OCCUPANCY MODELLING: STUDY DESIGN AND MODELS FOR DATA COLLECTED ALONG TRANSECTS

Gurutzeta Guillera-Arroita

A THESIS SUBMITTED

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY IN

THE SUBJECT OF STATISTICS

School of Mathematics, Statistics and Actuarial Science

University of Kent

July 2012

ABSTRACT

Occupancy, defined as the proportion of sites occupied by a species, is a state variable of interest in ecology and conservation. When modelling species occupancy it is crucial to account for the detection process, as most species can remain undetected at sites where present. This is usually achieved by carrying out separate repeat visits to each sampling site but other methods are sometimes used, such as surveying spatial subunits within each sampling site, or even collecting detection data continuously along a transect, during a single visit. This thesis deals with two aspects of occupancy modelling: (i) we explore issues related to the design of occupancy studies, including the trade-off in survey effort allocation between sites and repeat visits, sample size determination and the impact of sampling with replacement in studies based on spatial replication; and (ii) we develop and evaluate new models to estimate occupancy from species detection data collected along transects, motivated by the analysis of a data set from a Sumatran tiger *Panthera tigris sumatrae* survey which followed this type of sampling protocol. The models we propose, which describe the detection process as a continuous point process, can account for clustering and/or abundance-induced heterogeneity in the detection process and represent a step forward with respect to current modelling approaches which involve data discretisation and two-stage ad hoc procedures.

ACKNOWLEDGEMENTS

First and foremost I would like to express my gratitude to my supervisors, Professor Byron Morgan and Professor Martin Ridout, for all their support and patience. They have encouraged me to think and work independently while always being there to guide me with useful advice. It has been inspirational to work with you, thank you!

I am also grateful to the Engineering and Physical Sciences Research Council (EPSRC), the National Centre for Statistical Ecology (NCSE) and the School of Mathematics, Statistics and Actuarial Science at the University of Kent for funding my PhD research. Thanks as well to Dr. David Borchers and Dr. Xue Wang for examining my thesis; to Matthew Linkie and Hariyo Wibisono for providing the Sumatran tiger survey data; to Iding, Maryati and Agung for welcoming me to Sumatra and arranging my visit to the forest in Kerinci; and to Joseph Smith for his hospitality and support during my trips to Indonesia.

Gracias to my friends for helping me remain relatively sane during this time. Thanks with all my heart to my family, for their care, encouragement and understanding. And finally, to José, for constant support, useful discussion at work... and for making my life great!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 The study of wildlife populations

Drawing inferences about the state of wildlife populations is of central interest for ecological studies as it allows the evaluation of scientific hypotheses concerning the behaviour of the system. Furthermore, in the conduct of wildlife management and biodiversity conservation, learning about the state of the population allows the assessment of whether management objectives are met and state-dependent decisions to be made. Wildlife monitoring provides a feedback link between implementation and management, a crucial element for the decision-making process (Possingham *et al.* 2001), which is particularly useful when framed within an adaptive management strategy (Salafsky, Margoluis & Redford 2001; McCarthy & Possingham 2007).

Unfortunately, at present biodiversity is being lost worldwide at a rate several orders of magnitude higher than the typical background extinction rate (Barnosky *et al.* 2011), comparable only to rates during mass extinction events. The loss and degradation of habitats, the overexploitation of natural resources, and the introduction of alien and invasive species are all factors that have driven many species to the brink of ex-

tinction, a situation which is only expected to worsen under the predicted effects of climate change (Millennium Ecosystem Assessment 2005). With an ever-growing need to protect, restore and/or manage wildlife species, communities and ecosystems, interest in the development and evaluation of informative and efficient wildlife monitoring tools is paramount.

1.1.1 Abundance and occupancy as state variables in wildlife studies

There are different state variables that can be used in the study of wildlife populations and the choice on which one to use is very much dependent on the objectives of the study (Yoccoz, Nichols & Boulinier 2001). The selected state variable should provide a useful characterization of the system and allow discrimination among relevant competing hypotheses, therefore yielding useful information to be fed back to management.

For a single species, abundance or population size is a commonly used state variable (Borchers, Buckland & Zucchini 2003). Methods for estimating abundance include well-developed techniques such as closed-population mark-recapture (Otis *et al.* 1978) and distance sampling (Buckland *et al.* 2001; Buckland *et al.* 2008). However, while the information on abundance provides a powerful characterization of the population, collecting data for its estimation in general requires substantial effort and can therefore become too time-consuming and costly when the scale of the study is large.

Occupancy, defined as the proportion of sites occupied by a species, can also be a useful state variable in the context of single-species studies (MacKenzie *et al.* 2006). Occupancy is a concept widely used in ecology. It is central for determining the geographic range of species, and it is one of the state variables considered in the criteria

for assessing the conservation status of species (IUCN 2001). Occupancy is also often used as the response variable in the modelling of habitat relationships (e.g. Fleishman *et al.* 2002; Reunanen *et al.* 2002) and is a fundamental concept in metapopulation studies (Hanski 1999), which focus on the investigation of occupancy dynamics in patchy populations.

Occupancy has been proposed as an informative state variable for large-scale monitoring programmes (MacKenzie *et al.* 2006, pp. 41-44). Given that there is an obvious relationship between abundance and occupancy (a site is occupied if site abundance is greater than zero), occupancy can be viewed as a surrogate of abundance. For abundant species, a decline in abundance might not be reflected as a decline in occupancy, however occupancy can be informative when monitoring rare species. One of the reasons why occupancy is an attractive state variable to work with is that its estimation generally requires less effort in data collection than that required in programmes aimed at estimating abundance, something particularly relevant when working at large geographical scales. Furthermore, since in general the data required for estimating species occupancy are relatively easy to collect, these surveys are well suited to be implemented as volunteer-based programmes (e.g. Kéry, Gardner & Monnerat 2010; Sewell, Beebee & Griffiths 2010) which can help in obtaining larger sample sizes. The relative ease of data collection can also facilitate the involvement of local people in monitoring activities through participatory monitoring programmes, initiatives that help promote the engagement of the local communities in conservation efforts (Danielsen, Burgess & Balmfort 2005).

1.1.2 The issue of imperfect detection

Once an appropriate state variable is chosen, attention has to be paid to how to estimate it in the most meaningful way. When studying wildlife populations a major issue to consider is that individuals often remain undetected even when present at a site (Yoccoz, Nichols & Boulinier 2001). This is true for most animal species and has been shown to be a potential issue even for sessile species such as plants (Chen *et al.* 2009). It is possible as well for the wrong species to be mistakenly recorded as the species of interest, thus yielding false positives in the data. For some species and types of surveys false positives can be of concern. For instance they have been shown to be an issue in call-based occurrence surveys of anuran and bird species (McClintock *et al.* 2010). However, since false positives are in general far less of a problem than false negatives, statistical models of wildlife populations often assume a detection process in which such misclassifications are not possible.

In abundance estimation, the issue of imperfect detection has long been recognized and dealt with. For instance, closed-population mark-recapture studies account for individual detectability by recording whether (naturally or artificially) marked individuals are recaptured in successive recapture attempts. These recaptures, which can be actual physical recaptures or simply resightings, provide information on the probability of capturing individuals (i.e. individual detectability) and thus help to obtain a better abundance estimate. In distance sampling, the method of accounting for imperfect detection is to model detectability as a function of distance, based on the recorded distances at which individuals are detected from the line transect or point of count.

Obviously, imperfect detection is also a problem for studies aimed at estimating occupancy. While surveys of empty sites necessarily yield non-detections (in the absence of false positive errors), surveys of occupied sites can lead both to detections and to non-detections. Therefore, if not accounted for, imperfect detection can induce negative bias in the estimation of occupancy. Furthermore, temporal variation in detection probability can confound the estimation of occupancy trends (Kéry *et al.* 2010), while its spatial variation may lead to incorrect inferences regarding habitat relationships (Tyre *et al.* 2003; Gu & Swihart 2004; MacKenzie 2006) and therefore to incorrect predictions about the distribution of the species (Kéry 2011).

Despite the problems associated with imperfect detection, this issue had received little attention in the context of estimating species occupancy until relatively recently. Its treatment was limited to two-step ad hoc approaches (Geissler & Fuller 1986; Azuma, Baldwin & Noon 1990; Nichols & Karanth 2002) until a model-based method that accounted explicitly for detection probability while estimating occupancy was proposed independently by MacKenzie *et al.* (2002) and Tyre *et al.* (2003). These methods are based on a discrete sampling protocol in which repeat detection/non-detection surveys are carried out at the sampling sites. A major advantage of the model-based approach over ad hoc techniques is that it provides a flexible framework to compare competing hypothesis about factors affecting occupancy and detectability. As formulated by MacKenzie *et al.* (2002), the model also allows for survey-specific variation in detectability to be accounted for.

Since its publication, the occupancy modelling approach of MacKenzie *et al.* (2002) and Tyre *et al.* (2003), to which for simplicity we refer in this thesis as the 'basic oc-

cupancy model', has been widely accepted by ecologists as a tool to study wildlife populations. This is reflected by the large and growing number of published studies that use or discuss this technique (Figure 1-1).

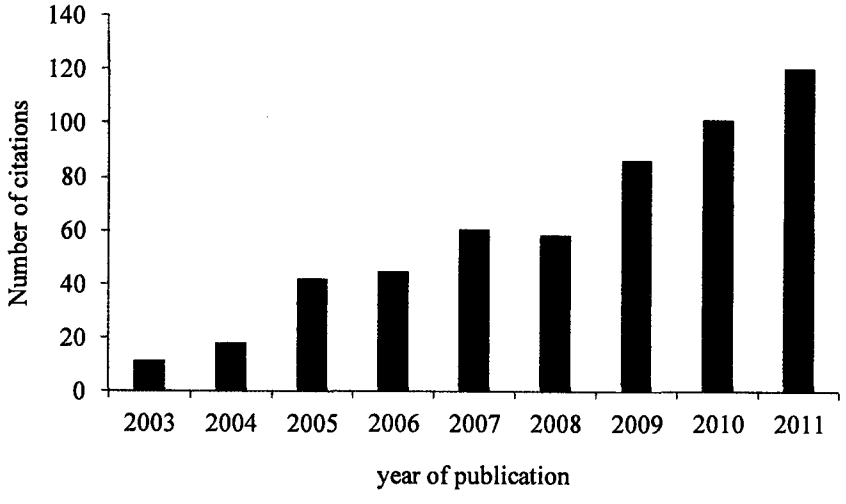


Figure 1-1 Number of new citations per year for MacKenzie *et al.* (2002), according to Web of Knowledge (© Thomson Reuters), accessed 02/April/2012.

The growing interest in this framework and its developments has been fuelled by the availability of several software packages for model fitting. These include: GUI[1]-based program MARK (White & Burnham 1999) and its R interface package, `R-Mark` (Laake & Rextad 2008), which were originally developed for fitting mark-recapture models (Williams, Nichols & Conroy 2002) but have been extended to include occupancy models; PRESENCE (Hines 2006), a GUI-based program specifically written to fit occupancy models; and finally the new R-package `Unmarked` (Fiske & Chandler 2011) which, based on a hierarchical model formulation, is being developed with the aim of providing a unified framework for modelling data from unmarked

[1] GUI: Graphical User Interface

individuals. All these packages are based on maximum-likelihood inference but the models can also be very easily implemented and fitted in the Bayesian framework using ready-made software packages for Markov chain Monte Carlo (MCMC) analysis such as WinBUGS (Lunn *et al.* 2000) or JAGS (Plummer 2003). WinBUGS code can be found for instance in Kéry & Schaub (2012, chapter 13).

1.2 Thesis motivation

This thesis deals with species occupancy modelling. Broadly, the work carried out in the thesis has two components: (i) we explore issues related to the design of occupancy studies and (ii) we develop new models to estimate occupancy from species detection data collected along transects, with the analysis of a data set from a Sumatran tiger survey as a motivating example. We concentrate on single-species (cf. co-occurrence models, e.g. MacKenzie, Bailey & Nichols 2004) and on single-season data (although we touch upon multiple-season studies when looking at study design). While the focus of the thesis is on occupancy as a state variable, our work links with the estimation of abundance from spatially-replicated repeated counts of unmarked individuals (Royle 2004b), as we will discuss later.

Study design is an important step in any statistical study involving sampling. It is only through careful study design that we can ensure good chances of obtaining meaningful results in the most efficient way. Properly addressing the design stage is particularly crucial for studies in ecological and conservation programmes, as resources are often very limited, particularly in developing countries where many of the world's most bio-diverse areas are located. Unfortunately, it is not rare for study design to be disregarded in wildlife monitoring programmes, which often renders the outcome of moni-

toring efforts to be largely uninformative (Yoccoz, Nichols & Boulinier 2001; Legg & Nagy 2006). Occupancy modelling is a relatively new area within wildlife monitoring so, although some work has been carried out previously addressing the design of occupancy surveys (Field, Tyre & Possingham 2005; MacKenzie & Royle 2005; Bailey *et al.* 2007), the investigation of design issues remains an active area of development.

The second area of research in this thesis is motivated by the fact that, in some cases, the collection of species detection data for occupancy estimation is carried out continuously along a transect, rather than using a discrete replicate sampling protocol with separate repeat visits to each sampling site. For instance, this type of survey has been applied recently to monitoring large carnivores, such as tigers in India (Hines *et al.* 2010) and Sumatra (data in this thesis), and snow leopards in Mongolia (McCarthy *et al.* 2010). Species detection data are also sometimes collected continuously during an interval of time, for instance in camera-trap surveys. Traditionally, such 'continuous' data have been analysed to estimate occupancy by discretizing the transect (or time interval) into shorter segments, assigning a '1' to each segment when there was at least one detection in the segment and a '0' otherwise, and then using an appropriate model from among those developed for discrete sampling protocols. In this thesis we develop new models that provide a more natural description for the detection process in such surveys, eliminating the need to divide transects into segments, which can be arbitrary and can lead to loss of information.

1.3 Sumatran tiger survey data

1.3.1 The Sumatran tiger

Tigers were once widely distributed in Asia, ranging from Turkey to eastern Russia, however today they only persist in less than 7% of their historic range. Tiger occupancy continues declining at present and, in fact, it has dropped dramatically in recent years (Sanderson *et al.* 2006; Dinerstein *et al.* 2007), with fewer than 3,500 individuals thought to remain in the wild. Overhunting has been the main driver of tiger decline across its range, while habitat loss and fragmentation have played locally an important role in some regions (Walston *et al.* 2010). Over the last century, tigers have disappeared from southwest and central Asia, from large areas of southeast and eastern Asia and from the Indonesian islands of Java and Bali, with several subspecies being wiped out in the process. Within Indonesia, tigers still persist on the island of Sumatra.

The Sumatran tiger *Panthera tigris sumatrae* is a subspecies of tiger that occurs only in Sumatra, in habitats ranging from lowland forest to sub-mountain and mountain forest, including some peat swamp forests. It is classified as Critically Endangered in the Red List of Threatened Species by the International Union for the Conservation of Nature (IUCN) due to the small size of its remaining population, estimated to be just a few hundred, and which appears to continue to be declining (Linkie *et al.* 2008). Sumatran tigers are under threat due to a combination of factors, including habitat destruction and degradation. Habitat loss in Sumatra is severe due to the expansion of oil palm and acacia plantations. In fact the island has one of highest rates of conversion from intact forest to non-forest in southeast Asia, with an average annual rate of deforestation of 2.1% (Uryu *et al.* 2010). Poaching of tigers is also a major problem as is the

depletion of their prey-base (Linkie *et al.* 2003). Given the critical situation of the Sumatran tiger, and the fact that it is such an iconic species, it is not surprising that several organizations, Indonesian and international, are working for its conservation and that large amounts of money are invested in trying to prevent its extinction.

Figure 1-2 Sumatran tiger *Panthera tigris sumatrae* captured by a camera trap at Kerinci-Seblat National Park (Photo: M. Linkie, Fauna & Flora International).

Since good data on the population of Sumatran tigers at the landscape-level were lacking, the recent 'National Tiger Recovery Plan' from the Indonesian Ministry of Forestry (Ministry of Forestry of Indonesia 2010) placed a strong emphasis on the development of a robust monitoring protocol to measure tiger population trends. To accomplish this, a partnership between non-governmental conservation organizations working across all the Sumatran tiger landscapes and the Indonesian Ministry of Forestry was established, resulting in the implementation of a joint Sumatra-wide survey. In

this thesis we used data from this survey as a motivating example to develop and explore statistical models for estimating species occupancy.

1.3.2 Island-wide tiger detection data set

In the period 2007-2009, eight wildlife conservation organisations (Fauna & Flora International, FFI; the Wildlife Conservation Society, WCS; the Durrell Institute of Conservation and Ecology from the University of Kent, DICE; the World Wildlife Fund, WWF; the Zoological Society of London, ZSL; the Leuser International Foundation, LIF; the Rhino Foundation of Indonesia and the Sumatran Tiger Protection and Conservation Foundation) partnered with the Indonesian Ministry of Forestry to conduct simultaneous tiger field surveys across Sumatran rainforests, following a common sampling protocol.

The aim of the joint survey initiative was to obtain baseline information about the Sumatran tiger population over the whole island. Sumatra is a large island, in fact the sixth largest in the world with a total area of 473,481 km^2. Despite large deforestation, in 2008-9, forests still covered an area of about 128,000 km^2 (Uryu *et al.* 2010). The rainforests in Sumatra are dense (Figure 1-3, left), and tigers are wide-ranging, cryptic and elusive. This, together with the fact that in Sumatra's rainforests tigers occur at low densities (Linkie *et al.* 2006), render direct observations of tigers extremely uncommon.

Obtaining indirect observations from camera-trapping is a survey method commonly used to study tiger populations in the wild (Linkie *et al.* 2010). In the case of tigers, camera-trapping provides information at the individual level given that tigers can be uniquely identified from their stripe patterns. This allows capture-recapture techniques

to be used to estimate tiger abundance and density (Karanth 1995; Karanth & Nichols 1998). However, while being a useful and powerful survey technique, camera-trapping is also very resource intensive and is therefore more suited for surveys carried out at relatively small scales.

Due to the large scale of the Sumatran tiger survey and the difficulties in detecting tigers, occupancy was chosen as a state variable for monitoring and the agreed sampling protocol consisted of footprint surveys. Despite claims that individual tigers can be identified from their footprints in some environments (e.g. Sharma, Jhala & Sawarkar 2005), in general the reliability of such identifications has been much questioned (Karanth *et al.* 2003). Consequently in this survey there was no attempt to identify individual tigers and the data collected reflect detection/non-detection at the species level.

Detecting tiger footprints is definitely easier than detecting tigers directly. However it still remains a challenging task given that only a small fraction of the forest floor can be inspected and that footprints are difficult to detect depending on the forest floor conditions (Figure 1-3, right). Even experienced surveyors can miss footprints present in a section assessed. Therefore, when estimating tiger occupancy from such data, it is crucial to model the detection process explicitly, so that false absences recorded at occupied sites can be accounted for. On the other hand, it is reasonable to expect false positives to be rare given that tiger footprints are easily distinguished from those of other sympatric carnivore species by their larger size. Hence, all the work developed in this thesis is under the assumption that the rate of false positives in the data is negligible.

Figure 1-3 Forest (left) and forest floor covered with leaf litter (right) at Kerinci-Seblat National Park

The agreed sampling protocol consisted of surveying 17 km x 17 km grid cells, and recording information on the location of tiger footprint detections along transects with a GPS (Figure 1-4). Within each of the grid cells, a team of four or five people surveyed forest trails, often along ridges. These routes were chosen given that, in the rugged Sumatran landscapes, tigers tend to use ridges to move around their territories and therefore the trails were more likely to contain tiger footprints. Since in practice only a very narrow strip of the forest floor is assessed in such surveys, all detections were considered to be made *on* the transect (i.e. the data do not consist of detections at different distances from the transect). Grid cell size was chosen based on the home-range size of adult male Sumatran tigers with the idea of allowing changes in the population of tigers to be better reflected as changes in the proportion of grid cells occupied, i.e.

as tigers disappear, cells become empty. For much greater cell sizes many individuals may disappear while a cell remains 'occupied' by the species.

Figure 1-4 Survey team members recording a tiger footprint detection at Kerinci-Seblat National Park (left) and close-up of a tiger footprint in the mud (right).

The surveys were conducted in all tiger habitat types, from sea-level peat swamp forests to the forest surrounding the volcanic peak of Mount Kerinci (Sumatra's highest point at 3,805 m above sea level) and focussed primarily on protected areas. Overall, a total of 13,511 km of transects were surveyed in 394 grid cells that covered tiger landscapes across all eight mainland Sumatran provinces (Figure 1-5). Tiger footprints were detected in 206 grid cells.

The data were provided for analysis after being processed to obtain a detection/non-detection history. To construct the history, transects were divided into segments of 1

km in length, assigning '1' to those segments containing at least one detection and '0' otherwise.

Figure 1-5 Map of Sumatra showing the grid cells used for the survey. The cells corresponding to Kerinci-Seblat national part are coloured in green.

1.3.3 Kerinci-Seblat tiger detection data set

A subset of the data corresponding to the surveys carried out in the forests in and around Kerinci-Seblat National Park, the largest national park in Sumatra (13,971 km^2), was made available in raw format. The data set included the geographical loca-

tion of each tiger footprint detection together with information on the actual route walked within each sampling site. This part of the data set was collected by Fauna and Flora International and the Durrell Institute of Conservation and Ecology during 2007 and 2008.

The surveys at Kerinci-Seblat covered a total distance of 2826.5 km within 89 sampling sites (17 km x 17 km grid cells). The distance surveyed per site varied from 1.8 km to 108.3 km, and was typically around 15-45 km (Figure 1-6). Within each cell the surveys were often made up of several transects which varied in length from 0.5 to 40.1 km, with most in the range 3-21 km. Tiger footprints were detected in 66 of the cells, which gives a naïve occupancy estimate (# sites with detection / # sites) of 0.74.

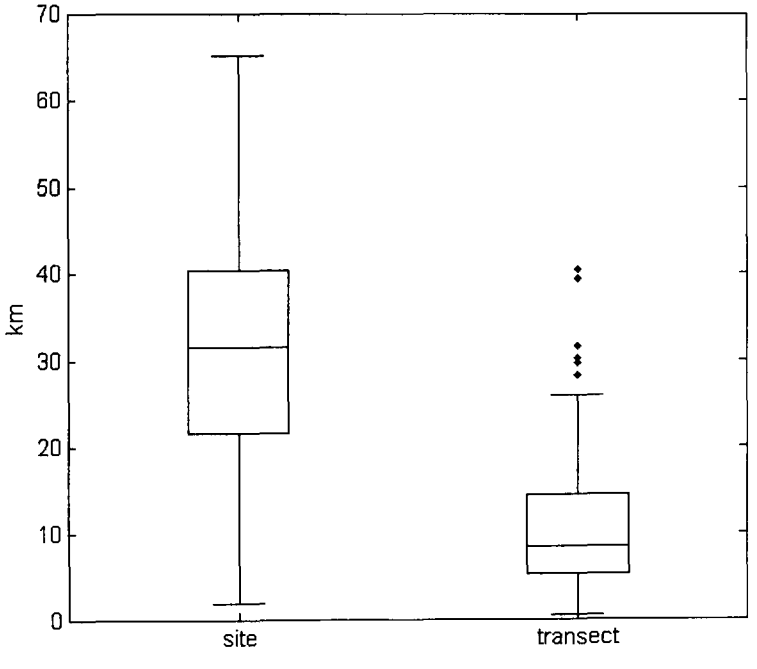


Figure 1-6 Survey lengths of individual transects and total distances walked within each 17x17 km sampling site in the tiger surveys at Kerinci-Seblat national park. In the plot the y-axis is limited for clarity to 70 km; for one of the cells a larger distance was surveyed (108.3 km).

1.4 General thesis methodology

The work in this thesis was carried out within the maximum-likelihood inference framework. Models were implemented in MATLAB (the Mathworks Inc., version 7.12 or earlier) unless otherwise stated. Maximum-likelihood estimates were obtained by numerical minimization of the negative log-likelihood function using the optimization routine *fminsearch*, which implements the Nelder-Mead simplex search algorithm (Nelder & Mead 1965). As *fminsearch* is an unconstrained optimization function, parameters were transformed to the logistic or log scale as appropriate prior to computation. The logit transformation was used for the probability parameters, such as occupancy and detection probabilities, which are constrained to the interval $[0, 1]$. The log transformation was used for those parameters that were constrained to $[0, \infty)$, such as detection rates or average abundance. When covariates were incorporated into the models, this was done on the transformed scale following a generalized linear model approach.

Standard errors were derived on the transformed scale (logistic or log) with the function *mlecov*, which returns an approximation to the asymptotic variance-covariance matrix of the maximum-likelihood estimators. This function computes a finite difference approximation to the negative hessian of the log-likelihood evaluated at the maximum-likelihood estimates (i.e. the observed information matrix; Morgan 2008, p. 78) and returns its inverse. Standard errors on the original parameter scale (probability or rate) were calculated from those obtained on the transformed scale using the delta method approximation (Davison 2003, pp. 33-35).

Model selection was carried out using the Akaike Information Criterion, or AIC (Akaike 1973; Burnham & Anderson 2002). The AIC is founded in information theory and is defined as

$$AIC = -2\mathcal{L}(\widehat{\boldsymbol{\theta}}|x) + 2K\,, \tag{1.1}$$

where $\mathcal{L}(\widehat{\boldsymbol{\theta}}|x)$ is the maximum value of the log-likelihood function given the observed data x (i.e. the value of the log-likelihood evaluated at the maximum-likelihood estimates $\widehat{\boldsymbol{\theta}}$) and K is the number of parameters in the model. The AIC provides a measure of the *relative* goodness of fit of a statistical model, with the smaller the AIC, the better the fit of the model. The AIC reflects the principle of parsimony, that is, the trade-off between under-fitting and over-fitting. The first term in (1.1) represents a measure of how well the model fits the data and can be reduced by introducing more parameters in the model. However, the second term acts as a penalty term, getting larger as more parameters are included. Hence, complicating the model results in a trade-off in terms of model support.

Since it is not the absolute value of the AIC that matters, but the differences between the AICs of the models included in the candidate set, normally the so-called ΔAICs are reported. The ΔAIC for model i is calculated as

$$\Delta AIC_i = AIC_i - AIC_{\min},$$

that is, the difference between the AIC of model i and the AIC of the best-fitting model in the candidate set (i.e. the lowest AIC). Of course, the best model in the set has a ΔAIC equal to zero.

A modification of the AIC for small sample sizes, based on a second-order bias correction (Hurvich & Tsai 1989), exists as follows

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}, \qquad \text{for } n > K+1,$$

where n is the sample size. AIC_c has an additional correction term which, when n is large with respect to K, is negligible. Therefore as the sample size increases the AIC_c converges to the AIC. As the AIC may perform poorly when there are too many estimated parameters in relation to the size of the sample, the use of the AIC_c variant is recommended in general (e.g. Anderson 2008, pp. 60-61). However, in the context of occupancy modelling, the question of what constitutes the effective sample size (e.g. number of sampling sites? number of detections? something else?) remains open (MacKenzie *et al.* 2006, p. 80). For this reason, throughout this thesis the AIC was used. An exception was the analysis of tiger data in section 2.3 for which the AIC_c was used, taking the number of sampling sites as the effective sampling size (as calculated by program MARK).

1.5 Thesis structure and contributions

This thesis is organized into four core chapters. Chapter 2 presents the main models that have been developed to date to study species occupancy while accounting for imperfect detection. All these models assume that the sampling protocol consists of replicate discrete surveys at a number of sites. We start the chapter by presenting in detail the basic occupancy model and exploring its properties, including the conditions that lead to boundary estimates. We then provide an overview of model extensions and related model developments that have been proposed in the literature, focusing on those that are particularly relevant for the rest of the work in this thesis. We conclude this chapter by applying the models to the analysis of the Sumatra-wide tiger data set.

In Chapter 3 we look at issues related to the design of occupancy studies based on the basic occupancy model. We look at the optimal allocation of survey effort into number of sampling sites and number of repeat visits per site, the determination of sample size to achieve a given power to detect a difference in occupancy between two samples, and the impact of sampling with replacement in occupancy studies based on spatial replication.

In Chapter 4 we propose and evaluate new occupancy models that are useful for sampling situations in which detection data are collected continuously along a transect (as in the Sumatran tiger survey) or interval of time. These models are based on describing the detection process as a point process rather than on artificially discretizing the data as was previously normally done when analyzing this kind of data. We start with a simple model that assumes independence among detections. We then relax the as-

sumption of independence to cover cases in which there is clustering in the detections, which we model using Markov-modulated Poisson processes (MMPPs).

In Chapter 5 we extend the models proposed in Chapter 4 to the case in which there is abundance-induced heterogeneity in the detection process. These models describe the species detection process as resulting from the superposition of individual point processes, and provide an estimate of population abundance. We first propose a model based on the assumption of independent detections and then present a model that accounts for clustering in the detections. Both in Chapter 4 and Chapter 5, we illustrate the utility of the models proposed by fitting them to the tiger data set from Kerinci-Seblat National Park.

Part of the work from this thesis has been published in the following papers:

- Sections 2.1.3 and 3.1 in *Methods in Ecology and Evolution* (Guillera-Arroita, Ridout & Morgan 2010)
- Section 2.3 in *PLoS ONE* (Wibisono *et al.* 2011)
- Section 3.4 in *Methods in Ecology and Evolution* (Guillera-Arroita 2011)
- Chapter 4 in *Journal of Agricultural, Biological, and Environmental Statistics* (Guillera-Arroita *et al.* 2011)
- Chapter 5 in *Methods in Ecology and Evolution* (Guillera-Arroita *et al.* 2012)

The work in Section 3.2 has been accepted for publication in *Methods in Ecology and Evolution*.

2 OCCUPANCY MODELS BASED ON DISCRETE SAMPLING

This chapter presents a review of models for the analysis of species occupancy data while accounting for imperfect detection based on discrete sampling protocols. The first part of the chapter, section 2.1, is devoted to describing in detail the basic occupancy model, including the model formulation, assumptions and properties of its maximum-likelihood estimators. This model, initially proposed by MacKenzie *et al.* (2002) and Tyre *et al.* (2003), has received wide acceptance as a tool among ecologists and provides the basis for model extensions that have been subsequently developed. The chapter then continues with section 2.2, which provides a general review of the models that have been suggested in the literature to relax some of the assumptions of the basic occupancy model. Emphasis is put on some extensions that are particularly relevant for the work developed in this thesis: (i) a model that relaxes the assumption of independence between adjacent replicates (Hines *et al.* 2010) and (ii) models for heterogeneous detection probabilities (Royle 2006), including a model extension to account for abundance-induced heterogeneity in detection probability (Royle &

Nichols 2003). We also discuss a related model that allows estimating abundance from replicated counts (Royle 2004b), as well as an extension of the basic single-season occupancy model to estimate the processes underlying occupancy dynamics (i.e. local extinction and colonization) from multiple-season data (MacKenzie *et al.* 2003). To conclude the chapter, section 2.3 provides an illustration of the application of the discussed single-season occupancy models to the analysis of the island-wide Sumatran tiger dataset. This work has been published in Wibisono *et al.* (2011), while some of the results in section 2.1.3 feature in Guillera-Arroita, Ridout & Morgan (2010).

2.1 Basic occupancy model

2.1.1 Detection/non-detection data

The occupancy model proposed by MacKenzie *et al.* (2002) and Tyre *et al.* (2003) assumes a sampling protocol in which a number of discrete replicate surveys, K, are carried out at a number of sampling sites, S, during a single sampling season, recording whether the species of interest was or was not detected at each of the individual surveys. The idea behind having replication within sites is to be able to estimate separately site occupancy probability (ψ) and species detection probability given occupancy (p). If no replication is used the quantity estimated by considering the proportion of sites where the species was detected would be the product ψp, therefore underestimating species occupancy if detection is not perfect ($p < 1$). In practice, the replication is often achieved by carrying out repeated surveys of the sampling sites at different points in time throughout the sampling season, but other kinds of replication involving a single visit to each site are possible (MacKenzie *et al.* 2006). For instance this can be achieved by having simultaneous independent surveys by different observ-

ers or simultaneous independent detection methods. Replication can also be achieved spatially by surveying different sectors (subunits) within each sampling site (Figure 2-1). If detectability is constant, a design based on replicate surveys within a single visit may be more efficient if there is considerable expense involved in accessing the sites. However, if detectability varies, e.g. daily, the resulting heterogeneity in detection probability among sites can induce bias in the estimation of occupancy (MacKenzie *et al.* 2006; Royle 2006). In this case a sampling protocol based on multiple visits would be a better option, as this allows each site to be surveyed under a range of conditions.

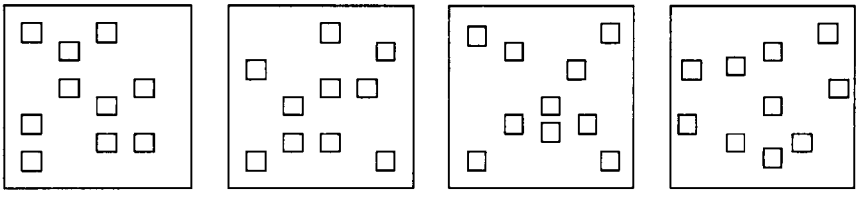


Figure 2-1 Four sampling sites and their sampling subunits in a hypothetical occupancy study with spatial replication.

The detection/non-detection history resulting from data collection is normally expressed as a matrix of '1's and '0's, denoting respectively whether the species was detected or not at each particular survey (Figure 2-2). This type of data is often referred to as 'presence/absence data'. However, since lack of detection cannot unequivocally be associated with species absence, a more appropriate term is 'detection/non-detection data'. Note that the number of replicates carried out at each sampling site, K_i, does not need to be the same necessarily and the model can readily accommodate missing observations.

site	replicates 1	2	3	4	5	...	K
1	0	1	0	1	1	...	0
2	0	0	0	1	1	...	1
3	0	0	0	0	0	...	1
4	0	1	1	0	0	...	0
5	0	0	0	1	0	...	0
6	0	0	0	0	0	...	0
7	0	0	0	0	1	...	0
8	1	1	0	1	0	...	0
9	0	0	0	0	0	...	0
...	...	...	...	...	...	...	...
S	0	1	1	1	1	...	0

Figure 2-2 Example of detection history data set where S sites are surveyed K times

2.1.2 Model formulation and assumptions

The model describes the detection data at sites occupied by the species as a series of independent Bernoulli trials with probability of success p. The occupancy status for each site is the outcome of a Bernoulli trial with probability ψ and is assumed to remain constant during the whole sampling season. Since the species can be absent from some of the sites the model is in fact a zero-inflated binomial model (Hall 2000), with $1-\psi$ determining the zero-inflation. The corresponding likelihood function given detection history h is

$$L(\psi, p|h) = \prod_{i=1}^{S} \{\psi p^{d_i}(1-p)^{K_i - d_i} + (1-\psi) I(d_i = 0)\}, \tag{2.1}$$

where d_i is the number of detections at site i and $I(\cdot)$ represents the indicator function, which takes the value one when the expression in brackets is true and zero otherwise. There are two possible explanations for site histories without detections: either (i) the

species was present at the site but was not detected at any of the repeat surveys or (ii) it was not present at the site. Note that the likelihood considers the data at each site as the particular sequence of detections/non-detections rather than as the number of detections, and therefore there are no associated combinatorial terms. While these terms do not involve the parameters, and consequently do not affect their estimation, the distinction above is relevant for the models to be comparable with those that allow for survey-specific detection probability (section 2.1.4).

In a hierarchical formulation (Royle & Dorazio 2008, p. 106) the model can be interpreted as the superposition of two linked stochastic processes,

$$\text{system process: } z_i \sim \text{Bernoulli}(\psi)$$
$$\text{observation process: } d_{ij} \sim \text{Bernoulli}(z_i p),$$

where z_i is the occupancy status for site i and d_{ij} is the outcome of replicate survey j at site i. The observation process describing the data is conditional on the system process, which models the ecological process and is in general the focus of inference.

In summary, the basic occupancy model makes the following assumptions:

(i) Sites are closed to changes in occupancy during the sampling season, that is, the occupancy status of each site remains the same for all the survey replicates. When replication is achieved spatially this assumption is violated if only part of the site is occupied by the species (see section 3.4);

(ii) Occupancy probability is constant across sites (or its variation is modelled by site covariates; see section 2.1.4);

(iii) Detection probability is constant across sites and replicates (or its variation is modelled by site/survey covariates; see section 2.1.4);

(iv) Detections of the species are independent, that is, whether the species is detected at a survey replicate or not does not depend on whether it was detected in other replicates;

(v) There are no false positives in the data, that is, there are no detections of other species misidentified as detections of the species of interest.

