.40	-1.36	-1.51	0.97	0.97
	θ_0	-0.60	-0.39	-0.68	0.83	0.72	-0.45	-0.70	0.84	0.71
	θ_1	1.00	1.32	1.13	0.66	0.53	1.23	1.08	0.59	0.40
	θ_2	-1.60	-1.22	-1.53	1.85	3.53	-1.50	-1.66	0.83	0.70
1000	β_0	0.50	1.09	0.53	3.26	10.88	0.84	0.59	1.18	1.50
	β_1	0.80	1.10	0.85	1.18	1.48	0.98	0.85	0.56	0.35
	β_2	-1.20	-1.46	-1.48	2.46	6.05	-1.32	-1.36	0.92	0.85
	θ_0	-0.60	-0.52	-0.64	0.51	0.26	-0.58	-0.70	0.49	0.24
	θ_1	1.00	1.09	1.00	0.34	0.13	1.06	0.97	0.34	0.12
	θ_2	-1.60	-1.48	-1.52	0.88	0.78	-1.51	-1.62	0.55	0.31

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.47	0.47	0.49	0.49
	$\overline{\delta}$	0.24	0.36	0.32	0.28
MPLE	$\overline{\psi}$	0.47	0.41	0.50	0.50
	$\overline{\delta}$	0.24	0.30	0.29	0.26

Logistic link for occupancy, Logistic link for detection, a discrete covariate is common. High probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	1.20	12.05	1.63	15.36	351.21	1.29	1.48	1.49	2.21
	β_1	1.60	15.90	1.77	22.21	692.78	1.45	1.24	1.27	1.62
	β_2	1.40	4.19	0.25	17.82	322.15	-0.47	-0.69	1.25	5.05
	θ_0	-0.60	3.21	-0.41	9.95	112.49	-0.57	-0.97	1.31	1.71
	θ_1	1.00	5.90	1.33	12.60	181.11	1.19	1.04	0.81	0.68
	θ_2	-1.60	-4.52	-1.79	10.77	123.36	-1.21	-1.26	0.60	0.51
500	β_0	1.20	3.87	1.21	10.90	124.72	1.86	1.65	1.44	2.49
	β_1	1.60	3.92	1.75	8.50	76.86	2.02	1.78	1.21	1.62
	β_2	1.40	2.74	1.74	4.54	22.18	0.89	0.63	1.46	2.37
	θ_0	-0.60	-0.45	-0.56	0.44	0.22	-0.71	-0.76	0.40	0.17
	θ_1	1.00	1.07	1.03	0.26	0.07	0.99	0.95	0.24	0.06
	θ_2	-1.60	-1.69	-1.63	0.49	0.24	-1.26	-1.23	0.31	0.21
1000	β_0	1.20	1.37	1.20	0.97	0.95	1.51	1.46	0.85	0.82
	β_1	1.60	1.83	1.67	0.78	0.66	1.80	1.74	0.67	0.49
	β_2	1.40	2.06	1.52	2.02	4.49	1.34	1.06	1.25	1.54
	θ_0	-0.60	-0.52	-0.59	0.29	0.09	-0.65	-0.67	0.26	0.07
	θ_1	1.00	1.03	1.00	0.16	0.03	0.99	0.97	0.15	0.02
	θ_2	-1.60	-1.69	-1.61	0.37	0.14	-1.47	-1.41	0.31	0.11

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.79	0.65	0.76	0.78
	$\overline{\delta}$	0.24	0.35	0.27	0.25
MPLE	$\overline{\psi}$	0.79	0.62	0.75	0.78
	$\overline{\delta}$	0.24	0.28	0.25	0.24

Logistic link for occupancy, Logistic link for detection, a continuous covariate is common. Low probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	-0.40	-0.20	-1.30	15.47	237.06	-1.42	-1.75	2.93	9.53
	β_1	0.60	14.01	2.42	21.72	646.89	1.32	1.23	0.85	1.24
	β_2	-1.40	-8.96	-4.14	19.78	444.60	-2.58	-1.04	6.34	41.19
	θ_0	-0.60	5.90	-0.79	18.36	376.13	-1.98	-2.05	3.91	17.02
	θ_1	1.00	3.44	0.90	21.13	447.98	1.33	1.27	0.72	0.62
	θ_2	-1.60	-10.40	-2.68	16.79	356.33	-1.82	-1.49	5.72	32.39
500	β_0	-0.40	1.33	0.49	3.55	15.44	0.53	0.38	2.40	6.54
	β_1	0.60	1.14	0.87	2.46	6.26	0.89	0.93	1.07	1.22
	β_2	-1.40	-3.12	-1.82	2.80	10.73	-2.07	-1.57	1.36	2.27
	θ_0	-0.60	0.30	-0.98	2.98	9.61	-0.35	-1.30	2.36	5.56
	θ_1	1.00	1.24	1.16	2.03	4.12	1.05	1.08	0.98	0.95
	θ_2	-1.60	-2.56	-1.70	2.45	6.87	-1.80	-1.48	1.12	1.29
1000	β_0	-0.40	0.81	-0.12	3.22	11.76	0.48	-0.03	2.31	6.05
	β_1	0.60	0.71	0.80	1.01	1.02	0.65	0.72	0.77	0.60
	β_2	-1.40	-2.49	-1.63	2.01	5.20	-2.05	-1.55	1.24	1.93
	θ_0	-0.60	0.20	-0.69	2.64	7.55	-0.25	-0.86	2.03	4.19
	θ_1	1.00	0.89	1.05	0.80	0.65	0.93	1.07	0.72	0.52
	θ_2	-1.60	-2.31	-1.53	1.83	3.84	-1.89	-1.43	1.06	1.20

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.27	0.36	0.39	0.37
	$\overline{\delta}$	0.22	0.35	0.28	0.28
MPLE	$\overline{\psi}$	0.27	0.30	0.40	0.38
	$\overline{\delta}$	0.22	0.24	0.27	0.27

Logistic link for occupancy, Logistic link for detection, a continuous covariate is common. Medium probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	0.50	3.93	-0.31	13.27	186.01	-0.23	-0.11	1.81	3.78
	β_1	0.80	15.89	2.64	20.54	645.61	1.26	1.21	0.52	0.47
	β_2	-1.20	-7.13	-2.17	18.98	391.82	-0.71	-0.73	1.14	1.52
	θ_0	-0.60	3.43	-0.28	10.67	128.96	-0.50	-1.18	1.66	2.73
	θ_1	1.00	3.92	0.86	13.04	176.84	1.27	1.29	0.49	0.31
	θ_2	-1.60	-5.22	-2.39	12.04	156.51	-1.44	-1.34	0.95	0.93
500	β_0	0.50	2.75	0.65	5.15	31.29	1.09	0.87	2.20	5.13
	β_1	0.80	2.19	1.14	6.35	41.89	1.22	1.09	0.85	0.89
	β_2	-1.20	-2.71	-1.46	6.33	41.98	-1.37	-0.90	1.47	2.16
	θ_0	-0.60	0.46	-0.81	2.36	6.66	-0.26	-1.24	1.71	3.01
	θ_1	1.00	0.85	1.03	0.57	0.34	0.94	1.07	0.49	0.24
	θ_2	-1.60	-2.42	-1.80	1.66	3.41	-1.82	-1.56	0.84	0.76
1000	β_0	0.50	3.23	0.99	8.11	72.61	1.34	1.20	2.16	5.31
	β_1	0.80	2.23	1.13	6.89	49.08	1.06	1.04	0.73	0.60
	β_2	-1.20	-3.31	-1.58	3.99	20.19	-1.80	-1.38	1.26	1.92
	θ_0	-0.60	0.12	-0.75	1.92	4.16	-0.31	-1.08	1.48	2.26
	θ_1	1.00	0.93	1.02	0.41	0.17	0.98	1.12	0.39	0.15
	θ_2	-1.60	-2.15	-1.68	1.33	2.06	-1.82	-1.56	0.78	0.65

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.47	0.42	0.50	0.50
	$\overline{\delta}$	0.22	0.36	0.30	0.28
MPLE	$\overline{\psi}$	0.47	0.43	0.53	0.53
	$\overline{\delta}$	0.22	0.32	0.28	0.26

Logistic link for occupancy, Logistic link for detection, a continuous covariate is common. High probability of occupancy - low probability of detection

			MLE				MPLE			
n	par	real	mean	median	se	mse	mean	median	se	mse
100	β_0	1.20	5.07	0.13	12.72	175.23	0.16	0.61	2.48	7.17
	β_1	1.60	15.22	2.88	21.37	637.47	1.22	1.14	0.55	0.44
	β_2	1.40	4.92	1.23	17.94	330.99	0.85	0.29	2.24	5.25
	θ_0	-0.60	2.11	0.11	7.18	58.34	-0.19	-0.55	1.16	1.50
	θ_1	1.00	0.89	0.80	3.77	14.06	1.17	1.18	0.46	0.24
	θ_2	-1.60	-3.69	-2.09	5.65	35.93	-1.73	-1.68	0.71	0.51
500	β_0	1.20	8.87	1.03	15.77	304.96	1.34	1.32	1.36	1.86
	β_1	1.60	11.07	2.14	17.55	394.70	1.77	1.62	1.03	1.09
	β_2	1.40	5.46	1.40	11.66	151.05	1.39	0.98	1.55	2.36
	θ_0	-0.60	-0.10	-0.46	0.99	1.22	-0.42	-0.66	0.75	0.59
	θ_1	1.00	0.95	0.96	0.36	0.13	1.07	1.11	0.31	0.10
	θ_2	-1.60	-1.93	-1.73	0.67	0.56	-1.76	-1.68	0.45	0.23
1000	β_0	1.20	6.38	1.06	13.51	207.56	1.52	1.25	1.47	2.24
	β_1	1.60	6.57	2.00	13.10	194.68	1.79	1.56	1.34	1.82
	β_2	1.40	4.71	1.85	10.06	111.09	1.62	1.17	1.49	2.25
	θ_0	-0.60	-0.27	-0.57	0.86	0.84	-0.49	-0.65	0.58	0.35
	θ_1	1.00	0.96	0.97	0.25	0.06	1.04	1.07	0.24	0.06
	θ_2	-1.60	-1.80	-1.67	0.58	0.37	-1.68	-1.62	0.33	0.11

Mean probability of occupancy and detection

			mean estimate		
		true	n=100	n=500	n=100
MLE	$\overline{\psi}$	0.78	0.59	0.72	0.74
	$\overline{\delta}$	0.22	0.35	0.27	0.25
MPLE	$\overline{\psi}$	0.78	0.61	0.75	0.76
	$\overline{\delta}$	0.22	0.30	0.24	0.23

Appendix C

Cluster sampling simulation results

Summary of simulated cases for the cluster sampling model using separate covariates, clusters of size 3 and low dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	3.39	-0.72	20.34	426.14	-0.20	-0.64	1.97	3.91
	β_1	0.90	10.04	1.36	27.41	831.09	1.52	1.10	1.35	2.20
	β_2	-1.20	-11.95	-1.89	38.45	1586.37	-1.32	-1.09	1.82	3.30
	θ_0	-0.50	3.08	-0.62	18.89	367.82	-0.81	-0.96	1.18	1.48
	θ_1	1.00	12.75	1.19	47.75	2406.95	1.24	0.95	0.91	0.88
	θ_2	-1.00	-4.83	-1.23	20.12	417.31	-0.80	-0.80	0.79	0.65
	γ	0.50	-0.80	0.31	25.43	644.90	-1.19	0.13	13.32	179.29

n=100	β_0	-0.40	1.62	-0.49	10.04	104.36	0.13	-0.50	2.03	4.40
	β_1	0.90	3.49	1.23	9.81	102.46	1.74	1.10	1.81	3.95
	β_2	-1.20	-3.19	-1.62	9.80	99.49	-1.70	-1.30	1.73	3.21
	θ_0	-0.50	0.29	-0.68	8.60	74.28	-0.65	-0.77	1.02	1.06
	θ_1	1.00	5.87	1.03	54.46	2974.99	1.10	0.98	0.68	0.47
	θ_2	-1.00	0.56	-1.09	25.61	655.03	-0.93	-0.92	0.69	0.48
	γ	0.50	-0.95	0.33	10.06	102.89	0.62	0.32	2.75	7.54
n=200	β_0	-0.40	-0.05	-0.30	1.25	1.68	-0.09	-0.32	1.19	1.52
	β_1	0.90	1.24	1.11	0.59	0.46	1.20	1.07	0.55	0.39
	β_2	-1.20	-1.63	-1.39	1.03	1.24	-1.53	-1.33	0.94	0.99
	θ_0	-0.50	-0.45	-0.57	1.01	1.02	-0.53	-0.64	0.84	0.71
	θ_1	1.00	1.12	1.03	0.52	0.29	1.08	1.01	0.44	0.20
	θ_2	-1.00	-1.15	-1.01	0.76	0.59	-1.05	-0.93	0.59	0.34
	γ	0.50	0.36	0.39	1.65	2.72	0.37	0.32	1.49	2.24
n=300	β_0	-0.40	-0.17	-0.39	0.85	0.77	-0.19	-0.40	0.84	0.74
	β_1	0.90	1.13	1.01	0.51	0.31	1.10	0.99	0.49	0.28
	β_2	-1.20	-1.53	-1.34	0.72	0.63	-1.47	-1.30	0.66	0.51
	θ_0	-0.50	-0.51	-0.59	0.61	0.37	-0.53	-0.60	0.61	0.37
	θ_1	1.00	1.05	1.03	0.28	0.08	1.04	1.02	0.27	0.08
	θ_2	-1.00	-1.07	-1.02	0.45	0.21	-1.03	-0.97	0.44	0.19
	γ	0.50	0.41	0.50	1.40	1.95	0.38	0.39	1.28	1.63

n=400	β_0	-0.40	-0.31	-0.42	0.60	0.37	-0.31	-0.43	0.61	0.38
	β_1	0.90	1.01	0.95	0.27	0.09	1.00	0.94	0.27	0.08
	β_2	-1.20	-1.36	-1.28	0.52	0.29	-1.33	-1.26	0.51	0.28
	θ_0	-0.50	-0.47	-0.48	0.42	0.17	-0.49	-0.50	0.42	0.18
	θ_1	1.00	1.05	1.03	0.22	0.05	1.04	1.02	0.21	0.05
	θ_2	-1.00	-1.05	-1.01	0.30	0.09	-1.02	-0.98	0.30	0.09
	γ	0.50	0.48	0.46	0.78	0.61	0.45	0.42	0.74	0.55
n=500	β_0	-0.40	-0.23	-0.36	0.61	0.40	-0.22	-0.36	0.67	0.48
	β_1	0.90	1.07	0.98	0.37	0.17	1.07	0.98	0.44	0.22
	β_2	-1.20	-1.40	-1.31	0.52	0.31	-1.39	-1.29	0.55	0.33
	θ_0	-0.50	-0.54	-0.54	0.40	0.16	-0.56	-0.56	0.41	0.17
	θ_1	1.00	1.02	0.98	0.21	0.04	1.01	0.98	0.21	0.04
	θ_2	-1.00	-1.02	-0.99	0.29	0.08	-0.99	-0.97	0.29	0.08
	γ	0.50	0.25	0.46	1.25	1.62	0.25	0.42	1.23	1.56