MacKenzie *et al.* (2006, pp. 104-108) provide some discussion on the impact that violating these assumptions has on the estimators. Model extensions have been proposed to address some cases in which these assumptions are not met (see section 2.2).

2.1.3 MLEs and estimator properties

In order to explore the general properties of the maximum-likelihood estimators (MLEs) for the basic occupancy model, a standard survey design with K surveys carried out in S sampling sites is assumed. The likelihood function corresponding to the constant probability occupancy model for a standard design can be rewritten in a compact form as follows

$$L(\psi, p|h) = \{\psi^{S_d} p^{d_T} (1-p)^{KS_d - d_T}\}(1-\psi p^{\star})^{S-S_d}, \tag{2.2}$$

where S_d is the number of sites where the species was detected at least once, $d_T = \sum_{i=1}^{S} d_i$ is the total number of detections in the detection history and $p^{\star} = 1 - (1-p)^K$ is the probability of detecting the species in at least one of the K surveys carried out at an occupied site. Note that (S_d, d_T) is a sufficient statistic as it summarizes

the detection history with no loss of information. MacKenzie *et al.* (2006, p. 95) point out that the analytical solution for the MLEs satisfies the equations

$$\hat{\psi} = \frac{S_d}{S\hat{p}^*}, \qquad \frac{\hat{p}}{\hat{p}^*} = \frac{d_T}{S_d K}, \tag{2.3}$$

that is, as $\hat{p}^*$ gets smaller, the estimate of occupancy $\hat{\psi}$ increases and the estimate of detection probability $\hat{p}$ decreases compared to the naïve estimates obtained assuming that the species was not missed at any of the occupied sites: $\hat{\psi}_{naive} = S_d/S$ and $\hat{p}_{naive} = d_T/(S_d K)$. The expressions in (2.3) can be easily derived using the parameterization proposed by Morgan, Revell & Freeman (2007) for simplifying the likelihood of site occupancy models. Setting $\theta = \psi p^*$, the probability that a site is occupied and the species is detected there, in (2.2) leads to

$$L(\theta, p|h) = \{\theta^{S_d}(1-\theta)^{S-S_d}\}\left[\left(\frac{p}{1-p}\right)^{d_T}\left\{\frac{(1-p)^K}{p^*}\right\}^{S_d}\right], \tag{2.4}$$

a factorization of the likelihood into two parts, each one only involving one of the parameters (θ or p). In the first part it is straightforward to see that $\hat{\theta} = S_d/S$, the proportion of cells where the species was detected, and that therefore the estimate of occupancy $\hat{\psi}$ satisfies (2.3). The MLE expression for detection probability can be easily derived by differentiating the second part of the likelihood function in (2.4) and setting this equal to zero.

Evaluating the performance of the model via simulations, MacKenzie *et al.* (2002) noted that, when working with small probabilities of detection, they sometimes obtained estimates of occupancy that tended to 1. Here we show how the detection histo-

ries that result in boundary occupancy estimates ($\widehat{\psi} = 1$) can be easily identified analytically. For this, let us start considering separately the part of the likelihood involving the parameter p, which we call here function h

$$h(p) = \left(\frac{p}{1-p}\right)^{d_T} \left\{\frac{(1-p)^K}{p^*}\right\}^{S_d}. \tag{2.5}$$

The condition in (2.3) fulfilled by the points of $h(p)$ with first derivative equal to zero can be rewritten for convenience as

$$x^K = xB + (1 - B), \tag{2.6}$$

where $x = 1 - p$, $B = KS_d/d_T$. Note that here 'hats' are removed from the notation, to recognize that not all the solutions of (2.6) lead to MLEs. Since p is a probability then $x \in [0, 1]$. Also, given that the number of sites where the species was detected cannot exceed the total number of detections (i.e. $S_d \leq d_T$), then $B \in [1, K]$. Regardless of B, (2.6) has a trivial solution for $x = 1$, that is, $p = 0$. Let $g(x) = x^K$ and $f(x) = xB + (1 - B)$. It is easy to see that (2.6) has a second real solution with $p > 0$ as long as $S_d \neq d_T$, that is, $B \neq K$ (Figure 2-3, b-d). Since there are two solutions, both of them cannot be maxima. Evaluating (2.5) at the trivial solution, applying l'Hôpital's rule, we have

$$\lim_{p\to 0} h(p) = \lim_{p\to 0} \frac{p^{d_T}}{(p^*)^{S_d}} = C \cdot \lim_{p\to 0} \frac{p^{d_T - S_d}}{(p^*)^0}, \tag{2.7}$$

where C is a constant. This expression is equal to zero if $d_T \neq S_d$ so, given that $h(p)$ cannot be smaller than zero, this solution must be a minimum of the function and

therefore the non-trivial solution is a maximum. When $d_T = S_d$ (Figure 2-3a), since there is only one solution to (2.6), then $p = 0$ must correspond to a maximum. This is easily verified considering that, in this case, the expression in (2.7) is larger than zero and that, for instance, $h(p)$ evaluated at $p = 1$ is zero, that is, the function is smaller for values other than $p = 0$.

As both ψ and p are probabilities, they are restricted to [0, 1]. However according to (2.3) $\hat{\psi}$ would take values larger than unity if $p^* < S_d/S$, or equivalently, considering that $p^* = pS_dK/d_T$, if $p < d_T/(SK)$, that is, $x > 1 - d_T/(SK)$. It can be easily seen (Figure 2-3) that $g(x)$ and $f(x)$ cross for $x \leq 1 - d_T/(SK)$, and that therefore (2.6) has a non-trivial solution not leading to $\psi > 1$, if

$$f\big(1 - d_T/(SK)\big) \geq g\big(1 - d_T/(SK)\big), \tag{2.8}$$

that is, if

$$\left(\frac{S - S_d}{S}\right) \geq \left(1 - \frac{d_T}{SK}\right)^K. \tag{2.9}$$

If the data do not fulfil this condition, then the expressions for the MLEs in (2.3) are not valid. The maximum of the likelihood function is in this case on the boundary $\hat{\psi} = 1$. Maximizing (2.2) evaluated at $\psi = 1$, we obtain that $\hat{p} = d_T/(SK)$.

The inequality in (2.9) indicates that the occupancy estimate hits the boundary when the proportion of sites where the species was not detected (left term) is smaller than the proportion of zeros in the history raised to the power of K (right term). This suggests

that boundary estimates may be an issue when working with small sample sizes and low probabilities, especially when the amount of replication is small.

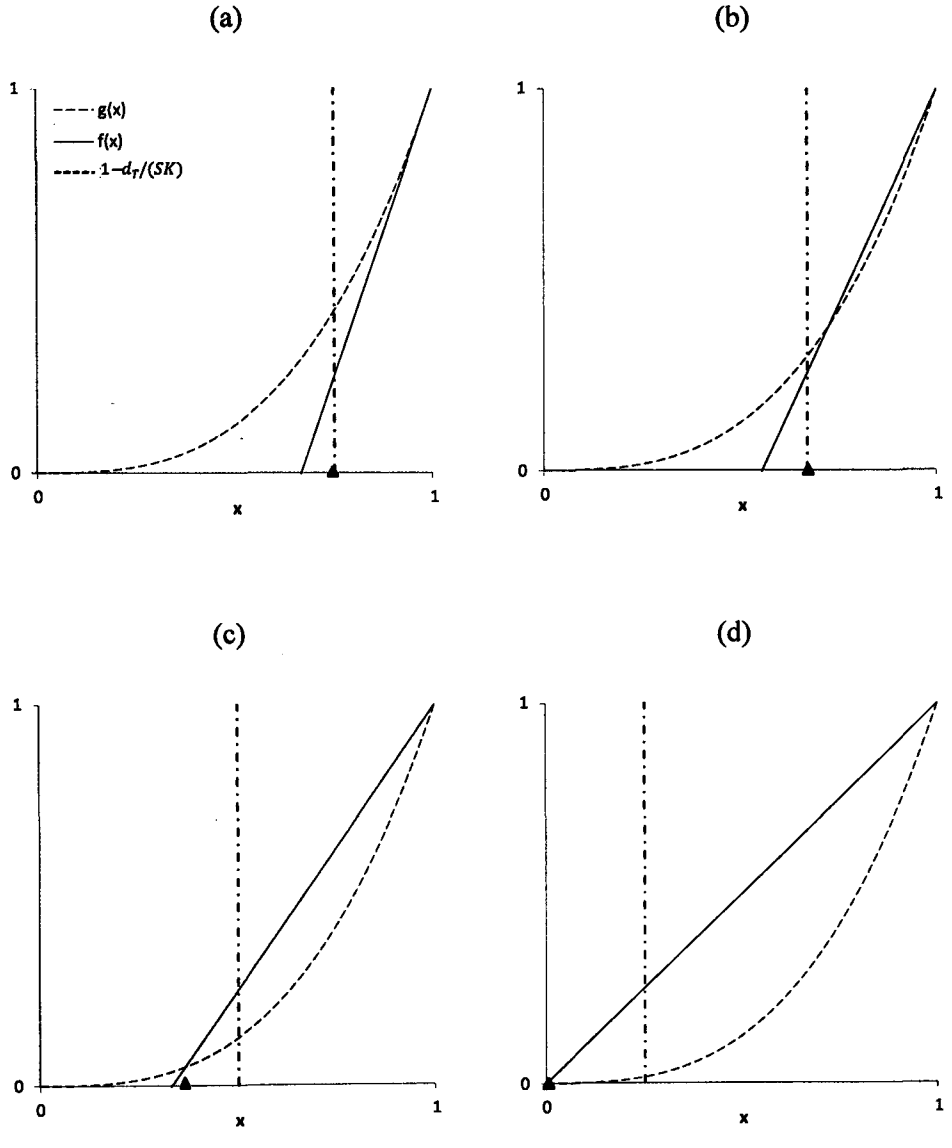


Figure 2-3 Condition for boundary occupancy estimate: scenarios (a) and (b) lead to boundary estimate, (c) and (d) lead to a non-boundary estimate. In (a) $d_T = S_d$ (i.e. $B = K$), while in (d) $d_T = S_d K$ (i.e. $B = 1$). This example was generated with $S = 100$, $S_d = 75, K = 3$ and (a) $d_T = 75$, (b) $d_T = 100$, (c) $d_T = 150$ and (d) $d_T = 225$. The black triangles in the x-axis represent the value corresponding to the actual MLE for detection probability. Note: $g(x) = x^K$, $f(x) = xB + (1 - B)$, $B = KS_d/d_T$ and $x = 1 - p$, with $x > 1 - d_T/(SK)$ being not feasible.

In summary, the MLEs of the basic occupancy model are such that:

(i) if $d_T = S_d$ (i.e. the species is detected at most once at any site with detections; Figure 2-3a) there is only one solution for (2.6), $p = 0$, which is a maximum of (2.5) and leads to a boundary occupancy estimate, $\hat{\psi} = 1$, and $\hat{p} = d_T/(SK)$.

(ii) if $d_T = S_d K$ (i.e. the species is detected at all survey replicates at all sites where detected; Figure 2-3d) there are two solutions for (2.6), the trivial solution, which is a minimum of (2.5), and a second one, which is the maximum, for $x = 0$, that is, $p = 1$. In this case the occupancy estimate coincides with the naive estimate, $\hat{\psi} = S_d/S$, and $\hat{p} = 1$, as given by (2.3).

(iii) if $\left(\frac{S-S_d}{S}\right) < \left(1 - \frac{d_T}{SK}\right)^K$ (Figure 2-3b) there are two solutions for (2.6), the trivial solution, which is a minimum of (2.5), and a second one, which is the maximum and leads to a boundary occupancy estimate, $\hat{\psi} = 1$, and $\hat{p} = d_T/(SK)$.

(iv) if $\left(\frac{S-S_d}{S}\right) \geq \left(1 - \frac{d_T}{SK}\right)^K$ (Figure 2-3c) there are two solutions for (2.6), the trivial solution, which is a minimum of (2.5), and a second one, which is the maximum and leads to an occupancy estimate which lies within the probability boundaries, determined as in (2.3).

A graphical representation of all the MLEs obtainable for a given design illustrates the issues resulting from small sample sizes and the effect that increasing the number of sites or replicates has on the quality of the estimates (Figure 2-4). Given a finite num-

ber of sites S and replicates K there is a finite number of histories that can be theoretically observed (i.e. 2^{KS} possible combinations of zeros and ones). Under the model with constant probabilities of occupancy and detectability, all those histories that share the same (S_d, d_T), which are sufficient statistics, produce the same estimates of occupancy and detection (2.3). This results in $(S+1)\{1+S(K-1)/2\}$ possible estimate points in the parameter space, represented as dots in the figure, with dotted lines connecting estimates for histories that share S_d, from $S_d = 1$ (bottom) to $S_d = S$ (top). Moving along the lines from right to left, dots correspond to histories with a decreasing d_T, from a maximum KS_d to a minimum S_d. At the right-most side of the graph, estimates correspond to the naive estimates and 'bend' upwards as detectability gets smaller.

When sample sizes are very small, there are only a few distinct detection histories that can be observed and, correspondingly, few possible parameter estimate values (Figure 2-4a). The parameter space is sparsely covered by the potential values for the MLEs, which indicates that the estimators are not precise, an effect more pronounced as probabilities of occupancy and detection get smaller. In fact there are no solutions covering the area corresponding to the lowest probabilities, which causes the estimator to be substantially biased in this region. As more samples are added to the study, the MLE solutions cover more of the probability space. Additional replication results in a better coverage of the area corresponding to low probabilities of detection (Figure 2-4b), while an increase in the number of sampling sites achieves a more even coverage in the area corresponding to high probabilities of detection (Figure 2-4c). When the amount of replication is large the MLEs coincide with the naïve estimates in most cases as p^* is close to unity, except for very low values of p.

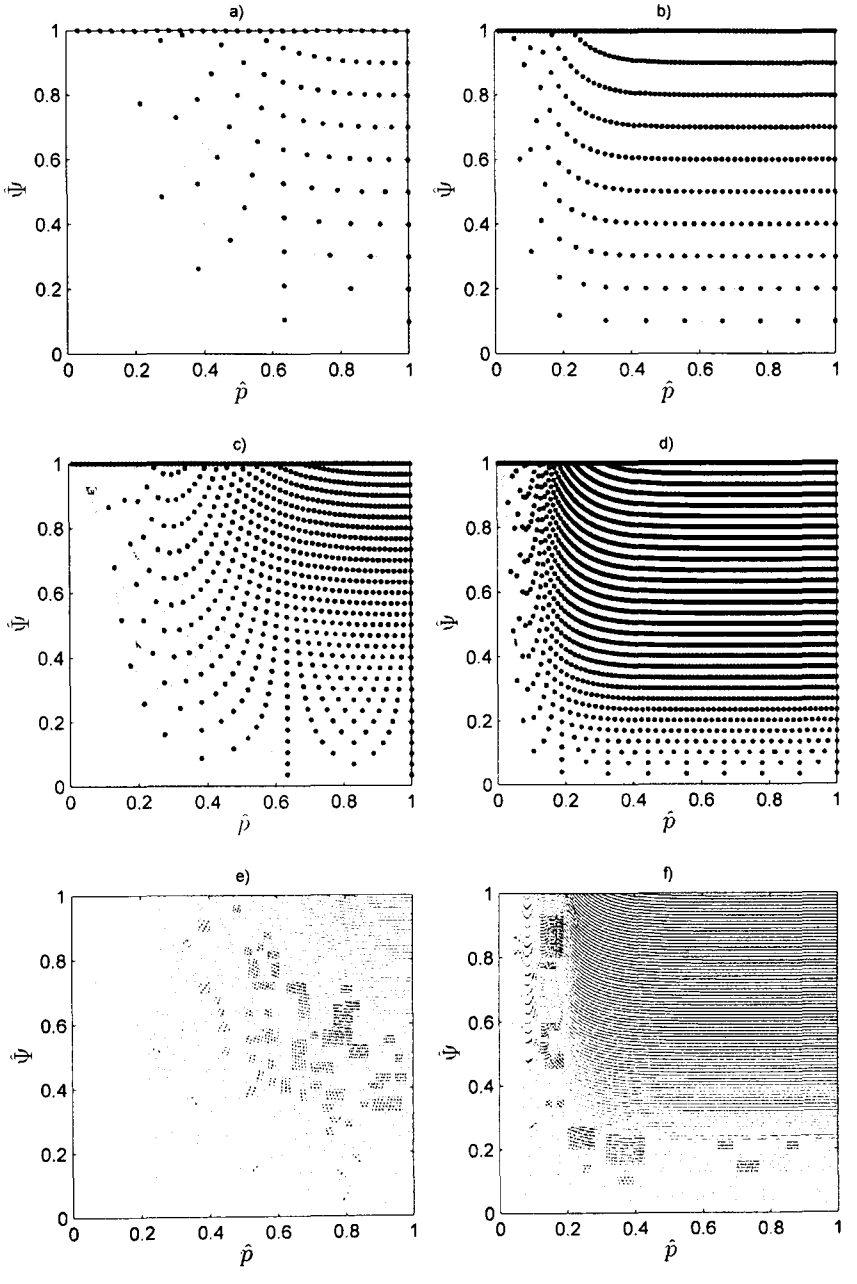


Figure 2-4 Maximum-likelihood estimates for all possible detection histories observable under a standard design with S and K (a) 10, 3, (b) 10, 9, (c) 30, 3, (d) 30, 9, (e) 100, 3, (f) 100, 9, with no assumptions made about true values of the parameters. Dotted lines connect estimates for histories that share S_d, from 1 (bottom) to S (top). For clarity (e) and (f) have been plotted without lines and using smaller markers.

Likelihood theory provides tools for approximating the properties of the MLEs when the sample size is large. The theory indicates that the estimators are asymptotically unbiased and normally distributed (Morgan 2008, p. 78). The asymptotic variance-covariance matrix of the maximum-likelihood estimators can be derived by inverting the expected Fisher information matrix, i.e. the expectation of the second derivative of the negative log-likelihood with respect to the parameters $\boldsymbol{\theta}$, which has elements

$$j_{ik} = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_k}\right],$$

where $\mathcal{L} = \log(L)$ is the log-likelihood function. MacKenzie and Royle (2005) provide the expression for the asymptotic variance of the occupancy estimator. Here we derive the remaining elements of the variance-covariance matrix.

For the model under consideration, the first derivative of the log-likelihood function, the scores vector, is

$$\begin{aligned}\mathbf{G} &= \left[\frac{\partial \mathcal{L}}{\partial \psi} \quad \frac{\partial \mathcal{L}}{\partial p}\right] \\ &= \left[\frac{S_d - S\psi p^*}{\psi(1-\psi p^*)} \quad \frac{d_T - KS_d p}{p(1-p)} - \frac{(S-S_d)\psi K(1-p)^{K-1}}{1-\psi p^*}\right].\end{aligned} \tag{2.10}$$

The elements of the observed information matrix $\mathbf{O}$, defined as minus the Hessian matrix of $\mathcal{L}$, are

$$\begin{aligned}
\mathbf{O}[1,1] &= -\frac{\partial^2 \mathcal{L}}{\partial \psi^2} = \frac{S_d}{\psi^2} + \frac{(S - S_d)p^{*2}}{(1 - \psi p^*)^2} \\
\mathbf{O}[1,2] &= -\frac{\partial^2 \mathcal{L}}{\partial \psi \partial p} = \frac{(S - S_d)K(1-p)^{K-1}}{(1 - \psi p^*)^2} \\
\mathbf{O}[2,2] &= -\frac{\partial^2 \mathcal{L}}{\partial p^2} = \frac{d_T}{p^2} + \frac{KS_d - d_T}{(1-p)^2} \\
&\quad - \frac{(S - S_d)\psi K(1-p)^{K-1}}{(1-p)(1 - \psi p^*)^2}(\psi p^* - 1 + K - \psi K).
\end{aligned} \tag{2.11}$$

Since the expectations for the data are

$$\begin{aligned}
\mathbb{E}[S_d] &= S\psi p^* \\
\mathbb{E}[d_T] &= S\psi pK,
\end{aligned}$$

the expected information matrix $\mathbf{I} = \mathbb{E}[\mathbf{O}]$ has elements

$$\begin{aligned}
\mathbf{I}[1,1] &= \frac{Sp^*}{\psi(1 - \psi p^*)} \\
\mathbf{I}[1,2] &= \frac{SK(1-p)^{K-1}}{(1 - \psi p^*)} \\
\mathbf{I}[2,2] &= \frac{SK\psi}{p(1-p)}\left\{1 - \frac{Kp(1-p)^{K-1}(1-\psi)}{1 - \psi p^*}\right\},
\end{aligned} \tag{2.12}$$

which in this case, given (2.3), are the same as those of the observed information matrix evaluated at the MLEs. Finally, the elements of the asymptotic variance-covariance matrix, $\mathbf{\Sigma} = \mathbf{I}^{-1}$, are given by

$$\boldsymbol{\Sigma}[1,1] = \mathrm{var}(\hat{\psi}) = \frac{\psi}{S}\left\{(1-\psi) + \frac{1-p^*}{p^* - Kp(1-p)^{K-1}}\right\},$$

$$\boldsymbol{\Sigma}[1,2] = \mathrm{cov}(\hat{\psi},\hat{p}) = \frac{-p}{S}\left\{\frac{1-p^*}{p^* - Kp(1-p)^{K-1}}\right\}, \tag{2.13}$$

$$\boldsymbol{\Sigma}[2,2] = \mathrm{var}(\hat{p}) = \frac{p(1-p)}{SK\psi}\left\{\frac{p^*}{p^* - Kp(1-p)^{K-1}}\right\}.$$

Looking at (2.13) it can be seen that, as p^* approaches unity, that is, as the probability of missing the species at occupied sites approaches zero,

(i) the variance of the occupancy estimator $\hat{\psi}$ tends to the variance dictated by the binomial proportion $\psi(1-\psi)/S$ and decreases as the number of sites increases (Figure 2-5);

(ii) the variance of the detection probability estimator $\hat{p}$ tends to $p(1-p)/(SK\psi)$ and decreases as the total effort (SK) increases regardless of whether it is spent on surveying more sites or more replicates within each site;

(iii) the covariance approaches zero.

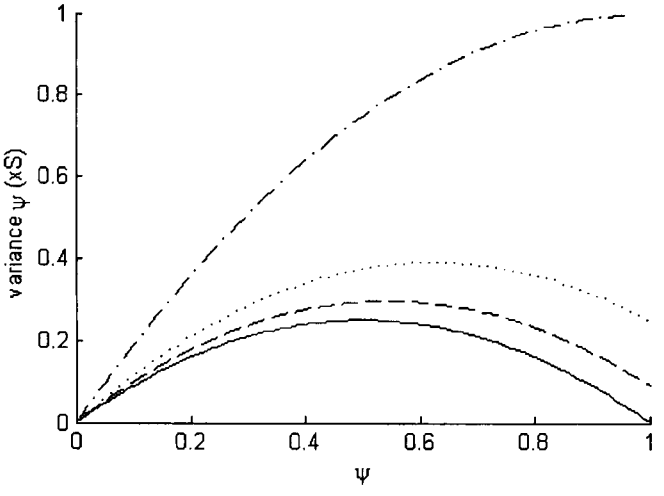


Figure 2-5 Asymptotic variance of the occupancy estimator $\times S$ for different levels of occupancy, given detection probability $p = 0.5$, S sampling sites and replication: $K = 2$ (dash-dot), $K = 3$ (dotted), $K = 4$ (dashed) and $K = 10$ (solid). For $K = 10$, the probability of missing the species at an occupied site is close to zero and therefore the variance is that of a binomial experiment.

Likelihood theory tells us that asymptotic approximations are good when the sample size is large enough, however it does not tell us how large it needs to be. Figure 2-6 illustrates graphically how the properties of the MLEs under the constant occupancy model depart from the asymptotic approximation for a combination of design parameter values that is realistic within the context of ecological studies ($SK = 168$ units of total effort). As might be expected, the difference between the approximate and actual estimator distributions is larger for low probabilities of occupancy and detection. For small probabilities of occupancy and detection the estimators have strong bias, with many of the detection histories resulting in boundary estimates (dots at the top left of the plot). As probabilities increase, the true distribution of the MLEs becomes closer to the bivariate normal distribution predicted by the asymptotic approximation.

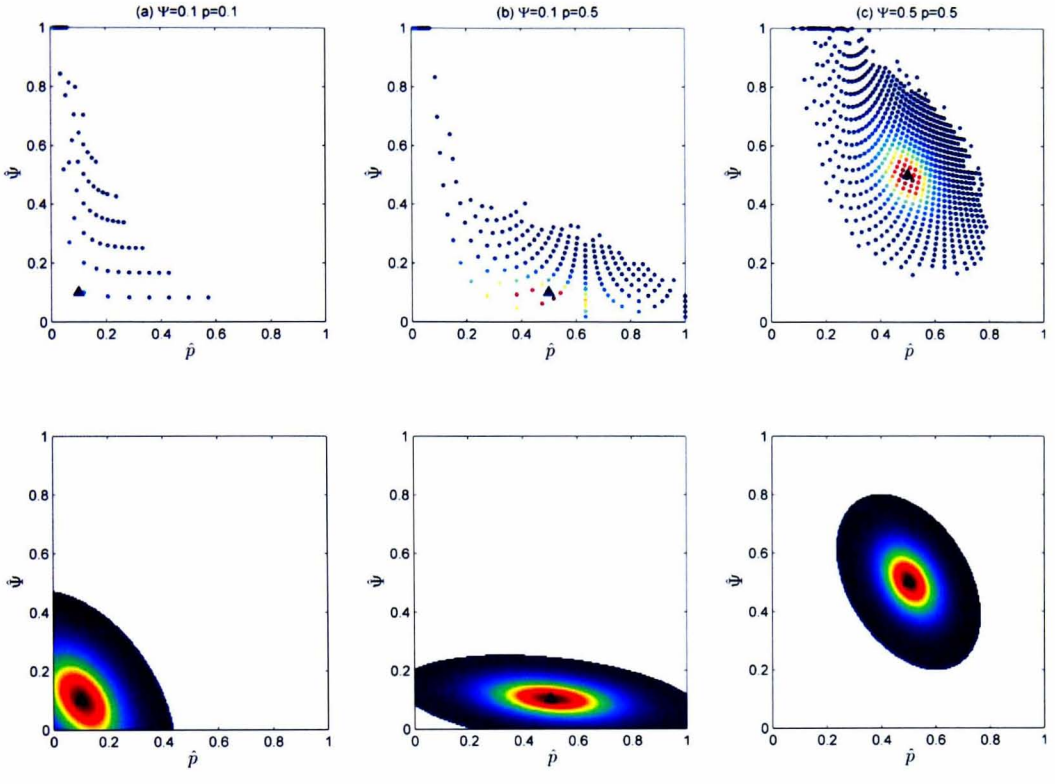


Figure 2-6 Actual and asymptotic distribution of the MLEs for different underlying probabilities of occupancy and detection (marked with a triangle) under a design with 168 units of total effort: (a) $S = 12$ sites and $K = 14$ replicates; (b) and (c) $S = 56$ sites and $K = 3$ replicates. The combination of S and K chosen for each case is the optimal design in terms of minimizing the variance of the occupancy estimator. Plots show part of the distribution that contains 0.999 probability. In the top row, no point that is excluded has higher probability that any of the points displayed (based on 10^6 simulated data sets).

As a means to reduce the bias in the occupancy estimator due to small sample size, Moreno & Lele (2010) propose the use of a penalized likelihood approach. They derive a heuristic penalty term which shrinks the occupancy MLE towards the naïve occupancy estimator when the sample size is small. In a simulation study they show that the mean and median of the maximum penalized likelihood estimator (MPLE) tends to

be closer to the true occupancy value than the MLE's, except when the probability of occupancy is large, when the MPLE can be biased.

2.1.4 Introducing covariates

The assumptions of constant occupancy and detectability are often not adequate. Commonly these quantities vary with diverse factors, some of which may be recorded in the field. The model structure discussed so far can be modified to incorporate covariates to describe how occupancy and detectability change with these factors, for instance following a generalized linear model framework. As both quantities are probabilities the logit function, which is bounded between 0 and 1, is an appropriate link function, and is commonly used.

The probability of occupancy can vary with site characteristics such as habitat type, elevation, climatic conditions or distance to some focal point (e.g. water source or human settlement). In fact, often the primary objective of the study is to assess these potential relationships. Using the logit link function, the probability of occupancy is expressed as a function of site-specific covariates as

$$\text{logit}(\psi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} \tag{2.14}$$

where $\beta_0 \ldots \beta_n$ are the regression coefficients and $x_{1i} \ldots x_{ni}$ are the values of the n covariates for site i. Site-specific covariates can vary from site to site, but remain constant within the sampling season. Under the assumption of closure, the occupancy status of sites is supposed to remain constant within the sampling season and, therefore, only site-specific covariates are appropriate for modelling occupancy probability.

The probability of detecting the species at an occupied site can also be modelled as a function of covariates. Here there are two types of covariates that may be considered: (i) site-specific and (ii) survey-specific. Examples of survey-specific covariates include time of day, weather conditions or observer skills. The probability of detecting the species at site i during survey j can be expressed as

$$\text{logit}(p_{ij}) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_n x_{ni} + \alpha_{n+1} y_{1ij} + \cdots + \alpha_{n+m} y_{mij} \qquad (2.15)$$

where $\alpha_0 \ldots \alpha_{n+m}$ are the regression coefficients, $x_{1i} \ldots x_{ni}$ are the values of the n site-specific covariates for site i and $y_{1ij} \ldots y_{mij}$ are the values of the m survey-specific covariates for site i and survey j. Note that, although for convenience we have used the same notation for the site-specific covariates in (2.14) and (2.15), these do not need to be the same. Obviously, when survey-specific covariates are incorporated, the data can no longer be summarized by the number of detection at each site, d_i, as in (2.1).

By using a logit link function for site occupancy, the modelling framework under consideration can be interpreted as a generalization of logistic regression analysis to account for uncertainty on the true or false nature of recorded absences. Under perfect detection (i.e. $p = 1$), the model for site occupancy reduces to the standard logistic regression.

2.2 Model extensions and related models

2.2.1 Overview

Since the publication of the basic occupancy model, several extensions have been proposed to relax some of its assumptions. Figure 2-7 presents an overview of the most relevant model extensions together with other related models also developed to analyze detection data of unmarked individuals, indicating the journal articles in which they were presented. The diagram includes the models developed in this thesis, thus setting them in the context of the relevant literature. In the diagram, model extensions are connected by solid lines, while dashed lines are used to highlight relationships between models. Different colours are used to indicate the type of data described by the models. This way we distinguish between the models that rely on discrete detection/non-detection data (in blue), and those developed in this thesis for detection data collected on a continuous interval of time/space (in orange). The diagram also includes models for data on repeated counts (displayed in green). Although these models are essentially different in that they use a different kind of data, and focus on the estimation of abundance, we will see how they are closely related to the work developed in this thesis.

In the next sections we provide some detail regarding those models in the diagram (marked with a star) that are particularly relevant to the work carried out in this thesis. Note that, while the diagram summarizes models for detection data of unmarked individuals, this summary is not exhaustive. For instance, the whole class of distance sampling techniques are excluded, as these are less directly related to our work.

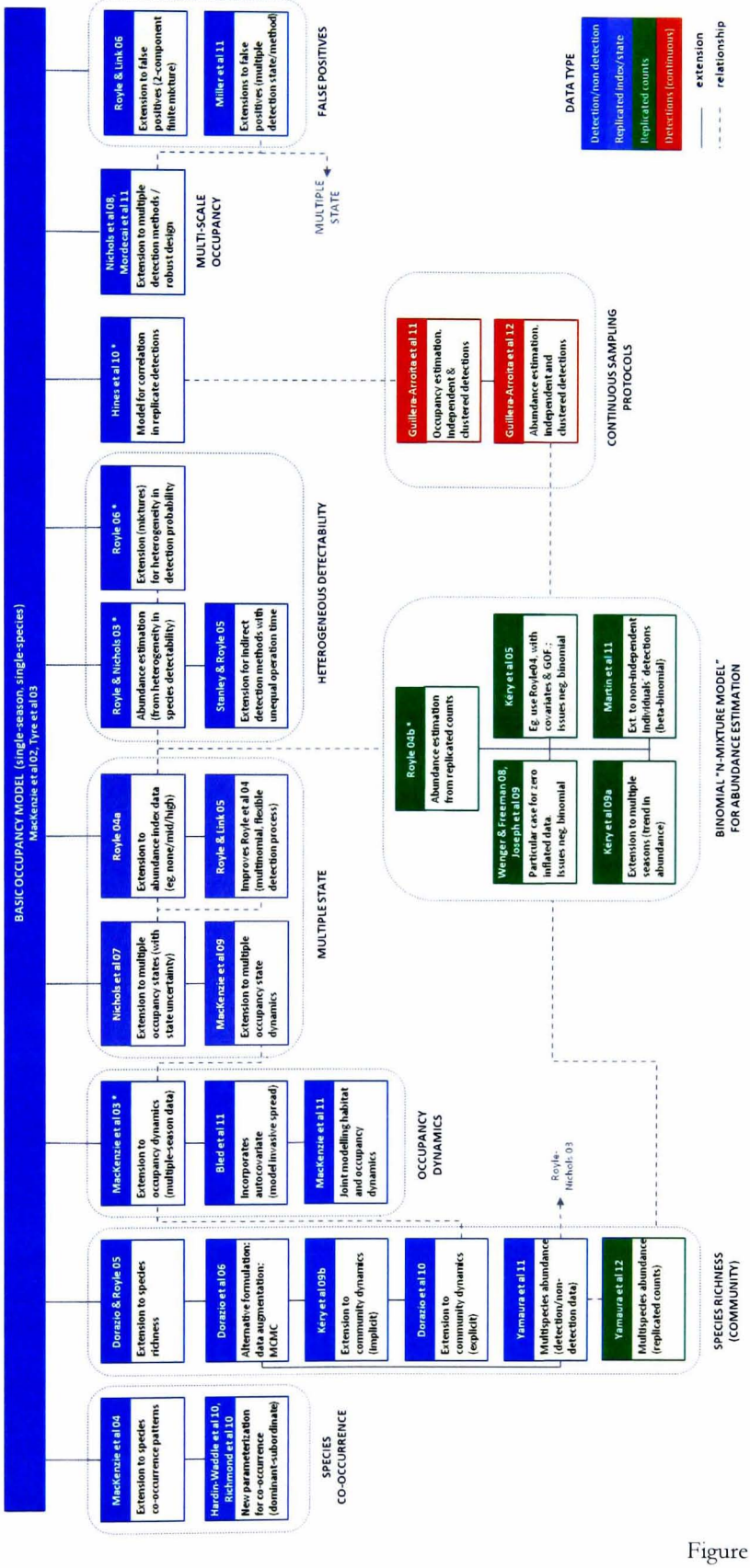

MODELS FOR DETECTION DATA OF UNMARKED INDIVIDUALS ACCOUNTING FOR IMPERFECT DETECTION
BASIC OCCUPANCY MODEL (single-season, single-species) Mackenzie et al 02, Tyre et al 03
SPECIES CO-OCCURRENCE
Mackenzie et al 04
Extension to species co-occurrence patterns
Hardin-Weddle et al 10, Richmond et al 10
New parameterization for co-occurrence (dominant-subordinate)
SPECIES RICHNESS (COMMUNITY)
Dorazio & Royle 05
Extension to species richness
Dorazio et al 06
Alternative formulation: data augmentation: MCMC
Kery et al 09b
Extension to community dynamics (implicit)
Dorazio et al 10
Extension to community dynamics (explicit)
Yamaura et al 11
Multispecies abundance (detection/non-detection data)
Yamaura et al 12
Multispecies abundance (replicated counts)
OCCUPANCY DYNAMICS
Royle-Nichols 03
Mackenzie et al 03 *
Extension to occupancy dynamics (multiple-season data)
Bled et al 11
Incorporates autocovariate (model invasive spread)
Mackenzie et al 11
Joint modelling habitat and occupancy dynamics
MULTIPLE STATE
Nichols et al 07
Extension to multiple occupancy states (with state uncertainty)
Mackenzie et al 09
Extension to multiple occupancy state dynamics
Royle 04a
Extension to abundance index data (eg. none/mid/high)
Royle & Link 05
Improves Royle et al 04 (multinomial, flexible detection process)
HETEROGENEOUS DETECTABILITY
Royle & Nichols 03 *
Abundance estimation (from heterogeneity in species detectability)
Stanley & Royle 05
Extension for indirect detection methods with unequal operation time
Royle 06 *
Extension (mixtures) for heterogeneity in detection probability
BINOMIAL "N-MIXTURE MODEL" FOR ABUNDANCE ESTIMATION
Royle 04b *
Abundance estimation from replicated counts
Wenger & Freeman 08, Joseph et al 09
Particular case for zero inflated data. Issues neg. binomial
Kery et al 05
Eg. use Royle04 with covariates & GOF.. Issues neg. binomial
Kery et al 09a
Extension to multiple seasons (trend in abundance)
Martin et al 11
Ext. to non-independent individuals' detections (beta-binomial)
CONTINUOUS SAMPLING PROTOCOLS
Guillera-Arroita et al 11
Occupancy estimation. Independent & clustered detections
Guillera-Arroita et al 12
Abundance estimation. Independent and clustered detections
Hines et al 10 *
Model for correlation in replicate detections
MULTI-SCALE OCCUPANCY
Nichols et al 08, Mordecai et al 11
Extension to multiple detection methods / robust design
MULTIPLE STATE
FALSE POSITIVES
Royle & Link 06
Extension to false positives (2-component finite mixture)
Miller et al 11
Extensions to false positives (multiple detection state/method)
DATA TYPE
Detection/non detection
Replicated index/state
Replicated counts
Detections (continuous)
extension
relationship

Figure 2-7

2.2.2 Clustered detections within sites

As noted in section 2.1.2, the basic occupancy model makes a 'closure assumption' that the occupancy status of each site is constant across replicates. When the sampling scheme is based on temporal replication (i.e. assessing the same location at different points in time), this requires site occupancy status to remain unchanged throughout the sampling season. For the closure assumption to hold when data are collected by spatial sampling within sites (Figure 2-1), all subunits in sites occupied by the species should be occupied. However this is not necessarily always true and the species may occupy the sites partially. We discuss in more detail issues related to the 'closure assumption' in section 3.4, but a key idea is that, if changes in availability among survey replicates occur completely at random, no bias is induced in the site occupancy estimator $(\hat{\psi})$. The occupancy estimator now reflects the probability that *part* of the site is occupied by the species. A similar situation arises for surveys based on signs if, at the time of survey, signs only occupy the sites partially (even if sites are fully occupied by the species). There has to be independence in the availability of signs for detection among survey replicates for the occupancy estimator to be unbiased.

However, the assumption of random changes in species availability among survey replicates (which leads to independent detections as assumed by the model) is not always appropriate. For instance, when spatial replicates are drawn from partially occupied sites, dependence may be induced by the way spatial subunits are chosen spatially, as in some scenarios the occupancy status of subunits that are close will tend to be correlated. This is something to consider when designing a study, however sometimes sampling designs that suffer from lack of replicate independence are preferred, e.g. due to logistics. This situation arises for instance in surveys carried out along transects, in

which replicates are defined as adjacent transect segments, such as in the Sumatran tiger data set. To account for the resulting dependence in the analysis of such detection data, Hines *et al.* (2010) propose a refinement of the basic occupancy model to incorporate first-order Markovian dependence between adjacent replicates. They applied this model to a tiger sign data set from India and showed that disregarding the dependence between consecutive replicates induces a negative bias in the occupancy estimator.

The model they present, to which they refer as the 'Markov process for segment occupancy model', defines the following two parameters for the probability that the species is available for detection at a survey replicate (i.e. segment)

(i) θ: probability that the species is available for detection at a survey replicate given it *was not* available for detection at the previous replicate,

(ii) θ': probability that the species is available for detection at a survey replicate given it *was* available for detection at the previous replicate.

The remaining two parameters of the model are: ψ, the probability of site occupancy, and p, the probability of detection given the species is present and available for detection at a survey replicate (denoted p_a in section 3.4). The detection process at occupied sites is described as a Hidden Markov model (Figure 2-8). If θ and θ' differ from each other, clustering is induced in the detections. It is reasonable to expect that $\theta' > \theta$, which would imply that it is more likely for the species to be available for detection at a survey replicate if it was available at the previous one.

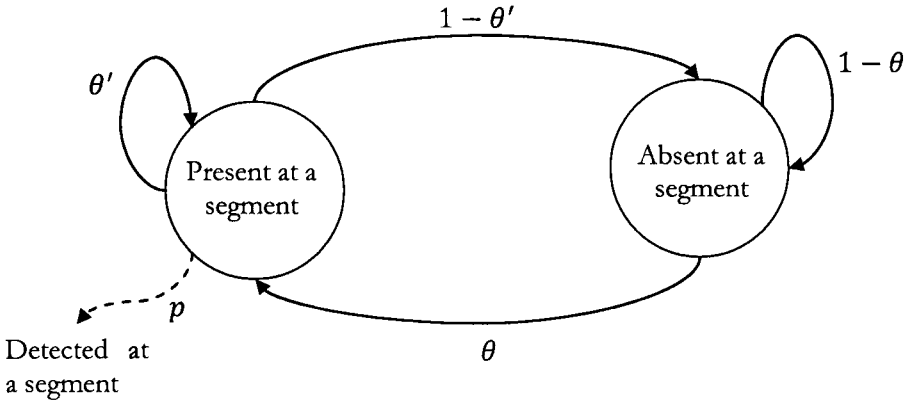


Figure 2-8 Hidden Markov chain describing the detection process at occupied sites for the 'Markov process for segment occupancy model'

It is only when there is Markovian dependence in the occupancy status of survey replicates (i.e. $\theta \neq \theta'$) that the three processes (site occupancy, availability for detection at a replicate and detectability at replicates where the species is available) become identifiable from data obtained following the standard sampling protocol (we discuss this further in section 3.4).

Note that, while the motivating scenario for the development of this model was a survey based on spatial replication (i.e. transect segments), the approach is also applicable to sampling situations based on temporal replication. In such studies, lack of independence can also be an issue if replicate surveys are carried out close in time with respect to the movement patterns of the species.

The likelihood for this model can be written as

$$L(\psi, \theta, \theta', p) = \prod_{i=1}^{S} \left[\psi \boldsymbol{\pi} \left\{ \prod_{t=1}^{K-1} \mathrm{diag}(\boldsymbol{P}_t) \boldsymbol{\Phi} \right\} \boldsymbol{P}_K + (1 - \psi) \mathrm{I}(d_i = 0) \right], \qquad (2.16)$$

where $\boldsymbol{\pi}$ is the vector of initial replicate occupancy, $\boldsymbol{P}_t = [0 \;\; p]$ if the replicate t resulted in detection and $[1 \;\; 1 - p]$ otherwise, $\mathrm{diag}(\boldsymbol{P}_t)$ is the matrix that has the elements of $\boldsymbol{P}_t$ as diagonal and $\boldsymbol{\Phi}$ is the transition matrix of the Markov chain given by

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 - \theta & \theta \\ 1 - \theta' & \theta' \end{bmatrix}.$$

While (2.16) assumes that the four parameters are constant, the model can be readily extended to accommodate site covariates in any of the parameters, and survey covariates in θ, θ' and p.

Worried about potential unidentifiability of the parameters, Hines *et al.* (2010) developed a second model as an approximation to the process generating the data, which they call the 'trap response model'. In this case the assumption is that the species is present in all survey replicates but that there is dependence in the detection from adjacent replicates. The model defines the following two parameters for the probability of detecting the species at a replicate

(i) p: probability that the species is detected at a survey replicate given it was not detected at the previous replicate,

(ii) p': probability that the species is detected at a survey replicate given it was detected at the previous replicate.

The detection process at occupied sites in this model is described as a Markov chain (Figure 2-9). While this model was proposed as an approximate description for the scenario of interest, in which the species is actually not available for detection at all replicates, Hines *et al.* (2010) point out that this model can be directly useful for other sampling situations.

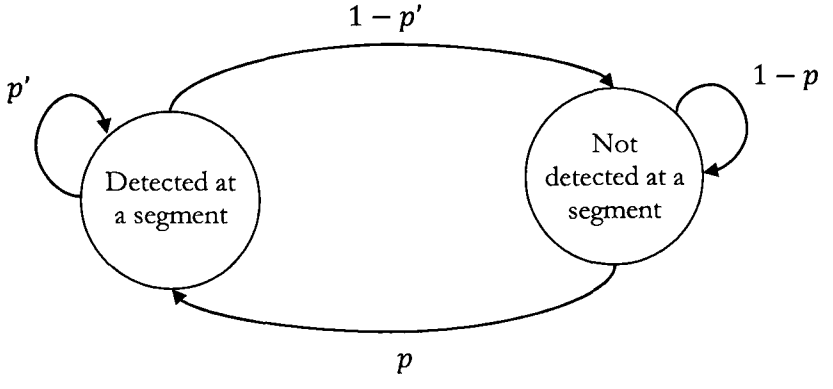


Figure 2-9 Markov chain describing the detection process at occupied sites for the 'trap response model'

The interpretation of both the 'Markov process for segment occupancy model' and the 'trap response model' as particular cases of a more general two-state Markov-modulated Bernoulli process model (2-MMBP) is discussed in section 4.2.2.

2.2.3 *Heterogeneity in detection probability*

Heterogeneity in detection probability can induce bias in the estimator of occupancy (Royle & Nichols 2003; MacKenzie *et al.* 2006; Royle 2006), with greater bias for higher levels of heterogeneity, low detection probability and small sample sizes. Minimizing heterogeneity through proper study design, plus collecting and incorporating relevant covariates into the model, is essential to obtain reliable occupancy estimates. However, sometimes some degree of heterogeneity remains which may be difficult or practically impossible to capture through the use of covariates. In such circumstances, mixture models can be used to account for unexplained variation in detection probability across sampling sites (Royle 2006). Using a finite (discrete) mixture on detection probability, the model in (2.1) is extended as

$$L(\psi, \boldsymbol{\theta}|h) = \prod_{i=1}^{S}\left[\psi \sum_{m=1}^{M}\{p_m^{d_i}(1-p_m)^{K_i-d_i}\Pr(p_m|\boldsymbol{\theta})\} + (1-\psi)I(d_i=0)\right], \tag{2.17}$$

where p_m are each of the M possible detectability values and $\Pr(p_m|\boldsymbol{\theta})$ is the probability mass function of a discrete distribution with parameters $\boldsymbol{\theta}$. Similarly, using a continuous mixture for detectability we have

$$L(\psi, \boldsymbol{\theta}|h) = \prod_{i=1}^{S}\left[\psi \int_0^1 \{p^{d_i}(1-p)^{K_i-d_i} f(p|\boldsymbol{\theta})\}dp + (1-\psi)I(d_i=0)\right], \tag{2.18}$$

where $f(p|\boldsymbol{\theta})$ is the probability density function of a continuous distribution. In general, maximizing (2.18) to obtain parameter estimates involves evaluating the integral, which can always be done numerically. However, a convenient choice for $f(p|\boldsymbol{\theta})$ that avoids the need for numerical integration is the beta distribution. Since a beta mixture of binomials results in a beta-binomial distribution (Johnson, Kemp & Kotz 2005, p. 374), the model in (2.18) results in a zero-inflated beta-binomial which has an explicit closed expression for its likelihood

$$
\begin{aligned}
&L(\psi, \alpha, \beta|h) \\
&= \prod_{i=1}^{S} \left[\psi \frac{\Gamma(K_i+1)}{\Gamma(d_i+1)\Gamma(K_i-d_i+1)} \frac{\Gamma(\alpha+d_i)\Gamma(K_i+\beta-d_i)}{\Gamma(\alpha+\beta+K_i)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right. \\
&\left. + (1-\psi) I(d_i=0) \right],
\end{aligned}
\tag{2.19}
$$

where α and β are the parameters of the beta mixing distribution and with $\Gamma(\cdot)$ the Gamma function.

Royle (2006) shows that in occupancy models it might not be possible to distinguish alternative mixture distributions from the data, a result in line with the findings of Link (2003) in the context of closed population mixture models for estimating population size. This 'identifiability problem' implies that models leading to considerably different occupancy estimates might receive identical support from the data. How large an effect the misspecification of the heterogeneity model has is related to the amount of heterogeneity and the mean detection probability. Royle (2006) cautions that cases with high levels of heterogeneity and low detection probabilities could be considered as situations in which occupancy cannot be reliably estimated.

2.2.4 *Abundance-induced heterogeneity in detection probability*

Differences in site abundance can be the main source of heterogeneity in species detectability in occupancy studies. For most systems it is reasonable to expect that the more individuals at a site, the easier it is to detect the species. In order to address this type of heterogeneity, Royle & Nichols (2003) propose a mixture model in which local abundance is the source of heterogeneity in detectability.

The key to the model is to consider that the species is detected at a site during a particular survey unless all the individuals present are missed. Therefore, assuming that all the individuals are equally and independently detectable during the survey, the probability of detecting the species at a site i in which n_i individuals are present is

$$p_i = 1 - (1 - r)^{n_i}, \tag{2.20}$$

where r is the probability of detecting an individual. Naturally site abundance (n_i) is unknown, so to account for its variability a mixture is used and inference is made on the parameters of the mixing distribution. Assuming that the number of individuals at a site remains constant during the whole survey period, the likelihood for the resulting model can be written as

$$L(r, \boldsymbol{\theta}|h) = \prod_{i=1}^{S} \left[\sum_{n_i=0}^{\infty} \{p_i^{d_i}(1 - p_i)^{K_i - d_i} \Pr(n_i|\boldsymbol{\theta})\} \right], \tag{2.21}$$

where p_i is a function of individual detectability and local abundance as indicated in (2.20) and $\Pr(n_i|\boldsymbol{\theta})$ is the mixing distribution describing local abundance with parameters $\boldsymbol{\theta}$. While the basic occupancy model assumes that the system is closed to oc-

cupancy changes, here the assumption is that the system is closed to changes in local abundance.

A natural choice for $\Pr(n_i|\boldsymbol{\theta})$ is the Poisson distribution which leads to the likelihood

$$L(r,\lambda|h) = \prod_{i=1}^{S}\left[\sum_{n_i=0}^{\infty}\left\{p_i^{d_i}(1-p_i)^{K_i-d_i}\frac{e^{-\lambda}\lambda^{n_i}}{n_i!}\right\}\right], \tag{2.22}$$

where λ is the parameter of the Poisson distribution. Although for simplicity here the parameters are assumed constant, the model can accommodate covariates. Site and survey covariates can be used to describe individual detectability r. Site covariates can also be used to describe mean local abundance λ. Covariates can be incorporated following a generalized linear model framework, as in section 2.1.4, via a logit link function for r and a log link function for λ.

The Poisson distribution assumes that individuals occur completely at random, that is, whether an individual occurs at a site or not is independent of whether others are present. This assumption can be relaxed using other mixture distributions such as the negative-binomial, which allows for overdispersion with respect to the Poisson, although it has been noted that this model is often difficult to fit (Royle & Nichols 2003). Allowing for zero-inflation has also been shown to be relevant for some data sets (Wenger & Freeman 2008). Underdispersion could be dealt with using a weighted Poisson distribution (e.g. Ridout & Besbeas 2004). Non-parametric finite mixtures are possible as well. In fact the basic occupancy model can be seen as a particular case of the abundance model with a finite mixing distribution for site abundance with only two support points, one of them at $n_i = 0$, with $\Pr(n_i = 0) = 1 - \psi$ and $\Pr(n_i =$

$n) = \psi$. While non-parametric distributions can be useful in some cases due to their flexibility, a key benefit of using a parametric distribution for abundance is the reduction in the number of parameters to be estimated, as well as the possibility to accommodate covariates.

Subject to the assumptions of the model being met, the mean of the estimated mixing distribution (e.g. λ for the Poisson) may be interpreted as an estimate of average site abundance. Although site occupancy is not a formal parameter in the formulation of the abundance model it can be immediately derived as the probability of having at least one individual given the abundance distribution (e.g. $\psi = \Pr(n_i > 0) = 1 - e^{-\lambda}$, if a Poisson is used).

Of course, the ability of the model to estimate abundance depends on whether species detectability and site abundance are indeed linked. For instance, it is obvious that inference about abundance is not possible from species detection/non-detection data if the species is so abundant that species detection at occupied sites is almost certain. It is also reasonable to expect that the extent to which the estimated abundance provides a good reflection of the actual abundance depends on whether there are other sources of heterogeneity in detectability that are not accounted for.

Royle (2006) notes that n_i does not necessarily need to be interpreted as site abundance and that this model can also be merely seen as an alternative way to accommodate heterogeneity in detection probability, with n_i being a generic random effect that induces variation in p.

2.2.5 *Abundance model for repeated counts: the binomial 'N-mixture model'*

Rather than just recording the detection/non-detection of the species, sometimes counts of detected individuals can be recorded in each of the replicate surveys of sampling sites (Figure 2-10). Given that counts contain more information than simple detection/non-detection records, we can expect abundance to be estimated more reliably from such data. For this purpose, Royle (2004b) proposes a model that describes the counts recorded at site i, c_{ij}, as binomially distributed $c_{ij} \sim \text{Bin}(n_i, r)$. As in the Royle-Nichols model, r represents the probability of detecting during a survey visit an individual present at the site and n_i is the number of individuals at the site, which is assumed to remain constant during the whole survey period. Since n_i is unknown, a mixture is used to account for its variability and inference is made on the parameters of the mixing distribution.

		replicates						
		1	2	3	4	5	...	*K*
site	1	0	1	0	2	1	...	0
	2	0	0	0	3	1	...	2
	3	0	0	0	0	0	...	1
	4	0	1	1	0	0	...	0
	5	0	0	0	4	0	...	0
	6	0	0	0	0	0	...	0
	7	0	0	0	0	1	...	0
	8	1	2	0	1	0	...	0
	9	0	0	0	0	0	...	0
	...	...	...	...	...	...	...	...
	S	0	2	1	2	1	...	0

Figure 2-10 Example of count history data set where S sites are surveyed K times

The likelihood can be written as

$$L(r, \boldsymbol{\theta}|h) = \prod_{i=1}^{S} \left(\sum_{n_i=\max(c_{ij})}^{\infty} \left[\left\{ \prod_{j=1}^{k} \binom{n_i}{c_{ij}} r^{c_{ij}} (1-r)^{n_i - c_{ij}} \right\} \Pr(n_i|\boldsymbol{\theta}) \right] \right), \quad (2.23)$$

where $\Pr(n_i|\boldsymbol{\theta})$ is the mixing distribution describing local abundance with parameters $\boldsymbol{\theta}$. Note that, by using a binomial distribution to describe the counting process, the model implicitly assumes that no individual is counted more than once within each survey visit. So, although we treat this as a model for detection data of unmarked individuals (e.g. in Figure 2-7), the model in fact implies that somehow the surveyor can distinguish individuals within a single survey (for instance due to the characteristics of the survey itself).

In the literature this model is often simply referred to as the 'N-mixture model'. This terminology highlights the fact that the mixing component is based on site abundance, however it can be confusing. For instance, the Royle-Nichols model is also based on such a mixture. In this thesis when referring to this 'N-mixture model' we indicate the type of data (i.e. 'repeated counts', c.f. detection/non-detection in Royle-Nichols). In places we also use the term 'binomial N-mixture model', to distinguish it from models we propose which use a different detection process.

2.2.6 Multiple-season occupancy model

Often, rather that estimating occupancy at a given point in time, there is interest in assessing how occupancy changes over time. In this case, the sampling protocol involves collecting detection/non-detection data at a number of sampling sites over several sampling seasons, with repeat surveys being carried out within each season. The de-

sign therefore involves two levels of replication with primary sampling periods (i.e. seasons) and secondary sampling periods nested within each primary period (i.e. repeat surveys).

One option for the analysis of multiple season data is to fit the basic (single-season) occupancy model in section 2.1 separately to each season's data set. In practice this implies an assumption of independence in the occupancy status of each site between seasons. However, depending on the species and time scale of the survey, there might be dependence in the occupancy status of sites between seasons. Multiple-season data can be analyzed by explicitly modelling the mechanisms underlying occupancy dynamics as a first-order Markov chain (MacKenzie *et al.* 2003). Each state of the two-state Markov chain represents the occupancy status of a given site, occupied or empty (Figure 2-11). A key assumption of the model is that the system is closed to changes in site occupancy within seasons but occupancy is allowed to change between seasons. Transitions between states are governed by the probabilities of colonization γ and local extinction ε, and only depend on the occupancy status at the previous time step. Since, due to imperfect detection, the state 'occupied' in the Markov chain is not perfectly observed, the model is a Hidden Markov model. As in the basic single-season occupancy model, detection/non-detection data from repeat surveys at occupied sites are modelled as a series of independent Bernoulli trials with probability p.

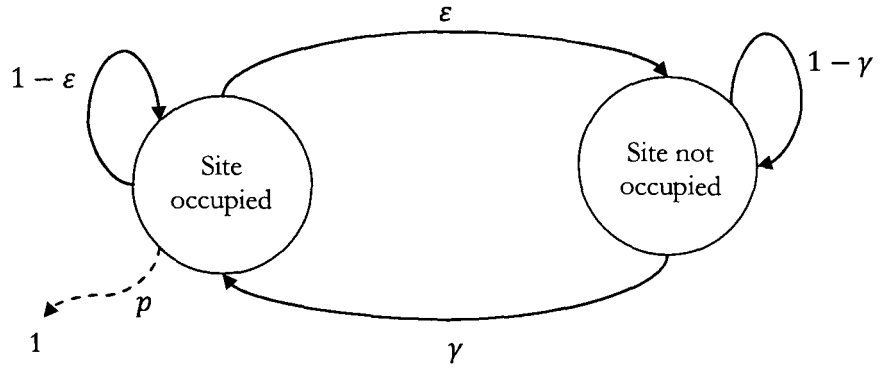


Figure 2-11 Hidden Markov chain in the multiple-season occupancy model.

The likelihood can be written as

$$L(\psi_1, \gamma, \varepsilon, p) = \prod_{i=1}^{S} \left[\boldsymbol{\phi}_0 \left\{ \prod_{t=1}^{T-1} \mathrm{diag}(\boldsymbol{p}_{h_i,t}) \boldsymbol{\phi}_t \right\} \boldsymbol{p}_{h_i,T} \right], \tag{2.24}$$

where S is the number of sampling sites, T is the number of seasons, ψ_1 is the initial occupancy probability, $\boldsymbol{\phi}_0 = [\psi_1 \quad 1 - \psi_1]$, $\boldsymbol{p}_{h_i,t}$ is a column vector with entries denoting the probability of observing the detection history $\boldsymbol{h}_{i,t}$ in site i and season t, conditional upon occupancy state, $\mathrm{diag}(\boldsymbol{p}_{h_i,t})$ is the diagonal matrix with the elements of $\boldsymbol{p}_{h_i,t}$ along its diagonal and $\boldsymbol{\phi}$ is the transition matrix of the Markov chain given by

$$\boldsymbol{\phi} = \begin{bmatrix} 1 - \varepsilon & \varepsilon \\ \gamma & 1 - \gamma \end{bmatrix}.$$

While (2.24) assumes that the parameters are constant, the model can be readily extended to accommodate site covariates in any of the parameters. Colonization and local extinction probabilities can also be expressed as a function of inter-season-specific covariates, and detection probability as a function of survey-specific characteristics.

2.3 Application example: analysis of Sumatran-wide tiger data set

As an illustration, here we present the application of the occupancy modelling framework discussed in this chapter to the analysis of the island-wide Sumatran tiger data set described in section 1.3.2.

2.3.1 Methods

To match the discrete sampling protocol assumed by the models used, in which a number of replicate surveys are conducted within each sampling site, a detection/non-detection history had been constructed by dividing transects into segments of 1 km in length, assigning '1' to those containing at least one detection and '0' otherwise. For the analysis, data were further collapsed into 5 km segments. This length was chosen to mitigate the dependence between consecutive replicates that, given tiger movement patterns, could be expected at smaller scales, but without compromising the results by the loss of data that would result when choosing a very coarse replicate length. To assess the robustness of the results to moderate changes in the definition of replicates, models were also run using different segment lengths (4 and 6 km).

Tiger detection/non-detection data were analyzed to estimate site occupancy ψ using the basic occupancy model and three of the extensions presented in section 2.2: the 'clustering model' (section 2.2.2), the 'beta-binomial model' (section 2.2.3) and the Royle-Nichols 'abundance model' (section 2.2.4). The 'clustering model' was considered relevant as lack of independence between detections in consecutive transect segments could potentially remain. The other two model types were included in the analysis to relax the assumption of constant detection probability across sites. Although the surveys followed a well-specified common sampling protocol, some heterogeneity in

species detectability could be expected given the large spatial scale of the survey and the fact that the survey period spanned over two years. For instance, different teams surveyed different areas and surveys were subject to different meteorological conditions, which could have an impact on how easily the species was detected. Most importantly tiger abundance is likely to change depending on site characteristics and therefore species detectability can be expected to vary, with more footprint detections at sites with more individuals. For the discrete mixture that describes species site abundance in the 'abundance model', a Poisson distribution and two of its generalizations to allow for zero-inflation and overdispersion (i.e. negative-binomial) were used.

We incorporated site covariates into the models to explore the potential influence of biophysical and anthropogenic threat factors on occupancy. Tiger site occupancy was expected to vary across Sumatra, given that the island has a diverse topography ranging from prey-rich lowlands to less productive rugged montane areas. Anthropogenic threats were also expected to be relevant predictors of tiger occupancy, with deforestation in particular being considered likely to be important. A set of nine potential explanatory variables was provided with the data set. These covariate data were extracted using ArcGIS v9.3 software (ESRI) from layers obtained from several sources (Table 2-1). Elevation, slope and distance covariates were extracted at a 30 m x 30 m resolution and a single value per site was obtained by averaging all the pixel values within each site.

Table 2-1 Set of candidate predictor variables considered for the analysis of the Sumatran tiger data set. Original data were obtained from the following sources: 1 – Digital Elevation Model from the Shuttle Radar Topography Mission (Rabus *et al.* 2003), 2 – Indonesian National Coordination Agency for Surveys and Mapping, 3 – Indonesian Ministry of Forestry, 4 – forest cover map (Uryu *et al.* 2010)

Covariate	Description
elevation[1]	average elevation of the site
slope[1]	average slope of the site
distance to roads[2]	average distance to nearest road
distance to settlement[2]	average distance to nearest settlement
protection status[3]	1: site mostly inside a protected area, 0: otherwise
distance to forest[4]	average distance to nearest forest patch
distance to forest edge[4]	average distance to forest edge from within the forest
forest cover[4]	percentage of the site covered by forest
deforestation[4]	forest area (ha) removed between 2000 and 2008

Candidate explanatory variables were standardized using a z-transformation. To assess for collinearity, Pearson correlations were calculated between the nine covariates. Two pairs of variables showed strong significant correlation: elevation and slope ($r = 0.80$, $p < 0.001$); forest cover and distance to forest ($r = -0.78$, $p < 0.001$). Covariates within each of these pairs were not included together within the same models.

The analysis was performed obtaining maximum-likelihood estimates by numerical maximization, using the R-package `RMark` 2.0.1 to run program MARK 6.1 for the basic occupancy and abundance (Poisson and negative binomial) models and our own MATLAB scripts for the clustering, beta-binomial and abundance (zero-inflated Pois-

son) models. Model selection was performed using AIC_c to compare model fit, with the effective sample size defined as the number of sampling sites. For the best model, individual site estimates (occupancy and abundance) were derived from the regression coefficient estimates ($\hat{\beta}$).

2.3.2 Results

Tiger footprints were detected in 206 out of 394 cells, which corresponds to a naïve occupancy estimate of 0.52. The model in the candidate set that best explained the observed data was an 'abundance model' based on a Poisson mixture distribution to describe site abundance and four covariates in the regression on its parameter λ: average distance to forest, elevation, recent deforestation and protected area status. This model had much stronger support than the constant model $\lambda(.)r(.)$ (ΔAIC_c = 60.7; Table 2-2) and was considerably better than the best competing model with one covariate less (model without protected area status; ΔAIC_c = 7.5). Adding one extra covariate only marginally improved model fit. The confidence interval for the corresponding regression coefficient (distance to road) included zero while the rest of the regression coefficients remained practically the same.

There was no support for zero-inflation in the abundance distribution. Allowing for zero-inflation in the best fitting model led to a 1.5-unit increase in AIC, with a low zero-inflation parameter estimate (0.06) and practically the same regression coefficients as before. Models that allowed for overdispersion either did not provide a better fit or failed to converge.

Table 2-2 Model selection for the Poisson abundance model: best model among those with a given number of parameters N_{par}. Replicate length 5 km.

Model	N_{par}	AIC	ΔAIC	$-2\mathcal{L}$
$\lambda(.)r(.)$	2	2187.9	60.7	2183.9
$\lambda(\text{dforest})r(.)$	3	2149.5	22.3	2143.5
$\lambda(\text{dforest} + \text{PA})r(.)$	4	2140.6	13.4	2132.5
$\lambda(\text{dforest} + \text{elev} + \text{defor})r(.)$	5	2134.7	7.5	2124.5
$\lambda(\text{dforest} + \text{PA} + \text{elev} + \text{defor})r(.)$	6	2127.2	0.0	2115.0
$\lambda(\text{dforest} + \text{PA} + \text{elev} + \text{defor} + \text{droad})r(.)$	7	2127.8	0.6	2113.5

Note - dforest: distance to forest, PA: protection status, elev: elevation, defor: deforestation and droad: distance to roads. $\mathcal{L}$ is the maximum value of the log-likelihood.

The regression coefficients for λ in the best model (Table 2-3) suggest that tiger site abundance, and therefore occupancy, was higher in habitat that was at lower elevation, closer to forest patches, with less recent forest clearance and within protected areas, results that are consistent with the initial expectations. The estimate and standard error (in brackets) for individual detection probability $\hat{r}$ was 0.13 (0.017) and the estimated $\hat{\lambda}$ averaged across the entire island was 1.5 (0.20). The averaged derived occupancy estimate was 0.72 (0.039). The maps in Figure 2-12 show the estimated $\hat{\lambda}$ and $\hat{\psi}$ for each of the sampling sites.

Table 2-3 Best model estimated regression coefficients and odds ratios for a one unit increase in each of the covariates with standard errors in brackets.

Covariate	Regression coefficient $\hat{\beta}$	Odds ratio $\widehat{OR} = e^{\hat{\beta}}$
elevation	-0.23 (0.073)	0.79 (0.058)
distance to forest	-0.63 (0.116)	0.53 (0.062)
deforestation	-0.28 (0.085)	0.76 (0.064)
protection status	0.39 (0.125)	1.48 (0.185)

The basic, clustering and beta-binomial models provided a considerably worse fit to the data than the abundance model. Twice the negative maximised log-likelihood for the corresponding saturated models (with all covariates in ψ) was respectively 50.0, 24.4 and 12.8 units worse than that for the best abundance model (which has four covariates). Therefore these models do not have support as good explanations for the observed data and there was no need to carry out covariate model selection within these model structures.

Moderate variations in the segment length used to define the spatial replicates did not lead to substantial changes in the results. The same model provided the best explanation for the data at 4 km and 6 km, and the support of the next highest ranked models remained consistent (Table 2-4).

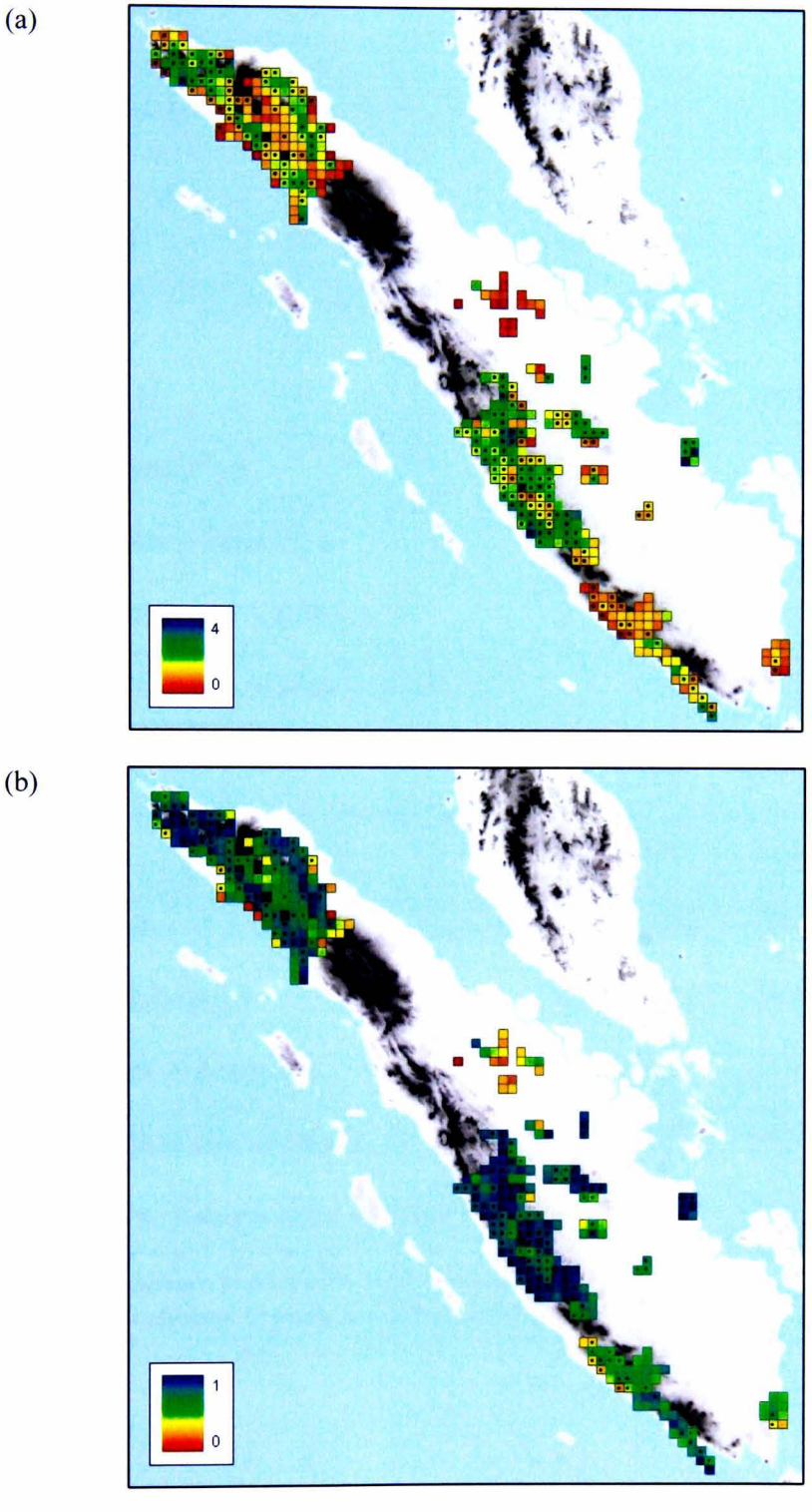


Figure 2-12 Site estimates for the best-fitting model: (a) estimated site average abundance and (b) derived site occupancy. Dots represent sites with detections.

Table 2-4 Model selection for the Poisson abundance model: best model among those with a given number of parameters N_{par}. Replicate length 4 km and 6 km.

Model (4 km)	N_{par}	AIC	ΔAIC	$-2\mathcal{L}$
$\lambda(.)r(.)$	2	2496.2	55.8	2492.2
λ(dforest)$r(.)$	3	2459.4	19.0	2453.3
λ(dforest + defor)$r(.)$	4	2452.3	11.9	2444.2
λ(dforest + elev + defor)$r(.)$	5	2445.5	5.1	2435.3
λ(dforest + PA + elev + defor)$r(.)$	6	2440.4	0.0	2428.2
λ(dforest + PA + elev + defor + droad)$r(.)$	7	2441.5	1.1	2427.2
Model (6 km)	N_{par}	AIC	ΔAIC	$-2\mathcal{L}$
$\lambda(.)r(.)$	2	1937.0	51.5	1933.0
λ(dforest)$r(.)$	3	1906.2	20.6	1900.1
λ(dforest + defor)$r(.)$	4	1896.3	10.8	1888.2
λ(dforest + PA + defor)$r(.)$	5	1890.7	5.2	1880.5
λ(dforest + PA + elev + defor)$r(.)$	6	1885.5	0.0	1873.3
λ(dforest + PA + elev + defor + droad)$r(.)$	7	1886.8	1.3	1872.5

Note - dforest: distance to forest, PA: protection status, elev: elevation, defor: deforestation and droad: distance to roads. $\mathcal{L}$ is the maximum value of the log-likelihood.