Summary of simulated cases for the cluster sampling model using a discrete covariate that is common, clusters of size 3 and low dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	2.02	-0.57	15.62	248.68	-0.23	-0.61	1.42	2.04
	β_1	0.90	7.26	1.28	22.27	533.74	1.42	0.91	1.46	2.40
	β_2	-1.20	-3.38	-1.79	26.46	701.43	-1.44	-1.37	1.40	2.00
	θ_0	-0.50	3.40	-0.54	16.71	293.12	-0.54	-0.53	1.02	1.05
	θ_1	1.00	10.56	1.36	32.11	1117.32	1.35	1.06	0.90	0.93
	θ_2	-1.00	-2.53	-1.20	15.21	232.56	-1.27	-1.18	1.94	3.83
	γ	0.50	-1.51	0.17	20.36	416.32	0.00	0.08	2.13	4.75

n=100	β_0	-0.40	3.66	-0.45	20.98	454.33	-0.11	-0.51	2.10	4.49
	β_1	0.90	4.77	1.36	17.73	327.63	1.63	1.21	1.41	2.50
	β_2	-1.20	-4.71	-1.30	28.63	828.14	-1.05	-1.24	1.48	2.19
	θ_0	-0.50	-0.12	-0.63	5.68	32.30	-0.58	-0.61	0.81	0.65
	θ_1	1.00	1.55	1.05	4.11	17.15	1.17	1.02	0.64	0.43
	θ_2	-1.00	-1.46	-1.32	3.97	15.88	-1.22	-1.24	0.67	0.50
	γ	0.50	1.42	0.50	14.90	221.85	0.65	0.32	2.33	5.43
n=200	β_0	-0.40	-0.22	-0.44	0.82	0.69	-0.23	-0.48	0.79	0.65
	β_1	0.90	1.23	1.05	0.66	0.55	1.17	0.99	0.60	0.43
	β_2	-1.20	-1.22	-1.49	1.89	3.55	-1.26	-1.35	1.05	1.10
	θ_0	-0.50	-0.49	-0.55	0.53	0.28	-0.50	-0.55	0.50	0.25
	θ_1	1.00	1.18	1.09	0.44	0.22	1.15	1.06	0.39	0.17
	θ_2	-1.00	-0.91	-1.03	1.20	1.45	-0.99	-1.11	0.82	0.67
	γ	0.50	0.49	0.55	1.66	2.74	0.55	0.41	1.24	1.53
n=300	β_0	-0.40	-0.34	-0.38	0.53	0.28	-0.32	-0.39	0.59	0.35
	β_1	0.90	1.16	1.04	0.50	0.32	1.13	0.99	0.49	0.29
	β_2	-1.20	-1.16	-1.33	1.32	1.73	-1.25	-1.30	0.81	0.66
	θ_0	-0.50	-0.46	-0.52	0.46	0.21	-0.47	-0.50	0.46	0.21
	θ_1	1.00	1.12	1.05	0.39	0.17	1.11	1.04	0.39	0.16
	θ_2	-1.00	-1.06	-1.03	0.77	0.59	-1.05	-1.07	0.60	0.36
	γ	0.50	0.43	0.60	1.36	1.84	0.47	0.51	1.15	1.31
n=400	β_0	-0.40	-0.34	-0.48	0.52	0.27	-0.34	-0.47	0.49	0.24
	β_1	0.90	1.10	1.01	0.40	0.20	1.07	0.99	0.38	0.17
	β_2	-1.20	-1.02	-1.06	1.01	1.04	-1.08	-1.09	0.78	0.62
	θ_0	-0.50	-0.49	-0.52	0.38	0.14	-0.49	-0.52	0.38	0.14
	θ_1	1.00	1.06	1.04	0.25	0.07	1.06	1.04	0.25	0.07
	θ_2	-1.00	-1.14	-1.19	0.73	0.54	-1.11	-1.19	0.59	0.36
	γ	0.50	0.54	0.55	0.94	0.89	0.54	0.47	0.79	0.62

n=500	β_0	-0.40	-0.29	-0.38	0.49	0.25	-0.30	-0.38	0.47	0.23
	β_1	0.90	1.08	0.95	0.37	0.17	1.06	0.95	0.36	0.16
	β_2	-1.20	-1.19	-1.26	0.88	0.78	-1.18	-1.25	0.75	0.56
	θ_0	-0.50	-0.53	-0.53	0.32	0.10	-0.52	-0.53	0.32	0.10
	θ_1	1.00	1.03	1.03	0.21	0.05	1.03	1.03	0.21	0.05
	θ_2	-1.00	-1.01	-1.05	0.65	0.43	-1.02	-1.06	0.57	0.33
	γ	0.50	0.51	0.54	1.01	1.01	0.48	0.50	0.95	0.89

Summary of simulated cases for the cluster sampling model using a continuous covariate that is common, clusters of size 3 and low dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	5.72	-0.44	24.90	654.34	0.01	-0.38	2.07	4.43
	β_1	0.90	8.75	1.44	27.87	834.07	1.60	1.43	1.12	1.74
	β_2	-1.20	-11.79	-2.46	33.43	1223.85	-1.61	-1.42	1.52	2.47
	θ_0	-0.50	0.40	-0.77	14.16	200.28	-0.78	-0.96	1.33	1.84
	θ_1	1.00	7.44	1.37	26.83	757.52	1.28	1.26	0.55	0.38
	θ_2	-1.00	-2.95	-1.29	17.98	325.57	-0.81	-0.82	0.87	0.78
	γ	0.50	-1.39	-0.12	35.55	1260.96	-0.03	-0.06	2.04	4.42
n=100	β_0	-0.40	3.21	-0.24	14.71	228.30	0.18	-0.50	2.38	5.99
	β_1	0.90	1.70	1.16	7.41	55.32	1.26	1.27	1.13	1.41
	β_2	-1.20	-5.77	-1.97	14.91	241.95	-2.20	-1.57	1.77	4.12
	θ_0	-0.50	-0.55	-0.90	6.11	37.19	-0.59	-1.04	1.44	2.07
	θ_1	1.00	4.39	1.31	22.59	519.16	1.33	1.26	0.76	0.69
	θ_2	-1.00	-0.79	-1.12	7.15	50.90	-1.02	-0.89	0.82	0.66
	γ	0.50	0.52	0.56	14.83	218.96	1.16	0.33	2.79	8.18

n=200	β_0	-0.40	1.50	-0.13	3.79	17.94	0.89	-0.10	2.51	7.91
	β_1	0.90	1.14	1.08	1.06	1.18	1.11	1.10	0.83	0.72
	β_2	-1.20	-3.03	-1.80	3.17	13.35	-2.24	-1.64	1.55	3.49
	θ_0	-0.50	-0.36	-0.91	1.47	2.18	-0.52	-1.00	1.38	1.89
	θ_1	1.00	1.16	1.14	0.51	0.29	1.07	1.14	0.47	0.22
	θ_2	-1.00	-1.38	-1.06	1.00	1.14	-1.16	-0.98	0.69	0.50
	γ	0.50	0.35	0.69	2.80	7.82	0.48	0.54	1.99	3.92
n=300	β_0	-0.40	1.01	-0.27	3.29	12.73	0.67	-0.34	2.60	7.88
	β_1	0.90	1.03	1.07	0.76	0.59	1.03	1.14	0.67	0.47
	β_2	-1.20	-2.80	-1.49	4.36	21.53	-2.13	-1.41	1.71	3.77
	θ_0	-0.50	-0.29	-0.67	1.32	1.79	-0.38	-0.66	1.28	1.64
	θ_1	1.00	1.10	1.11	0.42	0.19	1.05	1.09	0.41	0.17
	θ_2	-1.00	-1.34	-1.08	0.84	0.81	-1.20	-1.02	0.65	0.46
	γ	0.50	0.67	0.58	4.18	17.41	0.43	0.47	1.81	3.25
n=400	β_0	-0.40	0.51	-0.33	3.00	9.80	0.14	-0.48	2.09	4.65
	β_1	0.90	1.02	1.03	0.58	0.35	1.02	1.07	0.60	0.37
	β_2	-1.20	-2.14	-1.29	2.47	6.96	-1.77	-1.24	1.35	2.12
	θ_0	-0.50	-0.05	-0.61	1.48	2.38	-0.17	-0.60	1.35	1.93
	θ_1	1.00	1.01	1.07	0.44	0.20	0.99	1.06	0.45	0.20
	θ_2	-1.00	-1.45	-1.09	1.00	1.19	-1.28	-1.05	0.73	0.61
	γ	0.50	0.71	0.55	1.82	3.35	0.58	0.44	1.20	1.45
n=500	β_0	-0.40	0.54	-0.30	2.59	7.54	0.18	-0.42	2.17	5.03
	β_1	0.90	0.97	0.99	0.57	0.33	0.98	1.01	0.57	0.33
	β_2	-1.20	-2.11	-1.34	2.03	4.92	-1.85	-1.27	1.53	2.75
	θ_0	-0.50	-0.24	-0.66	1.36	1.92	-0.33	-0.64	1.22	1.52
	θ_1	1.00	1.05	1.06	0.41	0.17	1.02	1.09	0.41	0.17
	θ_2	-1.00	-1.32	-1.02	0.81	0.76	-1.18	-1.00	0.55	0.34
	γ	0.50	0.56	0.56	1.63	2.66	0.71	0.47	1.52	2.35

Summary of simulated cases for the cluster sampling model using separate covariates, clusters of size 6 and low dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	-4.66	-0.39	18.89	372.66	-0.06	-0.39	3.03	9.23
	β_1	0.90	5.68	1.60	13.30	198.60	1.96	1.39	1.95	4.89
	β_2	-1.20	-5.83	-2.07	16.09	278.65	-1.84	-1.70	1.78	3.56
	θ_0	-0.50	0.14	-0.89	8.53	72.77	-0.90	-1.03	1.23	1.67
	θ_1	1.00	2.35	0.96	11.54	134.02	1.06	0.88	0.81	0.65
	θ_2	-1.00	-1.64	-0.98	5.46	30.06	-0.92	-0.87	0.69	0.48
	γ	0.50	3.62	0.70	13.81	199.34	0.67	0.52	1.80	3.23
n=100	β_0	-0.40	-0.20	-0.39	1.05	1.14	-0.23	-0.43	1.05	1.12
	β_1	0.90	1.13	1.00	0.57	0.38	1.10	0.99	0.54	0.34
	β_2	-1.20	-1.50	-1.31	0.86	0.83	-1.42	-1.23	0.83	0.73
	θ_0	-0.50	-0.45	-0.53	0.71	0.50	-0.48	-0.56	0.71	0.50
	θ_1	1.00	1.13	1.09	0.44	0.21	1.11	1.07	0.41	0.18
	θ_2	-1.00	-1.10	-0.99	0.57	0.33	-1.03	-0.91	0.54	0.30
	γ	0.50	0.31	0.47	0.94	0.91	0.34	0.43	0.81	0.68
n=200	β_0	-0.40	-0.31	-0.42	0.56	0.32	-0.36	-0.45	0.54	0.29
	β_1	0.90	1.02	0.94	0.33	0.12	0.99	0.93	0.30	0.10
	β_2	-1.20	-1.33	-1.27	0.47	0.23	-1.28	-1.24	0.45	0.21
	θ_0	-0.50	-0.49	-0.47	0.40	0.16	-0.48	-0.47	0.40	0.16
	θ_1	1.00	1.05	1.04	0.24	0.06	1.06	1.04	0.24	0.06
	θ_2	-1.00	-1.05	-1.05	0.27	0.08	-1.03	-1.04	0.27	0.07
	γ	0.50	0.46	0.48	0.37	0.14	0.46	0.47	0.33	0.11

n=300	β_0	-0.40	-0.35	-0.43	0.49	0.24	-0.35	-0.43	0.52	0.27
	β_1	0.90	0.98	0.93	0.28	0.08	0.98	0.93	0.28	0.08
	β_2	-1.20	-1.26	-1.21	0.39	0.15	-1.25	-1.20	0.39	0.15
	θ_0	-0.50	-0.50	-0.48	0.31	0.10	-0.50	-0.48	0.32	0.10
	θ_1	1.00	1.02	1.01	0.15	0.02	1.02	1.01	0.16	0.02
	θ_2	-1.00	-1.02	-1.00	0.22	0.05	-1.01	-0.99	0.22	0.05
	γ	0.50	0.48	0.50	0.32	0.11	0.48	0.49	0.32	0.10
n=400	β_0	-0.40	-0.30	-0.37	0.46	0.22	-0.31	-0.37	0.46	0.22
	β_1	0.90	0.98	0.95	0.21	0.05	0.98	0.95	0.21	0.05
	β_2	-1.20	-1.30	-1.27	0.32	0.11	-1.29	-1.26	0.32	0.11
	θ_0	-0.50	-0.52	-0.51	0.34	0.12	-0.52	-0.51	0.35	0.12
	θ_1	1.00	1.00	1.01	0.15	0.02	1.00	1.01	0.15	0.02
	θ_2	-1.00	-1.00	-0.98	0.23	0.05	-0.99	-0.98	0.23	0.05
	γ	0.50	0.51	0.50	0.20	0.04	0.51	0.50	0.20	0.04
n=500	β_0	-0.40	-0.38	-0.40	0.29	0.09	-0.39	-0.41	0.29	0.09
	β_1	0.90	0.94	0.91	0.17	0.03	0.94	0.91	0.17	0.03
	β_2	-1.20	-1.23	-1.21	0.23	0.05	-1.23	-1.20	0.23	0.05
	θ_0	-0.50	-0.50	-0.49	0.25	0.06	-0.51	-0.49	0.25	0.06
	θ_1	1.00	1.02	1.02	0.12	0.02	1.02	1.02	0.12	0.02
	θ_2	-1.00	-1.01	-1.01	0.18	0.03	-1.00	-1.00	0.18	0.03
	γ	0.50	0.50	0.50	0.16	0.03	0.50	0.50	0.16	0.03

Summary of simulated cases for the cluster sampling model using a discrete covariate that is common, clusters of size 6 and low dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	-5.10	-0.69	18.44	360.39	-0.97	-0.76	2.70	7.59
	β_1	0.90	4.96	1.83	11.05	137.90	2.03	1.57	2.14	5.82
	β_2	-1.20	-4.55	-1.57	20.70	437.40	-1.60	-1.56	2.37	5.77
	θ_0	-0.50	-0.86	-0.93	0.66	0.57	-0.97	-1.06	0.72	0.73
	θ_1	1.00	1.02	0.91	0.49	0.24	0.97	0.83	0.46	0.21
	θ_2	-1.00	-1.23	-1.33	0.95	0.95	-1.26	-1.33	0.62	0.44
	γ	0.50	4.43	1.17	13.09	185.91	1.41	1.16	1.87	4.31
n=100	β_0	-0.40	0.02	-0.32	1.83	3.50	-0.11	-0.34	1.08	1.25
	β_1	0.90	1.43	1.06	1.72	3.22	1.29	1.01	1.02	1.19
	β_2	-1.20	-1.19	-1.28	2.14	4.55	-1.18	-1.19	1.06	1.11
	θ_0	-0.50	-0.54	-0.65	0.54	0.29	-0.54	-0.64	0.52	0.27
	θ_1	1.00	1.09	1.00	0.36	0.14	1.07	0.99	0.36	0.13
	θ_2	-1.00	-0.97	-1.11	1.03	1.06	-1.04	-1.12	0.65	0.42
	γ	0.50	0.35	0.50	1.29	1.68	0.33	0.46	1.05	1.12
n=200	β_0	-0.40	-0.35	-0.46	0.60	0.36	-0.36	-0.47	0.57	0.32
	β_1	0.90	1.06	0.97	0.37	0.16	1.05	0.96	0.35	0.14
	β_2	-1.20	-1.05	-1.09	0.89	0.81	-1.06	-1.07	0.74	0.56
	θ_0	-0.50	-0.52	-0.54	0.38	0.14	-0.52	-0.55	0.37	0.14
	θ_1	1.00	1.05	1.02	0.25	0.06	1.05	1.01	0.25	0.06
	θ_2	-1.00	-1.08	-1.13	0.68	0.47	-1.09	-1.12	0.58	0.34
	γ	0.50	0.57	0.52	0.29	0.09	0.56	0.52	0.28	0.08

n=300	β_0	-0.40	-0.31	-0.37	0.62	0.39	-0.33	-0.38	0.58	0.34
	β_1	0.90	1.04	0.96	0.38	0.16	1.03	0.95	0.35	0.14
	β_2	-1.20	-1.24	-1.30	0.69	0.48	-1.23	-1.30	0.62	0.39
	θ_0	-0.50	-0.54	-0.52	0.30	0.09	-0.53	-0.52	0.30	0.09
	θ_1	1.00	1.03	1.01	0.19	0.04	1.04	1.02	0.19	0.04
	θ_2	-1.00	-0.98	-1.02	0.55	0.31	-0.99	-1.02	0.51	0.26
	γ	0.50	0.49	0.53	0.56	0.31	0.51	0.52	0.47	0.22
n=400	β_0	-0.40	-0.38	-0.43	0.33	0.11	-0.38	-0.42	0.33	0.11
	β_1	0.90	0.96	0.92	0.21	0.05	0.96	0.91	0.21	0.05
	β_2	-1.20	-1.16	-1.19	0.49	0.24	-1.17	-1.19	0.45	0.20
	θ_0	-0.50	-0.49	-0.50	0.25	0.06	-0.49	-0.49	0.25	0.06
	θ_1	1.00	1.03	1.02	0.16	0.02	1.03	1.02	0.16	0.02
	θ_2	-1.00	-1.05	-1.06	0.42	0.18	-1.05	-1.06	0.40	0.16
	γ	0.50	0.52	0.51	0.18	0.03	0.52	0.51	0.18	0.03
n=500	β_0	-0.40	-0.37	-0.41	0.26	0.07	-0.37	-0.41	0.26	0.07
	β_1	0.90	0.95	0.92	0.16	0.03	0.95	0.92	0.16	0.03
	β_2	-1.20	-1.20	-1.19	0.42	0.18	-1.20	-1.20	0.40	0.16
	θ_0	-0.50	-0.52	-0.53	0.19	0.04	-0.52	-0.53	0.19	0.04
	θ_1	1.00	1.01	1.01	0.13	0.02	1.01	1.01	0.13	0.02
	θ_2	-1.00	-1.03	-1.04	0.37	0.13	-1.03	-1.04	0.35	0.12
	γ	0.50	0.53	0.51	0.15	0.02	0.52	0.51	0.15	0.02

Summary of simulated cases for the cluster sampling model using a continuous covariate that is common, clusters of size 6 and low dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	-1.53	-0.73	12.07	146.27	-0.78	-1.20	3.02	9.21
	β_1	0.90	3.82	1.36	10.71	122.56	1.77	1.48	1.55	3.13
	β_2	-1.20	-5.02	-2.50	9.44	103.33	-2.43	-2.11	2.21	6.34
	θ_0	-0.50	-0.80	-1.22	1.40	2.05	-1.07	-1.51	1.22	1.80
	θ_1	1.00	1.37	1.24	0.83	0.82	1.26	1.22	0.53	0.34
	θ_2	-1.00	-1.22	-0.96	1.09	1.23	-0.87	-0.80	0.69	0.49
	γ	0.50	4.08	1.30	11.41	142.31	1.46	1.35	1.69	3.76
n=100	β_0	-0.40	0.34	-0.47	3.33	11.58	0.26	-0.60	2.31	5.76
	β_1	0.90	1.22	1.17	0.90	0.91	1.13	1.18	0.70	0.55
	β_2	-1.20	-2.13	-1.43	2.16	5.51	-1.78	-1.31	1.39	2.27
	θ_0	-0.50	0.01	-0.62	2.55	6.71	-0.23	-0.66	1.44	2.12
	θ_1	1.00	1.55	1.05	7.01	49.13	1.02	1.07	0.80	0.64
	θ_2	-1.00	-1.30	-1.08	1.26	1.68	-1.14	-1.01	0.65	0.43
	γ	0.50	0.58	0.62	1.35	1.82	0.52	0.57	0.83	0.69
n=200	β_0	-0.40	0.32	-0.15	2.13	5.05	-0.09	-0.34	1.54	2.47
	β_1	0.90	1.01	1.04	0.59	0.36	1.01	1.04	0.53	0.29
	β_2	-1.20	-1.90	-1.43	1.67	3.27	-1.61	-1.35	1.06	1.30
	θ_0	-0.50	-0.40	-0.81	1.22	1.49	-0.45	-0.85	1.14	1.29
	θ_1	1.00	1.04	1.05	0.35	0.12	1.01	1.07	0.35	0.12
	θ_2	-1.00	-1.22	-0.97	0.75	0.61	-1.13	-0.94	0.58	0.35
	γ	0.50	0.59	0.61	0.83	0.69	0.66	0.62	0.61	0.40

n=300	β_0	-0.40	-0.59	-0.37	3.74	13.97	-0.56	-0.44	2.33	5.45
	β_1	0.90	0.99	0.89	1.33	1.78	0.97	0.90	0.93	0.86
	β_2	-1.20	-3.67	-1.42	12.87	170.99	-1.73	-1.43	1.84	3.65
	θ_0	-0.50	-0.67	-0.87	1.06	1.15	-0.72	-0.91	1.26	1.64
	θ_1	1.00	1.07	1.16	0.29	0.09	1.08	1.16	0.28	0.08
	θ_2	-1.00	-1.10	-0.96	0.57	0.33	-1.07	-0.91	0.83	0.69
	γ	0.50	1.86	0.66	5.44	31.32	1.09	0.70	1.35	2.17
n=400	β_0	-0.40	-0.11	-0.41	1.40	2.03	-0.16	-0.43	1.37	1.94
	β_1	0.90	0.98	1.02	0.41	0.18	0.99	1.01	0.41	0.17
	β_2	-1.20	-1.51	-1.30	0.93	0.96	-1.47	-1.27	0.86	0.81
	θ_0	-0.50	-0.23	-0.52	1.13	1.34	-0.25	-0.52	1.06	1.18
	θ_1	1.00	0.97	1.02	0.30	0.09	0.96	1.02	0.31	0.10
	θ_2	-1.00	-1.25	-1.00	0.67	0.51	-1.20	-1.00	0.53	0.32
	γ	0.50	0.57	0.53	0.39	0.16	0.57	0.52	0.40	0.16
n=500	β_0	-0.40	-0.29	-0.33	0.93	0.87	-0.32	-0.35	0.97	0.94
	β_1	0.90	1.01	1.02	0.34	0.12	1.02	1.03	0.33	0.12
	β_2	-1.20	-1.36	-1.20	0.46	0.24	-1.35	-1.18	0.47	0.24
	θ_0	-0.50	-0.20	-0.60	1.08	1.25	-0.20	-0.58	1.02	1.11
	θ_1	1.00	0.94	0.96	0.29	0.09	0.93	0.97	0.30	0.09
	θ_2	-1.00	-1.23	-1.02	0.67	0.50	-1.20	-1.01	0.55	0.34
	γ	0.50	0.54	0.51	0.20	0.04	0.53	0.50	0.21	0.04

Summary of simulated cases for the cluster sampling model using separate covariates, clusters of size 3 and large dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	6.08	-0.71	39.67	1608.01	-0.15	-0.68	2.13	4.59
	β_1	0.90	6.99	1.24	24.22	620.42	1.35	0.96	1.60	2.76
	β_2	-1.20	-10.17	-1.76	35.21	1314.05	-1.39	-1.16	1.60	2.57
	θ_0	-0.50	1.86	-0.65	17.34	304.64	-0.81	-0.82	1.90	3.68
	θ_1	1.00	9.74	1.22	30.74	1016.49	1.30	1.02	0.96	1.01
	θ_2	-1.00	-3.28	-1.23	15.63	248.19	-0.70	-0.75	1.64	2.76
	γ	1.00	2.97	0.75	29.81	888.07	-0.49	0.31	13.89	194.09
n=100	β_0	-0.40	2.42	-0.32	13.48	188.78	0.13	-0.37	2.18	5.01
	β_1	0.90	4.35	1.15	16.83	293.81	1.59	1.05	1.66	3.21
	β_2	-1.20	-3.22	-1.68	10.00	103.60	-1.67	-1.32	1.60	2.78
	θ_0	-0.50	0.69	-0.66	11.15	125.12	-0.63	-0.84	1.03	1.08
	θ_1	1.00	5.36	1.09	34.31	1190.27	1.23	1.03	0.98	1.01
	θ_2	-1.00	-2.32	-1.01	10.79	117.49	-0.98	-0.86	0.72	0.52
	γ	1.00	1.81	1.11	8.41	70.98	1.11	0.73	2.39	5.70
n=200	β_0	-0.40	-0.21	-0.51	1.45	2.12	-0.26	-0.54	1.20	1.45
	β_1	0.90	1.20	0.97	0.94	0.98	1.15	0.93	0.76	0.64
	β_2	-1.20	-1.58	-1.36	1.23	1.66	-1.45	-1.27	0.98	1.01
	θ_0	-0.50	-0.35	-0.44	0.91	0.85	-0.42	-0.48	0.79	0.63
	θ_1	1.00	1.16	1.06	0.53	0.31	1.12	1.04	0.44	0.21
	θ_2	-1.00	-1.19	-1.09	0.80	0.67	-1.09	-1.02	0.65	0.43
	γ	1.00	1.00	0.99	1.41	1.97	0.94	0.88	1.23	1.52

n=300	β_0	-0.40	-0.15	-0.47	1.93	3.76	-0.29	-0.49	0.86	0.74
	β_1	0.90	1.18	0.97	1.24	1.60	1.08	0.96	0.51	0.29
	β_2	-1.20	-1.50	-1.22	1.31	1.80	-1.38	-1.19	0.67	0.47
	θ_0	-0.50	-0.47	-0.50	0.70	0.50	-0.50	-0.53	0.66	0.43
	θ_1	1.00	1.08	1.03	0.29	0.09	1.07	1.02	0.28	0.09
	θ_2	-1.00	-1.07	-1.00	0.51	0.26	-1.02	-0.96	0.44	0.19
	γ	1.00	0.94	0.99	1.45	2.10	0.99	0.91	1.03	1.06
n=400	β_0	-0.40	-0.33	-0.42	0.56	0.32	-0.33	-0.42	0.57	0.33
	β_1	0.90	1.01	0.97	0.28	0.09	1.00	0.97	0.28	0.09
	β_2	-1.20	-1.34	-1.27	0.45	0.22	-1.31	-1.24	0.45	0.21
	θ_0	-0.50	-0.47	-0.50	0.41	0.17	-0.49	-0.51	0.42	0.18
	θ_1	1.00	1.03	1.00	0.20	0.04	1.02	0.98	0.20	0.04
	θ_2	-1.00	-1.06	-1.02	0.30	0.10	-1.04	-0.99	0.30	0.09
	γ	1.00	1.02	0.93	0.63	0.39	0.99	0.90	0.62	0.38
n=500	β_0	-0.40	-0.30	-0.42	0.67	0.45	-0.32	-0.43	0.63	0.40
	β_1	0.90	1.01	0.95	0.31	0.11	0.99	0.94	0.29	0.09
	β_2	-1.20	-1.37	-1.24	0.55	0.33	-1.34	-1.22	0.54	0.31
	θ_0	-0.50	-0.49	-0.50	0.38	0.14	-0.49	-0.49	0.38	0.15
	θ_1	1.00	1.04	1.02	0.19	0.04	1.04	1.01	0.19	0.04
	θ_2	-1.00	-1.04	-1.02	0.27	0.07	-1.02	-1.01	0.27	0.07
	γ	1.00	1.00	1.03	0.78	0.61	0.95	1.00	0.75	0.57

Summary of simulated cases for the cluster sampling model using a discrete covariate that is common, clusters of size 3 and large dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	3.29	-0.65	21.27	463.59	-0.25	-0.68	1.59	2.55
	β_1	0.90	7.20	1.25	22.57	546.22	1.28	0.94	1.17	1.51
	β_2	-1.20	-7.70	-1.47	32.40	1086.49	-1.61	-1.48	1.50	2.41
	θ_0	-0.50	7.11	-0.45	26.49	756.16	-0.34	-0.56	1.21	1.49
	θ_1	1.00	15.09	1.56	43.90	2116.07	1.50	1.22	1.08	1.40
	θ_2	-1.00	-7.37	-1.44	33.32	1144.95	-1.33	-1.17	2.06	4.34
	γ	1.00	0.90	0.79	20.65	424.09	0.64	0.47	1.87	3.62
n=100	β_0	-0.40	0.87	-0.35	6.12	38.86	0.02	-0.45	1.69	3.03
	β_1	0.90	2.80	1.25	7.74	63.26	1.55	1.12	1.27	2.03
	β_2	-1.20	-1.48	-1.68	7.65	58.26	-1.30	-1.37	1.46	2.14
	θ_0	-0.50	1.10	-0.62	13.70	189.44	-0.48	-0.62	0.87	0.76
	θ_1	1.00	3.77	1.14	19.20	374.31	1.26	1.07	0.84	0.78
	θ_2	-1.00	-2.21	-1.12	17.22	296.63	-1.15	-1.20	0.86	0.76
	γ	1.00	0.89	1.11	5.37	28.74	0.86	0.72	1.97	3.89
n=200	β_0	-0.40	-0.21	-0.44	1.48	2.21	-0.25	-0.50	1.24	1.56
	β_1	0.90	1.28	1.02	1.01	1.17	1.22	1.00	0.79	0.72
	β_2	-1.20	-1.35	-1.41	1.55	2.41	-1.32	-1.33	0.96	0.93
	θ_0	-0.50	-0.43	-0.50	0.60	0.37	-0.44	-0.51	0.56	0.32
	θ_1	1.00	1.23	1.13	0.52	0.32	1.19	1.11	0.47	0.26
	θ_2	-1.00	-0.99	-1.05	1.02	1.04	-1.03	-1.10	0.76	0.57
	γ	1.00	1.22	1.19	1.40	2.01	1.18	1.08	1.06	1.15

n=300	β_0	-0.40	-0.30	-0.44	0.67	0.46	-0.28	-0.44	0.73	0.54
	β_1	0.90	1.15	0.98	0.58	0.40	1.15	0.96	0.60	0.42
	β_2	-1.20	-1.14	-1.18	1.10	1.21	-1.17	-1.17	0.86	0.74
	θ_0	-0.50	-0.48	-0.50	0.41	0.16	-0.49	-0.50	0.41	0.17
	θ_1	1.00	1.08	1.02	0.32	0.11	1.07	1.01	0.32	0.10
	θ_2	-1.00	-1.04	-1.18	0.90	0.81	-1.05	-1.17	0.71	0.51
	γ	1.00	1.29	1.17	0.87	0.84	1.22	1.10	0.88	0.81
n=400	β_0	-0.40	-0.29	-0.42	0.64	0.42	-0.31	-0.43	0.62	0.39
	β_1	0.90	1.08	1.00	0.38	0.18	1.06	0.99	0.38	0.17
	β_2	-1.20	-1.25	-1.31	0.96	0.92	-1.24	-1.28	0.80	0.64
	θ_0	-0.50	-0.52	-0.52	0.39	0.15	-0.51	-0.54	0.38	0.14
	θ_1	1.00	1.07	1.05	0.28	0.09	1.07	1.04	0.28	0.08
	θ_2	-1.00	-1.03	-1.10	0.67	0.45	-1.04	-1.12	0.55	0.30
	γ	1.00	1.11	1.02	0.84	0.71	1.05	0.96	0.83	0.69
n=500	β_0	-0.40	-0.25	-0.40	0.67	0.46	-0.29	-0.41	0.59	0.36
	β_1	0.90	1.07	0.96	0.37	0.17	1.04	0.95	0.36	0.15
	β_2	-1.20	-1.23	-1.24	0.81	0.65	-1.21	-1.23	0.64	0.41
	θ_0	-0.50	-0.52	-0.54	0.34	0.12	-0.51	-0.53	0.33	0.11
	θ_1	1.00	1.05	1.00	0.25	0.07	1.05	1.01	0.25	0.07
	θ_2	-1.00	-1.03	-1.09	0.57	0.32	-1.04	-1.10	0.50	0.25
	γ	1.00	1.05	1.01	1.02	1.03	1.04	0.96	0.85	0.73

Summary of simulated cases for the cluster sampling model using a continuous covariate that is common, clusters of size 3 and large dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	6.41	-0.71	33.86	1186.98	-0.30	-1.05	2.45	5.98
	β_1	0.90	7.43	1.33	24.84	656.52	1.45	1.45	1.09	1.48
	β_2	-1.20	-7.34	-1.75	24.75	647.16	-1.29	-1.07	1.53	2.33
	θ_0	-0.50	-0.24	-0.71	17.12	291.49	-0.69	-0.82	1.36	1.88
	θ_1	1.00	10.50	1.45	30.29	1002.76	1.23	1.23	0.71	0.55
	θ_2	-1.00	-3.12	-1.27	26.07	680.30	-0.75	-0.79	0.95	0.96
	γ	1.00	3.95	0.67	33.64	1134.19	0.76	0.36	2.10	4.45
n=100	β_0	-0.40	0.61	-0.74	10.81	117.19	-0.33	-1.14	2.28	5.20
	β_1	0.90	2.41	1.07	12.81	165.59	1.04	1.14	1.04	1.09
	β_2	-1.20	-4.13	-1.62	15.14	236.59	-1.66	-1.34	1.50	2.43
	θ_0	-0.50	2.32	-0.75	18.40	344.66	-0.39	-0.79	1.44	2.08
	θ_1	1.00	4.22	1.27	20.17	415.05	1.26	1.23	0.78	0.67
	θ_2	-1.00	-1.41	-1.20	13.32	176.75	-0.97	-0.89	0.86	0.73
	γ	1.00	3.20	1.58	14.56	215.83	1.71	1.01	2.63	7.40
n=200	β_0	-0.40	0.67	-0.61	4.02	17.20	0.13	-0.71	2.37	5.86
	β_1	0.90	1.25	1.17	0.98	1.08	1.18	1.20	0.88	0.85
	β_2	-1.20	-2.39	-1.38	3.58	14.13	-1.80	-1.26	1.60	2.91
	θ_0	-0.50	0.22	-0.34	2.79	8.28	-0.14	-0.29	1.38	2.02
	θ_1	1.00	1.92	1.08	11.99	143.85	1.02	1.05	0.57	0.32
	θ_2	-1.00	-1.74	-1.17	2.27	5.68	-1.28	-1.08	0.76	0.66
	γ	1.00	1.56	1.15	2.52	6.62	1.12	0.90	1.66	2.74

n=300	β_0	-0.40	0.29	-0.44	2.71	7.80	0.01	-0.53	2.01	4.21
	β_1	0.90	1.05	1.03	0.68	0.48	1.03	1.03	0.63	0.41
	β_2	-1.20	-2.03	-1.30	2.34	6.15	-1.66	-1.22	1.34	1.98
	θ_0	-0.50	-0.11	-0.63	1.47	2.29	-0.21	-0.61	1.26	1.65
	θ_1	1.00	1.11	1.06	0.64	0.42	1.05	1.06	0.45	0.20
	θ_2	-1.00	-1.40	-1.05	0.93	1.03	-1.25	-1.02	0.68	0.52
	γ	1.00	1.32	1.12	1.07	1.24	1.19	1.03	0.91	0.85
n=400	β_0	-0.40	0.26	-0.20	2.22	5.33	0.03	-0.25	1.69	3.03
	β_1	0.90	1.03	1.04	0.61	0.39	1.01	1.06	0.53	0.29
	β_2	-1.20	-1.95	-1.49	1.65	3.28	-1.73	-1.40	1.07	1.42
	θ_0	-0.50	-0.27	-0.74	1.21	1.51	-0.31	-0.72	1.09	1.22
	θ_1	1.00	1.02	1.05	0.35	0.12	0.98	1.05	0.37	0.13
	θ_2	-1.00	-1.29	-1.05	0.78	0.68	-1.17	-1.03	0.54	0.32
	γ	1.00	1.34	1.23	1.15	1.43	1.27	1.12	1.09	1.26
n=500	β_0	-0.40	-0.07	-0.48	1.70	2.98	-0.22	-0.59	1.54	2.38
	β_1	0.90	1.01	1.00	0.45	0.21	1.01	1.02	0.44	0.20
	β_2	-1.20	-1.60	-1.31	1.20	1.58	-1.51	-1.24	1.05	1.20
	θ_0	-0.50	-0.14	-0.53	1.24	1.66	-0.18	-0.46	1.14	1.39
	θ_1	1.00	0.98	1.00	0.35	0.12	0.96	0.99	0.36	0.13
	θ_2	-1.00	-1.32	-1.05	0.78	0.70	-1.23	-1.04	0.59	0.39
	γ	1.00	1.32	1.13	0.94	0.98	1.31	1.05	1.05	1.19