2.3.3 Goodness-of-fit

As a goodness-of-fit check we ran a test based on parametric bootstrapping. The test was run for the best ranking model using the function *parboot* in package `Unmarked` (version 0.9-3). This function simulates data sets according to the given model and, for each simulation, refits the model and evaluates a user-specified fit-statistic. Finally it compares the observed fit-statistic with the sampling distribution obtained from the simulations. As a fit-statistic we used three metrics of discrepancy:

(i) Sum of squared residuals: $m_1 = \sum_i \sum_j (O_{ij} - E_{ij})^2$,

(ii) Pearson's chi-square: $m_2 = \sum_i \sum_j (O_{ij} - E_{ij})^2 / E_{ij}$,

(iii) Freeman-Tukey chi-square: $m_3 = \sum_i \sum_j (\sqrt{O_{ij}} - \sqrt{E_{ij}})^2$,

where the O_{ij}'s are the observed data, each representing the outcome of survey j at site i, and E_{ij} are the expected probabilities of observing a '1' in survey j at site i, given the estimated parameters.

The results from running the test with 100 simulations did not indicate evidence of lack-of-fit with any of the three fit-statistics (Figure 2-13). The corresponding p-values were 0.69, 0.99 and 0.62.

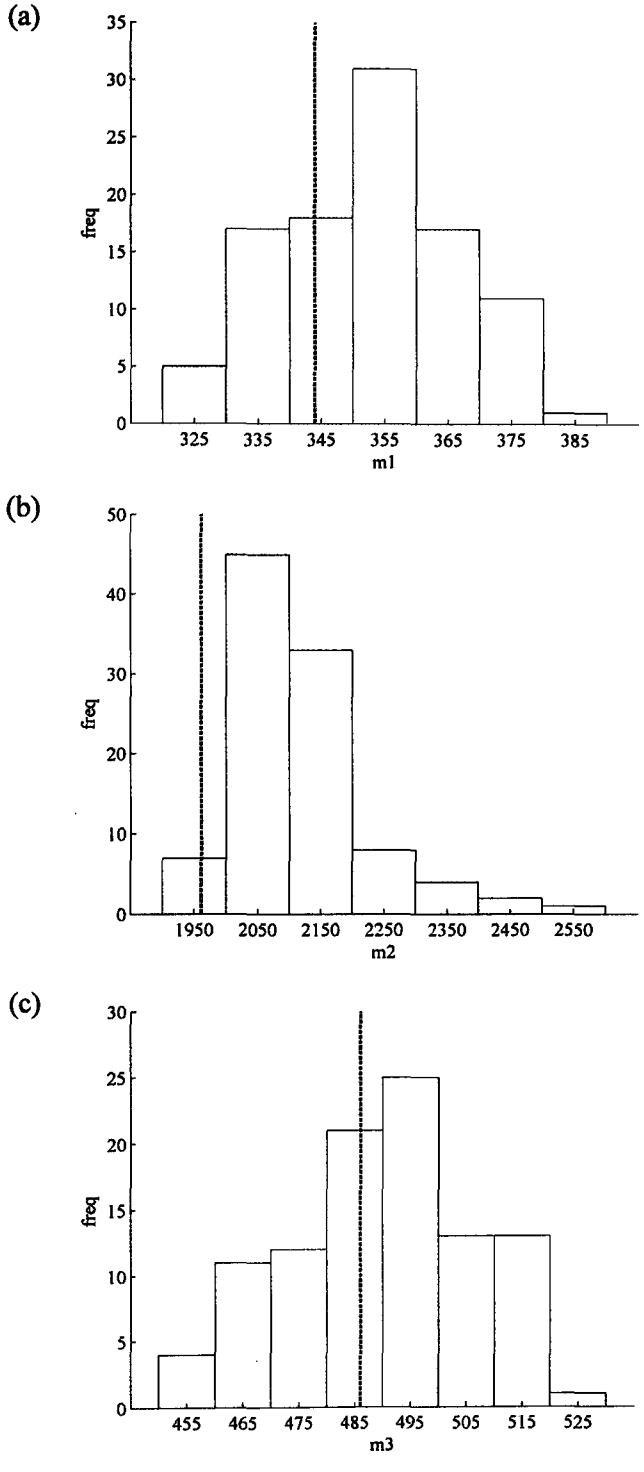


Figure 2-13 Model fit assessment by parametric bootstrapping based on three discrepancy metrics: (a) m_1: sum of squared residuals, (b) m_2: Pearson's chi-square, (c) m_3: Freeman-Tukey chi-square. The dotted line is the observed statistic. The histogram represents the sampling distribution based on 100 simulations.

2.3.4 Discussion

Understanding how underlying model assumptions are met is essential to the correct interpretation of the estimates obtained. The 'abundance model' that was ranked top in this analysis assumes that differences in abundance are the only source of heterogeneity in site-specific detection probabilities; otherwise bias may be induced in the estimators. In this study, a well-defined protocol was developed and implemented to minimize heterogeneity in detection probability. Nevertheless, some residual, unmodelled heterogeneity may still have remained, e.g. detecting footprints is easier in wetter substrates and surveys were conducted over both wet and dry seasons. The model also assumes that tiger site abundance closely fits a Poisson distribution. Tigers are territorial so it could be expected that their distribution exhibits some degree of underdispersion. Finally, as explained in section 2.2.4, the model is based on a functional dependence between species detectability p_i and the number of individuals n_i at a site of the form $p_i = 1 - (1 - r)^{n_i}$ where r is the individual detectability. This relationship implies that all individuals within each site are equally detectable at any replicate survey (i.e. transect segment here), that is, the system is closed to changes in abundance. In our survey, sites were larger than female tiger territories, which might have little overlap (Karanth & Nichols 2002). The assumption of equal detectability in any replicate would therefore be violated, as not all of the tiger territories in the site would overlap with each transect segment. However, a relationship of increased detection probability with increased abundance can still be expected, with greater tiger numbers resulting in an increase in the area within a site covered by their territories, where the species is detectable. When there are few individuals per site and low detectability, the binomial expansion of $(1 - r)^{n_i}$ shows that the relationship assumed by the model is

well approximated by a linear function, $p_i = rn_i$, which would be compatible with a scenario of site coverage proportional to abundance. Therefore, although we are cautious about interpreting our λ estimates as absolute numbers, we believe they provide a valuable tool to assess differences in relative tiger abundance across the landscape and their relationship to environmental and anthropogenic factors.

We ran a test to assess how our best ranking model fits the data. It is important to highlight here that goodness-of-fit is an area that has been explored very little to date in the context of occupancy modelling. To our knowledge, the only work that addresses this issue is MacKenzie and Bailey (2004), which explores the performance of a goodness-of-fit test for the basic occupancy model. The approach followed in their paper is also based on parametric bootstrapping but is different from the one used here in that it computes a fit-statistic from the observed and expected frequencies of encounter histories. One difficulty with this type of approach is that it can be difficult to implement when dealing with large histories and when missing values or continuous covariates are present. The method included in package `Unmarked`, which we use in our analysis, is much easier to implement. However there has not been any study formally evaluating the performance of this method and the various fit statistics that can be used for the different kinds of occupancy models, so this remains an area that requires further development.

3 STUDY DESIGN FOR THE BASIC OCCUPANCY MODEL

To ensure that studies provide meaningful results, and that therefore valuable monitoring resources are not wasted, it is critical to pay attention to survey design (Yoccoz, Nichols & Boulinier 2001; Legg & Nagy 2006). It is not only important to design the study so that what are considered biologically significant results can indeed be detected, but also to ensure that this is achieved in an efficient way.

In this chapter we examine various aspects related to study design for the basic occupancy model. The first three sections deal broadly with the same topic: the determination of how much survey effort is required and how this effort is best allocated, that is, how to choose the number of sampling sites S and the number of replicate surveys K. First, in section 3.1 we address study design assuming that the design target is set in terms of estimator quality. The main focus of this section is on survey effort trade-offs. We discuss optimal replication recommendations derived from large-sample properties of the estimators, which we extend to consider cases with detection probability as part of the design criteria. The need to use simulations for design when the sample size is small is also illustrated. This work has been published in Guillera-Arroita *et al.* (2010).

Second, in section 3.2 we address study design from the point of view of achieving a given power to detect a difference in occupancy between two samples, for instance changes between two points in time. The focus of the section is on determining the number of sites than need to be surveyed. We assume that the amount of replication is already decided upon, however we also explore how this choice affects the required sample size. An approximate formula for sample size determination is derived and its performance is assessed via simulations. The performance of alternative significance tests is also explored in this context. This work has been accepted for publication in *Methods in Ecology and Evolution.*

Third, in section 3.3 we revisit the work in section 3.1 formally acknowledging that the initial information on the parameter estimates, which is required as input for study design, is usually very uncertain. We discuss Bayesian and sequential design ideas, methods that provide more robust designs in the face of poor initial parameter estimates. In particular we evaluate the performance of a two-stage design. We show how such an approach can provide increased efficiency compared to a single-stage design and we explore how the optimal allocation of effort into the two stages changes depending on the prior parameter knowledge.

Finally, to complete the chapter, in section 3.4 we discuss a somewhat different design aspect. We review the assumption of closure made by the basic occupancy model and show that the general recommendation of sampling with replacement given by Kendall and White (2009) for occupancy studies based on spatial replication in which the species occupies the sites partially is not always adequate. This work has been published in Guillera-Arroita (2011).

3.1 Optimal replication

3.1.1 Background

Several papers have addressed the issue of study design in the context of occupancy modelling, and in particular the allocation of survey effort into number of sampling sites and amount of replication within sites. MacKenzie *et al.* (2002), Tyre *et al.* (2003) and Field *et al.* (2005) provide some guidance on the choice of number of replicate surveys based on simulations. MacKenzie and Royle (2005) present the first detailed investigation of this subject, deriving recommendations based on the analytic consideration of the large-sample properties of the occupancy estimator for different survey designs and cost functions. Bailey *et al.* (2007) describe a software tool (GENPRES) developed for exploring design trade-offs for different occupancy models, using either analytic approximations or simulations.

In this section we explore the issue of study design for the basic occupancy model when the design target is set in terms of estimator quality, focusing on survey effort allocation trade-offs. We discuss corresponding design recommendations based on large-sample properties of the estimators, extending existing guidelines for the case in which detection probability is a parameter of interest. We also provide an overview of the complete design procedure, illustrating the need to use simulations for design when sample size is small. Throughout this section (and likewise in sections 3.2 and 3.1) we assume a standard survey design and that all the individual surveys involve the same cost. Consequently our constraint when assessing optimality is total survey effort, that is, the product of the number of sampling sites and the number of replicate surveys per sites ($E = SK$).

3.1.2 Design recommendations based on asymptotic approximations

Design recommendations can be derived for a standard survey design using the asymptotic variance-covariance expressions in (2.13) which describe analytically how estimator precision changes with changes in design parameters. These expressions assume that both occupancy and detection probabilities are constant in time and space. Although in practice this simplification may not always be reasonable, it is convenient in order to provide general study design guidelines. Note also that these expressions are a function of ψ and p. Therefore in order to identify an optimal design it is necessary to assume some values for the parameters to be estimated.

In the sampling protocol under consideration there are different ways to allocate the total survey effort available: one may survey more sites with less replication per site or vice versa. For a fixed total effort E, increasing replication has two opposing effects on the variance of $\hat{\psi}$: (i) it decreases the additional term introduced to the binomial proportion variance due to imperfect detection, as the more replication the less likely it is to miss the species at occupied sites and the more precisely detection probability is estimated; (ii) it increases the binomial proportion variance due to the reduction in number of sampling sites. Therefore there is an optimal amount of replication to carry out with respect to the variance of the occupancy estimator. For instance, if $\psi = 0.5$ and $p = 0.4$ the variance is minimized if each site is surveyed four times (Figure 3-1).

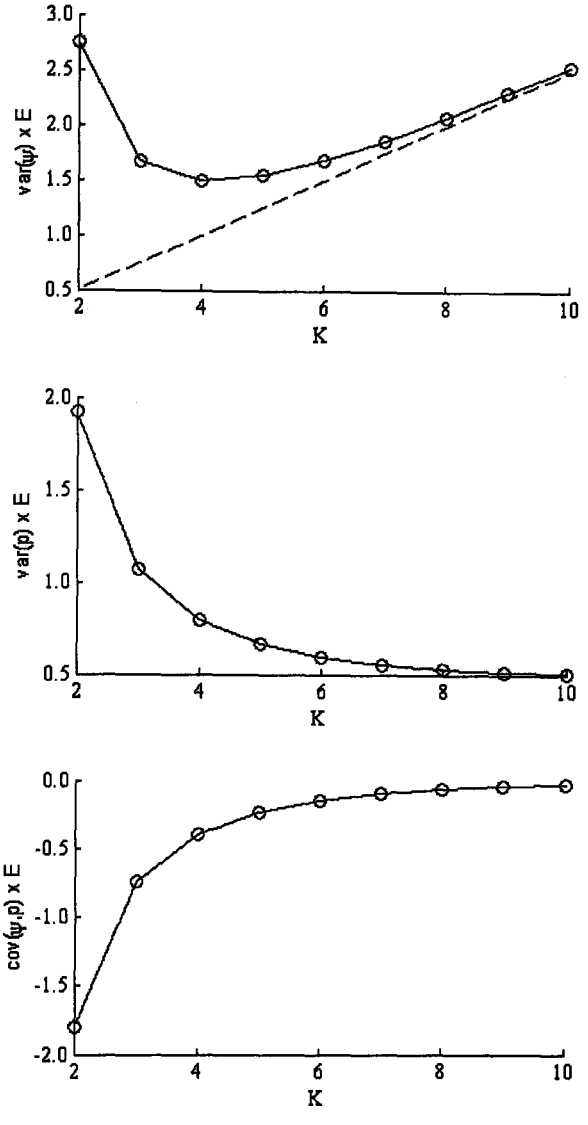


Figure 3-1 Asymptotic variance and covariance of the occupancy and detectability estimators as a function of the number of replicates K given a fixed effort $E = SK$, $\psi = 0.5$ and $p = 0.4$. The dashed line in the first plot shows the variance of the occupancy estimator under perfect species detection ($p = 1$).

The optimal replication results are independent of the total effort employed (assuming the sample size is large). This can be seen immediately by substituting $S = E/K$ in (2.13), as E only scales the variance-covariance expressions

$$\begin{aligned}
\mathrm{var}(\hat{\psi}) &= \frac{1}{E}\psi K\left\{(1-\psi)+\frac{1-p^*}{p^*-Kp(1-p)^{K-1}}\right\},\\
\mathrm{var}(\hat{p}) &= \frac{1}{E}\frac{p(1-p)}{\psi}\left\{\frac{p^*}{p^*-Kp(1-p)^{K-1}}\right\},\\
\mathrm{cov}(\hat{\psi},\hat{p}) &= -\frac{1}{E}pK\left\{\frac{1-p^*}{p^*-Kp(1-p)^{K-1}}\right\}.
\end{aligned} \tag{3.1}$$

MacKenzie and Royle (2005) provide a table with the optimal number of replicate surveys to carry out per site to minimize the variance of the occupancy estimator, as a function of the assumed values for ψ and p (Table 3-1a in page 79). Some general observations can be drawn from these results (also illustrated by Figure 3-2). On the one hand, the higher the detection probability the lower the optimal replication, which is expected as fewer visits are necessary to establish with reasonable certainty whether sites are occupied or not. When p is high the optimal design is always $K = 2$ and any effort invested in extra replication ($K > 2$) is 'wasted' with the variance increasing as S is reduced, as dictated by the variance of a binomial proportion. On the other hand, the higher the occupancy probability the higher the optimal replication is. The extra variance due to imperfect detection has more impact for high occupancy probabilities (Figure 2-5) and therefore it is more relevant to have replication in these cases. This leads to the general recommendation of surveying more sites less intensively for rare species and fewer sites more intensively for common species. It is useful when considering the results in Table 3-1 to realize that, although the optimal amount of replication is quite high when occupancy is high and detectability is low, the shape of the variance curve is in these cases rather flat around the maximum, so moderate departures from the optimal replication are not very critical (Figure 3-2). It is also worth noting that, in general, using less replication than the optimal has more impact than vice versa.

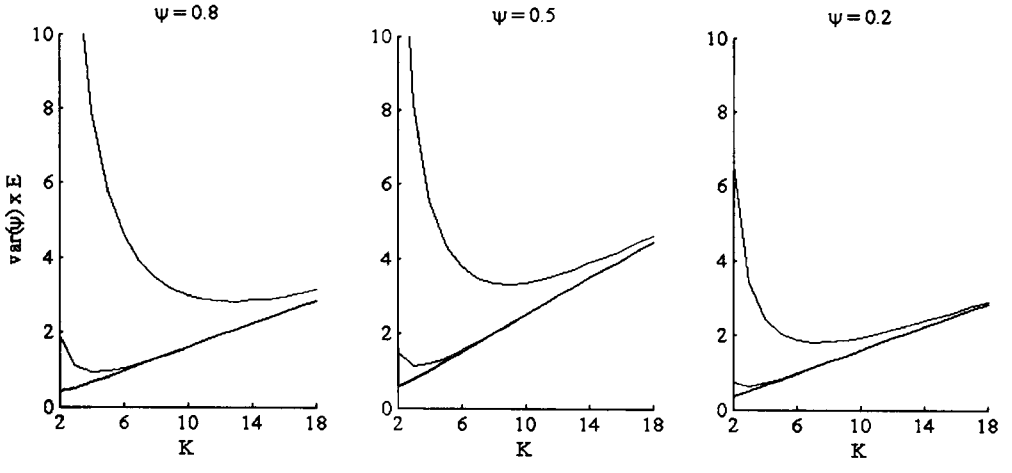


Figure 3-2 Asymptotic variance of the occupancy estimator as a function of the number of replicates K given a fixed effort E, and different levels of occupancy and detection probability (black p = 0.8, green p = 0.5, red p = 0.2).

Recommendations can also be produced after incorporating the variance of $\hat{p}$ as part of the design criterion, which is useful when detectability is itself a parameter of interest. While for many studies the primary object of inference is the probability of occupancy, with the probability of detection being regarded merely as a nuisance parameter, there are circumstances when the latter is a quantity of interest in its own right. For instance, this is the case when there is interest in evaluating the performance of detection methods (e.g. Mortelliti & Boitani 2008). Detectability may also be of interest when it reflects some important characteristic of the ecological system. For example it could be associated with reproduction (Best & Petersen 1982). Detectability estimates provide information on the number of times that a site needs to be visited before stating with a given degree of certainty whether the species of interest is present or absent at that particular location. This information can be especially relevant in the context of environmental impact assessments. Under these scenarios there is interest in obtaining a precise estimate of detection probability.

Like the variance of $\hat{\psi}$, the variance of $\hat{p}$ also starts by decreasing as more replication is added to the design but then, as p^* approaches unity, the variance of $\hat{p}$ tends to a constant level as then it is dictated by the total amount of effort, no matter whether the effort is spent on additional sites or replicates (Figure 3-1). This indicates that including p as part of the design criterion will lead to recommendations with more replicates to survey per site.

There are different criteria that can be used for optimal design when there is more than one parameter of interest. For a discussion of the merits of the different methods see Atkinson and Donev (1992, p. 106). One common approach is to minimize the trace of the variance-covariance matrix, that is, the sum of the variances of the parameters. This is called A-optimality. Alternatively, D-optimality minimizes the determinant of the variance-covariance matrix, that is, the area of the elliptical confidence region determined by the multivariate normal distribution of the estimators assuming large sample. Note that only considering the variance of $\hat{p}$ as a criterion for design would suggest that the best option is to sample one site and dedicate all the effort to further replication within this site. It is obvious that this is not an appropriate approach, as there is no guarantee that the species would be present in the site and therefore that informative data on its detectability can be collected. The asymptotic variance of $\hat{p}$ always decreases with K, however as K increases to such levels that the number of sites becomes very low, the asymptotic approximations are no longer suitable. Of course, if the aim of the study is solely to estimate detectability, then the obvious approach would be to dedicate all the survey effort to sample sites where the species is known to exist.

The optimal number of replicate surveys to be carried out at each sampling site using the A-optimality and D-optimality criteria for design is presented in Table 3-1b and Table 3-1c. Broadly the patterns are similar to those of Table 3-1a but, as expected, the optimal number of replicates increases, driven by the variance of $\hat{p}$. The largest changes are observed for low probabilities of occupancy and low probabilities of detection respectively. As happens when we consider the variance of the occupancy estimator only, the optimal number of replicate surveys in these two cases is determined by the parameter values (ψ and p) irrespective of the survey total effort (E).

Given (3.1) it is evident that that the optimal number of replicates is the same regardless of whether the study is designed to (i) minimize survey effort for a target estimator variance (measured through any of the three criteria discussed above) or (ii) minimize estimator variance for a target survey effort. Once the amount of replication is decided, then the number of sites to survey is to be determined. If the study is designed to minimize estimator variance for a given survey effort E, then the number of sites is derived as $S = E/K$. On the other hand, if the study is designed to minimize the survey effort needed to achieve a target estimator variance, the number of sites is derived using the expressions in (2.13). The following section provides a diagram showing the complete design procedure, including the consideration of small sample sizes and consequent need to use simulations as a tool for design.

Table 3-1 Optimum number of replicate surveys to be carried out at each sampling site for a standard design with constant per-survey costs when the criterion for design is based on minimizing (a) the variance of $\hat{\psi}$ (MacKenzie and Royle 2005), (b) the sum of variances of $\hat{\psi}$ and $\hat{p}$ (A-optimality) and (c) the determinant of the variance-covariance matrix (D-optimality); based on estimator properties that assume large sample sizes.

		ψ								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(a)										
p	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
(b)										
p	0.1	19	16	17	17	19	20	23	27	34
	0.2	13	10	9	9	9	10	11	13	16
	0.3	10	7	7	6	6	7	7	8	10
	0.4	8	6	5	5	5	5	5	6	7
	0.5	7	5	4	4	4	4	4	5	6
	0.6	6	4	4	3	3	3	3	4	4
	0.7	5	4	3	3	3	3	3	3	4
	0.8	4	3	3	2	2	2	2	2	3
	0.9	3	2	2	2	2	2	2	2	2
(c)										
p	0.1	19	19	20	21	23	24	27	30	36
	0.2	9	10	10	11	11	12	13	14	17
	0.3	6	6	7	7	7	8	8	9	11
	0.4	5	5	5	5	5	6	6	7	8
	0.5	4	4	4	4	4	4	5	5	6
	0.6	3	3	3	3	3	4	4	4	5
	0.7	3	3	3	3	3	3	3	3	4
	0.8	2	2	2	2	2	2	3	3	3
	0.9	2	2	2	2	2	2	2	2	2

3.1.3 Small sample size considerations and design procedure

Small sample sizes are not uncommon in ecological studies. In particular they are frequently encountered in surveys linked to conservation projects, as these often have limited resources and tend to focus on rare species. Pilot studies, by their nature, also tend to deal with relatively small amounts of data. Designing an occupancy study based on large-sample approximations is not appropriate if the intended sample size is small, especially when dealing with rare and elusive species, as then the probabilities of occupancy and detection are low. Under these circumstances, the actual quality of the estimators may be very different from that predicted by the asymptotic expressions (Figure 2-6) and the design identified as optimal using large sample approximations may not be the best available (as illustrated in an example below). In these cases the most appropriate method for designing a study relies on the use of simulations.

Figure 3-3 illustrates the design procedure for occupancy surveys when the design target is set in terms of estimator quality. Study design should start with a clear statement of the project requirements regarding the quality of the estimators (e.g. maximum allowed variance) and total survey effort available. With this in mind the design can be made to either (A) maximize the quality of the estimators or (B) minimize the total effort employed. Initial values for the parameters to be estimated need to be assumed. These can be based on the results of a pilot study, on studies carried out for the same or similar species in comparable circumstances or on expert opinion (see section 3.3 for methods to formally address the uncertainty in the initial estimates). The first issue to address is whether the sample size can be considered large enough to base the choice of design parameters on asymptotic approximations. If the total effort available is large and the probabilities of occupancy and detectability are expected to be rela-

tively high, the design can be based safely on these approximations. Otherwise, a simulation study is required, in which the actual quality of the estimators is evaluated for different design parameters. To assist in this process, we developed a software tool (SODA) that runs an automated search for a suitable design, given the assumptions and requirements specified by the user (Appendix A.1). Once a candidate design is identified, either through asymptotic approximations or simulations, it is necessary to verify whether it fulfils the project requirements. If it does, the study can proceed to data collection. Otherwise, if no suitable design is found, the objectives and constraints of the project need to be reconsidered: can more resources be allocated to this study? Could less precise estimates still be informative for the purpose of the study? If the answer to these questions is negative the study should not continue as it would be a waste of resources that could be used elsewhere. If the project objectives or constraints are redefined, a new design should be sought given the new requirements.

Example

As an illustration of the design process let us assume that: (i) our target is for the occupancy estimator to be approximately unbiased with a maximum SE of 0.075, i.e. maximum root mean square error (RMSE) = 0.075, (ii) the maximum effort that can employed in the study, $E_{\max}$, is 350 and (iii) the probabilities of occupancy and detectability are thought to be $\psi \approx 0.2$ and $p \approx 0.3$. To start with we can look at the recommendations derived from the asymptotic properties of the estimators to identity the optimal number of replicates to be used, in this case $K = 5$ (Table 3-1a). Let us first assume that the priority is to minimize the variance of ψ (option A in Figure 3-3). In this case we will make use of the total available effort and the number of sites to be surveyed is derived as $S = E/K = 70$. We should now evaluate the variance of the occupancy es-

timator under this design ($K = 5$ and $S = 70$) to verify whether it is within the target. From (2.13) we get that $\text{var}(\hat{\psi}) = 0.0033$, which gives a SE of 0.057. Assuming large sample size, the estimator is unbiased, so its RMSE is also 0.057. This is within the target set ($0.057 < 0.075$) so the design seems good. In order to verify that the approximations made were appropriate we would now run simulations. From 50,000 simulations we estimate that the actual SE of the occupancy estimator (0.070) is higher than predicted by the approximation (0.057), although still within the project target, so the design could be kept. However, given that the approximation was not very accurate it may be worth exploring other combinations of parameters as there is no guarantee of the optimality of the chosen design. For instance, a design with $K = 6$ and $S = 58$ would be a better choice (Table 3-2).

Let us now repeat the process assuming that the priority is on minimizing the total effort E (option B in Figure 3-3). Now the number of sites to be surveyed is derived from the expression of the asymptotic variance of the occupancy estimator in (2.13), setting $\text{var}(\hat{\psi})$ to the maximum allowed (0.075^2), which gives $S = 41$. The total effort required for this design (205) is within the target that our project set (350) so the design seems good. However, simulations show that the occupancy estimator has some bias and large variance; its RMSE (0.1391) is almost twice the maximum RMSE allowed by the project (0.075), which renders this design unsuitable. The asymptotic approximation is poor for the sample size so it is best to choose the design via simulations. By exploring different combinations of K and S we can identify the design that fulfils the variance target with minimum effort. In this case, $K = 7$ and $S = 43$ would be a good choice. Note that the number of replicates (7) differs from the optimal num-

ber suggested by the asymptotic approximations (5) and the total effort required is substantially larger (301 vs. 205).

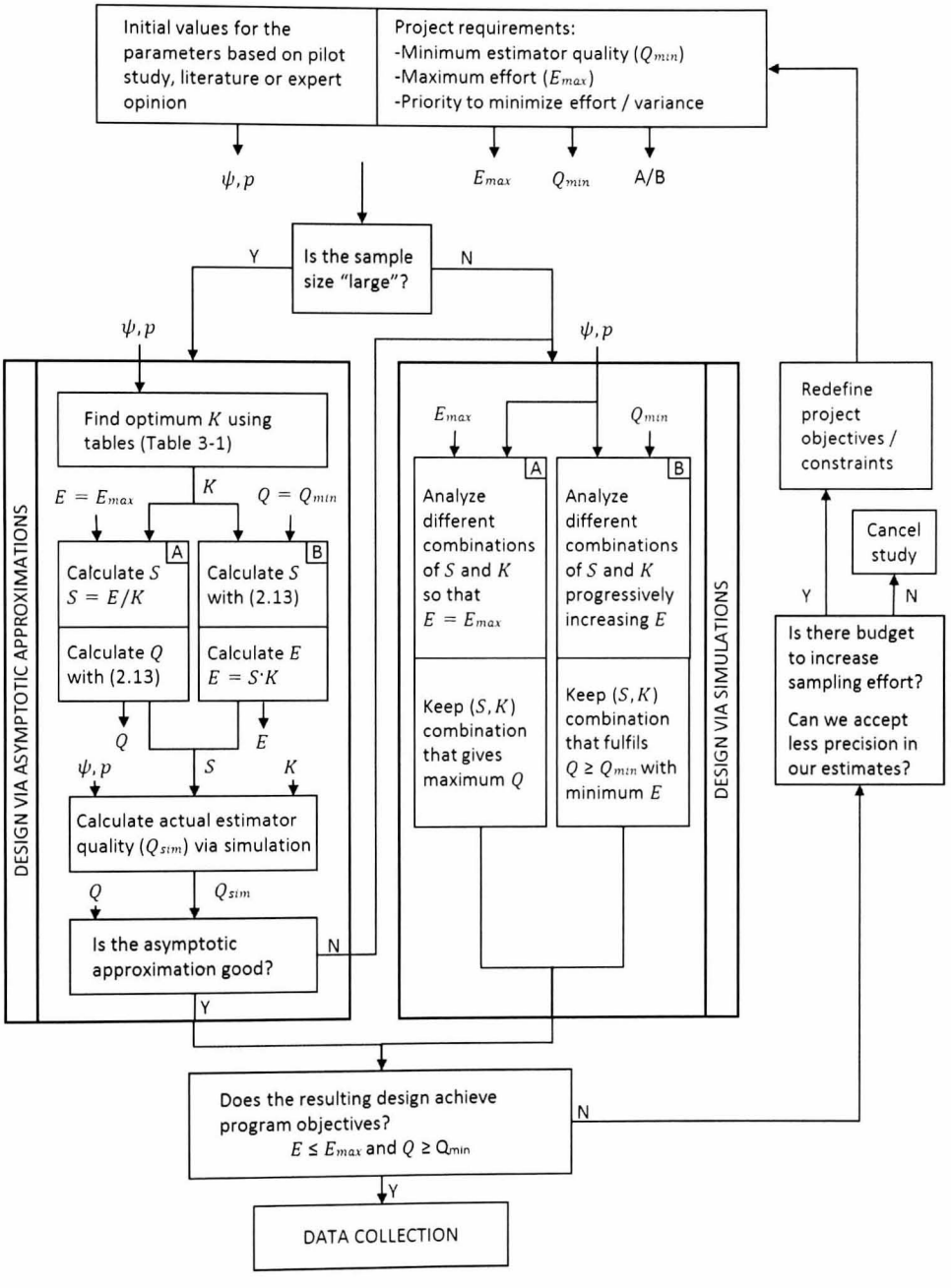


Figure 3-3 Occupancy survey design procedure when the design target is set in terms of estimator quality. Priority can be given to maximize estimator quality (A) or minimize total effort (B). A standard survey design and constant survey cost is assumed. Shaded boxes represent decision stages.

Table 3-2 Actual and asymptotic root mean-squared errors for $\hat{\psi}$ and $\hat{p}$ under different study designs assuming underlying probabilities $\psi = 0.2$ and $p = 0.3$.

	K					
	4	5	6	7	8	9
$E \approx 250$						
S	62	50	42	36	31	28
aRMSE $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	6.9/9.6	6.8/8.6	6.9/7.9	7.1/7.5	7.5/7.3	7.8/7.1
RMSE $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	12.6/10.1	10.6/9.3	9.6/8.7	9.3/8.4	9.6/8.2	9.6/8.0
RMSE* $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	9.3/9.7	8.2/9.0	7.7/8.4	7.5/8.1	7.7/8.0	7.9/7.7
Boundary estimates	1.1%	0.7%	0.5%	0.5%	0.5%	0.5%
$E \approx 300$						
S	75	60	50	43	37	33
aRMSE $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	6.3/8.7	6.2/7.9	6.3/7.3	6.6/6.9	6.9/6.7	7.2/6.5
RMSE $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	10.1/9.2	8.4/8.4	7.8/7.8	7.5/7.5	7.9/7.4	8.1/7.2
RMSE* $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	8.2/8.9	7.2/8.2	6.9/7.7	6.7/7.4	7.0/7.3	7.2/7.1
Boundary estimates	0.5%	0.3%	0.2%	0.2%	0.2%	0.3%
$E \approx 350$						
S	87	70	58	50	43	39
aRMSE $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	5.8/8.1	5.7/7.3	5.9/6.8	6.1/6.4	6.4/6.2	6.6/6.0
RMSE $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	8.3/8.5	7.0/7.6	6.6/7.2	6.7/6.9	6.9/6.7	7.1/6.6
RMSE* $\hat{\psi}$ / $\hat{p}$ ($\times 10^2$)	7.4/8.4	6.5/7.6	6.3/7.2	6.3/6.9	6.5/6.6	6.6/6.5
Boundary estimates	0.2%	0.1%	0.1%	0.1%	0.1%	0.1%
A-opt criterion ($\times 10^3$)	14.1	10.7	9.5	9.2	9.3	9.3
D-opt criterion ($\times 10^{-5}$)	3.28	2.21	1.95	1.96	2.04	2.05

E is the total survey effort and K the amount of replication. Asymptotic root mean-squared error (aRMSE) was obtained analytically. Actual root mean-squared error (RMSE) was estimated from 5000 simulations. The frequency of boundary estimates ($\hat{\psi} = 1$) and the actual root mean squared error after removing these (RMSE*) are shown. For $E = 350$, the sum of the mean-squared errors (A-optimality criterion) and the determinant of the MSE matrix (D-optimality criterion) are also reported.

3.2 Power analysis

3.2.1 Background

Rather than simply making inference about species occupancy at a given point in time, area or habitat type, studies often ultimately aim to make inference about potential differences in occupancy. The interest might be in assessing differences temporally (e.g. has occupancy changed since the last survey?) or spatially (e.g. is occupancy different in these two areas or habitat types?). These types of question can be relevant in various applications, from theoretical ecological studies to more applied impact assessments.

In this section we reconsider the design of occupancy studies when the interest is in assessing whether there are differences in occupancy between two samples, e.g. two points in time, areas or habitat types. The design procedure discussed in section 3.1 is based on targets set in terms of estimator precision: once the amount of replication is decided upon, sample size is determined either as the minimum number of sites to achieve this estimator precision target or as many sites as allowed by the available effort (and checking whether the target is met). Alternatively, the criterion for selecting the size of the sample can be expressed in terms of power, that is, the probability that the study will detect a significant difference in occupancy, given that the true difference is of a given size (Cohen 1988). A particularly beneficial aspect of power analysis is that it requires an explicit consideration of what constitutes a biologically significant result, allowing us to determine whether a given design renders our study a good chance of producing statistically significant results when the actual effect size is biologically significant.

While simulations provide a tool for power analysis, they can be time-consuming. Closed formulae can sometimes be derived to determine more easily the sample size required to achieve a given power. The development and performance evaluation of such formulae for a test comparing two independent binomial proportions has received a lot of attention in the literature (e.g. Cochran & Cox 1957 p.27; Fleiss 1973 p.30; Casagrande, Pike & Smith 1978; Walters 1979; Fleiss, Tytun & Ury 1980; Ury & Fleiss 1980; Dobson & Gebski 1986; Gordon & Watson 1996; Vorburger & Munoz 2006). These formulae are routinely used in different areas, such as the design of clinical trials (Donner 1984). However, since they assume that the outcome of the experiment, whether success or failure, is always observed without error, they are not applicable for occupancy studies, except for the unusual case in which species detection is perfect or enough replicate surveys are carried out to ensure its detection is practically certain.

To our knowledge, sample size formulae for models that account for imperfect detection when comparing two independent binomial proportions have not been proposed or evaluated to date. In this section we address this problem. We provide an approximate expression to calculate power and derive a closed-formula that allows the number of sites that need to be sampled to be determined, while accounting for species detectability. Using this expression we examine how the required sample size changes depending on the allocation of survey effort between number of sites and number of replicate visits and thus revisit the issue of optimal replication addressed in section 3.1 from the point of view of minimizing the variance of the estimator in single-season studies. Since the derived sample size formula involves asymptotic approximations its performance needs to be assessed, as this is essential to understand its applicability.

For this we ran simulations and checked how the resulting sample sizes compare to those indicated by the formula. In connection to this we evaluate the performance of various significance tests. In the context of studies that assess occupancy changes in time, we also address the case in which Markovian dependence is assumed in the occupancy status of sites between seasons, and illustrate the utility of our results when designing to detect a trend in multiple-season studies.

3.2.2 Power expression

A formula to assess the power to detect a difference in occupancy that would be achieved under a given study design and underlying probabilities of occupancy and detection can be derived by considering the properties of the estimators. Here we assume again a standard sampling design in which K replicate surveys are carried out at S sampling sites, and constant probabilities of occupancy and detection. Under the assumption of large sample size, the maximum-likelihood estimator of occupancy is unbiased and normally distributed $\hat{\psi} \sim N(\psi, \sigma^2)$, where σ^2 is the asymptotic variance of the occupancy estimator in (2.13). As discussed before, σ^2 has the form of the variance of a binomial proportion with an extra term $F = (1-p^*)/\{p^* - Kp(1-p)^{K-1}\}$ introduced by the imperfect detection. The term F is a function of detectability p and the number of replicate surveys per site K, and tends to zero as p^* tends to 1.

Let ψ_1 and ψ_2 be the true underlying occupancy probabilities in the two samples. For a significance level α (type I error), the critical region for a 2-tailed test is bounded by $\pm z_{\alpha/2}\sqrt{\sigma_1^2+\sigma_2^2}$. Consequently, the power to detect a difference in occupancy, that is, the probability of observing a difference that falls within the critical region, is

$$G = 1 - \beta = \left\{1 - \Phi\left(\frac{z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2} - (\psi_1 - \psi_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)\right\} + \Phi\left(\frac{-z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2} - (\psi_1 - \psi_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right), \quad (3.2)$$

where $\sigma_i^2 = \psi_i(1 - \psi_i + F_i)/S_i \quad i \in \{1, 2\}$, β is the probability of type II error and $\Phi(x)$ the value of the cumulative distribution function for the standard normal distribution at x.

Let R be the proportional difference in occupancy, so that $\psi_2 = \psi_1(1 - R)$, with $R > 0$ representing a decline, and $R < 0$ an increase. Note that $R \in [(\psi_1 - 1)/\psi_1, 1]$ to ensure that $\psi_2 \in [0, 1]$. The plot of G as a function of effect size (R here) is known as the 'power curve' of the test (Figure 3-4). All power curves pass through $(0, \alpha)$ since an effect of magnitude zero corresponds to the null hypothesis which by definition is rejected with probability α. As the effect size increases, the probability of rejecting the null hypothesis increases. For a given effect size, power increases as the number of sampling sites increases (Figure 3-4a). Power also increases with the number of replicate surveys (Figure 3-4b), approaching the power expected for a binomial experiment with perfect detection as p^* tends to one (but note that here increasing replication implies an increase in total effort; see section 3.2.4 for fixed survey effort). A similar behaviour takes place for increases in detection probability, with power saturating as p tend to one (Figure 3-4c). The larger the initial occupancy probability ψ_1, the larger the power to detect a given proportional difference R, as this translates into a larger absolute occupancy difference $\psi_1 - \psi_2$ (Figure 3-4d).

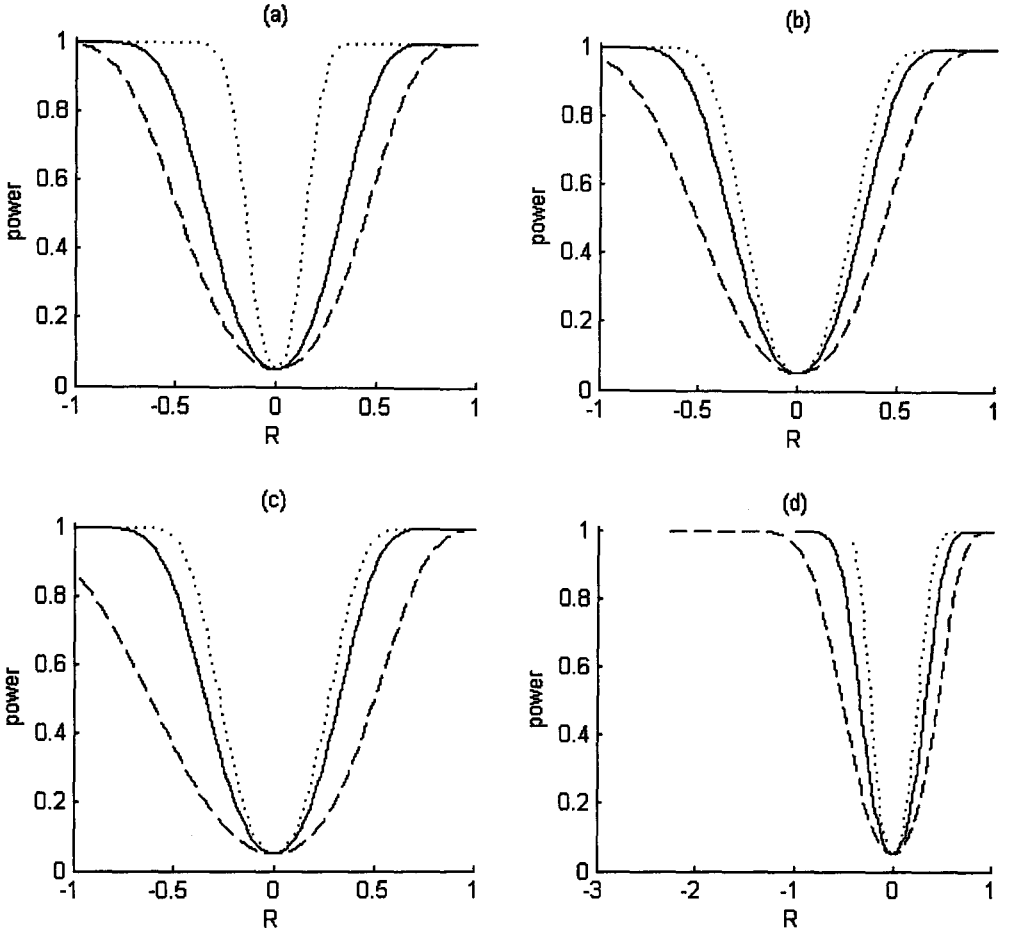


Figure 3-4 Power curves for testing a difference in occupancy between two samples. In all four panels the solid line represents a reference case with $\psi_1 = 0.5$ and $p = 0.5$, $K = 3$, $S = 100$ for both survey periods. In each panel one of these parameters is changed: (a) $S = 50$ (dash), $S = 500$ (dot); (b) $K = 2$ (dash), $K = 6$ (dot); (c) $p = 0.3$ (dash), $p = 1.0$ (dot); (d) $\psi_1 = 0.3$ (dash), $\psi_1 = 0.7$ (dot). Significance level set to $\alpha = 0.05$. Note $R < 0$ represents cases in which there is an increase in occupancy probability.

3.2.3 *Sample size formula*

Equation (3.2) can be solved numerically to determine the number of sites that need to be surveyed to achieve a given power. However, an approximation is possible that gives a convenient expression in closed form. Without loss of generality it can be assumed that $\psi_1 - \psi_2 > 0$. In this case the second term in equation (3.2) can be considered negligible as it represents the probability of detecting an apparent increase when in reality there is a decline. This probability will be small and corresponds to cases in which an incorrect inference about the sign of the occupancy difference would have been made. The power to detect can therefore be written now as

$$G = 1 - \beta \cong 1 - \Phi\left(\frac{z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2} - (\psi_1 - \psi_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right). \tag{3.3}$$

From (3.3), assuming that the same number of sites are to be surveyed in both occasions, and considering that by definition $1 - \beta = \Phi(z_\beta)$, and by symmetry $\Phi(x) = 1 - \Phi(-x)$, the number of sites S that have to be surveyed to achieve a given power to detect a difference in occupancy can be derived as a function of the significance level and effect size, given ψ_1, p_1, p_2, K_1 and K_2, as

$$S = \left(\frac{z_{\alpha/2}\sqrt{f_1 + f_2} + z_\beta\sqrt{f_1 + f_2}}{\psi_1 - \psi_2}\right)^2 = (f_1 + f_2)\left(\frac{z_{\alpha/2} + z_\beta}{\psi_1 - \psi_2}\right)^2 \tag{3.4}$$

where $f_i = \psi_i(1 - \psi_i + F_i), i \in \{1, 2\}$.

3.2.4 Survey effort allocation trade-off

The trade-off in survey allocation was discussed in section 3.1, in terms of the optimal amount of replication to minimize estimator variance (or other related criteria). Obviously, this trade-off can also be appreciated when looking at survey design from the point of view of the power to detect an occupancy difference between two samples. Figure 3-5 explores this and shows how the amount of total survey effort $E = KS$ needed to detect a 50% occupancy decline with power = 0.8 changes depending on how much replication is used. The same general observations as in Figure 3-2 can be made. When detectability is high the design requiring minimum effort is always that with $K = 2$. As p decreases, the optimum K increases (and so does the required minimum effort). The higher the initial occupancy, the larger the optimum K. Choosing K smaller than the optimum has a much greater impact than otherwise.

In fact, for the scenarios under consideration in Figure 3-5, which assume that detectability is the same in both seasons, the value of K that minimizes the total survey effort required for achieving the target power largely corresponds to that which minimizes the variance of the occupancy estimator for the first survey season. There is only some slight discrepancy in more extreme cases of very low detectability together with high initial occupancy, as then the optimal K to minimize the variance of ψ_1 differs more from that required to minimize the variance of ψ_2. In these cases the optimal K in terms of power was slightly lower, as we were assessing a decline, and thus the optimal K to minimize the variance of ψ_2 would be lower than that of ψ_1.

Finally note in Figure 3-5b that, as ψ_1 decreases, effort increases. This is because the smaller the ψ_1, the smaller is the absolute occupancy difference to be detected.

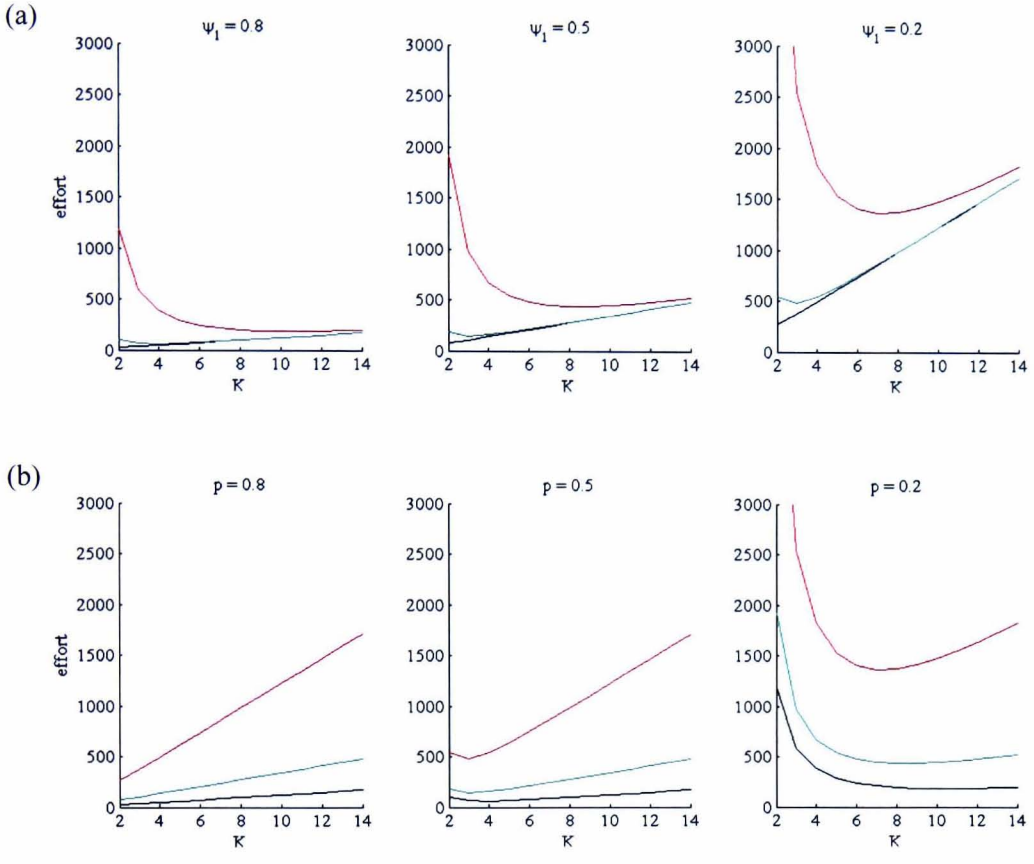


Figure 3-5 Minimum survey effort ($E = SK$) to achieve 80% power to detect a 50% decline in occupancy, for varying K and different scenarios of initial occupancy ψ_1 and detectability p (α = 0.05). In both seasons p and K are kept the same. Top panels (a) display cases with the same ψ_1 (black p = 0.8, green p = 0.5, red p = 0.2) while lower panels (b) compare cases with the same p (black ψ_1 = 0.8, green ψ_1 = 0.5, red ψ_1 = 0.2).

3.2.5 Testing for significance in occupancy differences

Various approaches can be used for testing the null hypothesis of no difference in occupancy between two samples ($\psi_1$and ψ_2). One possibility is to determine significance based on a z-test. Since the *estimator* of occupancy $\hat{\psi}$ is unbiased and normally distributed with mean ψ and variance σ^2, the random variable $D = \hat{\psi}_1 - \hat{\psi}_2$ is normally distributed $D \sim N(\psi_1 - \psi_2, \sigma_1^2 + \sigma_2^2)$. Under the null hypothesis of no differ-

ence $D \sim N(0, \sigma_1^2 + \sigma_2^2)$ so, for a given significance level α, the critical region for a two-tailed test is bounded by $\pm z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2}$, where $z_{\alpha/2}$ is the upper $100\alpha/2$-percentage point for the standard normal distribution. Consequently, a difference would be considered significant if

$$\frac{|\hat{\psi}_1 - \hat{\psi}_2|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} > z_{\alpha/2},$$

where $\hat{\psi}_i$ are the maximum-likelihood *estimates* of occupancy and $\hat{\sigma}_i$ are the corresponding estimated standard errors, $i \in \{1, 2\}$. This is in fact a Wald test (Morgan 2008, p. 101) and is equivalent to assessing whether the confidence interval for the estimate of the difference in occupancy $\hat{\psi}_1 - \hat{\psi}_2$ includes zero. In practice, the likelihood maximization is usually done via a logistic reparameterization. The test could in principle also be performed on the logistic scale, considering a difference significant if

$$\frac{|\hat{\beta}_1 - \hat{\beta}_2|}{\sqrt{\hat{\sigma}_{\beta 1}^2 + \hat{\sigma}_{\beta 2}^2}} > z_{\alpha/2},$$

where $\hat{\beta}_i = \text{logit}(\hat{\psi}_i)$ are the parameter estimates on the logistic scale and $\hat{\sigma}_{\beta i}$ are their corresponding estimated standard errors.

Another possibility is to carry out a likelihood-ratio test (Morgan 2008, p. 80), which compares the fit of two models, where one (the null) is a special case of the other (the alternative). The test involves fitting both models and is based on the ratio of their maximum-likelihood values. The null model here would be a model with a common parameter for occupancy across both periods ($\psi_1 = \psi_2 = \psi$)

$$L(\psi, p_1, p_2|h) = \left\{\psi^{S_{d_1}} p_1^{d_{T_1}} (1-p_1)^{K_1 S_{d_1} - d_{T_1}}\right\} (1-\psi p_1^*)^{S_1 - S_{d_1}}$$
$$\times \left\{\psi^{S_{d_2}} p_2^{d_{T_2}} (1-p_2)^{K_2 S_{d_2} - d_{T_2}}\right\} (1-\psi p_2^*)^{S_2 - S_{d_2}},$$

while the alternative would allow for different occupancy parameters for each of the two periods. In practice this involves fitting two separate models, one to each data set. A difference would be considered significant at the α level if

$$-2\mathcal{L}_0 + 2\mathcal{L}_A > \chi^2_{\alpha:1},$$

where $\mathcal{L}_0$ and $\mathcal{L}_A$ are the maximum log-likelihood values for the null and alternative models respectively and $\chi^2_{\alpha:1}$ is the upper 100α-percentage point for the chi-square distribution with one degree of freedom.

A third option is to perform a score test (Morgan 2008, pp. 102-103), which would render an occupancy change to be significant at the α significance level if

$$\mathbf{G}'\mathbf{J}^{-1}\mathbf{G} > \chi^2_{\alpha:1},$$

where $\mathbf{G}$ is the scores vector and $\mathbf{J}$ is the information matrix for the alternative model evaluated at the MLEs of the null model. The score test may be performed using either the expected or the observed information matrix. In general the former is preferred as it has been shown that it outperforms the latter in some applications (e.g. Catchpole & Morgan 1996) and that the use of the observed information can be problematic in some others (e.g. Morgan, Palmer & Ridout 2007). Since the alternative model is equivalent to analyzing the two occupancy data sets separately, we have that

$$\mathbf{G} = (\mathbf{G}_1 \quad \mathbf{G}_2), \mathbf{O} = \begin{pmatrix} \mathbf{O}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{O}_2 \end{pmatrix} \text{ and } \mathbf{I} = \begin{pmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 \end{pmatrix},$$

where $\mathbf{G}_i$, $\mathbf{O}_i$ and $\mathbf{I}_i$ are the basic occupancy model scores vector and information matrices for time i, $i \in \{1, 2\}$, given by (2.10), (2.11) and (2.12).

3.2.6 Performance of significance tests

The Wald, likelihood-ratio and score tests are all based on asymptotic approximations and are asymptotically equivalent under the null hypothesis. However, for small sample sizes, the tests can produce contradictory results and an obvious question arises: which test to use? There are some general properties of the tests that can influence this choice. On the one hand the likelihood-ratio test has the disadvantage that it requires fitting both the null and alternative models, while the other tests require fitting only one (alternative model for Wald test; null model for score test). However, since the models under consideration are rather simple, this feature is not crucial for this application. On the other hand, the likelihood-ratio test and the score test based on the expected information matrix have an interesting property: they are invariant under a reparameterization, unlike the Wald test. Of course, in any case, the most important consideration is the actual performance of the tests when the sample size is small. In order to assess this we ran simulations comparing the power achieved by each of them for increasing effect size R. We started with zero effect, to check whether the size of the tests was close to the nominal significance level. We first assessed a scenario with $\psi_1 = 0.8$, $p = 0.5$, $S = 50$ and $K = 3$, used as a reference case. We then simulated variations of this scenario, by changing one parameter at a time (Table 3-3).

Table 3-3 Scenarios simulated to assess the performance of significance tests for differences in occupancy probability. The parameter value that is changed with respect to the reference case is underlined.

Scenario	ψ_1	p	S	K	p^*
Case 1: Reference case	0.8	0.5	50	3	0.875
Case 2: Increase p	0.8	<u>0.8</u>	50	3	0.992
Case 3: Increase S	0.8	0.5	<u>250</u>	3	0.875
Case 4: Decrease ψ_1	<u>0.4</u>	0.5	50	3	0.875
Case 5: Increase K	0.8	0.5	50	<u>7</u>	0.992
Case 6: Increase K	0.8	0.5	50	<u>10</u>	0.999

Note: $p^* = 1 - (1 - p)^K$

In all scenarios 5000 simulations were run, which should provide sufficiently precise power estimates (SE $= \sqrt{0.5(1 - 0.5)/5000} = 0.007$ for power = 0.5, the most demanding case). In each simulation, maximum-likelihood estimates and corresponding standard errors were obtained on the logistic scale. Standard errors on the probability scale were obtained using the delta-method approximation. Significance was determined at three significance levels ($\alpha = 0.05$, 0.10 and 0.20) via the following tests:

(i) Wald test (probability and logistic scale),

(ii) Likelihood-ratio test,

(iii) Score test with expected information matrix,

(iv) Score test with observed information matrix (probability and logistic scale).

Note that the score test based on the expected information matrix is invariant to a reparameterization, but this property does not hold when the observed information matrix is used (Boos 1992).

Power for each scenario was calculated as the proportion of simulations in which a significant decline was detected. The simulation results for the reference case show that the performance of the tests is very different (Figure 3-6). In particular the following observations can be made:

(i) Score tests based on the observed information matrix behave badly on both the probability and logistic scales, with power markedly decreasing as R increases;

(ii) The score test based on the expected information matrix has very similar performance to the likelihood-ratio test (although there is some divergence for high values of R);

(iii) The Wald test has considerably lower power on the logistic scale than on the probability scale;

(iv) The Wald test on the probability scale has higher power than the likelihood-ratio and score tests. The difference in their performance depends on the significance level and is higher for lower α.

If detectability increases the tests agree more, all showing very similar performance in case 2 for $R < 0.5$ (Figure 3-7). However the score test based on the observed information matrix still behaves oddly, with marked power decrease for higher effect sizes. The rest of the results (Appendix A.2) show that, as might be expected, increasing the sample size (cases 3, 5 and 6) also results in more similar test performance except, once again, for the score-observed test. Interestingly, while this problem practically disappears when increasing the number of replicates from 3 to 10 (case 6), it is still

evident in a scenario with a 5-fold increase in the number of sites (case 3). The results from a scenario with lower initial occupancy (case 4) lead to the same observations as those discussed for the reference case, now of course with the power being lower.

Overall, in our simulations the Wald test on the probability scale performs better than any of the other tests evaluated. The test shows higher power than other tests while having the right size, which suggests it is a good choice for this type of study. However, if performed on the logistic scale, the Wald test shows considerably decreased power in some scenarios (case 1 and 4). Indeed, it has been shown that the Wald test can produce misleading results when working with discrete probability distributions under certain parameterizations (Vaeth 1985). Hauck and Donner (1977) showed for instance that the test has an aberrant behaviour when testing the equality of two proportions on the logistic scale, losing power as the difference between them increases. Our results suggest evidence of an aberrant behaviour when dealing with the comparison of two proportions (e.g. occupancy values) under imperfect detection.

Both the likelihood-ratio and the score tests show inferior performance compared to the Wald test on the probability scale. In particular, the score test based on the observed information matrix displays an aberrant behaviour for some of the scenarios investigated. In fact, the test statistics are often negative in these cases (compare Figures 3-8 and 3-9). Such behaviour has been previously observed for zero-inflated Poisson models (Morgan, Palmer & Ridout 2007), where negative test statistics are obtained in a test assessing whether zero-inflation is different from zero. This is in fact a similar case to ours as our test assesses whether zero-inflation differs in two zero-inflated binomial samples.

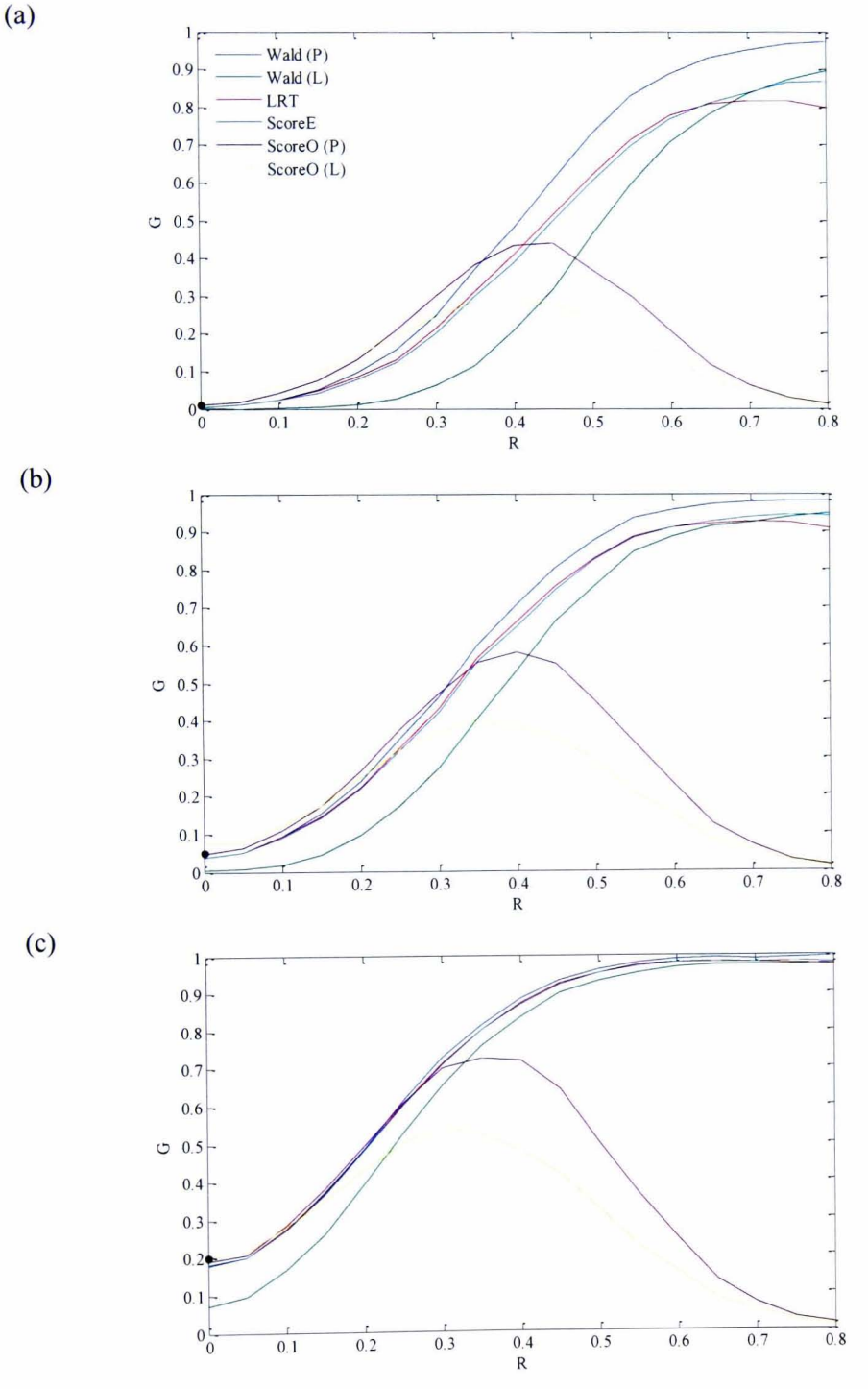


Figure 3-6 Power to detect an occupancy decline for different significance tests in simulation case 1 (black dot on y-axis indicates significance level)

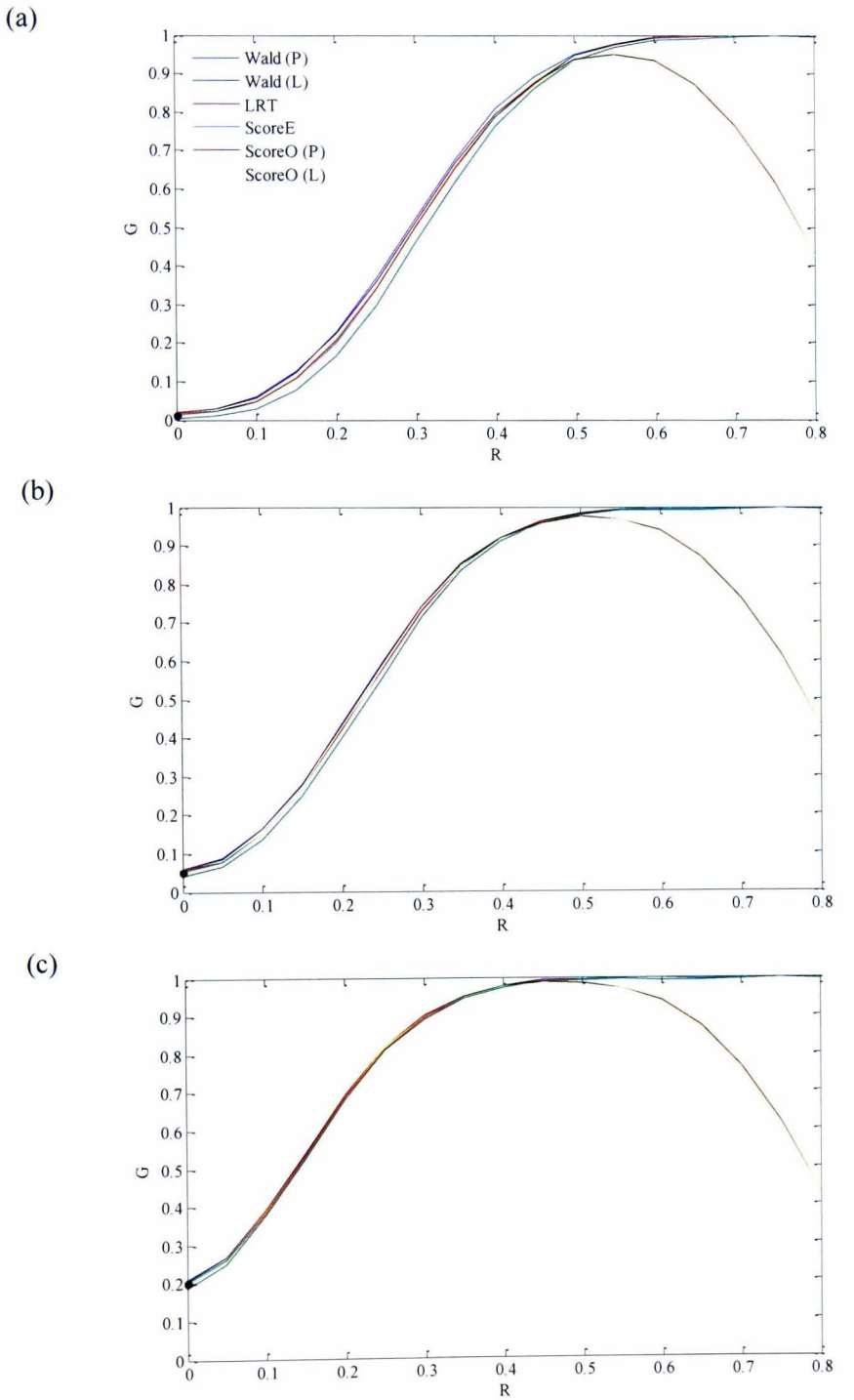


Figure 3-7 Power to detect an occupancy decline for different significance tests in simulation case 2 (black dot on y-axis indicates significance level)

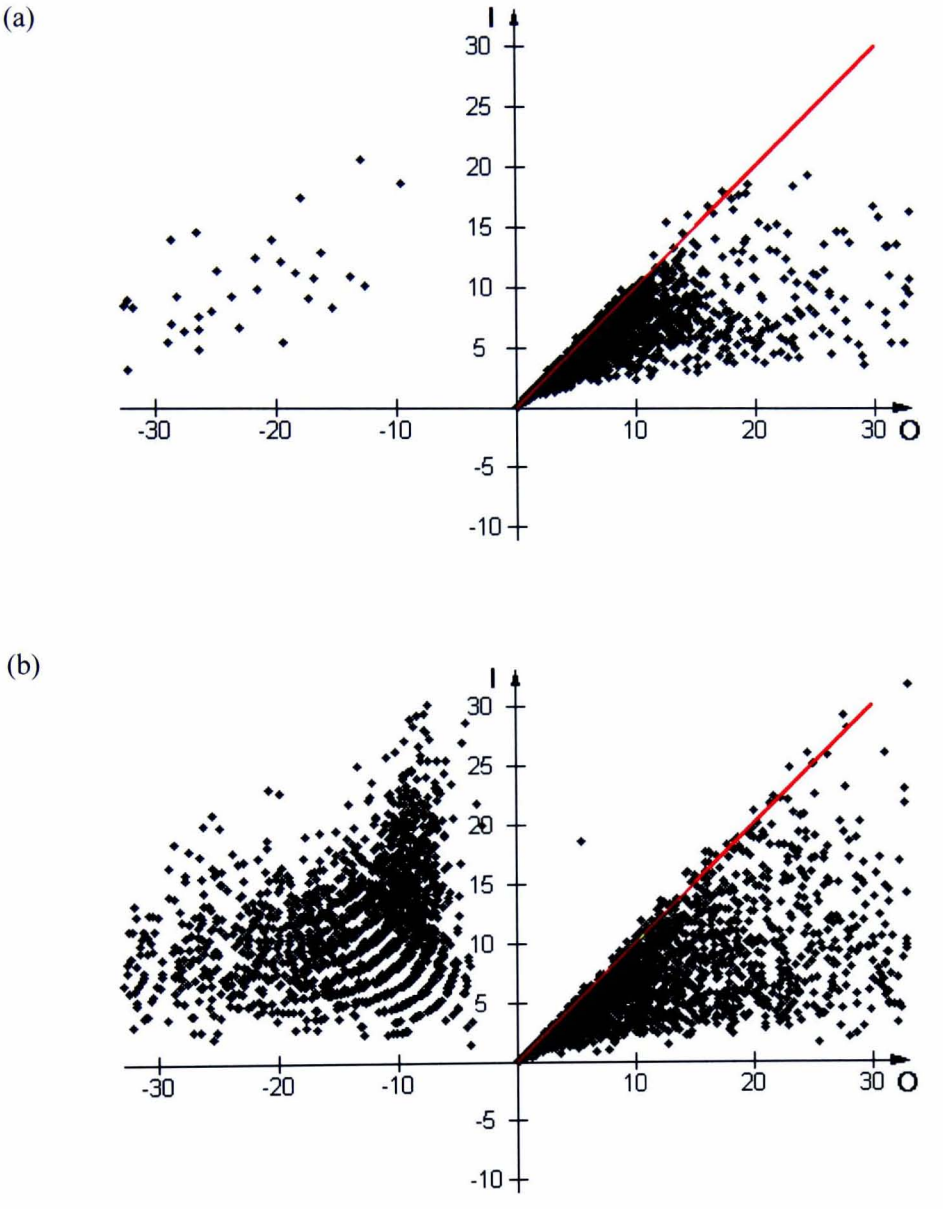


Figure 3-8 Scatter plot of the score test statistic based on the expected information matrix (I) with respect to the score test statistic based on the observed information matrix on the probability scale (O) for simulation case 1 ($\psi_1 = 0.8$, $p = 0.5$, $S = 50$, $K = 3$), for effect size (a) $R = 0.25$ and (b) $R = 0.5$. The threshold to determine significance at the $\alpha = 0.05$ level is 3.84.

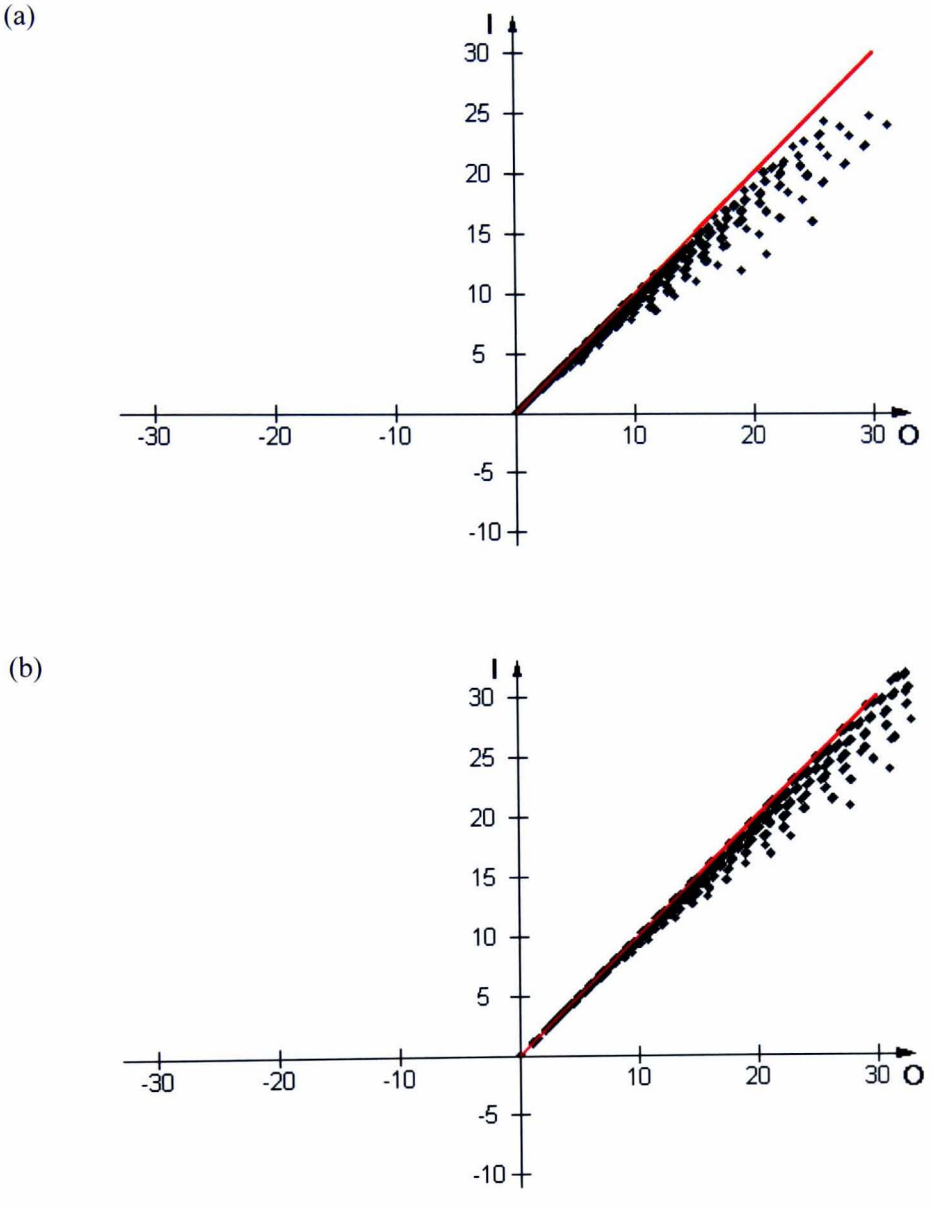


Figure 3-9 Scatter plot of the score test statistic based on the expected information matrix (I) with respect to the score test statistic based on the observed information matrix on the probability scale (O) for simulation case 6 ($\psi_1 = 0.8$, $p = 0.5$, $S = 50$, $K = 10$), with effect size (a) $R = 0.25$ and (b) $R = 0.5$. The threshold to determine significance at the $\alpha = 0.05$ level is 3.84.