Summary of simulated cases for the cluster sampling model using separate covariates, clusters of size 6 and large dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	1.39	0.16	11.04	124.52	0.35	0.01	2.26	5.64
	β_1	0.90	1.60	1.02	3.86	15.28	1.14	0.96	1.03	1.11
	β_2	-1.20	-2.83	-1.52	9.56	93.66	-1.57	-1.24	1.73	3.12
	θ_0	-0.50	-0.50	-0.65	0.79	0.62	-0.64	-0.77	0.71	0.52
	θ_1	1.00	1.11	1.03	0.64	0.42	1.04	0.98	0.47	0.22
	θ_2	-1.00	-1.10	-1.04	0.57	0.33	-0.97	-0.93	0.50	0.25
	γ	1.00	1.51	1.17	2.99	9.16	1.21	1.15	1.20	1.47
n=100	β_0	-0.40	-0.25	-0.35	0.90	0.84	-0.32	-0.40	0.93	0.87
	β_1	0.90	1.00	0.98	0.35	0.13	1.00	0.97	0.38	0.15
	β_2	-1.20	-1.41	-1.29	0.88	0.81	-1.36	-1.25	0.84	0.73
	θ_0	-0.50	-0.49	-0.52	0.36	0.13	-0.53	-0.55	0.36	0.13
	θ_1	1.00	1.04	1.02	0.22	0.05	1.03	1.00	0.22	0.05
	θ_2	-1.00	-1.04	-1.02	0.29	0.09	-0.99	-0.98	0.29	0.08
	γ	1.00	1.09	1.09	0.32	0.11	1.09	1.09	0.29	0.09
n=200	β_0	-0.40	-0.30	-0.33	0.52	0.28	-0.28	-0.36	0.74	0.55
	β_1	0.90	0.95	0.95	0.23	0.05	0.95	0.95	0.23	0.05
	β_2	-1.20	-1.25	-1.22	0.36	0.13	-1.23	-1.21	0.37	0.13
	θ_0	-0.50	-0.51	-0.51	0.21	0.04	-0.53	-0.52	0.21	0.04
	θ_1	1.00	1.01	1.01	0.13	0.02	1.01	1.00	0.13	0.02
	θ_2	-1.00	-1.01	-1.00	0.19	0.04	-0.99	-0.98	0.19	0.04
	γ	1.00	1.02	1.03	0.21	0.04	1.02	1.02	0.18	0.03

n=300	β_0	-0.40	-0.30	-0.38	0.54	0.30	-0.33	-0.39	0.52	0.27
	β_1	0.90	0.93	0.92	0.19	0.04	0.93	0.92	0.19	0.04
	β_2	-1.20	-1.16	-1.18	0.32	0.11	-1.15	-1.17	0.32	0.11
	θ_0	-0.50	-0.50	-0.52	0.17	0.03	-0.51	-0.53	0.17	0.03
	θ_1	1.00	1.03	1.02	0.12	0.01	1.03	1.01	0.12	0.01
	θ_2	-1.00	-1.01	-0.99	0.17	0.03	-1.00	-0.99	0.17	0.03
	γ	1.00	1.00	1.01	0.11	0.01	1.00	1.01	0.11	0.01
n=400	β_0	-0.40	-0.36	-0.40	0.33	0.11	-0.37	-0.41	0.33	0.11
	β_1	0.90	0.92	0.91	0.15	0.02	0.92	0.91	0.15	0.02
	β_2	-1.20	-1.21	-1.23	0.30	0.09	-1.21	-1.23	0.30	0.09
	θ_0	-0.50	-0.50	-0.52	0.14	0.02	-0.51	-0.52	0.14	0.02
	θ_1	1.00	1.01	1.00	0.09	0.01	1.01	1.00	0.09	0.01
	θ_2	-1.00	-1.01	-1.02	0.15	0.02	-1.01	-1.01	0.15	0.02
	γ	1.00	1.01	1.01	0.08	0.01	1.01	1.01	0.08	0.01
n=500	β_0	-0.40	-0.41	-0.42	0.28	0.08	-0.41	-0.42	0.28	0.08
	β_1	0.90	0.90	0.89	0.13	0.02	0.90	0.89	0.13	0.02
	β_2	-1.20	-1.22	-1.22	0.26	0.07	-1.22	-1.22	0.26	0.07
	θ_0	-0.50	-0.51	-0.52	0.13	0.02	-0.51	-0.53	0.13	0.02
	θ_1	1.00	1.01	1.00	0.09	0.01	1.01	1.00	0.09	0.01
	θ_2	-1.00	-0.99	-0.98	0.15	0.02	-0.99	-0.98	0.15	0.02
	γ	1.00	1.01	1.01	0.07	0.01	1.01	1.01	0.07	0.01

Summary of simulated cases for the cluster sampling model using a discrete covariate that is common, clusters of size 6 and large dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	0.10	-0.22	4.93	24.45	-0.34	-0.49	1.34	1.79
	β_1	0.90	1.38	0.98	3.85	15.00	1.11	1.03	0.83	0.73
	β_2	-1.20	-1.01	-1.54	6.27	39.15	-1.40	-1.29	1.33	1.80
	θ_0	-0.50	-0.01	-0.52	6.75	45.55	-0.53	-0.56	0.58	0.34
	θ_1	1.00	1.79	1.06	8.87	78.95	1.11	1.02	0.55	0.31
	θ_2	-1.00	-1.50	-0.97	6.87	47.14	-1.03	-1.02	0.51	0.26
	γ	1.00	1.43	1.19	2.27	5.29	1.24	1.19	0.63	0.46
n=100	β_0	-0.40	-0.31	-0.33	0.93	0.87	-0.39	-0.43	0.90	0.81
	β_1	0.90	0.97	0.97	0.39	0.16	0.98	0.98	0.38	0.15
	β_2	-1.20	-1.23	-1.30	0.96	0.92	-1.20	-1.23	0.82	0.67
	θ_0	-0.50	-0.49	-0.54	0.35	0.12	-0.50	-0.54	0.32	0.10
	θ_1	1.00	1.05	1.01	0.27	0.07	1.04	0.99	0.25	0.06
	θ_2	-1.00	-0.99	-1.00	0.54	0.29	-1.00	-1.04	0.43	0.18
	γ	1.00	1.06	1.08	0.26	0.07	1.05	1.08	0.22	0.05
n=200	β_0	-0.40	-0.41	-0.44	0.50	0.25	-0.44	-0.45	0.47	0.22
	β_1	0.90	0.94	0.92	0.22	0.05	0.94	0.92	0.22	0.05
	β_2	-1.20	-1.28	-1.30	0.59	0.35	-1.27	-1.27	0.54	0.30
	θ_0	-0.50	-0.50	-0.52	0.20	0.04	-0.50	-0.52	0.19	0.04
	θ_1	1.00	1.01	0.99	0.14	0.02	1.01	0.99	0.14	0.02
	θ_2	-1.00	-1.01	-1.00	0.30	0.09	-1.01	-1.01	0.27	0.07
	γ	1.00	1.04	1.05	0.13	0.02	1.04	1.06	0.13	0.02

n=300	β_0	-0.40	-0.28	-0.37	0.54	0.30	-0.30	-0.38	0.52	0.28
	β_1	0.90	0.92	0.92	0.18	0.03	0.92	0.92	0.18	0.03
	β_2	-1.20	-1.19	-1.21	0.44	0.19	-1.18	-1.20	0.41	0.17
	θ_0	-0.50	-0.51	-0.50	0.14	0.02	-0.51	-0.51	0.14	0.02
	θ_1	1.00	1.02	1.02	0.12	0.01	1.02	1.01	0.12	0.01
	θ_2	-1.00	-1.02	-1.04	0.24	0.06	-1.02	-1.05	0.23	0.05
	γ	1.00	1.02	1.03	0.10	0.01	1.02	1.04	0.10	0.01
n=400	β_0	-0.40	-0.41	-0.42	0.42	0.18	-0.42	-0.43	0.41	0.17
	β_1	0.90	0.91	0.91	0.15	0.02	0.92	0.91	0.15	0.02
	β_2	-1.20	-1.14	-1.19	0.41	0.17	-1.13	-1.17	0.40	0.16
	θ_0	-0.50	-0.50	-0.50	0.12	0.01	-0.50	-0.50	0.12	0.01
	θ_1	1.00	1.00	1.00	0.08	0.01	1.00	1.00	0.08	0.01
	θ_2	-1.00	-1.03	-1.02	0.21	0.05	-1.04	-1.02	0.21	0.04
	γ	1.00	1.02	1.02	0.08	0.01	1.02	1.02	0.08	0.01
n=500	β_0	-0.40	-0.39	-0.37	0.29	0.08	-0.39	-0.37	0.29	0.08
	β_1	0.90	0.92	0.91	0.14	0.02	0.92	0.91	0.14	0.02
	β_2	-1.20	-1.23	-1.22	0.31	0.10	-1.23	-1.22	0.31	0.09
	θ_0	-0.50	-0.51	-0.51	0.10	0.01	-0.51	-0.51	0.10	0.01
	θ_1	1.00	1.00	1.00	0.08	0.01	1.00	1.00	0.08	0.01
	θ_2	-1.00	-1.00	-1.02	0.18	0.03	-1.01	-1.02	0.18	0.03
	γ	1.00	1.01	1.02	0.07	0.01	1.02	1.02	0.07	0.01

Summary of simulated cases for the cluster sampling model using a continuous covariate that is common, clusters of size 6 and large dependence.

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=50	β_0	-0.40	0.10	-0.49	12.31	151.14	-0.43	-0.74	2.05	4.19
	β_1	0.90	1.62	1.16	6.15	38.10	1.26	1.16	1.21	1.59
	β_2	-1.20	-2.09	-1.33	7.04	50.07	-1.31	-1.13	1.76	3.10
	θ_0	-0.50	-0.45	-0.62	0.93	0.86	-0.63	-0.72	0.80	0.64
	θ_1	1.00	1.15	1.13	0.51	0.28	1.10	1.08	0.39	0.16
	θ_2	-1.00	-1.22	-1.07	0.71	0.56	-1.02	-0.95	0.51	0.26
	γ	1.00	2.34	1.27	5.35	30.29	1.48	1.28	1.21	1.69
n=100	β_0	-0.40	0.08	-0.17	1.63	2.89	-0.09	-0.25	1.26	1.67
	β_1	0.90	1.01	1.02	0.65	0.44	1.03	1.03	0.61	0.38
	β_2	-1.20	-1.43	-1.17	1.64	2.74	-1.27	-1.11	1.15	1.33
	θ_0	-0.50	-0.27	-0.54	0.96	0.98	-0.35	-0.59	0.84	0.73
	θ_1	1.00	1.02	1.02	0.34	0.12	1.00	1.02	0.32	0.10
	θ_2	-1.00	-1.22	-1.04	0.68	0.51	-1.11	-0.99	0.50	0.26
	γ	1.00	1.19	1.10	0.68	0.49	1.13	1.11	0.43	0.20
n=200	β_0	-0.40	-0.38	-0.43	0.59	0.35	-0.42	-0.45	0.57	0.32
	β_1	0.90	0.89	0.90	0.32	0.10	0.89	0.90	0.30	0.09
	β_2	-1.20	-1.33	-1.28	0.56	0.32	-1.32	-1.26	0.56	0.33
	θ_0	-0.50	-0.45	-0.53	0.50	0.25	-0.49	-0.55	0.40	0.16
	θ_1	1.00	1.02	1.02	0.21	0.04	1.02	1.02	0.19	0.04
	θ_2	-1.00	-1.07	-1.01	0.33	0.12	-1.04	-1.00	0.25	0.06
	γ	1.00	1.04	1.05	0.18	0.03	1.05	1.05	0.18	0.03

n=300	β_0	-0.40	-0.26	-0.37	0.57	0.34	-0.27	-0.38	0.57	0.34
	β_1	0.90	0.88	0.93	0.30	0.09	0.89	0.93	0.29	0.08
	β_2	-1.20	-1.37	-1.34	0.41	0.20	-1.36	-1.34	0.41	0.19
	θ_0	-0.50	-0.53	-0.55	0.27	0.07	-0.53	-0.55	0.27	0.07
	θ_1	1.00	1.01	1.01	0.16	0.02	1.01	1.00	0.15	0.02
	θ_2	-1.00	-1.00	-1.01	0.17	0.03	-0.99	-1.01	0.17	0.03
	γ	1.00	1.04	1.04	0.12	0.02	1.04	1.04	0.13	0.02
n=400	β_0	-0.40	-0.21	-0.37	0.65	0.46	-0.22	-0.37	0.64	0.44
	β_1	0.90	0.88	0.89	0.25	0.06	0.89	0.89	0.25	0.06
	β_2	-1.20	-1.29	-1.26	0.43	0.20	-1.29	-1.25	0.44	0.20
	θ_0	-0.50	-0.51	-0.50	0.19	0.04	-0.51	-0.51	0.19	0.04
	θ_1	1.00	1.01	1.01	0.13	0.02	1.00	1.00	0.13	0.02
	θ_2	-1.00	-1.02	-1.01	0.15	0.02	-1.02	-1.01	0.15	0.02
	γ	1.00	1.02	1.02	0.13	0.02	1.02	1.02	0.13	0.02
n=500	β_0	-0.40	-0.41	-0.43	0.30	0.09	-0.42	-0.44	0.30	0.09
	β_1	0.90	0.90	0.91	0.18	0.03	0.91	0.91	0.18	0.03
	β_2	-1.20	-1.20	-1.20	0.30	0.09	-1.19	-1.19	0.30	0.09
	θ_0	-0.50	-0.48	-0.51	0.18	0.03	-0.49	-0.51	0.18	0.03
	θ_1	1.00	1.00	1.01	0.11	0.01	1.00	1.01	0.11	0.01
	θ_2	-1.00	-1.02	-1.01	0.13	0.02	-1.01	-1.01	0.13	0.02
	γ	1.00	1.00	1.00	0.09	0.01	1.00	1.00	0.09	0.01

Appendix D

Multiple Species simulation results

Summary of simulation results for the multi species model case 1: $\{0.40, 0.40, 0.10\}$ **with 2 species and assuming that the probability of detecting one species is independent of the presence/absence of the other.**

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=150	β_{11}	0.50	0.97	13.85	36.50	1496.82	0.04	-0.06	1.24	1.83
	β_{12}	-1.00	-1.64	-11.44	28.23	897.81	-0.91	-0.92	0.51	0.26
	β_{21}	0.50	0.70	10.57	31.89	1108.31	-0.22	-0.04	1.85	3.69
	β_{22}	-1.00	-1.49	-7.54	19.32	412.36	-0.97	-1.36	1.49	2.34
	γ	-2.50	-3.40	-17.15	44.72	2194.37	-1.91	-2.00	4.58	21.03
	θ_{11}	-1.50	-1.66	-1.13	5.51	30.20	-1.51	-1.55	0.69	0.47
	θ_{12}	-1.60	-1.70	-3.26	13.04	171.15	-1.68	-1.77	0.66	0.46
	θ_{21}	-1.50	-1.36	-0.62	13.59	183.66	-1.33	-1.33	1.25	1.57
	θ_{22}	-1.20	-1.41	-8.81	32.03	1073.57	-1.44	-2.08	2.60	7.49

n=300	β_{11}	0.50	0.71	3.00	12.31	156.27	0.47	0.68	1.20	1.46
	β_{12}	-1.00	-1.24	-2.29	5.55	32.15	-1.07	-1.16	0.51	0.29
	β_{21}	0.50	0.85	3.67	16.23	270.82	0.57	0.93	1.97	4.04
	β_{22}	-1.00	-1.34	-2.84	9.73	97.09	-1.15	-1.35	1.10	1.31
	γ	-2.50	-3.30	-5.71	11.62	143.93	-2.72	-2.94	2.44	6.08
	θ_{11}	-1.50	-1.47	-1.48	0.53	0.28	-1.44	-1.46	0.48	0.23
	θ_{12}	-1.60	-1.57	-1.71	0.78	0.62	-1.58	-1.68	0.64	0.41
	θ_{21}	-1.50	-1.67	-1.14	3.92	15.32	-1.58	-1.51	0.66	0.43
	θ_{22}	-1.20	-1.28	-2.53	11.00	121.62	-1.26	-1.43	0.72	0.57
n=600	β_{11}	0.50	0.69	0.76	0.83	0.74	0.60	0.71	0.79	0.66
	β_{12}	-1.00	-1.09	-1.14	0.37	0.16	-1.09	-1.11	0.36	0.14
	β_{21}	0.50	0.62	0.75	1.04	1.13	0.59	0.72	1.00	1.05
	β_{22}	-1.00	-1.10	-1.18	0.42	0.20	-1.09	-1.16	0.39	0.18
	γ	-2.50	-2.81	-2.96	1.30	1.89	-2.74	-2.89	1.25	1.71
	θ_{11}	-1.50	-1.60	-1.56	0.27	0.07	-1.59	-1.55	0.26	0.07
	θ_{12}	-1.60	-1.68	-1.72	0.37	0.15	-1.68	-1.72	0.37	0.15
	θ_{21}	-1.50	-1.51	-1.48	0.38	0.14	-1.52	-1.48	0.38	0.14
	θ_{22}	-1.20	-1.21	-1.28	0.36	0.13	-1.20	-1.27	0.36	0.13

Summary of simulation results for the multi species model case 2: $\{0.39, 0.40, 0.22\}$ **with 2 species and assuming that the probability of detecting one species is independent of the presence/absence of the other.**

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=150	β_{11}	-0.50	-0.36	2.48	21.00	445.52	-1.37	-1.46	1.03	1.96
	β_{12}	-1.50	-2.17	-8.78	22.21	541.60	-1.19	-1.30	0.68	0.50
	β_{21}	-0.50	-0.32	-0.62	68.15	4597.76	-1.37	-1.35	1.59	3.22
	β_{22}	-1.20	-1.31	-15.11	39.56	1743.07	-0.93	-1.13	1.12	1.25
	γ	-0.50	-1.17	-0.93	56.07	3112.25	1.57	1.63	2.44	10.43
	θ_{11}	-1.50	-1.61	-0.77	9.23	84.95	-1.56	-1.61	0.61	0.39
	θ_{12}	-1.60	-1.69	-2.99	9.58	92.70	-1.61	-1.88	0.82	0.74
	θ_{21}	-1.50	-1.70	-0.90	7.70	59.11	-1.59	-1.56	0.90	0.81
	θ_{22}	-1.20	-1.34	-3.39	16.91	287.78	-1.29	-1.74	1.47	2.43
n=300	β_{11}	-0.50	-0.40	-0.21	2.70	7.29	-0.67	-0.80	0.87	0.83
	β_{12}	-1.50	-1.72	-2.31	2.94	9.22	-1.50	-1.61	0.76	0.58
	β_{21}	-0.50	-0.43	-0.16	2.49	6.26	-0.68	-0.75	0.93	0.93
	β_{22}	-1.20	-1.32	-1.79	2.56	6.81	-1.18	-1.23	0.51	0.25
	γ	-0.50	-0.54	-0.84	4.49	20.08	-0.15	0.10	1.65	3.07
	θ_{11}	-1.50	-1.54	-1.59	0.44	0.20	-1.50	-1.53	0.40	0.16
	θ_{12}	-1.60	-1.52	-1.73	0.85	0.74	-1.54	-1.70	0.64	0.41
	θ_{21}	-1.50	-1.51	-1.50	0.55	0.30	-1.49	-1.47	0.56	0.31
	θ_{22}	-1.20	-1.28	-1.41	0.65	0.46	-1.26	-1.41	0.63	0.44
n=600	β_{11}	-0.50	-0.47	-0.45	0.63	0.39	-0.48	-0.49	0.62	0.38
	β_{12}	-1.50	-1.60	-1.71	0.51	0.30	-1.58	-1.68	0.48	0.26
	β_{21}	-0.50	-0.44	-0.40	0.69	0.48	-0.46	-0.43	0.67	0.44
	β_{22}	-1.20	-1.32	-1.37	0.50	0.28	-1.30	-1.35	0.48	0.25
	γ	-0.50	-0.57	-0.63	1.03	1.07	-0.53	-0.56	1.02	1.03
	θ_{11}	-1.50	-1.53	-1.54	0.28	0.08	-1.52	-1.54	0.29	0.08
	θ_{12}	-1.60	-1.62	-1.68	0.33	0.11	-1.63	-1.68	0.32	0.11
	θ_{21}	-1.50	-1.48	-1.49	0.34	0.12	-1.48	-1.48	0.35	0.12
	θ_{22}	-1.20	-1.23	-1.27	0.24	0.06	-1.24	-1.27	0.24	0.06

Summary of simulation results for the multi species model case 3: $\{0.63, 0.66, 0.43\}$ **with 2 species and assuming that the probability of detecting one species is independent of the presence/absence of the other.**

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=150	β_{11}	1.80	2.14	12.22	37.14	1473.93	0.52	0.35	1.40	4.03
	β_{12}	1.00	1.23	5.86	16.84	304.40	0.64	0.59	0.45	0.37
	β_{21}	1.90	2.83	14.66	36.84	1506.36	0.86	0.87	2.76	8.63
	β_{22}	1.50	1.91	8.09	20.38	454.69	1.37	1.75	1.66	2.78
	γ	-1.50	-1.94	-6.48	27.71	784.98	-0.19	0.75	3.01	14.03
	θ_{11}	-1.50	-1.60	-1.59	0.50	0.26	-1.52	-1.49	0.49	0.24
	θ_{12}	-1.60	-1.68	-1.87	0.78	0.68	-1.73	-1.89	0.71	0.58
	θ_{21}	-1.50	-1.43	-1.45	0.49	0.24	-1.46	-1.48	0.48	0.23
	θ_{22}	-1.20	-1.28	-1.33	0.44	0.21	-1.25	-1.30	0.45	0.21
n=300	β_{11}	1.80	1.81	5.79	23.35	555.71	1.51	1.53	1.11	1.30
	β_{12}	1.00	1.05	2.40	8.24	69.21	0.95	0.94	0.40	0.16
	β_{21}	1.90	2.15	6.07	23.33	556.01	1.86	2.11	3.28	10.68
	β_{22}	1.50	1.86	3.20	8.20	69.44	1.67	1.93	2.09	4.49
	γ	-1.50	-1.71	-4.80	18.62	353.95	-1.28	-1.26	1.56	2.48
	θ_{11}	-1.50	-1.41	-1.49	0.28	0.08	-1.40	-1.46	0.28	0.08
	θ_{12}	-1.60	-1.65	-1.72	0.39	0.16	-1.67	-1.74	0.38	0.16
	θ_{21}	-1.50	-1.50	-1.43	0.49	0.24	-1.49	-1.44	0.45	0.20
	θ_{22}	-1.20	-1.22	-1.37	0.93	0.88	-1.23	-1.36	0.84	0.72
n=600	β_{11}	1.80	1.98	2.39	1.35	2.14	1.90	2.19	1.13	1.42
	β_{12}	1.00	1.13	1.25	0.47	0.28	1.10	1.18	0.41	0.20
	β_{21}	1.90	2.19	2.53	1.50	2.61	2.14	2.34	1.30	1.87
	β_{22}	1.50	1.72	1.80	0.60	0.45	1.67	1.73	0.56	0.36
	γ	-1.50	-2.02	-2.10	1.30	2.03	-1.89	-1.93	1.13	1.45
	θ_{11}	-1.50	-1.50	-1.51	0.19	0.04	-1.49	-1.51	0.19	0.03
	θ_{12}	-1.60	-1.58	-1.62	0.22	0.05	-1.58	-1.63	0.22	0.05
	θ_{21}	-1.50	-1.55	-1.52	0.23	0.05	-1.55	-1.52	0.23	0.05
	θ_{22}	-1.20	-1.26	-1.28	0.23	0.06	-1.26	-1.28	0.22	0.06

Summary of simulation results for the multi species model case 4: $\{0.78, 0.78, 0.58\}$ **with 2 species and assuming that the probability of detecting one species is independent of the presence/absence of the other.**

			MLE				MPLE			
		true	mean	median	se	mse	mean	median	se	mse
n=150	β_{11}	1.00	2.55	7.26	34.94	1248.13	-1.34	-0.84	1.16	4.74
	β_{12}	1.20	1.71	8.16	20.06	446.90	1.11	1.12	0.68	0.46
	β_{21}	1.00	1.42	7.93	50.43	2566.27	-0.48	0.09	2.11	5.23
	β_{22}	-1.20	-1.50	-8.50	23.47	598.69	-1.42	-1.54	1.23	1.62
	γ	1.00	1.77	11.28	52.53	2836.99	3.43	3.61	2.70	14.04
	θ_{11}	-1.50	-1.52	-1.56	0.32	0.11	-1.48	-1.47	0.34	0.12
	θ_{12}	-1.60	-1.67	-1.72	0.48	0.24	-1.70	-1.73	0.44	0.21
	θ_{21}	-1.50	-1.44	-1.42	0.49	0.25	-1.50	-1.47	0.46	0.21
	θ_{22}	-1.20	-1.33	-1.49	0.91	0.91	-1.28	-1.41	0.83	0.72
n=300	β_{11}	1.00	2.59	6.05	20.41	438.02	-0.25	-0.09	1.57	3.62
	β_{12}	1.20	1.32	3.36	10.63	116.50	1.24	1.34	0.63	0.42
	β_{21}	1.00	1.66	7.35	26.27	723.73	0.77	1.66	4.52	20.65
	β_{22}	-1.20	-1.43	-4.23	13.53	190.32	-1.70	-2.20	2.45	6.97
	γ	1.00	0.78	0.38	6.43	41.29	2.68	2.44	2.34	7.50
	θ_{11}	-1.50	-1.52	-1.54	0.23	0.06	-1.48	-1.50	0.23	0.05
	θ_{12}	-1.60	-1.66	-1.67	0.28	0.09	-1.69	-1.71	0.28	0.09
	θ_{21}	-1.50	-1.52	-1.52	0.23	0.05	-1.53	-1.53	0.23	0.05
	θ_{22}	-1.20	-1.28	-1.30	0.30	0.10	-1.26	-1.29	0.30	0.09
n=600	β_{11}	1.00	1.44	2.72	3.54	15.36	0.78	0.79	1.44	2.10
	β_{12}	1.20	1.23	1.33	0.49	0.25	1.22	1.29	0.40	0.16
	β_{21}	1.00	1.37	4.50	16.18	271.53	1.23	1.68	3.42	12.05
	β_{22}	-1.20	-1.16	-2.25	8.48	72.21	-1.24	-1.67	1.85	3.60
	γ	1.00	0.55	-0.50	3.77	16.30	1.12	1.32	1.79	3.28
	θ_{11}	-1.50	-1.50	-1.50	0.15	0.02	-1.50	-1.49	0.15	0.02
	θ_{12}	-1.60	-1.65	-1.66	0.19	0.04	-1.66	-1.67	0.19	0.04
	θ_{21}	-1.50	-1.53	-1.49	0.18	0.03	-1.53	-1.50	0.18	0.03
	θ_{22}	-1.20	-1.23	-1.25	0.18	0.03	-1.22	-1.24	0.18	0.04

Appendix E

Derivation of the partial likelihood

Let $\underline{X}_i^U$ and $\underline{W}_i^U$ for $i = 1, ..., N$ denote the data sets corresponding to the used points and $\underline{X}_j^A$ and $\underline{W}_j^A$ for $j = 1, ..., m$ denote a random sample from the available distribution. Let $r = \frac{N}{N+M}$ the likelihood function can be written as:

$$L\left(\underline{\beta},\underline{\theta}/\underline{X}^U,\underline{W}^U,\underline{X}^A,\underline{W}^A\right) = \prod_{i=1}^{N} f^U\left(\underline{X}_i^U, \underline{W}_i^U\right) \prod_{J=1}^{M} f^A\left(\underline{X}_i^A, \underline{W}_i^A\right)$$

where

$$f^U\left(\underline{X}_i^U, \underline{W}_i^U\right) = \frac{\pi\left(\underline{X}_i^U, \underline{\beta}\right) \delta\left(\underline{W}_i^U, \underline{\theta}\right) f^A\left(\underline{X}_i^U, \underline{W}_i^U\right)}{\int\int \pi\left(\underline{X}_i^U, \underline{\beta}\right) \delta\left(\underline{W}_i^U, \underline{\theta}\right) f^A\left(\underline{X}_i^U, \underline{W}_i^U\right) dxdz} = \frac{\pi\left(\underline{X}_i^U, \underline{\beta}\right) \delta\left(\underline{W}_i^U, \underline{\theta}\right) f^A\left(\underline{X}_i^U, \underline{W}_i^U\right)}{P\left(\underline{\beta}, \underline{\theta}\right)}$$

hence,

$$L\left(\underline{\beta},\underline{\theta}/\underline{X}^U,\underline{W}^U,\underline{X}^A,\underline{W}^A\right) = \prod_{i=1}^{N} \frac{\pi\left(\underline{X}_i^U, \underline{\beta}\right) \delta\left(\underline{W}_i^U, \underline{\theta}\right) f^A\left(\underline{X}_i^U, \underline{W}_i^U\right)}{P\left(\underline{\beta}, \underline{\theta}\right)} \prod_{J=1}^{M} f^A\left(\underline{X}_i^A, \underline{W}_i^A\right)$$

$$
\begin{aligned}
L\left(\underline{\beta},\underline{\theta}/\underline{X}^U,\underline{W}^U,\underline{X}^A,\underline{W}^A\right) &= \prod_{i=1}^{N}\frac{\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)f^A\left(\underline{X}_i^U,\underline{W}_i^U\right)}{P\left(\underline{\beta},\underline{\theta}\right)}}{r\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)f^A\left(\underline{X}_i^U,\underline{W}_i^U\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)\,f^A\left(\underline{X}_i^A,\underline{W}_i^A\right)} \\
&\quad \prod_{j=1}^{M}\frac{f^A\left(\underline{X}_j^U,\underline{W}_j^U\right)}{r\frac{\pi\left(\underline{X}_j^A,\underline{\beta}\right)\delta\left(\underline{W}_j^A,\underline{\theta}\right)f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)\,f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)} \\
&\quad \prod_{i=1}^{N} r\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)f^A\left(\underline{X}_i^U,\underline{W}_i^U\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)\,f^A\left(\underline{X}_i^U,\underline{W}_i^U\right) \\
&\quad \prod_{j=1}^{M} r\frac{\pi\left(\underline{X}_j^A,\underline{\beta}\right)\delta\left(\underline{W}_j^A,\underline{\theta}\right)f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)\,f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)
\end{aligned}
$$

where

$$
\prod_{i=1}^{N}\frac{\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)f^A\left(\underline{X}_i^U,\underline{W}_i^U\right)}{P\left(\underline{\beta},\underline{\theta}\right)}}{r\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)f^A\left(\underline{X}_i^U,\underline{W}_i^U\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)\,f^A\left(\underline{X}_i^A,\underline{W}_i^A\right)}=\prod_{i=1}^{N}\frac{\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)}{P\left(\underline{\beta},\underline{\theta}\right)}}{r\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)}
$$

and

$$
\begin{aligned}
\prod_{j=1}^{M}\frac{f^A\left(\underline{X}_j^U,\underline{W}_j^U\right)}{r\frac{\pi\left(\underline{X}_j^A,\underline{\beta}\right)\delta\left(\underline{W}_j^A,\underline{\theta}\right)f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)\,f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)} &= \prod_{j=1}^{M}\frac{1}{r\frac{\pi\left(\underline{X}_j^A,\underline{\beta}\right)\delta\left(\underline{W}_j^A,\underline{\theta}\right)}{P\left(\underline{\beta},\underline{\theta}\right)}+(1-r)} \\
&= \frac{1}{(1-r)^M}\prod_{j=1}^{M}\frac{1}{r\pi\left(\underline{X}_j^A,\underline{\beta}\right)\delta\left(\underline{W}_j^A,\underline{\theta}\right)+(1-r)\,P\left(\underline{\beta},\underline{\theta}\right)}
\end{aligned}
$$

So the full likelihood can be expressed as a product of the following two terms:

$$
PL\left(\underline{\beta},\underline{\theta},\alpha\right)=\prod_{i=1}^{N}\frac{r\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)}{r\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)+(1-r)\,\alpha}\prod_{j=1}^{M}\frac{(1-r)\,\alpha}{r\pi\left(\underline{X}_j^A,\underline{\beta}\right)\delta\left(\underline{W}_j^A,\underline{\theta}\right)+(1-r)\,\alpha}
$$

where $\alpha=P\left(\underline{\beta},\underline{\theta}\right)$, and the remainder term

$$RT = \frac{1}{r^N(1-r)^M} \prod_{i=1}^{N} r\frac{\pi\left(\underline{X}_i^U,\underline{\beta}\right)\delta\left(\underline{W}_i^U,\underline{\theta}\right)f^A\left(\underline{X}_i^U,\underline{W}_i^U\right)}{P\left(\underline{\beta},\underline{\theta}\right)} + (1-r)\,f^A\left(\underline{X}_i^U,\underline{W}_i^U\right)$$
$$\prod_{j=1}^{M} r\frac{\pi\left(\underline{X}_j^A,\underline{\beta}\right)\delta\left(\underline{W}_j^A,\underline{\theta}\right)f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)}{P\left(\underline{\beta},\underline{\theta}\right)} + (1-r)\,f^A\left(\underline{X}_j^A,\underline{W}_j^A\right)$$

The results in Gilbert et al (1999) show that maximizing $PL\left(\underline{\beta},\underline{\theta},\alpha\right)$ with respect to $\left(\underline{\beta},\underline{\theta},\alpha\right)$ under the restriction $0 < \alpha \leq 1$, leads to the same estimators asymptotically as those obtained by maximizing the full likelihood.

Appendix F

RSPF simulation results

Summary of simulation results for the RSPF using 500 data sets using the logit link for the detection probability and separate covariates.