3.2.7 Sample size formula performance

In order to verify the performance of the formula in (3.4) we again ran simulations and compared the survey effort required to achieve a given power to detect an occupancy decline as indicated by both approaches under various scenarios. We explored different values of initial occupancy (ψ_1= 0.2, 0.5 and 0.8), detection probability (p = 0.2, 0.5 and 0.8), and effect size (R = 0.5 and 0.3) under a range of replication levels (K = 2 to 6), using a significance level α of 0.05. We simulated scenarios with increasing levels of total survey effort in steps of 10%, starting from the survey effort indicated by the formula to achieve a power of 0.7, and stopping when the power achieved in the simulations was larger than 0.9. As before, we ran 5000 simulations per scenario. For each simulation, we assessed significance according to three methods: Wald tests on the probability and logistic scales, and a likelihood-ratio test. Score tests were not included in these simulations but, according to the results in section 3.2.6, we can expect the score test based on the expected information matrix to perform similarly to the likelihood ratio test, while the test based on the observed information matrix is not interesting due to its aberrant behaviour. We also ran simulations to verify the size of the tests for the scenarios described above with the design suggested by the formula for R = 0.5 and power = 0.8.

Figure 3-10 compares the simulation results with curves obtained using the sample size formula for R = 0.5 and power = 0.8. In all cases, the sample size determined by the closed-formula was similar to that indicated by the simulations, which tended to suggest somewhat higher sampling effort. Consistent with the results obtained in section 3.2.6, the simulations indicated that a greater sampling effort was required when significance was assessed with a likelihood-ratio test compared to the Wald test on the

probability scale. The results from the Wald test on the logistic scale were very similar to those from the Wald test on the probability scale when the probability of occupancy was not high. When ψ_1= 0.8 (and therefore sample sizes were small) the Wald test on the logistic scale appeared to have considerably lower power, thus requiring greater sampling efforts to detect a change. The size of the Wald test on the probability scale and the likelihood-ratio test was reasonably close to the nominal significance level (Figure 3-11).

The discrepancy between the formula and simulation results is due to the underlying assumptions not being perfectly met for some of the sample sizes under consideration. For instance, for the case of $\psi_1 = 0.8$, $p = 0.5$ and $K = 2$ in Figure 3-10, the asymptotic variance for the occupancy estimator for $\psi_2 = 0.4$ when 78 sites are surveyed is 0.0081, while the true variance according to simulations is about 50% higher (0.0123). This underestimation explains, at least partly, why the sample size required according to the formula (78 sites) was smaller than that suggested by the simulations (96 sites). The formula expected more precise occupancy estimators for such conditions, so that therefore it would be easier to detect differences between them. It should thus be kept in mind that the formula's outcome represents a lower bound and that, in some situations, more effort might be needed in the study. The sample size simulation results for $R = 0.3$ (Figure 3-12) show more agreement between the formula and the simulations, which is expected as detecting a smaller effect implies larger sample sizes, and thus less discrepancy with the large-sample approximations.

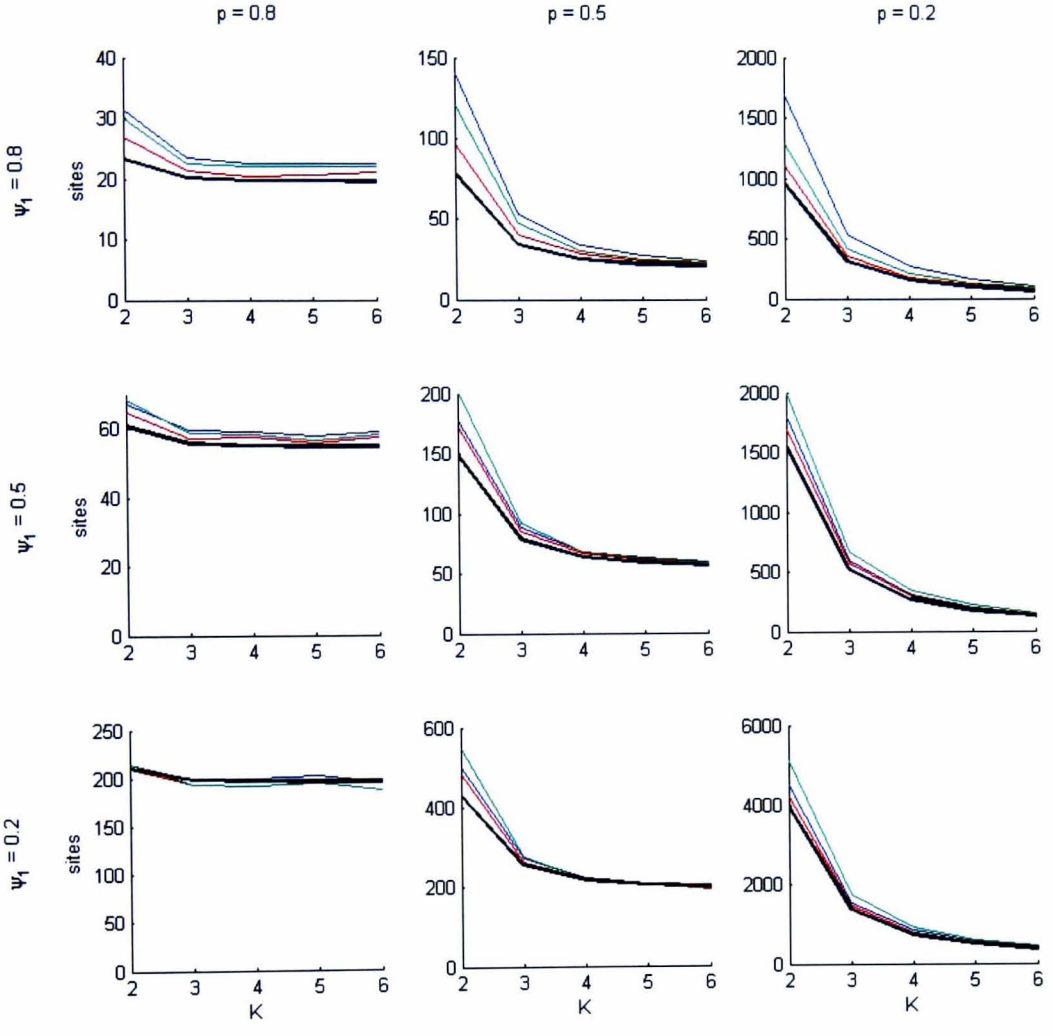


Figure 3-10 Number of sampling sites required to achieve 80% power to detect an occupancy decline $R = 0.5$ as indicated by the formula (thick black line) and simulations (red: Wald test on probability scale; blue: Wald test on logistic scale; green: likelihood-ratio test), for varying levels of replication (K survey visits per site), and different scenarios of initial occupancy (ψ_1) and detection probability (p). Significance level was set to $\alpha = 0.05$.

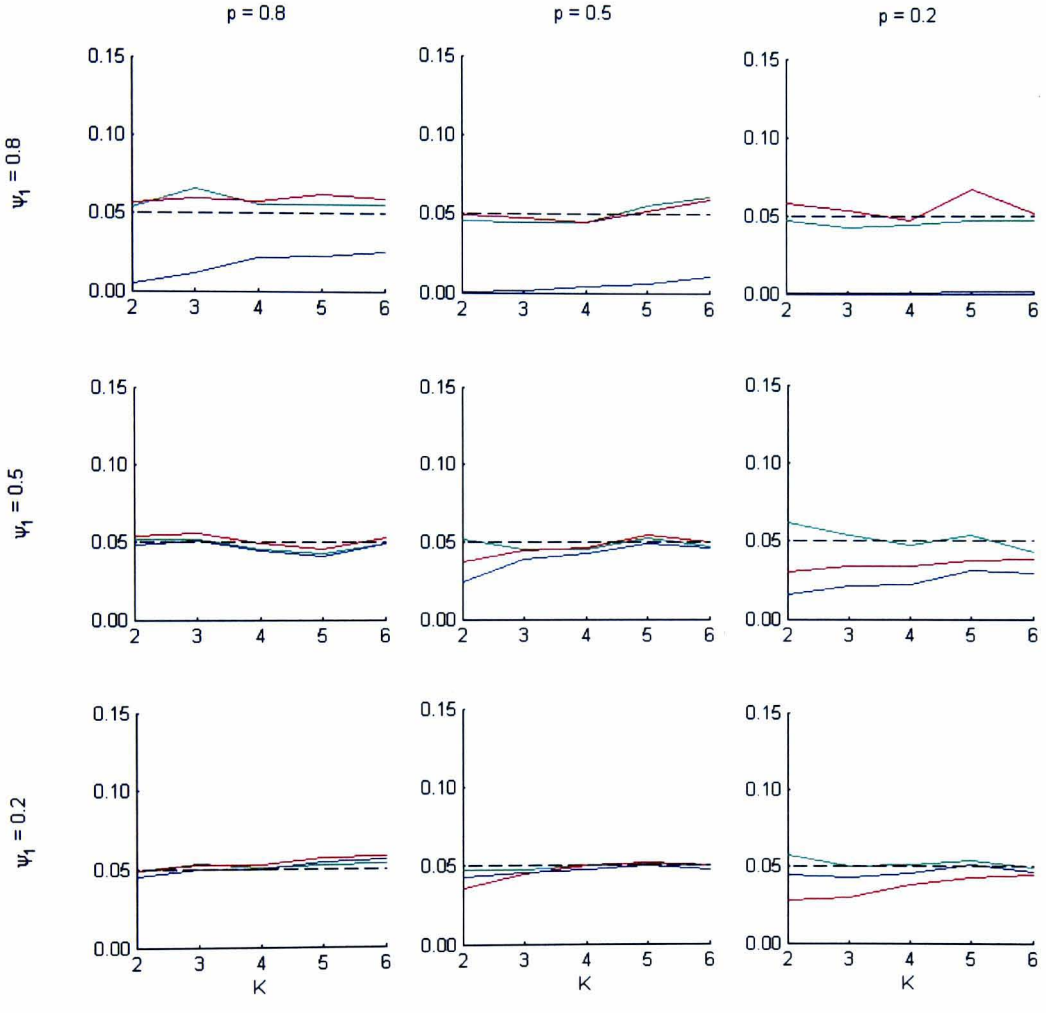


Figure 3-11 Size achieved by the tests in the simulations (red: Wald test on probability scale; blue: Wald test on logistic scale; green: likelihood-ratio test) for the designs indicated by the formula to achieve a power = 0.8 to detect a decline in occupancy $R = 0.5$, for varying levels of replication (K survey visits per site) and different scenarios of initial occupancy (ψ_1) and detection probability (p). The black dashed line represents the nominal significance level ($\alpha = 0.05$).

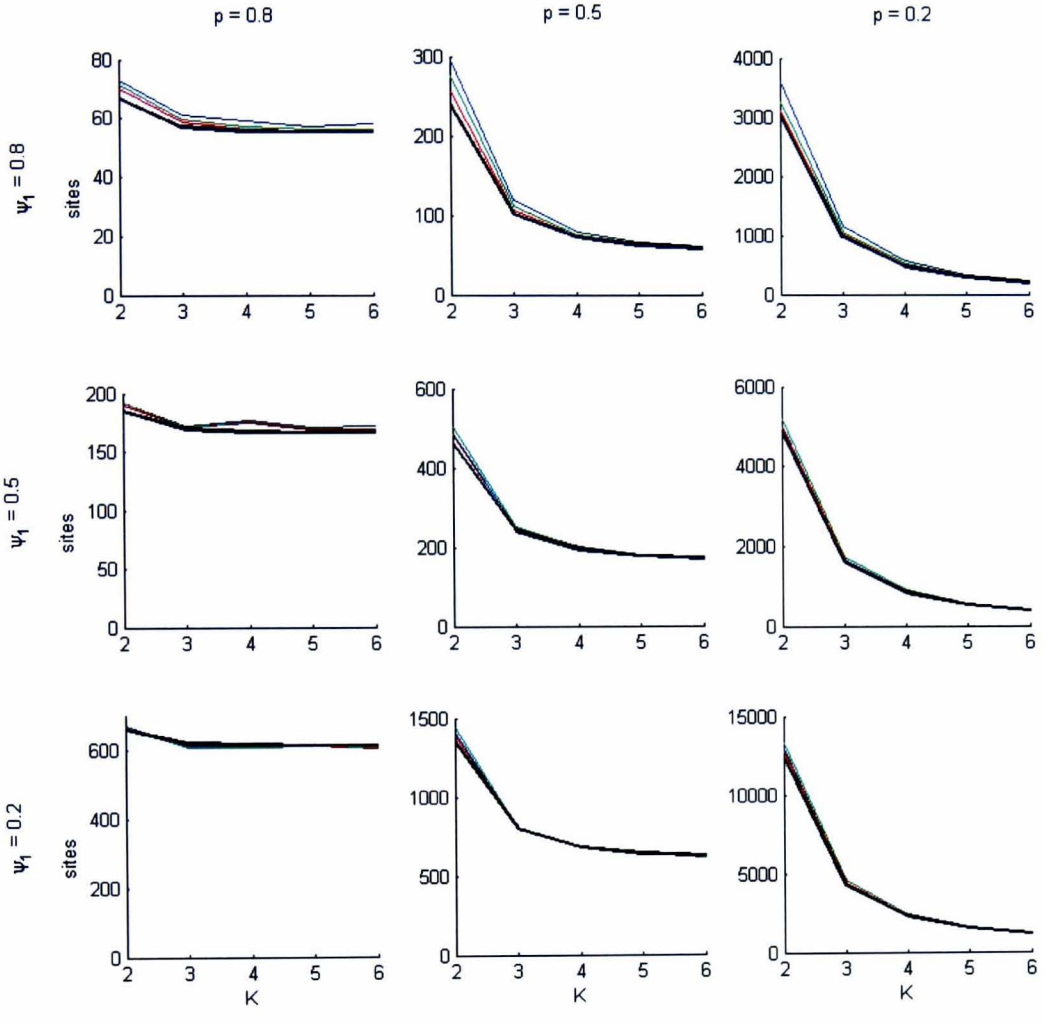


Figure 3-12 Number of sampling sites required to achieve 80% power to detect an occupancy decline $R = 0.3$ as indicated by the formula (thick black line) and simulations (red: Wald test on probability scale; blue: Wald test on logistic scale; green: likelihood-ratio test), for varying levels of replication (K survey visits per site), and different scenarios of initial occupancy (ψ_1) and detection probability (p). Significance level was set to $\alpha = 0.05$.

3.2.8 Applicability to cases with Markovian dependence in site occupancy status

We have so far implicitly assumed that the two samples that want to be compared can be considered independent, i.e. the occupancy status of the sites in one sample does not depend on the other sample. Two samples are independent if they consist of different sampling sites, for instance when comparing occupancy between two geographical locations or habitat types. Studies assessing changes between two points in time may also sample different sites, although often the same sites are sampled in both seasons. Even so, the assumption of independence can still be valid, for example when dealing with species that display a low degree of site fidelity or when the time elapsed between seasons is sufficiently long so that the changes can effectively be considered random.

For those cases with dependence in the occupancy status of sites between seasons, data can be analyzed with a model that explicitly describes the process underlying occupancy dynamics as a first-order Markov chain (MacKenzie *et al.* 2003), as we described in section 2.2.6. The MLEs of the occupancy estimators in the two-season Markovian model have the same expression and variance as when assuming independence (Appendix A.4). However the covariance is no longer zero and the variance of the difference in occupancy $\widehat{D}$ is $\text{var}(\widehat{D}) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\text{cov}(\hat{\psi}_1, \hat{\psi}_2)$. The covariance is larger than zero when the site occupancy status is positively correlated ($1 - \varepsilon > \gamma$). This means that the sample size formula in (3.4) provides a conservative design in case such dependence is present. The formula can be extended to the Markovian case as

$$S = (f_1 + f_2 - 2f_c)\left(\frac{z_{\alpha/2} + z_\beta}{\psi_1 - \psi_2}\right)^2,$$

where $f_c = \psi_1\{(1 - \varepsilon) - \psi_2\}$.

3.2.9 Applicability to multiple-season studies with more than two seasons

Very little has been published regarding the design of multiple-season occupancy studies. Two exceptions are MacKenzie (2005) and Field, Tyre & Possingham (2005), which explore scenarios involving a linear trend in occupancy on the logistic scale under the assumption of independence between seasons. MacKenzie (2005) presents a brief exploration showing how the coefficient of variation of the estimated trend parameter decreases as more seasons are added to the study (keeping the interval between seasons constant). Field, Tyre & Possingham (2005) use power analysis simulations to explore the optimal allocation of effort under a scenario involving three seasons and exponentially increasing survey costs.

In this section we have considered a scenario with two survey seasons. When the target number of seasons is moderately larger than two, sample size determination based on the change expected between the first and last season can still provide some useful (conservative) guidance, as we can normally expect the power not to be radically different to that obtained considering the trend across all seasons. To illustrate this we ran a small simulation study, assuming a scenario with a declining trend in occupancy (linear on the logit scale) over four seasons, so that $\text{logit}(\psi_i) = \beta_1 + \beta_t * (i-1)$, $i = 1, \ldots, 4$. We explored different scenarios of initial occupancy ($\psi_1 = 0.2$, 0.5 and 0.8) and overall decline between seasons one and four ($R_4 = 0.5$, 0.3 and 0.15), with $p = 0.5$ and $K = 3$. We run 5000 simulations for each scenario and estimated the power to detect a significant trend (i.e. $\hat{\beta}_t$ significantly different from zero in a Wald test) for the following sampling situations: (a) S sites surveyed in all four seasons, (b) S sites surveyed in the two first seasons, (c) S sites surveyed in the first and last seasons, (d) $2S$ sites surveyed in the two first seasons and (e) $2S$ sites surveyed in the first

and last seasons. In all cases, $S = 200$. Note that (a), (d) and (e) involve the same total survey effort ($4S$), while cases (b) and (c) involve half the amount ($2S$).

The results show that the power estimated when only the first and last seasons are considered is generally similar to that obtained from considering all four seasons (Table 3-4, compare columns a and c). In fact, if a linear trend is indeed expected and the focus is on estimating overall change, concentrating all the survey effort in the first and last seasons provides a more powerful design (Table 3-4, compare columns a and e). Note nevertheless that a design with various sampling seasons may be more robust for estimating a trend when there is noise because of variations from season to season on top of the trend and allows detecting departures from linearity.

Table 3-4 Power to detect a declining trend in occupancy (linear on the logit scale) when the trend is estimated based on (a) four seasons, (b,d) first and second season and (c,e) first and last season. In (a-c) $S = 200$ sites; in (d,e) all survey effort was concentrated in two seasons (i.e. 400 sites per season). β_t is the rate of seasonal change on the logit scale, corresponding to a proportional decline R_4 between seasons 1 and 4. Different scenarios of initial occupancy ψ_1 and R_4 are assessed with $K = 3$, $p = 0.5$, $\alpha = 0.05$. Computed from 5000 simulations.

		Seasons	1,2,3,4	1,2	1,4	1,2	1,4
		Nr. of sites	S	S	S	2S	2S
R_4	ψ_1	β_t	(a)	(b)	(c)	(d)	(e)
0.5	0.8	-0.597	1.000	0.412	1.000	0.724	1.000
	0.5	-0.366	0.995	0.314	0.992	0.565	1.000
	0.2	-0.270	0.707	0.138	0.669	0.249	0.927
0.3	0.8	-0.382	0.981	0.158	0.970	0.327	1.000
	0.5	-0.206	0.748	0.136	0.705	0.223	0.945
	0.2	-0.143	0.294	0.074	0.275	0.101	0.491
0.15	0.8	-0.211	0.523	0.062	0.456	0.118	0.781
	0.5	-0.101	0.254	0.060	0.233	0.088	0.409
	0.2	-0.066	0.099	0.048	0.097	0.058	0.162

3.3 Bayesian design and sequential methods

3.3.1 Background

In section 3.1 we have examined the issue of survey effort allocation in single-season occupancy studies. As we pointed out, in order to determine the optimum amount of replication to be used in such studies, knowledge about the parameters is needed as the information matrix is a function of their values. However, the values of the parameters are evidently not known. In fact, given that these are the object of estimation, it is reasonable to expect that there will be a considerable degree of uncertainty regarding their values. Designing a study based on poor initial estimates may translate into significant loss in estimator performance, as we illustrate below.

In this section we revisit the issue of survey effort allocation, exploring methods to obtain more robust designs in the face of uncertain initial parameter estimates, using the variance of the occupancy estimator as a criterion for design. One way of formally dealing with uncertainty in the initial parameter estimates is to follow a 'Bayesian experimental design' approach (Chaloner & Verdinelli 1995). This method involves specifying a prior distribution for the unknown parameters and selecting a design that optimizes the expectation of the chosen design criterion with regard to this prior. Another approach is to use 'sequential methods'. These involve dividing the experiment into a number of stages, with each subsequent stage being designed after updating the initial estimates based on the data collected up to that stage (e.g. Abdelbasit & Plackett 1983). The key idea is that the design can be steered towards the optimal design as we learn about the parameter values. Bayesian and sequential ideas can of course be combined within a single design problem (e.g. Ridout 1995).

3.3.2 Impact of poor initial estimates

The impact of poor initial estimates on the quality of the occupancy estimator can be assessed by comparing its asymptotic variance when the amount of replication (K) is chosen based on the initial estimates (ψ_0, p_0), with that obtained if the true parameter values (ψ, p) were known.

Table 3-5 shows such an assessment for three scenarios of initial estimates:

(a) $\psi_0 = 0.8, p_0 = 0.3,\ K = 8,$

(b) $\psi_0 = 0.5, p_0 = 0.5,\ K = 3,$

(c) $\psi_0 = 0.3, p_0 = 0.8,\ K = 2,$

and all combinations of true parameter values with ψ and p ranging from 0.2 to 0.8 in steps of 0.1, with K chosen to minimize the variance of $\hat{\psi}$ (Table 3-1a in page 79). As expected, the impact of having poor estimates is larger the further the initial estimates are from the true estimates, with the performance deterioration being generally larger the more different the chosen K is compared to the optimum K according to the true parameter values. For instance, when $\psi_0 = 0.5\ p_0 = 0.5\ (K = 3)$, performance loss is much greater if $\psi = 0.8\ p = 0.2$, which corresponds to an optimum amount of replication $K_{opt} = 8$, than if $\psi = 0.2\ p = 0.8$ for which $K_{opt} = 2$. Also, the results suggest that when p_0 is correct there is relatively little effect of mispecifying ψ_0, compared to the effect of mispecifying p_0 when ψ_0 is correct. Note that, assuming large sample size, these results are independent of the total survey effort E, as E is only a scaling factor in the expression of the asymptotic variance of the occupancy estimator (3.1).

Table 3-5 Robustness of the 'optimal' occupancy study design to poor initial estimates for different scenarios of true occupancy and detectability (ψ, p) and three scenarios of initial assumed parameter values (ψ_0, p_0), marked in bold font. The values shown are the ratio between the asymptotic variance of $\hat{\psi}$ that would be obtained if the design would be based on the true parameter values and that chosen based on the assumed initial parameter values. The design criterion is to minimize the asymptotic variance of ψ.

(a)

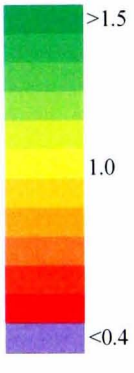
		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	1.00	1.00	1.00	0.99	0.97	0.92	0.82
	0.3	0.82	0.85	0.88	0.91	0.95	0.98	**1.00**
	0.4	0.64	0.66	0.68	0.72	0.77	0.82	0.89
p	0.5	0.49	0.51	0.53	0.56	0.60	0.64	0.71
	0.6	0.39	0.41	0.43	0.45	0.47	0.50	0.56
	0.7	0.31	0.32	0.33	0.34	0.36	0.40	0.44
	0.8	0.27	0.27	0.28	0.28	0.29	0.30	0.33

(b)

		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	0.53	0.49	0.45	0.41	0.36	0.30	0.23
	0.3	0.81	0.77	0.73	0.67	0.60	0.52	0.41
	0.4	0.99	0.96	0.93	0.89	0.85	0.76	0.64
p	0.5	1.00	1.00	1.00	**1.00**	1.00	0.95	0.86
	0.6	0.92	0.95	1.00	1.00	1.00	1.00	1.00
	0.7	0.79	0.80	0.82	0.85	0.89	0.96	1.00
	0.8	0.71	0.72	0.72	0.74	0.75	0.78	0.84

(c)

		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	0.27	0.25	0.23	0.20	0.17	0.14	0.11
	0.3	0.46	0.43	0.40	0.35	0.31	0.26	0.20
	0.4	0.69	0.64	0.59	0.54	0.49	0.41	0.32
p	0.5	0.88	0.84	0.80	0.75	0.70	0.60	0.48
	0.6	1.00	1.00	1.00	0.95	0.89	0.80	0.70
	0.7	1.00	1.00	1.00	1.00	1.00	1.00	0.92
	0.8	1.00	**1.00**	1.00	1.00	1.00	1.00	1.00



3.3.3 *Bayesian design*

Here we provide an example of optimal replication determination following a Bayesian experimental design approach. We characterized the initial knowledge about the parameters using either independent uniform priors or a bivariate logit-normal prior, and identified the value of K that minimizes the expected asymptotic variance of the occupancy estimator over that prior

$$\mathbb{E}[\mathrm{var}(\hat{\psi}_1)] = \iint \frac{K\psi_0}{E}\left\{(1-\psi_0) + \frac{1-p_0^*}{p_0^* - Kp_0(1-p_0)^{K-1}}\right\} \cdot f(\psi_0, p_0)\, d\psi_0\, dp_0, \tag{3.5}$$

where $f(\psi_0, p_0)$ denotes the probability density function of the prior. The integration was performed numerically using the in-built MATLAB function *dblquad*, which evaluates a double integral over a rectangle.

As expected, the resulting optimum K is an intermediate value compared to those that would have been selected based on particular parameter values within the prior (Figure 3-13). Of course, the choice of prior has an impact. If the prior is too concentrated then there is the risk of decreased performance if the true parameter values happen to lie outside its range, as we have seen in Table 3-5 for the extreme case of a single initial estimate value. On the other hand, if the prior is too dispersed then this can result in efficiency loss as the design is no longer specifically constructed for the underlying scenario. In the following section we leave Bayesian design and move to a different approach (sequential design), but in section 3.3.5 we will see how both methods can be integrated.

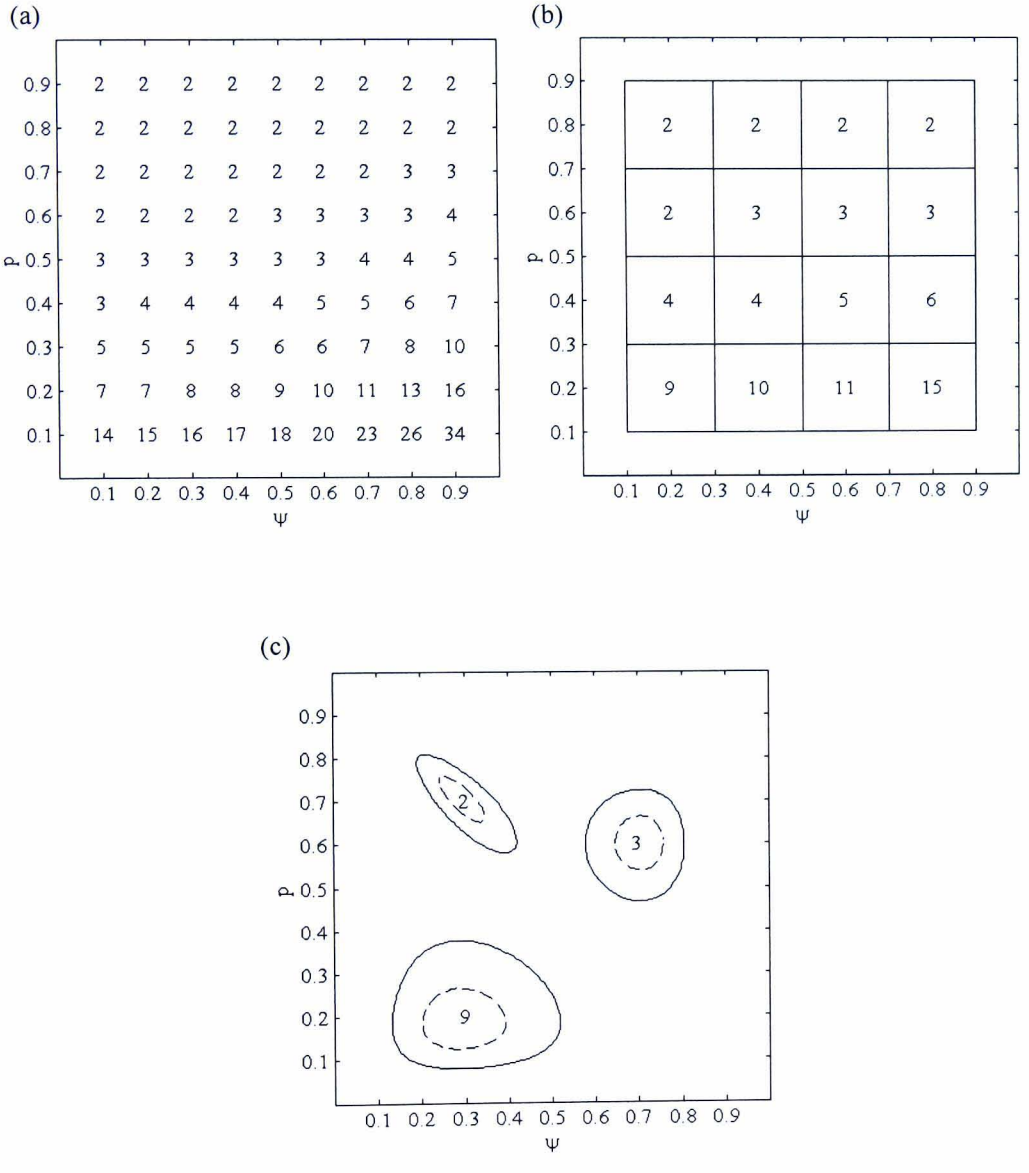


Figure 3-13 Examples of Bayesian optimal design to minimize the variance of $\hat{\psi}$. The numbers represent the optimal K for each case. The initial estimate information used is: (a) a single value (each point in the grid; this case is included for reference), (b) a uniform prior (each square) and (c) a bivariate logit-normal. In (c) the three examples correspond to (i) $\boldsymbol{\mu}$ = [logit(0.3), logit(0.2)], $\boldsymbol{\Sigma}$ = [0.15, 0; 0, 0.15], (ii) $\boldsymbol{\mu}$ = [logit(0.7), logit(0.6)], $\boldsymbol{\Sigma}$ = [0.05, 0; 0, 0.05], and (iii) $\boldsymbol{\mu}$ = [logit(0.3), logit(0.7)], $\boldsymbol{\Sigma}$ = [0.05, -0.04; -0.04, 0.05], where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively the vector of means and the variance-covariance matrix on the logit scale; contours enclose 0.5 and 0.95 probability. The prior can represent an educated guess about the parameter values or the estimates resulting from a previous study.

3.3.4 Two-stage sequential design

Here we consider a two-stage sequential design for occupancy studies, in which $100Z\%$ of the total survey effort (E) is employed in the first stage (or pilot phase) and the remaining effort is used in the second stage. For simplicity we assume that the effort corresponding to the second stage is allocated to new sampling sites (i.e. sites different from those sampled in stage 1). The optimal amount replication for stage 2 (K_2), based on the updated parameter information, might be different from that identified for stage 1 (K_1) based on the initial parameter estimates (Figure 3-14).

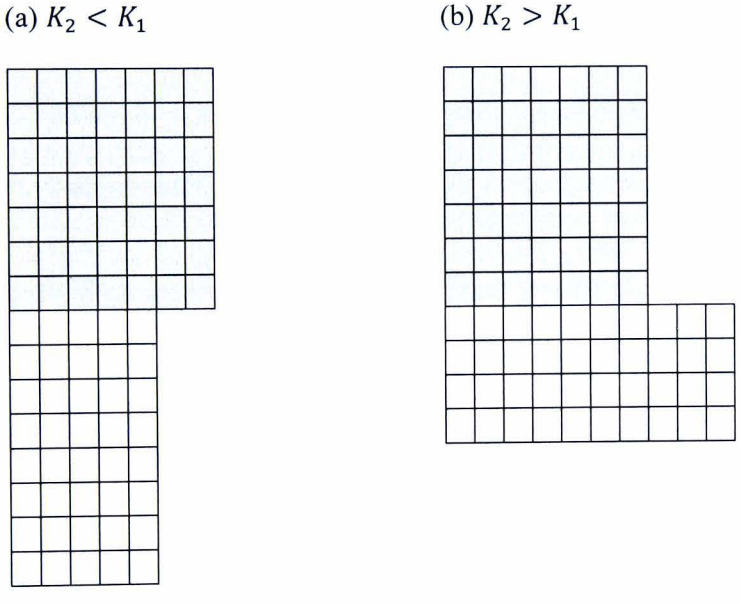


Figure 3-14 Hypothetical examples of a two-stage design where each row represents a sampling site and each square within a row a survey visit. Stage 1 is represented by grey squares and Stage 2 by white squares. In (a) Stage 2 uses less replication than in Stage 1, while in (b) it is the other way around.

We consider the following two-stage procedure:

(i) Based on the initial point estimates (ψ_0, p_0), choose the design for stage 1, i.e. identify the optimal K_1, and from this $S_1 = \text{round}[EZ/K_1]$;

(ii) Collect data (summarized by S_{d_1}, d_{T_1}) and analyze them to obtain the estimates $(\hat{\psi}_1, \hat{p}_1)$;

(iii) Based on the updated parameter knowledge (i.e. estimates $\hat{\psi}_1, \hat{p}_1$), choose the design for stage 2, i.e. identify the optimal K_2, and from this $S_2 = \text{floor}[E(1-Z)/K_2]$;

(iv) Collect data (summarized by S_{d_2}, d_{T_2}) and analyze the full data set, from both stages, to obtain MLEs $(\hat{\psi}_2, \hat{p}_2)$.