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.16	0.14	0.14	0.08	0.01	—	—	—	—
		β_0	-0.30	-0.55	-0.35	1.04	1.13	-2.35	-0.33	7.16	55.32
		β_1	0.70	0.71	0.69	0.18	0.03	0.72	0.71	0.20	0.04
		β_2	-1.20	-1.23	-1.23	0.31	0.10	-1.23	-1.20	0.35	0.13
		θ_0	-0.60	-0.61	-0.58	0.65	0.42	-1.68	-0.66	7.71	60.48
		θ_1	-1.00	-1.11	-1.04	0.42	0.19	-1.04	-1.02	0.34	0.12
		θ_2	1.00	1.03	0.98	0.37	0.13	1.04	0.99	0.41	0.17
	naive	α	0.16	0.29	0.31	0.14	0.04	—	—	—	—
		β_0	-0.30	-0.88	-0.31	2.52	6.69	-2.01	-0.28	6.53	45.41
		β_1	0.70	0.73	0.72	0.19	0.04	0.72	0.72	0.19	0.04
		β_2	-1.20	-1.24	-1.20	0.34	0.12	-1.23	-1.20	0.33	0.11

1000	full	α	0.16	0.15	0.15	0.07	0.00	—	—	—	—
		β_0	-0.30	-0.38	-0.28	0.89	0.78	-3.40	-0.33	32.93	1091.69
		β_1	0.70	0.75	0.74	0.18	0.03	0.71	0.71	0.17	0.03
		β_2	-1.20	-1.26	-1.23	0.30	0.09	-1.22	-1.17	0.31	0.09
		θ_0	-0.60	-0.67	-0.68	0.53	0.28	-1.48	-0.67	7.00	49.61
		θ_1	-1.00	-1.05	-1.02	0.26	0.07	-1.02	-1.01	0.28	0.08
		θ_2	1.00	1.04	1.05	0.25	0.07	1.02	1.00	0.34	0.11
	naive	α	0.16	0.29	0.30	0.12	0.03	—	—	—	—
		β_0	-0.30	-0.64	-0.32	1.75	3.17	-1.79	-0.35	6.03	38.50
		β_1	0.70	0.72	0.70	0.16	0.03	0.71	0.70	0.17	0.03
		β_2	-1.20	-1.22	-1.18	0.29	0.09	-1.21	-1.18	0.29	0.09
2000	full	α	0.16	0.19	0.20	0.09	0.01	—	—	—	—
		β_0	-0.30	-0.34	0.10	3.03	9.18	-1.85	-0.37	6.14	40.03
		β_1	0.70	0.78	0.78	0.18	0.04	0.71	0.71	0.15	0.02
		β_2	-1.20	-1.35	-1.32	0.41	0.19	-1.20	-1.19	0.27	0.07
		θ_0	-0.60	-0.52	-0.46	0.66	0.44	-0.98	-0.63	2.64	7.10
		θ_1	-1.00	-1.09	-1.07	0.27	0.08	-0.99	-0.98	0.25	0.06
		θ_2	1.00	1.08	1.07	0.35	0.13	0.99	0.98	0.31	0.09
	naive	α	0.16	0.29	0.30	0.11	0.03	—	—	—	—
		β_0	-0.30	-0.67	-0.29	2.02	4.21	-1.57	-0.35	5.49	31.73
		β_1	0.70	0.72	0.71	0.15	0.02	0.71	0.71	0.15	0.02
		β_2	-1.20	-1.22	-1.21	0.27	0.07	-1.20	-1.19	0.27	0.07

Summary of simulation results for the RSPF using 500 data sets using the logit link for the detection probability and a discrete common covariate associated with the parameters β_2 and θ_2 .

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.14	0.13	0.12	0.09	0.01	—	—	—	—
		β_0	-0.30	-0.50	-0.45	1.06	1.14	-2.89	-0.34	13.13	178.82
		β_1	0.70	0.72	0.69	0.23	0.05	0.74	0.70	0.23	0.05
		β_2	-1.20	-1.25	-1.15	0.62	0.39	-4.91	-1.26	34.17	1178.95
		θ_0	-0.60	-0.59	-0.58	1.20	1.43	-5.38	-0.67	35.09	1251.50
		θ_1	-1.00	-1.13	-1.09	0.55	0.32	-1.03	-1.02	0.34	0.12
		θ_2	1.00	1.07	1.08	0.70	0.50	4.70	1.14	34.26	1184.90
	naive	α	0.14	0.34	0.36	0.17	0.07	—	—	—	—
		β_0	-0.30	-0.86	-0.32	2.67	7.43	-2.65	-0.34	7.96	68.72
		β_1	0.70	0.81	0.79	0.25	0.08	0.80	0.78	0.25	0.07
		β_2	-1.20	-0.71	-0.62	0.68	0.70	-0.72	-0.59	1.61	2.81
1000	full	α	0.14	0.13	0.12	0.08	0.01	—	—	—	—
		β_0	-0.30	-0.39	-0.21	0.88	0.78	-1.79	-0.42	6.39	42.95
		β_1	0.70	0.75	0.73	0.16	0.03	0.73	0.70	0.19	0.04
		β_2	-1.20	-1.28	-1.28	0.47	0.23	-2.84	-1.23	6.28	42.09
		θ_0	-0.60	-0.71	-0.62	0.67	0.45	-2.85	-0.60	8.85	83.30
		θ_1	-1.00	-1.00	-0.98	0.21	0.04	-1.03	-1.03	0.29	0.09
		θ_2	1.00	1.04	1.02	0.56	0.31	2.64	1.11	6.26	41.81
	naive	α	0.14	0.35	0.37	0.15	0.07	—	—	—	—
		β_0	-0.30	-0.80	-0.27	2.87	8.45	-1.97	-0.31	6.39	43.48
		β_1	0.70	0.81	0.80	0.20	0.05	0.80	0.79	0.21	0.05
		β_2	-1.20	-0.67	-0.63	0.26	0.35	-0.65	-0.62	0.26	0.37

2000	full	α	0.14	0.19	0.19	0.10	0.01	—	—	—	—
		β_0	-0.30	-0.11	0.10	1.45	2.15	-3.11	-0.38	13.46	188.65
		β_1	0.70	0.82	0.80	0.22	0.06	0.70	0.69	0.17	0.03
		β_2	-1.20	-1.32	-1.21	1.00	1.00	-2.40	-1.28	7.83	62.60
		θ_0	-0.60	-0.47	-0.29	1.26	1.60	-2.49	-0.65	8.08	68.80
		θ_1	-1.00	-1.12	-1.10	0.26	0.08	-1.02	-1.01	0.25	0.06
		θ_2	1.00	1.05	0.96	1.08	1.16	2.23	1.09	7.78	61.93
	naive	α	0.14	0.34	0.36	0.14	0.06	—	—	—	—
		β_0	-0.30	-0.75	-0.31	2.15	4.82	-2.86	-0.31	7.87	68.41
		β_1	0.70	0.79	0.77	0.19	0.04	0.78	0.77	0.19	0.04
		β_2	-1.20	-0.66	-0.65	0.25	0.35	-0.62	-0.62	0.24	0.39

Summary of simulation results for the RSPF using 500 data sets using the logit link for the detection probability and a continuous common covariate associated with the parameters β_1 and θ_1 .

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.13	0.13	0.08	0.12	0.01	—	—	—	—
		β_0	-0.30	-0.67	-0.81	1.66	2.86	-9.75	-1.26	44.14	2033.76
		β_1	0.70	0.60	0.68	0.54	0.30	-1.38	0.54	13.54	187.34
		β_2	-1.20	-1.32	-1.14	0.65	0.43	-1.47	-1.10	2.60	6.84
		θ_0	-0.60	-0.73	-0.66	0.93	0.88	-6.67	-1.03	35.51	1295.65
		θ_1	-1.00	-0.94	-1.08	0.75	0.56	1.19	-0.86	13.59	189.20
		θ_2	1.00	1.84	0.97	4.64	21.99	1.11	0.81	2.40	5.78
	naive	α	0.13	0.49	0.61	0.26	0.20	—	—	—	—
		β_0	-0.30	3.28	2.17	8.49	84.73	9.73	3.45	20.94	538.32
		β_1	0.70	0.17	0.17	0.25	0.34	0.17	0.17	0.22	0.34
		β_2	-1.20	-5.58	-2.63	6.11	56.44	-14.72	-3.61	14.64	396.73

1000	full	α	0.13	0.13	0.10	0.11	0.01	—	—	—	—
		β_0	-0.30	-0.71	-0.56	1.30	1.84	-7.07	-1.38	24.89	664.11
		β_1	0.70	0.59	0.67	0.47	0.23	-0.44	0.55	6.97	49.82
		β_2	-1.20	-1.22	-1.11	0.39	0.15	-1.22	-1.06	0.54	0.29
		θ_0	-0.60	-0.67	-0.52	0.87	0.76	-4.98	-1.15	30.01	917.88
		θ_1	-1.00	-0.94	-1.09	0.66	0.43	0.24	-0.94	7.02	50.79
		θ_2	1.00	1.25	1.06	1.19	1.46	1.16	0.80	2.59	6.74
	naive	α	0.13	0.53	0.65	0.26	0.22	—	—	—	—
		β_0	-0.30	5.05	5.01	7.96	91.90	11.22	26.40	22.22	625.28
		β_1	0.70	0.16	0.16	0.20	0.33	0.14	0.15	0.18	0.34
		β_2	-1.20	-6.98	-5.19	5.47	63.23	-16.83	-26.60	14.78	462.33
2000	full	α	0.13	0.26	0.28	0.12	0.03	—	—	—	—
		β_0	-0.30	0.21	0.65	4.54	20.83	-4.69	-1.11	9.89	116.83
		β_1	0.70	0.49	0.61	0.49	0.28	-0.15	0.58	1.50	2.96
		β_2	-1.20	-1.64	-1.55	0.93	1.05	-1.28	-1.06	1.52	2.33
		θ_0	-0.60	-0.06	0.02	0.94	1.19	-2.86	-1.00	6.32	44.93
		θ_1	-1.00	-0.81	-1.03	0.80	0.67	-0.03	-0.93	1.66	3.70
		θ_2	1.00	1.59	1.37	1.38	2.26	0.96	0.82	0.46	0.21
	naive	α	0.13	0.55	0.66	0.23	0.23	—	—	—	—
		β_0	-0.30	5.51	6.42	7.25	86.12	12.15	26.30	21.23	604.61
		β_1	0.70	0.18	0.19	0.19	0.30	0.16	0.17	0.17	0.33
		β_2	-1.20	-7.00	-6.72	5.27	61.42	-16.89	-26.45	14.71	462.34

Summary of simulation results for the RSPF using 500 data sets using the complement log log link for the detection probability and separate covariates.

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.19	0.17	0.17	0.09	0.01	—	—	—	—
		β_0	-0.30	-0.53	-0.41	1.09	1.22	-2.85	-0.23	8.62	80.56
		β_1	0.70	0.71	0.70	0.19	0.04	0.73	0.73	0.19	0.04
		β_2	-1.20	-1.24	-1.15	0.36	0.13	-1.23	-1.21	0.35	0.12
		θ_0	-0.60	-0.63	-0.63	0.33	0.11	-0.70	-0.66	0.82	0.68
		θ_1	-1.00	-1.00	-1.01	0.25	0.06	-1.02	-0.99	0.27	0.07
		θ_2	1.00	0.99	1.00	0.28	0.08	1.02	0.99	0.30	0.09
	naive	α	0.19	0.30	0.32	0.14	0.03	—	—	—	—
		β_0	-0.30	-0.97	-0.25	2.78	8.16	-2.38	-0.22	7.25	56.85
		β_1	0.70	0.73	0.73	0.18	0.03	0.73	0.73	0.18	0.03
		β_2	-1.20	-1.24	-1.21	0.34	0.11	-1.23	-1.20	0.34	0.11
1000	full	α	0.19	0.18	0.19	0.08	0.01	—	—	—	—
		β_0	-0.30	-0.43	-0.24	0.98	0.97	-1.76	-0.26	5.87	36.53
		β_1	0.70	0.73	0.71	0.16	0.03	0.73	0.72	0.17	0.03
		β_2	-1.20	-1.25	-1.24	0.31	0.10	-1.25	-1.23	0.31	0.10
		θ_0	-0.60	-0.61	-0.64	0.27	0.07	-0.74	-0.61	1.48	2.21
		θ_1	-1.00	-1.05	-1.02	0.18	0.04	-1.02	-1.02	0.20	0.04
		θ_2	1.00	1.04	1.03	0.29	0.08	1.02	0.99	0.29	0.08
	naive	α	0.19	0.31	0.32	0.12	0.03	—	—	—	—
		β_0	-0.30	-0.50	-0.22	1.83	3.39	-1.52	-0.26	5.61	32.91
		β_1	0.70	0.74	0.73	0.17	0.03	0.73	0.72	0.17	0.03
		β_2	-1.20	-1.27	-1.24	0.30	0.09	-1.25	-1.24	0.30	0.09

2000	full	α	0.19	0.23	0.23	0.09	0.01	—	—	—	—
		β_0	-0.30	0.01	0.12	1.09	1.28	-1.84	-0.34	6.27	41.54
		β_1	0.70	0.80	0.79	0.19	0.05	0.71	0.71	0.16	0.02
		β_2	-1.20	-1.38	-1.34	0.41	0.20	-1.21	-1.20	0.28	0.08
		θ_0	-0.60	-0.58	-0.59	0.29	0.09	-0.67	-0.63	0.69	0.48
		θ_1	-1.00	-1.04	-1.01	0.20	0.04	-1.01	-1.00	0.18	0.03
		θ_2	1.00	1.04	1.02	0.27	0.08	1.01	0.99	0.22	0.05
	naive	α	0.19	0.30	0.31	0.11	0.02	—	—	—	—
		β_0	-0.30	-0.50	-0.31	1.34	1.82	-1.43	-0.31	5.09	27.14
		β_1	0.70	0.72	0.70	0.15	0.02	0.71	0.70	0.15	0.02
		β_2	-1.20	-1.22	-1.21	0.25	0.06	-1.21	-1.20	0.27	0.07

Summary of simulation results for the RSPF using 500 data sets using the complement log log link for the detection probability and a discrete common covariate associated with the parameters β_2 and θ_2 .

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.17	0.16	0.15	0.09	0.01	—	—	—	—
		β_0	-0.30	-0.51	-0.40	1.15	1.35	-3.36	-0.42	9.59	101.20
		β_1	0.70	0.74	0.70	0.22	0.05	0.72	0.69	0.21	0.04
		β_2	-1.20	-1.26	-1.17	0.52	0.27	-1.84	-1.24	3.49	12.54
		θ_0	-0.60	-0.57	-0.57	0.50	0.25	-1.48	-0.63	3.79	15.09
		θ_1	-1.00	-1.10	-1.06	0.26	0.08	-1.05	-1.05	0.24	0.06
		θ_2	1.00	1.07	1.05	0.48	0.24	1.66	1.10	3.54	12.93
	naive	α	0.17	0.33	0.35	0.18	0.06	—	—	—	—
		β_0	-0.30	-1.33	-0.39	3.28	11.81	-3.32	-0.41	8.50	81.25
		β_1	0.70	0.79	0.76	0.25	0.07	0.79	0.77	0.24	0.07
		β_2	-1.20	-0.66	-0.60	0.34	0.40	-0.64	-0.58	0.33	0.42

1000	full	α	0.17	0.16	0.17	0.08	0.01	0.00	0.00	0.00	0.00
		β_0	-0.30	-0.40	-0.20	0.91	0.83	-2.11	-0.26	8.51	75.53
		β_1	0.70	0.73	0.74	0.15	0.02	0.73	0.71	0.18	0.03
		β_2	-1.20	-1.30	-1.22	0.51	0.27	-1.67	-1.25	2.69	7.45
		θ_0	-0.60	-0.66	-0.58	0.52	0.27	-1.20	-0.59	3.19	10.51
		θ_1	-1.00	-1.03	-1.01	0.18	0.03	-1.03	-1.02	0.22	0.05
		θ_2	1.00	1.07	1.08	0.47	0.22	1.46	1.07	2.64	7.19
	naive	α	0.17	0.36	0.37	0.14	0.06	—	—	—	—
		β_0	-0.30	-0.72	-0.26	2.34	5.65	-1.94	-0.27	6.53	45.19
		β_1	0.70	0.81	0.79	0.20	0.05	0.80	0.79	0.21	0.05
		β_2	-1.20	-0.66	-0.63	0.27	0.37	-0.63	-0.60	0.28	0.40
2000	full	α	0.17	0.21	0.21	0.11	0.01	—	—	—	—
		β_0	-0.30	-0.19	0.06	2.00	4.01	-2.36	-0.35	13.97	198.88
		β_1	0.70	0.80	0.78	0.20	0.05	0.70	0.69	0.16	0.02
		β_2	-1.20	-1.42	-1.24	1.68	2.87	-1.83	-1.25	3.54	12.93
		θ_0	-0.60	-0.63	-0.47	1.69	2.84	-1.39	-0.66	3.78	14.91
		θ_1	-1.00	-1.04	-1.03	0.18	0.04	-1.01	-1.00	0.18	0.03
		θ_2	1.00	1.07	0.95	1.61	2.60	1.65	1.08	3.55	13.01
	naive	α	0.17	0.34	0.36	0.14	0.05	—	—	—	—
		β_0	-0.30	-0.77	-0.31	2.19	5.00	-1.70	-0.34	5.80	35.50
		β_1	0.70	0.78	0.77	0.18	0.04	0.77	0.77	0.18	0.04
		β_2	-1.20	-0.66	-0.62	0.24	0.35	-0.62	-0.60	0.23	0.39

Summary of simulation results for the RSPF using 500 data sets using the complement log log link for the detection probability and a continuous common covariate associated with the parameters β_1 and θ_1 .