The likelihood for the resulting two-stage design is the product of the contributions from the two stages

$$L(\psi, p) = \left\{\psi^{S_{d_1}} p^{d_{T_1}} (1-p)^{K_1 S_{d_1} - d_{T_1}}\right\}\left(1 - \psi p^*_{K_1}\right)^{S_1 - S_{d_1}}$$
$$\times \left\{\psi^{S_{d_2}} p^{d_{T_2}} (1-p)^{K_2 S_{d_2} - d_{T_2}}\right\}\left(1 - \psi p^*_{K_2}\right)^{S_2 - S_{d_2}},$$

where $p^*_{K_1} = 1 - (1-p)^{K_1}$ and $p^*_{K_2} = 1 - (1-p)^{K_2}$. Given the expressions for the expected information matrix for a single-stage design (2.12), the elements of the information matrix for the two-stage design are

$$
\begin{aligned}
\mathbf{I}[1,1] &= \frac{S_1 p_{K_1}^*}{\psi(1-\psi p_{K_1}^*)} + \frac{S_2 p_{K_2}^*}{\psi(1-\psi p_{K_2}^*)}, \\
\mathbf{I}[1,2] &= \frac{S_1 K_1 (1-p)^{K_1-1}}{(1-\psi p_{K_1}^*)} + \frac{S_2 K_2 (1-p)^{K_2-1}}{(1-\psi p_{K_2}^*)}, \\
\mathbf{I}[2,2] &= \frac{S_1 K_1 \psi}{p(1-p)} \left\{ 1 - \frac{K_1 p (1-p)^{K_1-1}(1-\psi)}{1-\psi p_{K_1}^*} \right\} \\
&\quad + \frac{S_2 K_2 \psi}{p(1-p)} \left\{ 1 - \frac{K_2 p (1-p)^{K_2-1}(1-\psi)}{1-\psi p_{K_2}^*} \right\},
\end{aligned}
\tag{3.6}
$$

and the asymptotic variance of the final occupancy estimator is $\mathrm{var}(\hat{\psi}_2) = \mathbf{I}^{-1}[1,1] = \mathbf{I}[2,2]/(\mathbf{I}[1,1] \cdot \mathbf{I}[2,2] - \mathbf{I}[1,2] \cdot \mathbf{I}[1,2])$. This expression is conditional on the outcome of the first stage $(\hat{\psi}_1, \hat{p}_1)$, as this determines the design for the second stage (K_2 and S_2). We can compute the unconditional variance of the occupancy estimator by taking the expectation of the variance over all the possible outcomes of the first stage

$$
\mathbb{E}[\mathrm{var}(\hat{\psi}_2)] = \sum_{\forall \hat{\psi}_1} \sum_{\forall \hat{p}_1} \mathrm{var}(\hat{\psi}_2 | \psi, p, K_1, K_2, S_1, S_2) \cdot \Pr(\hat{\psi}_1, \hat{p}_1 | \psi, p, S_1, K_1), \tag{3.7}
$$

This expression is ultimately a function of the initial estimates ψ_0, p_0, the true parameter values ψ, p, and of E and Z.

To assess the efficiency of the two-stage design with respect to a single-stage design we computed the ratio between the expected asymptotic variance in (3.7) and the asymptotic variance for a single-stage design, as in (2.13), for the same scenarios of initial estimates and true parameter values as in section 3.3.2. Assuming large sample size, we considered that the MLEs resulting from the first stage in the two-stage design $(\hat{\psi}_1, \hat{p}_1)$ were distributed according to a bivariate logit-normal with mean [logit(ψ),

logit(p)] and variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{L}}$ on the logit scale (and therefore we replaced the double summation in the expectation by a double integration). We derived $\boldsymbol{\Sigma}_{\mathbf{L}}$ from (2.13) using the delta method. For each value of $(\hat{\psi}_1, \hat{p}_1)$ we identified the optimum K_2 that minimizes the variance of $\hat{\psi}_2$ and finally computed the actual variance of $\hat{\psi}_2$ given the true parameter values. These efficiency calculations were carried out in MATLAB, and the integration was carried out numerically using *dblquad.*

Table 3-6 shows the efficiency of the two-stage design for a scenario of total effort $E = 1000$ and with an equal allocation of effort between the two stages (i.e. $Z = 0.5$). It can be immediately seen than the sequential approach has great potential for performance improvement with respect to the single-stage design (green cells). Of course, the efficiency of the two-stage design is greater the more different the initial estimates are with respect to the real parameter values (or, rather, the greater the difference in the associated K's is), as these are the cases when there is a greater loss in performance in the single-stage design. The two-stage design can result in performance loss if the initial estimates lead to the 'correct' optimum K. In this example there is only ever very marginal loss but we will later see how this can be more pronounced in other scenarios. Table 3-7 shows the efficiency of the two-stage design with respect to the ideal design scenario, i.e. a single-stage design based on the true parameter values. This table, which is equal to the element-wise product of Tables 3-5 and 3-6, shows that the two-stage design can considerably approach the best achievable performance. Of course there can still be some loss with respect to the ideal case, as indicated by the red cells in the tables, but it is evident that the two-stage design is significantly more robust to poor initial estimates than the single-stage design (evaluated in Table 3-5).

Table 3-6 Efficiency of the two-stage design with respect to the single-stage design for different scenarios of true occupancy and detectability (ψ, p) and three scenarios of initial assumed parameter values (ψ_0, p_0), marked in bold font. The values shown are the ratio between the expected asymptotic variance of $\hat{\psi}$ for the single-stage design and the asymptotic variance of $\hat{\psi}$ for the two-stage design. The design criterion is to minimize the asymptotic variance of ψ. $E = 1000$ and $Z = 0.5$.

(a)

		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	0.96	0.97	0.97	0.98	0.99	1.03	1.09
	0.3	1.14	1.12	1.09	1.05	1.02	1.00	**0.98**
	0.4	1.41	1.36	1.30	1.25	1.19	1.13	1.06
p	0.5	1.72	1.66	1.58	1.49	1.41	1.32	1.22
	0.6	1.99	1.94	1.88	1.80	1.69	1.55	1.43
	0.7	2.21	2.18	2.14	2.08	2.01	1.89	1.70
	0.8	2.37	2.36	2.34	2.31	2.27	2.21	2.09

(b)

		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	1.57	1.69	1.81	1.95	2.14	2.40	2.83
	0.3	1.14	1.19	1.25	1.33	1.43	1.58	1.83
	0.4	0.99	1.01	1.04	1.07	1.12	1.20	1.34
p	0.5	0.98	0.98	0.98	**0.99**	1.00	1.03	1.09
	0.6	1.07	1.06	1.03	1.01	0.99	0.99	1.00
	0.7	1.16	1.15	1.13	1.11	1.09	1.05	1.00
	0.8	1.21	1.20	1.20	1.19	1.18	1.15	1.12

(c)

		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	2.71	2.98	3.24	3.53	3.90	4.42	5.33
	0.3	1.81	1.95	2.09	2.27	2.49	2.82	3.36
	0.4	1.34	1.41	1.50	1.60	1.75	1.95	2.31
p	0.5	1.09	1.13	1.18	1.24	1.33	1.46	1.68
	0.6	0.99	1.00	1.02	1.05	1.09	1.16	1.29
	0.7	0.99	0.99	0.99	0.99	0.99	1.01	1.07
	0.8	1.00	**1.00**	1.00	1.00	1.00	1.00	1.00

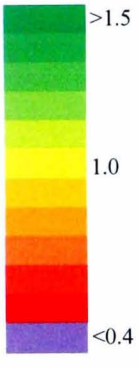


Table 3-7 Efficiency of the two-stage design with respect to the ideal case (i.e. single-stage design with known parameter values) for different scenarios of true occupancy and detectability (ψ, p) and three scenarios of initial assumed parameter values (ψ_0, p_0), marked in bold font. The values shown are the ratio between the asymptotic variance of $\hat{\psi}$ for the ideal design and the expected asymptotic variance of $\hat{\psi}$ for the two-stage design. The design criterion is to minimize the asymptotic variance of ψ. $E = 1000$ and $Z = 0.5$.

(a)

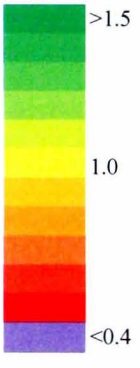
		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	0.96	0.97	0.97	0.97	0.96	0.94	0.90
	0.3	0.94	0.95	0.96	0.96	0.97	0.98	**0.98**
	0.4	0.90	0.89	0.89	0.90	0.92	0.92	0.94
p	0.5	0.84	0.84	0.83	0.83	0.85	0.85	0.87
	0.6	0.77	0.79	0.82	0.81	0.79	0.77	0.80
	0.7	0.68	0.69	0.70	0.71	0.73	0.76	0.75
	0.8	0.64	0.64	0.64	0.65	0.66	0.67	0.69

(b)

		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	0.83	0.83	0.82	0.80	0.76	0.72	0.65
	0.3	0.92	0.92	0.91	0.89	0.86	0.82	0.76
	0.4	0.98	0.97	0.97	0.96	0.95	0.91	0.86
p	0.5	0.98	0.98	0.98	**0.99**	1.00	0.97	0.94
	0.6	0.99	1.01	1.03	1.01	0.99	0.99	1.00
	0.7	0.91	0.92	0.93	0.95	0.97	1.01	1.00
	0.8	0.86	0.86	0.87	0.87	0.89	0.90	0.93

(c)

		ψ						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
	0.2	0.73	0.74	0.73	0.71	0.68	0.64	0.58
	0.3	0.84	0.84	0.83	0.80	0.77	0.73	0.66
	0.4	0.92	0.91	0.89	0.87	0.85	0.80	0.74
p	0.5	0.95	0.95	0.94	0.93	0.93	0.88	0.82
	0.6	0.99	1.00	1.02	1.00	0.97	0.93	0.90
	0.7	0.99	0.99	0.99	0.99	0.99	1.01	0.98
	0.8	1.00	**1.00**	1.00	1.00	1.00	1.00	1.00



In the above example we have assumed an equal allocation of effort between the two stages ($Z = 0.5$). However we can expect the efficiency of the two-stage design to vary with Z. In fact, we can expect that there is an optimum Z given that the larger the value of Z, the more effort is dedicated to improve the knowledge of the parameter values during the first stage, but then less effort remains to profit from that improved knowledge through a better designed second stage. Figure 3-15 displays the efficiency of the two-stage design with respect to the single-stage design as a function of Z for the same three cases of initial estimates as above and four scenarios of true parameter values, ψ and p,

(t1) $\psi = 0.3, p = 0.8, K_{opt} = 2,$ (t3) $\psi = 0.3, p = 0.3, K_{opt} = 5,$

(t2) $\psi = 0.8, p = 0.3, K_{opt} = 8,$ (t4) $\psi = 0.8, p = 0.8, K_{opt} = 2,$

where K_{opt} indicates the replication that would have been the optimal in each case. As Z increases all the curves tend to unity, which is expected given that the case $Z = 1$ is in fact a single-stage design. Some of the curves also tend to bend downwards as Z gets small, and therefore there is an optimal Z although, as might be expected, its value varies depending on the scenario. Note that in Figure 3-15a the cases (t1) and (t4) (blue and green curves) keep high efficiency even for low Z. This is because in these cases K_{opt} takes the same value (= 2) in a relatively large region around (ψ, p) and therefore they are less affected by an increase in the variance of the estimates from stage 1. Of course, when working with smaller efforts, we can expect all the curves to bend downwards more prominently for low Z, as then the pilot stage involves a smaller sample size which would be less informative or even misleading.

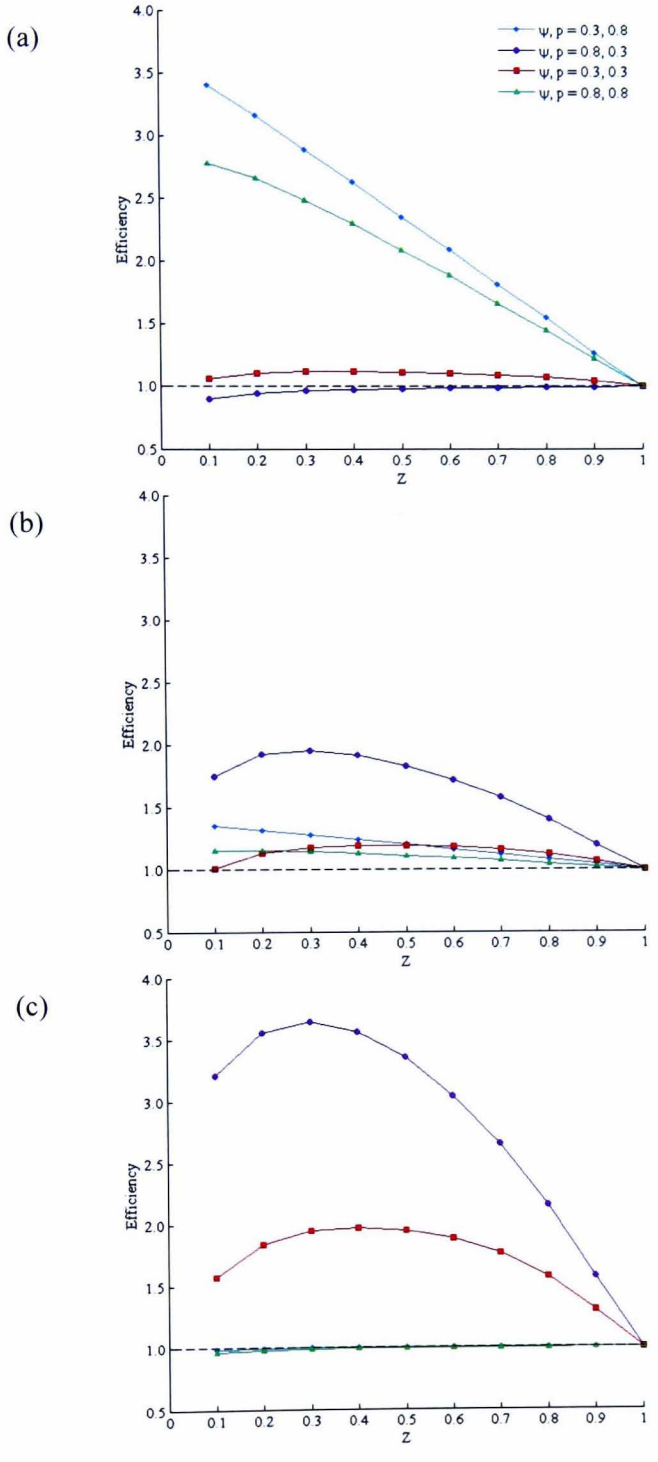


Figure 3-15 Efficiency of the two-stage design with respect to the single-stage design for varying effort allocation between the stages (Z), for four scenarios of true occupancy and detectability (ψ, p) and three scenarios of initial assumed parameter values: (a) $\psi_0 = 0.8$, $p_0 = 0.3$, (b) $\psi_0 = 0.5$, $p_0 = 0.5$, (c) $\psi_0 = 0.3$, $p_0 = 0.8$. The design criterion is to minimize the asymptotic variance of ψ. $E = 1000$.

Since the above calculations rely on large-sample approximations, we ran simulations to assess the actual performance improvement achieved for the same scenarios, first with survey effort $E = 1000$, and then reducing it to $E = 500$ and 250. For simplicity, we selected the optimal level of replication for each stage based on the large-sample variance expressions but, in practice, this could have been also obtained via simulations. In the event of obtaining a history with all zeros in the first stage, we used $\psi_1 = 0.2$ and $p_1 = 0.2$ as updated parameter values. The efficiency of the two-stage procedure was measured as the ratio between the mean square error (MSE) of its occupancy estimator and that obtained from a single-stage design, from 5000 simulations.

The results (Figure 3-16) show that for $E = 1000$ the actual efficiency of the two-stage design is in general very similar to that calculated by numerical evaluation of (3.7), and therefore we can also expect these to match for larger effort. The main difference is in case (t3) in Figure 3-16c (red line), where the MSE of the single-stage design is underestimated by the large-sample approximations. As the effort decreases, the differences between the simulations and the numerical evaluation become more noticeable but the patterns remain broadly the same. The downwards bending of the efficiency curves for low Z is more prominent as we expected, and this effect is more evident in the simulations, which suggests that in these cases the large-sample approximations underestimate the MSE of the stage 1 estimator in the two-stage design. Once again there are noticeable differences in Figure 3-16c for case (t3) (red line), as well as for (t2) (purple line); in these two cases K_{opt} is larger than the one suggested by (ψ_0, p_0). The differences are essentially driven by the inaccuracy of the large-sample approximation for the single-stage design (leading to an under/overestimation of its

MSE). It is interesting to note that the large-sample approximation of the MSE for the two-stage design is fairly accurate in these cases.

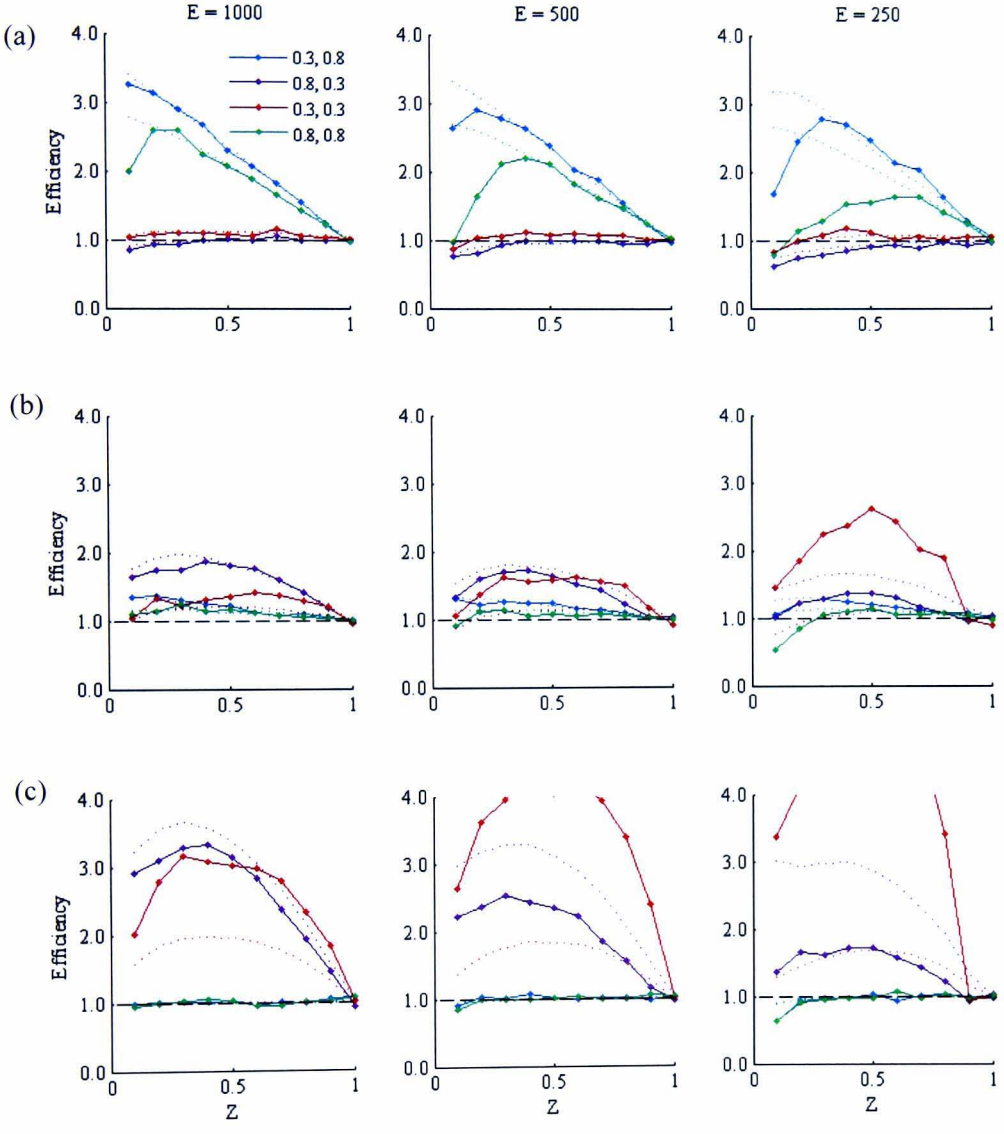


Figure 3-16 Efficiency of the two-stage design with respect to the single-stage design obtained from 5000 simulations for effort levels $E = 1000$, 500 and 250. Dotted lines show the corresponding results derived from large-sample calculations for comparison. The graphs display the efficiency achieved for varying effort allocation between the stages (Z), for four scenarios of true occupancy and detectability (ψ, p) and three scenarios of initial assumed parameter values: (a) $\psi_0 = 0.8$, $p_0 = 0.3$, (b) $\psi_0 = 0.5$, $p_0 = 0.5$, (c) $\psi_0 = 0.3$, $p_0 = 0.8$. The design criterion is to minimize the asymptotic variance of ψ. In the graphs we kept the same scale on the Y-axis for comparison purposes, but note that in some cases there are values larger than 4.

The differences between the simulations and the large-sample calculations suggest that, when E is of the order of a few hundred units, there is value in addressing study design via simulations. Finally note that, for simplicity, in our simulations the boundary estimates from the first stage were not treated in any particular way, although in practice one would be wary about directly using these to inform the design of stage 2 as they can lead to extreme values for K_2.

3.3.5 Optimal determination of Z for the two-stage design

We have assumed up to now that the decision on how to allocate the survey effort into the two stages (i.e. the value of Z) was already made. In this section we illustrate how Bayesian ideas can be combined into the sequential design approach to identify the optimal choice of Z, given the initial knowledge of the parameter values ψ_0, p_0. Here we assume that we have a prior distribution characterizing this knowledge.

The target is to choose a design that provides the best overall performance and thus to identify the value of Z that provides the lowest expected occupancy estimator variance, integrated over the prior. Assuming large-sample results again, this quantity is computed as

$$\mathbb{E}[\mathrm{var}(\hat{\psi}_2)] = \int_{p_0}\int_{\psi_0}\int_{\hat{p}_1}\int_{\hat{\psi}_1} \mathrm{var}(\hat{\psi}_2|\psi_0, p_0, K_1, K_2, S_1, S_2) \cdot f(\hat{\psi}_1, \hat{p}_1|\psi_0, p_0, S_1, K_1) \cdot f(\psi_0, p_0)\, d\hat{\psi}_1 d\hat{p}_1 d\psi_0 dp_0,$$

and is a function of E, Z and the prior $f(\psi_0, p_0)$.

Figure 3-17 displays the results of such computation for a variety of scenarios of prior knowledge and two levels of effort $E = 500$ and 1000. In all cases we assumed a uniform prior for simplicity. For the numerical integration we used MATLAB's function *triplequad* to integrate with respect to $\hat{\psi}_1$, $\hat{p}_1$and ψ_0 and then used the function *trapz* to integrate with respect to p_0 with step 0.01. The design for the first stage (i.e. K_1) was obtained following a Bayesian approach (as in section 3.3.3). However our approach is not fully Bayesian in that the design for the second stage (i.e. K_2) was based solely on the point estimates from the first stage. This was due to implementation issues (namely limitations in the nesting of integration functions in MATLAB), but does not need to be so. Note also that the choice of K_1 is made without consideration of the two-stage nature of the design. One could instead do a two-dimensional optimization, i.e. evaluating the combination of K_1 and Z most suitable given the prior.

It is obvious that if our prior consists of a single point (Figure 3-17, bottom right), the optimal choice for Z is 1, as we are implicitly suggesting that we do not have uncertainty regarding the initial values. In such an unrealistic case the best design is of course to do a single-stage design. The difference in performance with the two-stage design is however not very large for a range of values of Z, as illustrated by the flatness of the curves. As the prior gets wider, and therefore we have more uncertainty, there is more benefit in using a two-stage design, as illustrated by the U-shaped curves. The location of the optimum depends on the case. It changes with the prior and also with the amount of total effort, with the optimum Z shifting to the right (higher values) as the total effort gets smaller.

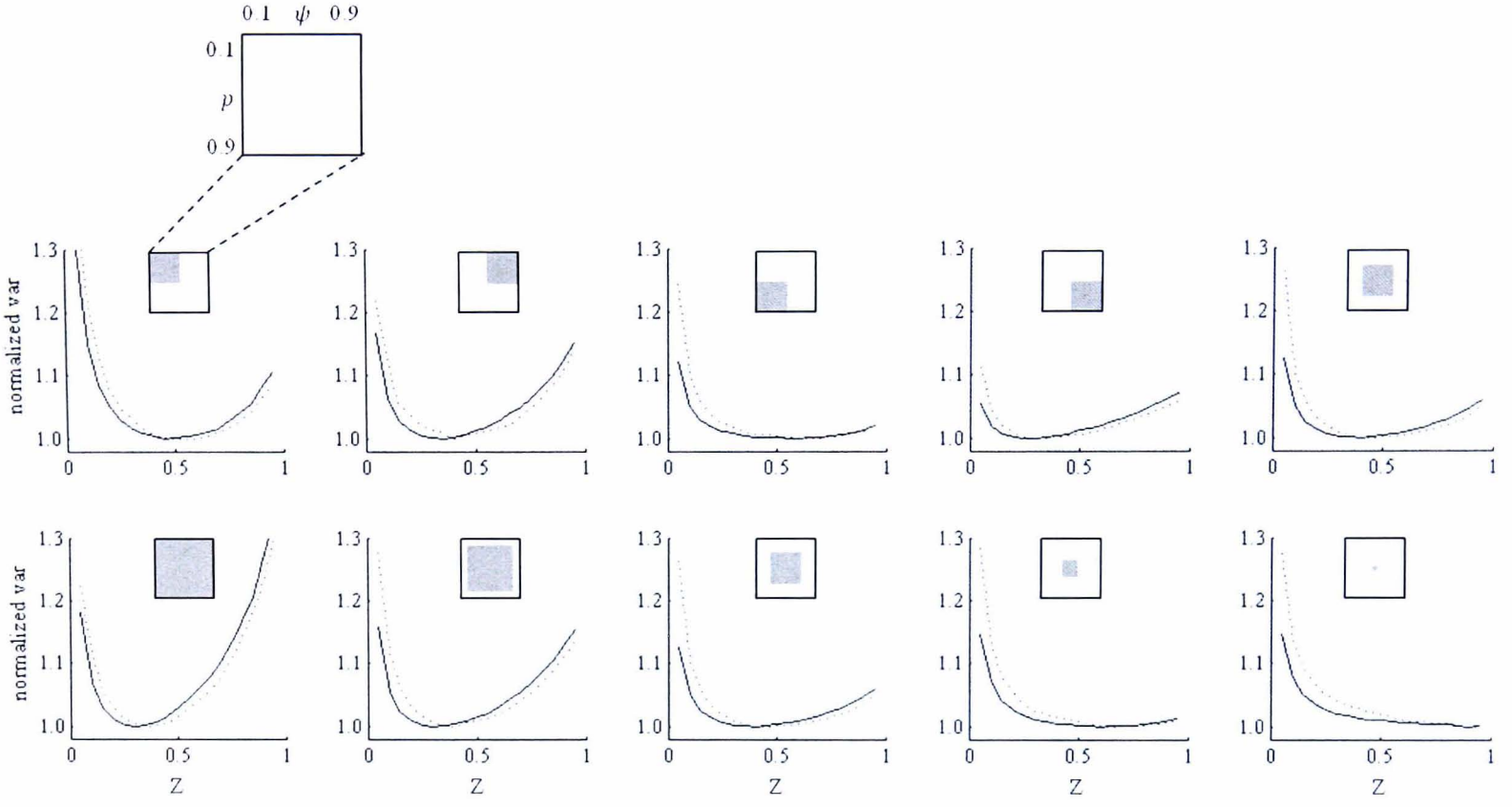


Figure 3-17 Determination of the optimal effort allocation in a two-stage design following a Bayesian approach. The curves represent the expected variance of the occupancy estimator, normalized to the minimum variance for each scenario, as a function of Z. Each plot represents a scenario of prior knowledge (uniform), as indicated by the inset. Two levels of total effort are displayed: $E = 500$ (dotted), 1000 (solid).

3.4 Sampling for spatial replication

3.4.1 Background

In section 2.2.2 we briefly touched on the issue of the 'closure assumption' made by the basic occupancy model, in connection with the description of the model proposed by Hines *et al.* (2010) to account for Markovian dependence in species availability between survey replicates. Here we come back to this issue to discuss the effect that sampling with replacement has on the independence among spatial replicates drawn from partially occupied sites, and consequently on the occupancy estimator.

As noted in section 2.2.2, when the sampling scheme is based on temporal replication (i.e. assessing the same location at different points in time), the 'closure assumption' requires site occupancy status to remain unchanged throughout the sampling season. MacKenzie *et al.* (2006) discuss the impact of violating this assumption. If the species is not always available at occupied sites, and availability varies non-randomly during the sampling period, then the occupancy estimator ($\hat{\psi}$) may be biased. However, if changes in availability among survey replicates occur completely at random, no bias is induced, although the interpretation of $\hat{\psi}$ somewhat changes, as it reflects the probability that the site is used by the species, rather than permanently occupied. Detection probability decreases as it is now composed of two elements: (i) the availability of the species at the site during a survey, that is, the probability that the species is using the site at the time of survey (θ), and (ii) the probability of detecting the species given it is available at the site during the survey (which we denote here by p_a to highlight that it is conditional on availability). The detection probability estimator ($\hat{p}$) would in fact be estimating the product θp_a. On the other extreme, if survey replicates are carried out

sufficiently close in time that there are no changes in availability between them, then the occupancy estimator would estimate the product $\psi\theta$ while the detection probability estimator would estimate p_a. Note that, although with the basic occupancy model it is not possible to estimate these three parameters (ψ, θ, p_a) separately, a sampling protocol based on the use of simultaneous independent detection methods (Nichols *et al.* 2008) or with surveys carried out at two temporal scales (so called 'robust design'), allows this to be done.

Kendall & White (2009) (from now on KW09) review the single-season occupancy model and its robustness to violations of the closure assumption, focusing on the case when data are collected by spatial subsampling within sites (Figure 2-1). In this situation the closure assumption is violated if the species is present at a fraction of sampling subunits within occupied sites. KW09 used simulations to evaluate the performance of the occupancy estimator under different sampling schemes in this situation. They considered sampling with and without replacement. Under sampling with replacement, previously assessed spatial subunits can be revisited in subsequent replicate surveys. Under sampling without replacement, spatial subunits are not revisited, that is, each replicate survey involves assessing a spatial subunit that has not been assessed previously. Therefore the number of replicate surveys cannot exceed the total number of spatial subunits per site. KW09 distinguish what they call 'exhaustive sampling' as a third sampling approach. In 'exhaustive sampling' all available spatial subunits are visited once. Based on their results they argue that absence of the species from a subset of spatial subunits may induce positive bias in the site occupancy estimator if locations are sampled without replacement and the species is not highly mobile. Consequently they give a general study design recommendation of sampling with replacement.

While raising an interesting issue, KW09 only consider the specific scenario where a constant fraction of spatial subunits is occupied within each occupied site (Figure 3-18 – Scenario A). Here we reexamine their recommendations considering a second scenario which arguably can often provide a more realistic description of ecological systems, whereby each spatial subunit at occupied sites has a given probability of being occupied by the species, regardless of the status of the other subunits (Figure 3-18 – Scenario B). This results in a variable number of occupied subunits within each occupied site, potentially ranging from zero to the number of subunits under consideration, as dictated by a binomial distribution. Like the scenario considered by KW09, the one proposed here describes a system with occupancy acting at two spatial scales, but without imposing the restriction of a fixed number of occupied subunits within each sampling site, as in KW09. This can often be a reasonable assumption as sampling sites are defined by the study and are therefore not intrinsically tied with a constant proportion of occupied subunits.

For example, let us assume that we sample forest patches (sites) looking for a bird species at certain landscape features (e.g. nest boxes), which define our spatial subunits. Scenario B can provide an appropriate description of the system, having a constant probability of each nest box (subunit) being occupied, rather than a fixed proportion of nest boxes being occupied in each site. Similarly, this scenario could be suitable to describe data collected during surveys based on tracks, if the species is equally likely to leave tracks at any spatial subunit within an occupied site, instead of the tracks always covering the same proportion of subunits within a site.

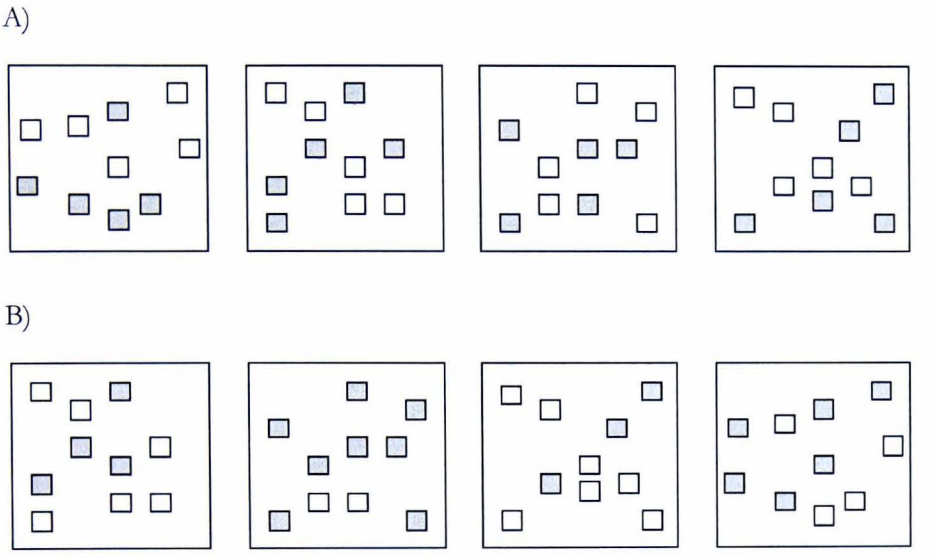


Figure 3-18 Four occupied sites and their sampling subunits in a hypothetical occupancy study with spatial replication. Shaded subunits indicate presence of the species. In scenario A) the species is present at a fixed proportion of subunits (5/10 = 0.5) in all occupied sites. Scenario B) was generated using a fixed probability (0.5) of subunit occupancy within occupied sites.

3.4.2 Simulation study

Simulations were used to investigate the occupancy estimator properties in the two scenarios described above under different sampling schemes, with and without replacement, comparing the results to those obtained by KW09. Since 'exhaustive sampling' is a particular case of sampling without replacement, it was treated as such and not separately. The following assumptions were made

(i) each replicate survey involves surveying one spatial subunit,

(ii) the timing of surveys and characteristics of the species are such that the occupancy status of each spatial subunit ('species availability' in KW09) does not change between replicate surveys, that is, replicate surveys are

conducted close in time compared to the speed of movement of the species or the decay rate of tracks in the case of surveys based on tracks,

(iii) the probability of detecting the species at an occupied subunit is independent of other surveys carried out in that subunit.

To be able to compare the results with those presented by KW09, the same parameter values were used, and then variations of these were explored. Sampling with and without replacement of 1000 sites each containing $N = 10$ spatial subunits was simulated. The number of occupied spatial subunits in occupied sites varied from 1 to 10, fixed for scenario A and on average for scenario B (i.e. 'a out of 10' occupied subunits in scenario A and a 'probability of subunit occupancy' of $a/10$ in scenario B). The number of replicate surveys per sampling site, K, varied from 5 to 10 (when sampling without replacement the latter corresponds to the 'exhaustive sampling' in KW09). Simulations were run with 1000 iterations, occupancy probability $\psi = 0.5$ and $p_a = 0.5$. To evaluate the performance of the occupancy estimator its RMSE was calculated. The mean of the estimator was also computed to produce plots that are directly comparable to those provided by KW09. To assess the relevance of sampling with or without replacement when the number of spatial subunits per site is large, the simulations were repeated with ten times more spatial subunits per site ($N = 100$), keeping the same average/fixed proportion of occupied subunits in occupied sites (number of occupied subunits a from 10 to 100 in steps on 10), and the same number of replicate surveys per sampling site (K from 5 to 10).

3.4.3 Results

The simulation results confirm that, assuming the system is well represented by scenario A, sampling without replacement increases the RMSE in the estimator of occupancy (Figure 3-19), introducing a positive bias (Figure 3-20), as discussed by KW09. However, when the simulated scenario is B, sampling with replacement is the approach that introduces bias, and therefore in this case it is best to sample without replacement. The bias induced is negative, thus underestimating occupancy. One way to interpret this is that, if the same subunit is sampled more than once, then there is no independence between the species availability status of those replicates, as it is the same. In an extreme case, if only one subunit is repeatedly sampled, then all the survey replicates have the same availability status and, as discussed in 3.4.1, the occupancy estimator estimates the product $\psi\theta$. When sampling with replacement, the induced lack of independence in the availability status of some replicates causes the occupancy estimator to estimate a value between ψ and $\psi\theta$, and so to be negatively biased.

As illustrated by the curves, scenario B without replacement (B1) and scenario A with replacement (A2) produce identical results in terms of properties of the occupancy estimator. Indeed both cases are mathematically equivalent (see Appendix A.3). This is not surprising as in practice both situations result in each sampling subunit drawn from the population having a probability of being occupied that is independent of the status of other subunits. In these two cases the observed RMSE value can be approximated by the asymptotic variance of the occupancy estimator in (2.13), using as detectability parameter p the product of the probability of subunit occupancy in occupied cells (a/N) and the probability of detection given species availability (p_a). The asymptotic variances of the occupancy estimator for (i) $a = 1, K = 5$, (ii) $a = 1, K = 10$, (iii) $a =$

10, $K = 5$, and (iv) $a = 10$, $K = 10$, are respectively 0.01738, 0.00373, 0.00027 and 0.00025 which, under the assumption of unbiasedness due to large sample size, correspond to RMSE values of 0.1318, 0.0610, 0.0164 and 0.0158. These values are close to those in Figure 3-19.