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.16	0.16	0.12	0.14	0.02	—	—	—	—
		β_0	-0.30	-0.55	-0.58	1.50	2.30	-5.09	-0.71	20.70	450.50
		β_1	0.70	0.61	0.69	0.59	0.35	0.44	0.69	1.37	1.94
		β_2	-1.20	-1.29	-1.18	0.44	0.20	-1.39	-1.11	1.97	3.90
		θ_0	-0.60	-0.80	-0.59	0.81	0.68	-1.83	-0.76	3.44	13.29
		θ_1	-1.00	-0.89	-1.04	0.68	0.48	-0.70	-1.05	1.46	2.21
		θ_2	1.00	1.08	0.98	0.40	0.16	1.05	0.96	1.05	1.11
	naive	α	0.16	0.43	0.50	0.28	0.15	—	—	—	—
		β_0	-0.30	1.54	0.88	8.06	68.17	7.74	2.60	21.87	541.99
		β_1	0.70	0.09	0.03	0.25	0.43	0.10	0.10	0.23	0.41
		β_2	-1.20	-4.57	-1.78	4.99	36.24	-13.88	-3.00	14.68	375.81
1000	full	α	0.16	0.13	0.11	0.10	0.01	—	—	—	—
		β_0	-0.30	-0.75	-0.61	1.24	1.72	-3.64	-0.79	9.37	98.81
		β_1	0.70	0.72	0.71	0.41	0.17	0.38	0.68	1.17	1.48
		β_2	-1.20	-1.21	-1.13	0.32	0.10	-1.29	-1.12	1.77	3.13
		θ_0	-0.60	-0.81	-0.61	0.72	0.56	-1.60	-0.72	2.85	9.11
		θ_1	-1.00	-1.04	-1.10	0.47	0.22	-0.63	-1.04	1.28	1.77
		θ_2	1.00	1.02	0.99	0.31	0.09	1.06	0.99	1.05	1.10
	naive	α	0.16	0.50	0.64	0.27	0.18	—	—	—	—
		β_0	-0.30	4.13	2.99	8.19	86.48	8.77	3.15	22.50	587.59
		β_1	0.70	0.13	0.13	0.21	0.37	0.10	0.11	0.19	0.39
		β_2	-1.20	-6.39	-3.34	5.42	56.20	-15.12	-3.60	14.78	411.79

2000	full	α	0.16	0.24	0.27	0.13	0.02	—	—	—	—
		β_0	-0.30	-0.72	0.49	12.26	150.17	-3.09	-0.52	13.17	180.82
		β_1	0.70	0.56	0.66	0.56	0.33	0.47	0.71	0.98	1.02
		β_2	-1.20	-1.50	-1.43	0.76	0.67	-1.27	-1.17	0.92	0.86
		θ_0	-0.60	-0.58	-0.43	1.77	3.14	-1.47	-0.70	2.55	7.25
		θ_1	-1.00	-0.74	-0.96	0.79	0.69	-0.74	-1.04	1.06	1.20
		θ_2	1.00	1.17	1.09	1.62	2.65	1.02	0.97	0.73	0.54
	naive	α	0.16	0.51	0.66	0.27	0.19	—	—	—	—
		β_0	-0.30	4.83	7.56	8.25	94.33	9.22	4.42	23.34	634.55
		β_1	0.70	0.13	0.14	0.20	0.36	0.10	0.11	0.17	0.39
		β_2	-1.20	-6.99	-7.67	5.46	63.29	-16.08	-4.81	15.01	446.26

Summary of simulation results for the RSPF using 500 data sets using the probit link for the detection probability and separate covariates.

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.16	0.14	0.13	0.07	0.01	—	—	—	—
		β_0	-0.30	-0.61	-0.53	0.97	1.04	-2.78	-0.29	10.41	114.23
		β_1	0.70	0.71	0.68	0.17	0.03	0.72	0.71	0.19	0.04
		β_2	-1.20	-1.20	-1.17	0.31	0.09	-1.24	-1.23	0.36	0.13
		θ_0	-0.60	-0.59	-0.61	0.22	0.05	-0.62	-0.62	0.23	0.05
		θ_1	-1.00	-1.04	-1.00	0.21	0.05	-1.03	-1.01	0.23	0.05
		θ_2	1.00	1.04	1.00	0.27	0.07	1.02	1.02	0.26	0.07
	naive	α	0.16	0.30	0.31	0.14	0.04	—	—	—	—
		β_0	-0.30	-0.80	-0.24	2.51	6.56	-2.03	-0.26	6.64	46.96
		β_1	0.70	0.73	0.71	0.19	0.04	0.72	0.71	0.19	0.03
		β_2	-1.20	-1.26	-1.22	0.35	0.13	-1.24	-1.21	0.34	0.12

1000	full	α	0.16	0.16	0.17	0.07	0.01	—	—	—	—
		β_0	-0.30	-0.31	-0.07	1.02	1.02	-2.20	-0.30	7.14	54.57
		β_1	0.70	0.74	0.70	0.21	0.04	0.72	0.71	0.18	0.03
		β_2	-1.20	-1.30	-1.27	0.33	0.12	-1.24	-1.23	0.33	0.11
		θ_0	-0.60	-0.61	-0.58	0.17	0.03	-0.62	-0.62	0.20	0.04
		θ_1	-1.00	-1.03	-1.00	0.18	0.03	-1.01	-1.00	0.18	0.03
		θ_2	1.00	1.04	1.02	0.22	0.05	1.02	1.01	0.21	0.05
	naive	α	0.16	0.30	0.32	0.12	0.04	—	—	—	—
		β_0	-0.30	-0.58	-0.25	1.78	3.24	-1.87	-0.26	6.26	41.56
		β_1	0.70	0.73	0.72	0.17	0.03	0.72	0.71	0.17	0.03
		β_2	-1.20	-1.26	-1.21	0.30	0.09	-1.24	-1.21	0.30	0.09
2000	full	α	0.16	0.20	0.19	0.09	0.01	—	—	—	—
		β_0	-0.30	0.06	0.05	1.11	1.35	-2.19	-0.38	6.32	43.44
		β_1	0.70	0.80	0.78	0.19	0.04	0.70	0.69	0.16	0.02
		β_2	-1.20	-1.39	-1.31	0.41	0.21	-1.20	-1.18	0.29	0.08
		θ_0	-0.60	-0.57	-0.58	0.18	0.03	-0.62	-0.62	0.18	0.03
		θ_1	-1.00	-1.05	-1.04	0.18	0.04	-1.01	-1.00	0.17	0.03
		θ_2	1.00	1.06	1.04	0.21	0.05	1.01	1.00	0.19	0.04
	naive	α	0.16	0.29	0.30	0.12	0.03	—	—	—	—
		β_0	-0.30	-0.61	-0.35	1.63	2.74	-1.91	-0.37	6.20	40.90
		β_1	0.70	0.71	0.69	0.14	0.02	0.69	0.68	0.15	0.02
		β_2	-1.20	-1.22	-1.20	0.26	0.07	-1.20	-1.19	0.26	0.07

Summary of simulation results for the RSPF using 500 data sets using the probit link for the detection probability and a discrete common covariate associated with the parameters β_2 and θ_2 .

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.13	0.13	0.13	0.07	0.00	—	—	—	—
		β_0	-0.30	-0.29	-0.19	0.95	0.89	-4.00	-0.44	28.48	823.21
		β_1	0.70	0.75	0.73	0.20	0.04	0.71	0.69	0.21	0.04
		β_2	-1.20	-1.30	-1.24	0.52	0.28	-1.38	-1.18	2.48	6.17
		θ_0	-0.60	-0.62	-0.60	0.33	0.11	-0.68	-0.61	0.82	0.68
		θ_1	-1.00	-1.02	-1.03	0.20	0.04	-1.01	-1.01	0.24	0.06
		θ_2	1.00	1.01	0.98	0.32	0.10	1.02	1.02	0.37	0.13
	naive	α	0.13	0.31	0.33	0.19	0.07	—	—	—	—
		β_0	-0.30	-1.79	-0.62	3.66	15.55	-4.66	-0.64	9.63	111.59
		β_1	0.70	0.78	0.74	0.27	0.08	0.77	0.73	0.26	0.07
		β_2	-1.20	-0.38	-0.32	0.34	0.79	-0.34	-0.30	0.29	0.82
1000	full	α	0.13	0.13	0.14	0.07	0.01	—	—	—	—
		β_0	-0.30	-0.49	-0.38	0.98	0.98	-3.06	-0.32	8.41	78.11
		β_1	0.70	0.72	0.71	0.17	0.03	0.70	0.69	0.18	0.03
		β_2	-1.20	-1.22	-1.16	0.35	0.12	-1.30	-1.26	0.47	0.23
		θ_0	-0.60	-0.55	-0.54	0.25	0.06	-0.65	-0.60	0.32	0.10
		θ_1	-1.00	-1.05	-1.05	0.17	0.03	-1.01	-1.01	0.20	0.04
		θ_2	1.00	1.01	0.99	0.27	0.07	1.07	1.07	0.29	0.09
	naive	α	0.13	0.32	0.35	0.17	0.06	—	—	—	—
		β_0	-0.30	-1.40	-0.51	2.93	9.78	-4.22	-0.57	9.08	97.63
		β_1	0.70	0.76	0.75	0.22	0.05	0.75	0.73	0.22	0.05
		β_2	-1.20	-0.39	-0.35	0.24	0.71	-0.35	-0.31	0.24	0.78

2000	full	α	0.13	0.19	0.18	0.08	0.01	—	—	—	—
		β_0	-0.30	-0.28	0.25	9.35	87.22	-2.67	-0.39	9.64	98.46
		β_1	0.70	0.84	0.83	0.22	0.07	0.71	0.70	0.16	0.03
		β_2	-1.20	-1.33	-1.29	0.41	0.18	-1.26	-1.21	0.43	0.19
		θ_0	-0.60	-0.51	-0.50	0.26	0.08	-0.64	-0.63	0.29	0.09
		θ_1	-1.00	-1.07	-1.05	0.17	0.03	-1.01	-1.00	0.17	0.03
		θ_2	1.00	1.00	0.99	0.25	0.06	1.04	1.04	0.24	0.06
	naive	α	0.13	0.32	0.34	0.16	0.06	—	—	—	—
		β_0	-0.30	-1.16	-0.57	2.39	6.46	-3.45	-0.63	8.20	76.98
		β_1	0.70	0.76	0.74	0.20	0.05	0.75	0.74	0.20	0.04
		β_2	-1.20	-0.38	-0.34	0.21	0.73	-0.32	-0.30	0.21	0.81

Summary of simulation results for the RSPF using 500 data sets using the probit link for the detection probability and a continuous common covariate associated with the parameters β_1 and θ_1 .

				Partial Likelihood				Full Likelihood			
used	model	par	true	mean	median	se	mse	mean	median	se	mse
500	full	α	0.12	0.13	0.11	0.12	0.01	—	—	—	—
		β_0	-0.30	-0.45	-0.41	1.60	2.55	-3.71	-0.20	10.00	111.40
		β_1	0.70	0.75	0.77	0.26	0.07	0.95	0.78	1.63	2.71
		β_2	-1.20	-1.38	-1.21	0.63	0.42	-1.63	-1.23	2.51	6.47
		θ_0	-0.60	-0.61	-0.59	0.41	0.17	-0.94	-0.62	2.45	6.13
		θ_1	-1.00	-1.07	-1.02	0.27	0.08	-1.02	-1.01	0.24	0.06
		θ_2	1.00	1.09	1.04	0.40	0.17	1.02	1.00	0.40	0.16
	naive	α	0.12	0.27	0.29	0.23	0.08	—	—	—	—
		β_0	-0.30	-2.71	-0.29	4.89	29.70	-5.41	-0.15	11.52	158.48
		β_1	0.70	-0.47	-0.33	0.47	1.58	-0.41	-0.33	0.39	1.39
		β_2	-1.20	-1.24	-1.18	0.40	0.16	-1.64	-1.26	3.30	11.05

1000	full	α	0.12	0.14	0.10	0.11	0.01	—	—	—	—
		β_0	-0.30	-0.19	-0.22	1.32	1.74	-2.70	-0.30	8.28	74.14
		β_1	0.70	0.81	0.79	0.25	0.07	0.82	0.77	0.46	0.23
		β_2	-1.20	-1.39	-1.25	0.43	0.22	-1.38	-1.22	1.21	1.48
		θ_0	-0.60	-0.69	-0.63	0.41	0.17	-0.75	-0.61	0.98	0.97
		θ_1	-1.00	-1.01	-1.00	0.19	0.03	-1.03	-1.02	0.21	0.04
		θ_2	1.00	1.02	0.99	0.31	0.09	1.02	1.01	0.33	0.11
	naive	α	0.12	0.29	0.29	0.22	0.08	—	—	—	—
		β_0	-0.30	-2.16	-0.28	4.65	25.06	-4.18	-0.22	9.38	102.78
		β_1	0.70	-0.48	-0.37	0.44	1.58	-0.38	-0.33	0.26	1.23
		β_2	-1.20	-1.25	-1.22	0.35	0.12	-1.28	-1.26	0.33	0.12
2000	full	α	0.12	0.25	0.23	0.16	0.04	—	—	—	—
		β_0	-0.30	2.26	0.72	25.19	639.90	-1.99	-0.31	6.49	44.96
		β_1	0.70	0.61	0.68	0.56	0.33	0.77	0.75	0.27	0.08
		β_2	-1.20	-3.56	-1.53	24.90	624.40	-1.36	-1.21	1.34	1.83
		θ_0	-0.60	-0.43	-0.42	0.50	0.28	-0.70	-0.61	0.60	0.37
		θ_1	-1.00	-1.03	-1.05	0.36	0.13	-1.03	-1.02	0.18	0.03
		θ_2	1.00	2.69	1.21	12.41	156.54	1.03	1.03	0.31	0.09
	naive	α	0.12	0.31	0.32	0.23	0.09	—	—	—	—
		β_0	-0.30	-1.94	-0.14	4.50	22.94	-4.20	-0.21	9.41	103.57
		β_1	0.70	-0.54	-0.39	0.51	1.79	-0.36	-0.33	0.18	1.16
		β_2	-1.20	-1.27	-1.25	0.66	0.43	-1.27	-1.27	0.30	0.10

Bibliography

[1] Proteus wildlife research consultants http://www.proteus.co.nz/home.html.

[2] Alan Agresti. *Categorical data analysis*. Book News Inc, 2003.

[3] H Resit Akcakaya, Gus Mills, and C Patrick Doncaster. The role of metapopulations in conservation. In D W Macdonald and K Service, editors, *Key topics in conservation biology*, chapter 5, pages 64–84. Blackwell publishing, 2007.

[4] H Resit Akcakaya and P Sjogren-Gulve. Population viability analysis in conservation planning: an overview. *Ecological Bulletins*, 48:9–21, 2000.

[5] David L Azuma, James A Baldwin, and Barry R Noon. Estimating the occupancy of spotted owl habitat areas by sampling and adjusting for bias. *USDA Forest Service Gen. Tech. Rep*, 1990.

[6] E Bayne, S R Lele, and P Solymos. Bias in the estimation of bird density and relative abundance when the closure assumption of multiple survey approach is violated: a simulation study. *submitted manuscript*, 2010.

[7] J Bengtsson. Interspecific competition increases local extinction rate in a metapopulation system. *Nature*, 340:713–715, 1989.

[8] J Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistical Society*, 36:192–236, 1974.

[9] B Bolker. *Ecological models and data in R*. Princenton University Press, NJ, 2008.

[10] James H Brown and Mark V Lomolino. *Biogeography*. Sinauer Associates Inc, 1998.

[11] K P Burnham and D R Anderson. *Model selection and Multimodel Inference: A practical information - theoretic approach*. Springer-Verlag, 2002.

[12] Mar Cabeza, Anni Arponen, Laura Jaattela, Heini Kujala, Astrid van Teeffelen, and Ilkka Hanski. Conservation planning with insects at three different spatial scales. *Ecography*, 33:54–63, 2010.

[13] P Caragea and M S Kaiser. Autologistic model with interpretable parameters. *Journal of agricultural, biological and environmental statistics*, 14(3):281–300, 2009.

[14] G Casella and E I George. Explaining the gibbs sampler. *The American Statistician*, 46:167–174, 1992.

[15] Jonathan M Chase, Peter A Abrams, James P Grover, Sebastian Diehl, Peter Chesson, Robert D Holt, Shane A Richards, Roger M Nisbet, and Ted J Case. The interaction between predation and competition: a review and synthesis. *Ecology letters*, 5:302–315, 2002.

[16] Martin L Cody and Jared Mason Diamond. *Ecology and Evolution of Communities*. The Belknap Press of Harvard University Press, 1975.

[17] Rosamonde R Cook and Ilkka Hanski. On expected lifetimes of small-bodies and large-bodies species of birds on islands. *The American Naturalist*, 145(2):307–315, 1995.

[18] Jared Mason Diamond. Assembly of species communities. In M L Cody and Jared Mason Diamond, editors, *Ecology and Evolution of Communities*. Harvard University Press, 1975.

[19] Jared Mason Diamond. The island dilemma: lessons of modern biogeographic studies for the design of natural reserves. *Biological conservation*, 7:129–146, 1975.

[20] Robert M Dorazio. On the choice of statistical models for estimating occurrence and extinction from animal surveys. *Ecology*, 88(11):2773–2782, 2007.

[21] Robert M Dorazio and J Andrew Royle. Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, 100(470):389–398, 2005.

[22] Robert M Dorazio and J Andrew Royle. The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics*, 61:874–876, 2005.

[23] Robert M Dorazio, J Andrew Royle, Bo Soderstrom, and A Glimskar. Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, 87(4):842–854, 2006.

[24] D Van Dorp and P F M Opdam. Effects of patch size, isolation and regional on forest bird communities. *Landscape Ecology*, 1:59–73, 1987.

[25] C M Downes and B T Collins. The canadian breeding bird survey 1967-2000. *Canadian Wildlife Service, National Wildlife Research Centre, Ottawa, ON.*, Progress notes No 219, 2003.

[26] Jerome A Dupuis and Carl James Schwarz. A bayesian approach to multistate jolly-seber capture-recapture model. *Biometrics*, 63:1015–1022, 2007.

[27] B Efron and R J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, 1993.

[28] Lenore Fahrig. Relative effect of habitat loss and fragmentation on population extinction. *Journal of wildlife managment*, 61(3):603–610, 1997.

[29] Jacqueline L Frair, Scott E Nielsen, Evelyn H Merrill, Subhash R Lele, Mark S Boyce, Robin H M Munro, Gordon B Stenhouse, and Hawthorne L Beyer. Removing gps collar bias in habitat selection studies. *Journal of Applied Ecology*, 41:201–212, 2004.

[30] P H Geissler and M R Fuller. Estimation of the proportion of area occupied by an animal species. *Proceedings of the section on survey research methods of the american statistical association*, pages 533–538, 1986.

[31] R K GodFrey and J W Wooten. *Aquatic and Wetland Plants of Southeastern United States*. The University of Georgia Press, 1981.

[32] P J Green. *Penalized likelihood*, volume 2. Encyclopedia of statistical sciences, 1996.

[33] P J Green and B W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. London, Chapman and Hall, 1994.

[34] W Gu and R K Swihart. Absent or undetected? effects of non-detection of species occurrence on wildlife-habitat models. *Biological conservation*, 116:195–203, 2004.

[35] J M Hammersley and P Clifford. Markov fields on finite graphs and latttices. *University of California, Berkeley and Oxford University*, 1971.

[36] Ilkka Hanski. Population dynamics of shrews on small islands accord to the equilibrium model. *Biological Journal of the Linnean Society*, 28:23–36, 1986.

[37] Ilkka Hanski. Inferences from ecological incidence functions. *The American Naturalist*, 139(3):657–662, 1992.

[38] Ilkka Hanski, Mikko Kuussaari, and Marko Nieminen. Metapopulation structure and migration in butterfly melitaea cinxia. *Ecology*, 75(3):747–762, 1994.

[39] Ilkka Hanski, Atte Moilanen, and Mats Gyllenberg. Minimum viable metapopulation size. *The American Naturalist*, 147(4), 1996.

[40] Ilkka Hanski, Atte Moilanen, Timo Pakkala, and Mikko Kuussaari. The quanttitative incidence function model and pesistence of an endangered butterfly metapopulation. *Conservation Biology*, 10(2):578–590, 1996.

[41] Ilkka Hanski and Daniel Simberloff. The metapopulation approach, its history, conceptual domain, and applications to conservation. In Ilkka Hanski and Michael E Gilpin, editors, *Metapopulation biology: ecology, genetics and evolution*, chapter 1. Academic Press, 1997.

[42] Ilkka Hanski and Chris D Thomas. Metapopulation dynamics and conservation: a spatially explicit model applied to butterflies. *Biological conservation*, 68:167–180, 1994.

[43] Susan Harrison and Andrew D Taylor. Empirical evidence for metapopulation dynamics. In Ilkka Hanski and Michael E Gilpin, editors, *Metapopulation biology: ecology, genetics and evolution*, chapter 2. Academic Press, 1997.

[44] F L He, J Zhou, and H Zhu. Autologistic regression model for the distribution of vegetation. *Journal of agricultural, biological and environmental statistics*, 8:205–222, 2003.

[45] James E Hines, James D Nichols, J Andrew Royle, Darryl I MacKenzie, A M Gopalaswamy, N Samba Kumar, and K U Karanth. Tigers on trail: occupancy modeling for cluster sampling. *Ecological applications*, 20(5):1456–1466, 2010.

[46] K A Hobson and E Bayne. Breeding bird communities in boreal forest of western canada: consequences of "unmixing" the mixed woods. *Condor*, 102:759–769, 2002.

[47] D G Horvitz and D J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47(260):663–685, 1952.

[48] F W Huffer and H Wu. Markov chain monte carlo for autologistic regression model with application to the distribution of plant species. *Biometrics*, 54:509–524, 1998.

[49] Mats E Johansson and Paul A Keddy. Intensity and asymmetry of competition between two plan pairs of different degrees of similarity: an experimental study on two guilds of wetland plants. *Oikos*, 60(1):27–34, 1991.

[50] Barry L Johnson and Karen H Hagerty. *Status and trends of selected resources of the Upper Mississippi River System. Technical report LTRMP 2008-T002.* U.S. Geological Survey, Upper Midwest Environmental Sciences Center, La Crosse, Wisconsin., http://pubs.usgs.gov/mis/LTRMP2008-T002/, 2008.

[51] C J Johnson, S E Nielsen, E H Merrill, T L McDonald, and M S Boyce. Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. *Journal of wildlife managment*, 70(2):347–357, 2006.

[52] G M Jolly. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52(1):225–247, 1965.

[53] William L Kendall. Robustness of closed capture-recapture methods to violations of the closure assumption. *Ecology*, 80(8):2517–2525, 1999.

[54] D S Klute, M K Lovallo, and W M Tzilkowski. Autologistic regression modeling of american woodcock habitat use with spatially dependent data. In J M Scott, P J Heglund, M L Morrison, J B Haufler, M G Raphael, W A Wall, and F B Samson, editors, *Predicting species occurrences: issues of accurancy and scale*, chapter Prediction species occurrences: issues of accuracy and scale, pages 335–343. Island Press, Washington, D.C., USA., 2002.

[55] C J Krebs. *Ecology: The experimental analysis of distribution and abundance (3rd Edition).* Harper and Row, New York, Usa, 1985.

[56] Roland H Lamberson, Robert McKelvey, Barry R Noon, and Curtis Voss. A dynamic analysis of nothern spotted owl viability in fragmented forest landscape. *Conservation Biology*, 6(4):505–512, 1992.

[57] P S Laplace. Sur les naissances, les mariages et les morts. *Historie de L'Academie Royale des Sciences*, page 693, 1783.

[58] A M Latimer, S Wu, A E Gelfand, and J A Silander. Building statistical models to analyze species distribution. *Ecological applications*, 16:33–50, 2006.

[59] S R Lele. A new method for estimation of resource selection probability function. *Journal of wildlife managment*, 73(1):122–127, 2009.

[60] S R Lele, B Dennis, and F Lutscher. Data cloning: easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. *Ecology letters*, 10:551–563, 2007.

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

[62] S R Lele, K Nadeem, and B Schmuland. Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of American Statistical Association*, 105(492):1617–1625, 2010.

[63] Richard Levins. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Journal of Entomological Society of America*, 15(3):237–240, 1969.

[64] Frederick C Lincoln. Calculating waterfowl abundance on basis of banding returns. *U.S. G.P.O*, 1930.

[65] William A Link. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130, 2003.

[66] D J Lunn, A Thomas, N Best, and D Spiegelhalter. Winbugs: a bayesian modeling framework: concepts structure and extensibility. *Statistics and computing*, 10:325–337, 2000.

[67] R H MacArthur and E O Wilson. *The Theory of Island Biogeography*. Princenton University Press, NJ, 1967.

[68] Darryl I MacKenzie. Was it there? dealing with imperfect detection for species presence/absence data. *Australian Statistical Publishing Association*, 47(1):65–74, 2005.

[69] Darryl I MacKenzie. What are the issues with presence-absence data for wildlife managers? *Journal of wildlife managment*, 69(3):849–860, 2005.

[70] Darryl I MacKenzie and Larissa L Bailey. Assessing the fit of site-occupancy models. *Journal of agricultural, biological and environmental statistics*, 9(3):300–318, 2004.

[71] Darryl I MacKenzie, Larissa L Bailey, and James D Nichols. Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology*, 73(546-555), 2004.

[72] Darryl I MacKenzie, James D Nichols, James E Hines, M G Knuston, and A D Franklin. Estimating site occupancy, colonization and local extinction when a species is detected imperfectly. *Ecology*, 84:2200–2207, 2003.

[73] Darryl I MacKenzie, James D Nichols, G B Lachman, S Droege, J A Royle, and C A Langtimm. Estimating site occupancy rates when detection probabilites are less than one. *Ecology*, 83:2248–2255, 2002.

[74] Darryl I MacKenzie and J Andrew Royle. Designing occupancy studies: general advice and allocating survey effort. *Journal of applied ecology*, 42:1105–1114, 2005.

[75] P MacTavish. Saskatchewan digital landcover mapping project. *Saskatchewan Research Council, Saskatoon, SK.*, 1995.

[76] A E Magurran. *Measuring biological diversity.* Oxford:Blackwell Science, 2004.

[77] B FJ Manly, L L McDonald, D L Thomas, T L McDonald, and W P Erickson. *Resource selection by animals: statistical design and analysis for field studies. Second Edition.* Kluwer Academic Publishers, 2002.

[78] A Moilanen. Implications of empirical data quality for metapopulation model parameter estimation and application. *Oikos*, 96:516–530, 2002.

[79] Atte Moilanen and Ilkka Hanski. Habitat destruction and competitive coexistence in a spatially realistic metapopulation model. *Journal of Animal Ecology*, 64(1):141–144, 1995.

[80] Atte Moilanen, Andrew T Smith, and Ilkka Hanski. Long-term dynamics in the metapopulation of the american pika. *The American Naturalist*, 152(4), 1998.

[81] Daniel Montoya. Habitat loss, dispersal, and the probability of extinction of tree species. *Communicative and Integrative Biology*, pages 146–147, 2008.

[82] M Moreno and S R Lele. Improved estimation of site occupancy using penalized likelihood. *Ecology*, 91:341–346, 2010.

[83] James D Nichols, L L Bailey, Allan F O'Connell, Neil W Talancy, Evan H Campbell Grant, Andrew T Gilbert, Elizabeth M Annand, Thomas P Husband, and James E Hines. Multi-scale occupancy estimation and modeling using multiple detection methods. *Journal of applied ecology*, 45:1321–1329, 2008.

[84] James D Nichols, James E Hines, Darryl I MacKenzie, Mark E Seamans, and R J Gutierrez. Occupancy estimation and modeling with multiple states and state uncertainty. *Ecology*, 88(6):1395–1400, 2007.

[85] Samuel C Nicol, Iadine Chades, Simon Linke, and Hugh P Possingham. Conservation decision-making in large state spaces. *Ecological modelling*, 221:2531–2536, 2010.

[86] Samuel C Nicol and Hugh P Possingham. Should metapopulation restoration strategies increase patch area or number of patches? *Ecological applications*, 20(2):566–581, 2010.

[87] Eugene P Odum and A J Potin. Population density of the underground and lasius flavus, as determined by tagging with p32. *Ecology*, 42:182–188, 1961.

[88] V I P Pajunen. Distributional dynamics of daphnia species in a rock-pool environment. *Annales Zoologici Fennici*, 23:131–140, 1986.

[89] G P Patil and C R Rao. Weighted distributions ans size biased sampling with applications to wildlife populations and human families. *Biometrica*, 34:179–189, 1978.

[90] Shirley Pledger and Murray Efford. Correction of bias due to heterogeneous capture probability in capture-recapture studies of open populations. *Biometrics*, 54:888–898, 1998.

[91] I Pokki. Distribution, demography and dispersal of the field vole microtus agrestis in the tvarmine arhipelago, finland. *Acta Zoologica Fennica*, 164:1–48, 1981.

[92] H Ronald Pulliam, John B Dunning, and Jianguo Liu. Population dynamics in complex landscapes: a case study. *Ecological applications*, 2:165–177, 1992.

[93] Jason D Riddle, Rua S Mordecai, Kenneth H Pollock, and Theodore R Simons. Effects of prior detections on estimates of detection probability, abundance and occupancy. *The Auk*, 127(1):94–99, 2010.

[94] C P Robert and G Casella. *Monte Carlo Statistical Methods*. New YorkL Springer Verlag, 1999.

[95] M L Rosenzweig and R H MacArthur. Graphical representation and stability conditions of predator-prey interactions. *The American Naturalist*, 97(895):209–223, 1963.

[96] Christopher T Rota, Robert J Fletcher, Robert M Dorazio, and Matthew G Betts. Occupancy estimation and the closure assumption. *Journal of applied ecology*, 46(6):1173–1181, 2009.

[97] J Andrew Royle. Site occupancy models with heterogeneous detection probabilities. *Biometrics*, 62:97–102, 2006.

[98] J Andrew Royle and Robert M Dorazio. *Hierarchical Modeling and Inference in Ecology: The analysis of data from populations, metapopulations and communities*. 2008.

[99] J Andrew Royle and Marc Kery. A bayesian state-space formulation of dynamics occupancy models. *Ecology*, 88(7):1813–1823, 2007.

[100] J Andrew Royle and William A Link. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841, 2006.

[101] J Andrew Royle and James D Nichols. Estimating abundance from repeated presence absence data or point counts. *Ecology*, 84:777–790, 2003.

[102] G A Sargeant, M A Sovada, C C Slivinski, and D H Johnson. Markov chain monte carlo estimation of species distribution: a case study of the swift fox in western kansas. *Journal of wildlife managment*, 69:483–497, 2005.

[103] J Schieck. Biased detection of bird vocalizations affects comparisons of bird abundance among forest habitats. *Condor*, 99:179–190, 1997.

[104] Carl James Schwarz. The jolly-seber model: more than just abundance. *Journal of agricultural, biological and environmental statistics*, 6(2):195–205, 2001.

[105] Carl James Schwarz and A Neil Arnason. A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics*, 52(3):860–873, 1996.

[106] Carl James Schwarz and A Neil Arnason. *Program MARK "A Gentle Introduction"*, chapter 13, pages 1–53. Online, 2006.

[107] Carl James Schwarz and Geoge A F Seber. Estimating animal abundance: review iii. *Statistical Science*, 14:427–456, 1999.

[108] G A F Seber. A note on the multiple-recapture census. *Biometrika*, 52(1):249–259, 1965.

[109] Andrew T Smith. The distribution and dispersal of pikas: consequences of insular population structure. *Ecology*, 55(5):1112–1119, 1974.

[110] Andrew T Smith. Temporal changes in insular populations of the pika ochotona princeps. *Ecology*, 61:8–13, 1980.

[111] P Solymos and M Moreno. Analyzing single visit site occupancy data with detection error. r package version 1.0-0, 2010.

[112] D P Tarver, J A Rogers, M J Mahler, and R L Razor. *Aquatic and Wetland Plants of Florida*. Florida Department of Natural Resources, 1986.

[113] R.D.C Team. R: a language and environment for statistical computig. *R foundation for statistical computing, Vienna, Austria*, 2008.

[114] C A Toft and T W Schoener. Abundance and diversity of orb spiders on 106 bahamic islands: biogeography at an intermidiate trophic level. *Oikos*, 41:411–426, 1983.

[115] J Hardin Waddle, Robert M Dorazio, Susan C Walls, Kenneth G Rice, Jeff Beauchamp, Melinda J Schuman, and Frank J Mazzotti. A new parameterization for estimating co-occurrence of interacting species. *Ecological applications*, 20(5):1467–1475, 2010.

[116] Byron K Williams, James D Nichols, and Michael J Conroy. *Analysis and Managment of Animal Populations*. Academic Press, 2002.

[117] B A Wintle and D C Bardos. Modeling species-habitat relationships with spatially autocorrelated observations data. *Ecological applications*, 16(5):1945–1958, 2006.

[118] H Wu and F W Huffer. Modeling the distribution of plant species using the autologitics regression model. *Environmental and Ecological Statistics*, 4(1):31–48, 1997.

[119] Weidong Wu, Raimo Heikkila, and Ilkka Hanski. Estimating the consequences of habitat fragmentation on extinction risk in dynamic landscapes. *Landscape Ecology*, 17:699–710, 2002.

[120] Yao Yin, Jennifer S Winkelman, and Heidi A Langrehr. *Long term resource monitoring procedures: aquatic vegetation monitoring. WI LTRMP 95-P002-7.* U.S. Geological Survey, Upper Midwest Environmental Sciences Center, La Crosse, Wisconsin., 2000.