Note that our results do not exhibit the downwards bending of the curves for case A observed by KW09 in their Figure 2b when the number of subunits occupied within occupied sites is low and few subunits are sampled without replacement. This bending would suggest that in these cases the bias of the estimator is smaller if fewer samples are collected, a result which is counter-intuitive. Instead the simulations shown here indicate that, as expected, it is always best to have a larger number of replicates. Occupancy estimates at the boundary of the parameter space ($\hat{\psi} = 1$) were always obtained for the case where there was only one occupied subunit at all occupied sites, that is, when there was at most one detection in the history per occupied site (i.e. $S_d = d_T$). Indeed these results match the theory regarding the conditions for obtaining boundary estimates under the basic occupancy model (section 2.1.3), thus supporting the validity of our simulations.

Based on these results it can be argued that the general recommendation given by Kendall & White (2009) regarding sampling with replacement in occupancy studies based on spatial replication should be taken with care as sampling with replacement can introduce estimation bias for scenarios that can be considered ecologically realistic. If the system is better described by a probability of subunit occupancy independent of the occupancy status of other subunits within the same site, then a recommendation of sampling without replacement would be more appropriate, as a negative bias may

be induced in the estimator of occupancy if sampling is done with replacement. It is advisable to take this into consideration when designing the sampling protocol rather than assuming that selecting samples with replacement is always the best strategy to follow. Sometimes sampling with replacement may be logistically costly or impractical; therefore it is important to assess whether it is indeed beneficial for the properties of the occupancy estimator.

The simulation results with 100 spatial subunits per site (Figure 3-21 and Figure 3-22) illustrate that in this case it makes little difference which of the two strategies is chosen. The choice of whether to sample with or without replacement is relevant when the sample size (i.e. the number of replicates per site, K) is large compared to the number of spatial subunits per sampling site (N). When the sampling site contains many subunits both approaches yield essentially the same result as, under sampling with replacement, the probability of sampling an already sampled subunit is small. It is often considered that the binomial distribution (which models the process of sampling with replacement) approximates well the hypergeometric distribution (which models the process of sampling without replacement) when the sample size is less than a tenth of the total population size (Wild & Seber 2000, p. 210), that is, in our case if $K < 0.1N$. Another issue that influences the relevance of the choice of whether to sample with replacement is the mobility of the species (or rate of decay of tracks in the case of track surveys) compared to the timing of the replicate surveys. In these simulations it was assumed that the occupancy status of each spatial subunit did not change between replicate surveys. As pointed out by KW09, when this is not the case and the occupancy status of spatial subunits is independent of their status in previous replicate surveys (e.g. for highly mobile species), both sampling strategies yield the same results.

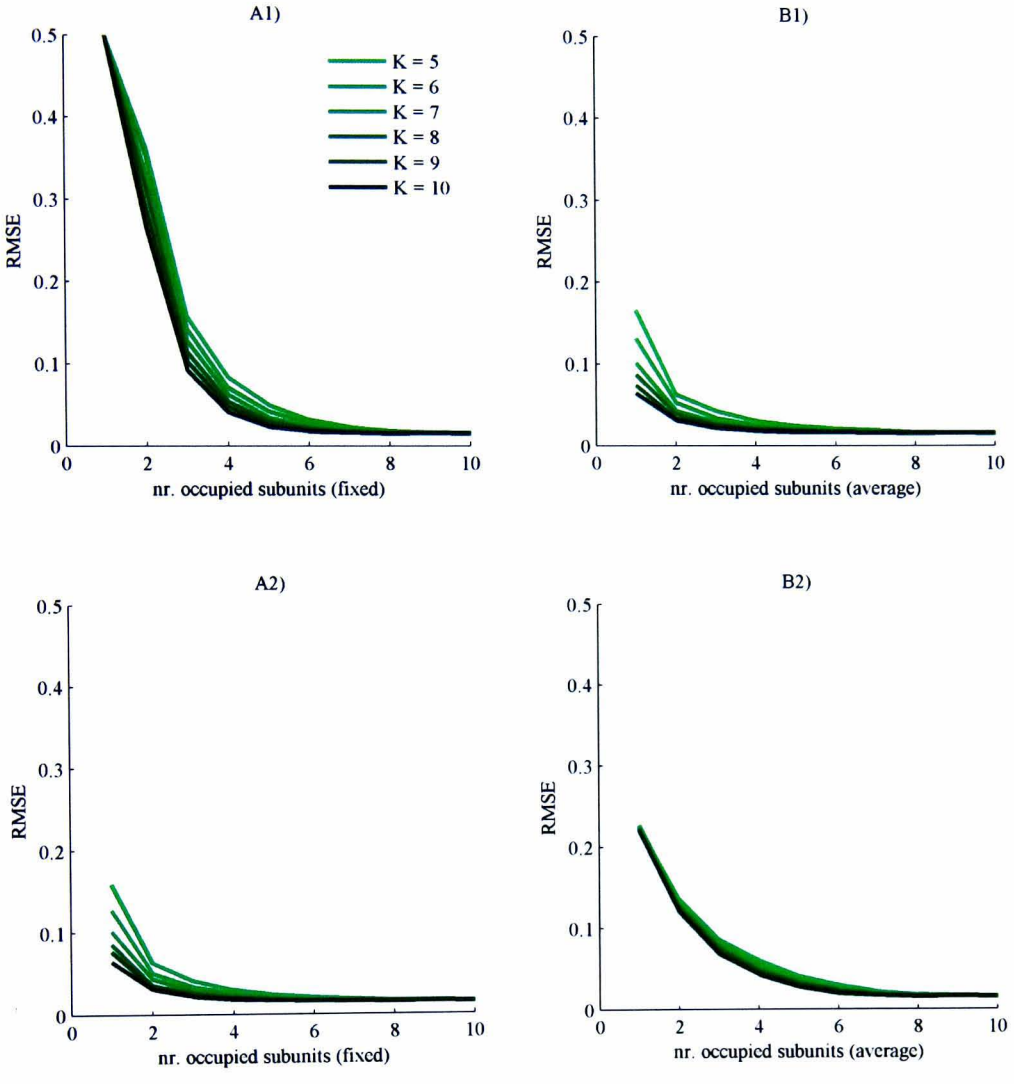


Figure 3-19 Root mean square error (RMSE) of the occupancy estimator as a function of the (fixed/average) number of occupied spatial subunits at occupied sites ($a = 1, 2, \ldots, 10$) for different numbers of spatial replicates ($K = 5, 6, \ldots, 10$), based on a simulation with probability of occupancy $\psi = 0.5$, detection probability $p_a = 0.5$, 1000 sites, $N = 10$ subunits per site and 1000 iterations. Two subunit occupancy scenarios are considered at occupied sites: A) the proportion of occupied subunits is fixed and B) the probability of a subunit being occupied is fixed. Two sampling approaches are evaluated: 1) without replacement and 2) with replacement.

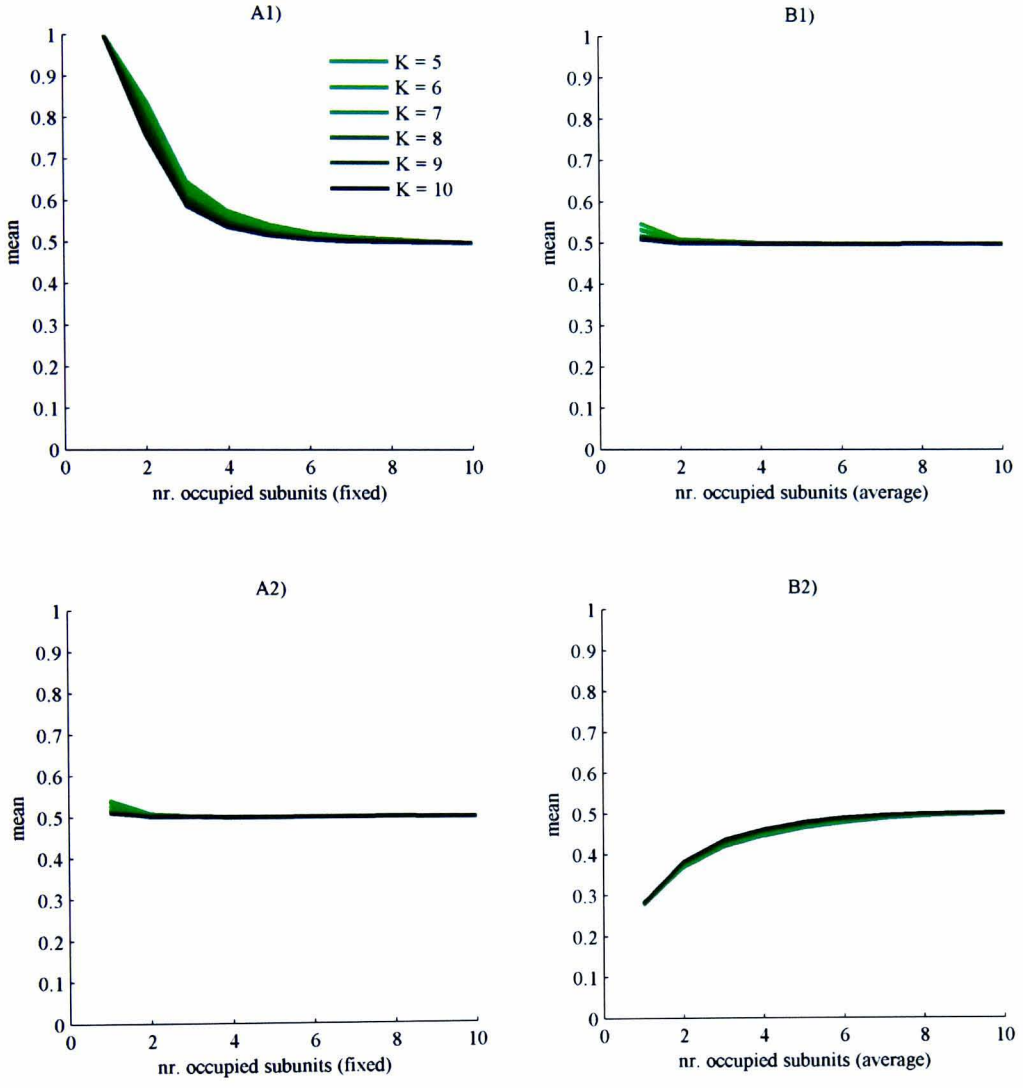


Figure 3-20 Mean of the occupancy estimator as a function of the (fixed/average) number of occupied spatial subunits at occupied sites (a = 1, 2, …, 10) for different numbers of spatial replicates (K = 5, 6, …, 10), based on a simulation with probability of occupancy ψ = 0.5, detection probability p_a = 0.5, 1000 sites, N =10 subunits per site and 1000 iterations. Two subunit occupancy scenarios are considered at occupied sites: A) the proportion of occupied subunits is fixed and B) the probability of a subunit being occupied is fixed. Two sampling approaches are evaluated: 1) without replacement and 2) with replacement.

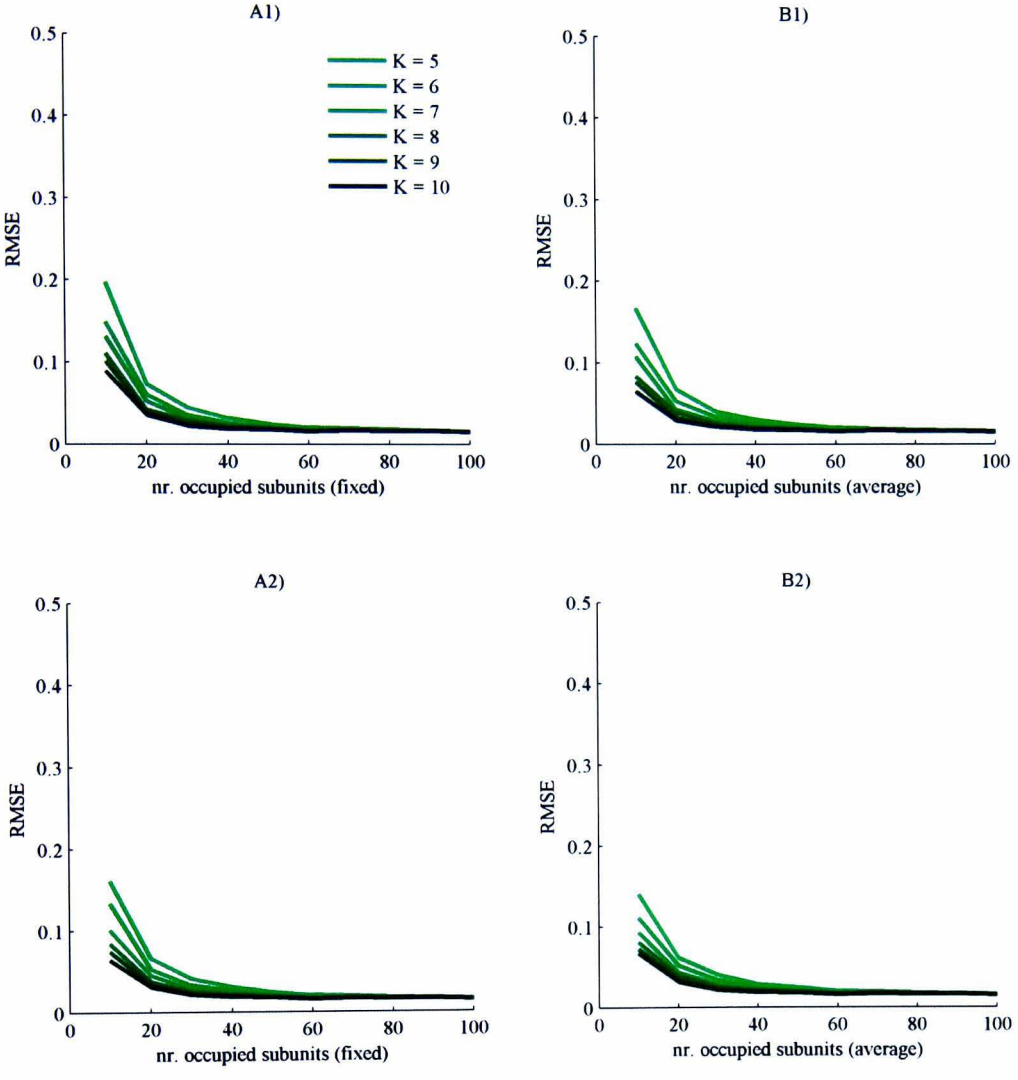


Figure 3-21. Root mean square error (RMSE) of the occupancy estimator as a function of the (fixed/average) number of occupied spatial subunits at occupied sites (a = 10, 20, …, 100) for different numbers of spatial replicates (K = 5, 6, …, 10), based on a simulation with probability of occupancy ψ = 0.5, detection probability p_a = 0.5, 1000 sites, N = 100 subunits per site and 1000 iterations. Two subunit occupancy scenarios are considered at occupied sites: A) the proportion of occupied subunits is fixed and B) the probability of a subunit being occupied is fixed. Two sampling approaches are evaluated: 1) without replacement and 2) with replacement.

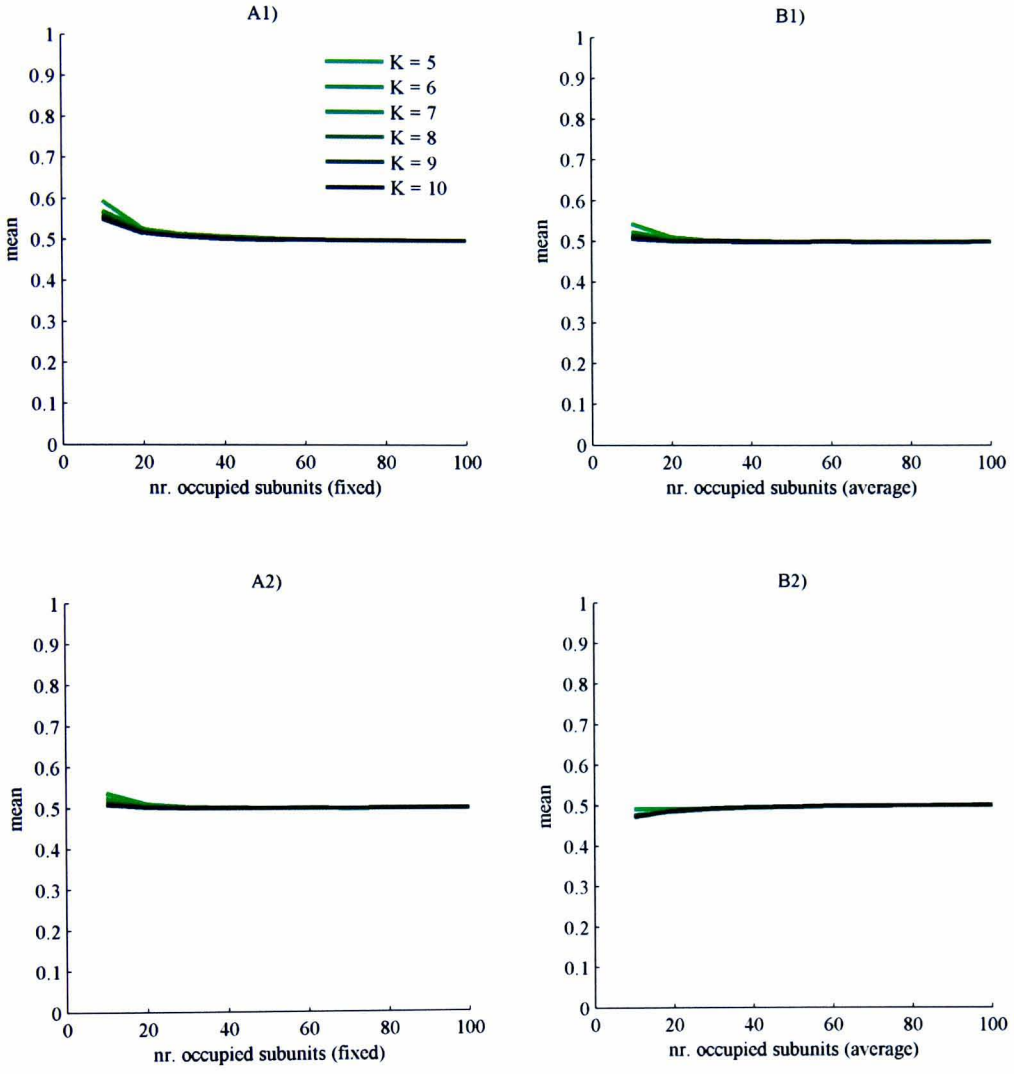


Figure 3-22. Mean of the occupancy estimator as a function of the (fixed/average) number of occupied spatial subunits at occupied sites (a = 10, 20, …, 100) for different numbers of spatial replicates (K = 5, 6, …, 10), based on a simulation with probability of occupancy ψ = 0.5, detection probability p_a = 0.5, 1000 sites, N = 100 subunits per site and 1000 iterations. Two subunit occupancy scenarios are considered at occupied sites: A) the proportion of occupied subunits is fixed and B) the probability of a subunit being occupied is fixed. Two sampling approaches are evaluated: 1) without replacement and 2) with replacement.

3.5 Discussion

In this chapter we have explored different aspects related to the design of occupancy studies. The main focus of the chapter has been on the determination of the number of sites and replicates to be used in the study, except for section 3.4, in which we have shown that a recent general recommendation of sampling with replacement in occupancy studies based on spatial replication is not appropriate for cases that are ecologically realistic.

We have started by considering a standard single-stage occupancy design and we have shown how, using asymptotic approximations, we can determine the optimal level of replication to be used as well as the number of sampling sites that need to be surveyed for the study to yield meaningful results. We have however also illustrated how the asymptotic approximations do not hold for samples sizes that are realistic within ecological studies and therefore highlighted the value of using simulations as a tool for design.

We have considered how Bayesian and sequential design ideas can be applied to the design of occupancy studies, in particular regarding the allocation of survey effort into number of sites and replicates. These methods result in designs that are more robust to poor initial parameter estimates. We have shown how a two-stage design can provide a substantial improvement in efficiency and we have explored how to allocate best the effort between the two stages when such a design is used. Although for this we have concentrated on the variance of the occupancy estimator as a criterion for design, other criteria that involve the detection probability estimator could have been used, as in section 3.1. While we have restricted our study to a two-stage design, more stages could

potentially be considered. Whether dividing the study into more stages is helpful will most likely depend strongly on the quality of the initial estimates. For instance, in the context of quantal response data experiments, Abdelbasit & Plackett (1983) show that higher-stage designs can be more efficient, but only when the initial estimates are rather poor. Our exploration has been largely based on (first-order) asymptotic approximations. Using second-order variance approximations can potentially lead to more accurate design recommendations for smaller sample sizes (e.g. Kalish 1990), so this is another possible route for future work. Finally, the effect of relaxing our assumption of previously sampled sites not being revisited in the second stage could also be explored.

When looking at the optimal allocation of effort between number of sites and number of replicate surveys (sections 3.1, 3.2.4 and 3.3), we have always assumed a standard survey design (single-stage or two-stage) and that all the individual surveys involve the same cost. Consequently we have assessed optimality with respect to total survey effort ($E = SK$ in the single-stage design; $E = S_1K_1 + S_2K_2$ in the two-stage design). However other survey designs or cost functions are of course possible. For instance, MacKenzie & Royle (2005) consider a scenario in which the first visit to each site is more costly. Field, Tyre & Possingham (2005) use a cost function that accounts for an exponentially increasing cost of adding new sites, to reflect a scenario in which sites with lowest access cost are chosen first (but note this sampling approach is not ideal as access cost and occupancy may not be independent). When other designs or cost functions are more appropriate, the exploration carried out in this chapter could be reproduced incorporating these.

4 OCCUPANCY MODELS BASED ON CONTINUOUS SAMPLING

In some cases, rather than using a discrete replicate sampling protocol, as assumed by the occupancy model framework of MacKenzie *et al.* (2002) and Tyre *et al.* (2003), species detection data are collected continuously along a transect or over an interval of time. One example of such a sampling protocol is the Sumatran tiger survey, in which the locations of tiger footprint detections along transects were recorded. This kind of data can be analyzed by discretizing the transect (or time interval) into segments, assigning a '1' or a '0' to each segment to indicate whether there was at least one detection in the segment, and then using an appropriate model from amongst those developed for discrete sampling protocols. For instance, Hines *et al.* (2010) model tiger footprint detections collected along transects in India by discretizing the transects into 1-km segments. In section 2.3 we provided another example of this approach with the analysis of the Sumatran tiger survey data.

In this chapter we present an alternative framework for modelling species occupancy when detection data are collected in a continuous sampling protocol. Our approach, based on describing the detection process as a continuous point process, provides a more natural description of the data and eliminates the need to divide the transect into

discrete segments, which can be arbitrary and may lead to increased bias in the estimator of occupancy and increased chances of obtaining estimates at the boundary of the parameter space. The chapter begins with a short introduction to the relevant theory of point processes in section 4.1. As a starting point, in section 4.2 we discuss a model based on a Poisson process, which is appropriate when the locations of species detections along the transect can be assumed to be independent. Since this condition is likely to be violated for some species, including the tiger, in section 4.3 we relax this assumption and propose a model in which detections are described as a clustered point process. The model is based on a two-state Markov-modulated Poisson process (2-MMPP) which provides a structure for modelling clustering in the occurrence of events, while being a convenient process to work with due to its mathematical tractability. To illustrate the application of these two models, in section 4.4 we fit them to the Kerinci Seblat tiger data set. The work in this chapter has been published in Guillera-Arroita *et al.* (2011).

4.1 Introduction to point processes

A natural framework for the description of data on point occurrences collected along a continuous axis is provided by the theory of point processes (for an introduction to the fundamental theory see for instance Cox & Isham 1980). A point process is a particular kind of stochastic process, in which a realization consists of a collection of points, each with a well-defined position. Here we limit the term 'point process' to processes on the real half-line $\mathbb{R}_+ = [0, \infty)$, compared to higher-dimensional processes, such as spatial point processes which are defined on the real plane $\mathbb{R}^2$. The real half-line often represents time, but can also represent distance from an origin along a line.

Point processes are frequently used as models for the occurrence of random events in varied fields such as seismology (e.g. Vere-Jones 1970), telecommunications (e.g. Lee & Fapojuwo 2005), safety and reliability engineering (e.g. Thompson 1988), neuroscience (e.g. Johnson 1996) or finance (e.g. Daykin, Pentikäinen & Pesonen 1994). In ecology, the collection of data on the location of species or individual detections along line transects provides a natural application for this framework. Indeed point processes have for instance been used to model detections along transects in the context of distance sampling studies (Skaug 2006). However, their potential within the occupancy modelling framework has not been explored to date.

Given that our motivating application is the modelling of detections along a transect, in this section we present the background material on point processes using the terms 'position' and 'distance' to refer to the locations of occurrences on the axis and the intervals between them. These terms could be replaced by 'time' when the axis has that interpretation.

4.1.1 Poisson process

The simplest point process of all is the Poisson process which assumes that events occur totally at random along the axis, that is, events are independently and uniformly distributed over the interval of interest. This process is characterized by a single constant λ, the intensity (or rate) of the Poisson process. Two properties can be derived from the above definition:

(i) the number of events in an interval of length L is Poisson distributed with parameter λL, with the numbers of events in disjoint intervals being independent;

(ii) the inter-event distances (i.e. distances until the next occurrence) are a sequence of independent exponentially distributed random variables with parameter λ, and mean $1/\lambda$.

The probability mass function for the count of occurrences, D, in an interval of length L is therefore

$$\Pr(D = d) = \frac{(\lambda L)^d e^{-\lambda L}}{d!}, \qquad d = 0, 1, 2, \ldots \tag{4.1}$$

On the other hand, the joint probability density of the inter-event distances $\boldsymbol{l}$ is

$$f_{\boldsymbol{l}}(\boldsymbol{l}) = \lambda e^{-\lambda l_1} \ldots \lambda e^{-\lambda l_d} e^{-\lambda l_{d+1}} = \lambda^d e^{-\lambda L}, \tag{4.2}$$

where $\boldsymbol{l} = \{l_1, \ldots, l_d, l_{d+1}\}$, with l_1 the distance from the beginning of the interval to the first event, l_{d+1} the distance from the last event to the end of the interval and $L = \sum_{i=1}^{d+1} l_i$.

The Poisson process is stationary (the probability distribution of the number of events in any interval only depends on the length of the interval) and memoryless (the number of events in any interval after position x is independent of the number of events before x), and it is of central importance in the theory of point process. It provides a natural starting point for constructing other more complex processes and it occurs in many limiting situations (Cox & Isham 1980, pp 47-48). An important result, analo-

gous to the Central Limit Theorem for random variables, is that the superposition of point processes is asymptotically a Poisson process (subject to the condition that the individual processes are such that no process dominates the rest).

Although the Poisson process is convenient for some scenarios, generalizations that relax the requirement of independence are often a more appropriate representation of reality. A first useful generalization is to allow the intensity of the process to be a function of the position, $\lambda(x)$, or, even more generally, to let the intensity at position x be a function of an observed position-dependent explanatory variable $z(x)$. Such a non-stationary process is called a non-homogenous (or inhomogeneous) Poisson process while, in contrast, the particular case with constant intensity described above is called a homogenous Poisson process. In a non-homogenous Poisson process the expected number of events in an interval I is

$$\lambda_I = \int_I \lambda(x)dx\,.$$

In this case the joint probability density function for the inter-event distances $\boldsymbol{l}$ in (4.2) generalizes to

$$f_{\boldsymbol{\ell}}(\boldsymbol{l}) = \left\{\prod_{j=1}^{d} \lambda(x_j)\right\} \exp\left(-\int_I \lambda(x)dx\right). \tag{4.3}$$

where x_j is the position of the j-th detection, i.e. $x_j = \sum_{i=1}^{j} l_i$.

4.1.2 Markov-modulated Poisson process

A further generalization of the Poisson process is the doubly-stochastic Poisson process (Cox & Isham 1980, pp. 70-75), also known as the Cox process. In this kind of process the position-varying intensity function $\lambda(x)$, rather than being deterministic as in previous examples, is itself the realization of an unobserved stochastic process $\Lambda(x)$ (hence the term doubly-stochastic).

A particular type of doubly-stochastic Poisson process is the Markov-modulated Poisson process (MMPP), where the intensity of the Poisson process is governed by an unobserved Markov process. A summary of properties of MMPPs can be found in Fischer & Meier-Hellstern (1993). In a n-MMPP, the intensity of the Poisson process alternates among n possible values, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_n\}$, each of them corresponding to one of the n states of the underlying Markov process. If $n = 1$, the MMPP reduces to an ordinary Poisson process.

To specify a MMPP the underlying Markov process needs to be specified. In a Markov process the intervals spent in each state, the holding times or sojourn times, are independent and exponentially distributed with parameter ξ_i. Given the exit rate from each state, ξ_i, and the transition probabilities between states in the embedded Markov chain, p_{ij}, the (infinitesimal) transition rates from state i to state j, μ_{ij}, are derived as $\mu_{ij} = \xi_i p_{ij}$. A Markov process can be characterized by its $n \times n$ infinitesimal generator matrix $\boldsymbol{Q}$, which has elements

$$q_{ij} = \begin{cases} \mu_{ij}, & i \neq j \\ -\sum_{j \neq i} q_{ij}, & i = j \end{cases}$$

with the diagonal elements q_{ii} ensuring that the sum of the elements in every row is zero.

Apart from $\boldsymbol{Q}$ and $\boldsymbol{\lambda}$, the state of the Markov process at $x = 0$ (the beginning of the transect) needs also to be provided in order to characterize a MMPP fully. Let $\boldsymbol{\pi}$ be the initial probability vector of the MMPP, that is, the probability of being in each of the n intensity states at $x = 0$. There are two stationary versions of the MMPP depending on the choice of $\boldsymbol{\pi}$ (Fischer & Meier-Hellstern 1993; Rydén 1994):

(i) Let $D(x)$ be the number of events in an interval $[0, x]$. For the counting process $\{D(x)\}$ to be stationary, $\boldsymbol{\pi}$ is chosen to be the stationary (or equilibrium) distribution of the underlying Markov process, $\boldsymbol{\pi}'$, which satisfies

$$\boldsymbol{\pi}'\boldsymbol{Q} = 0. \tag{4.4}$$

In this case the MMPP is said to be environment-stationary (or time-stationary).

(ii) Let X_k be the state of the Markov chain at the time of event k and ℓ_k the distance between events k and $k-1$. The sequences $\{X_k\}$ and $\{\ell_k\}$ are stationary if $\boldsymbol{\pi}$ is chosen to be the stationary vector of the transition probability matrix of the Markov chain embedded at the events, $\boldsymbol{\pi}^*$. The vector $\boldsymbol{\pi}^*$ is

$$\boldsymbol{\pi}^* = \frac{\boldsymbol{\pi}'\boldsymbol{\Lambda}}{\boldsymbol{\pi}'\boldsymbol{\lambda}^T}, \tag{4.5}$$

where $\boldsymbol{\pi}'$ is given by (4.4) and $\boldsymbol{\Lambda}$ is the diagonal matrix $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$. In this case the MMPP is started at an event time and is said to be interval-stationary.

As we will see later, the vectors $\boldsymbol{\pi}'$ and $\boldsymbol{\pi}^*$ can be used as a specification of the initial probability vector of the MMPP when constructing the likelihood function. The vector $\boldsymbol{\pi}'$ is appropriate if the realizations of the MMPP start at a random point, while $\boldsymbol{\pi}^*$ is relevant when the realizations start at an event.

Let us now define $\boldsymbol{F}(l)$ to be the matrix with entries

$$F_{ij}(l) = \Pr(\ell_k \leq l, X_k = j | X_{k-1} = i),$$

that is, the probability that, given the previous detection was generated under state i, the current one occurs at a distance smaller than l and is generated under state j. The derivative of $\boldsymbol{F}(l)$, $\boldsymbol{f}(l)$, is given by (Meier-Hellstern 1987; Rydén 1994)

$$\boldsymbol{f}(l) = \boldsymbol{F}'(l) = \exp(\boldsymbol{C}l)\boldsymbol{\Lambda}, \tag{4.6}$$

where $\boldsymbol{C} = \boldsymbol{Q} - \boldsymbol{\Lambda}$. Note that here 'exp' denotes the matrix exponential function, which for a square matrix $\boldsymbol{X}$ is defined as the following convergent power series

$$\exp(\boldsymbol{X}) = \sum_{k=0}^{\infty} \frac{1}{k!} \boldsymbol{X}^k.$$

From (4.6) the joint probability density of the inter-event distances $\boldsymbol{l}$ for an n-MMPP can be written as

$$f_{\boldsymbol{l}}(\boldsymbol{l}|\boldsymbol{\lambda},\boldsymbol{\mu}) = \boldsymbol{\pi}\exp(\boldsymbol{C}l_1)\,\boldsymbol{\Lambda}\ldots\exp(\boldsymbol{C}l_d)\boldsymbol{\Lambda}\exp(\boldsymbol{C}l_{d+1})\,\boldsymbol{e}, \tag{4.7}$$

where $\boldsymbol{\pi}$ is the initial probability vector of the MMPP and $\boldsymbol{e}^T = [1,1,\ldots,1]_{n\times 1}$. For intervals with no events the only element of $\boldsymbol{l}$ is L, the length from the beginning to the end of the transect and (4.7) takes the value

$$f_{\boldsymbol{l}}(\boldsymbol{l}|\boldsymbol{\lambda},\boldsymbol{\mu}) = \boldsymbol{\pi}\exp(\boldsymbol{C}L)\boldsymbol{e}. \tag{4.8}$$

Although perhaps most naturally seen as a doubly-stochastic Poisson process, note that the MMPP can also be interpreted as a hidden Markov model and as a Markov renewal process (Fischer & Meier-Hellstern 1993; Rydén 1994).

MMPPs are useful for modelling time-varying (or position-varying) intensity rate processes and are applied in a variety of areas. They are widely used in telecommunications where they are often applied to model packetized voice and data streams in communication networks such as those from Internet traffic (e.g. Heffes & Lucantoni 1986; Muscariello *et al.* 2005). MMPPs have been applied to financial problems (e.g. Takada, Sumita & Takahashi 2011) and have also been utilized to model environmental data including applications related to air pollution exposure (e.g. Ramesh 1995) or earthquake occurrence (e.g. Lu 2012). In ecology, the 2-MMPP has been proposed to model the detection of individuals along a transect in the context of abundance estimation with distance sampling (Skaug 2006).

In this chapter we use a 2-MMPP to describe clustered species detection data collected along a transect in the context of occupancy modelling. According to a 2-MMPP, species detections take place at two different rates λ_1 and λ_2, and the interval spent surveying in each of these two states is stochastic and exponentially distributed with parameters: μ_{12}, the switching intensity from λ_1 to λ_2, and μ_{21}, the switching intensity from λ_2 to λ_1 (Figure 4-1). The 2-MMPP is sometimes referred to as a Switched Poisson process (SPP) and, often, the particular case with one detection rate equal to zero is referred to as an Interrupted Poisson process (IPP).

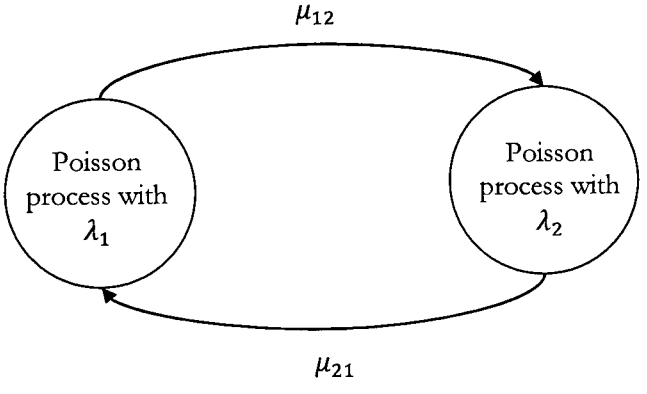


Figure 4-1 State transition graph for a two-state Markov-modulated Poisson process (2-MMPP)

The structure of the 2-MMPP induces clustering in the arrival of events, with the interval of time spent in the high-detection rate corresponding to event bursts and the degree of clustering increasing as the difference between the two intensity rates of the Poisson processes increases, as illustrated in Figure 4-2. Note here as well that the realizations of a homogenous Poisson process sometimes display some apparent clustering, which is a consequence of its property of total randomness. Due to its capability to

accommodate actual clustering, the 2-MMPP may be, for instance, an adequate representation for transect survey data from a species that only partially covers occupied sampling sites or when surveying for tracks along trails that individuals of the species may follow intermittently.

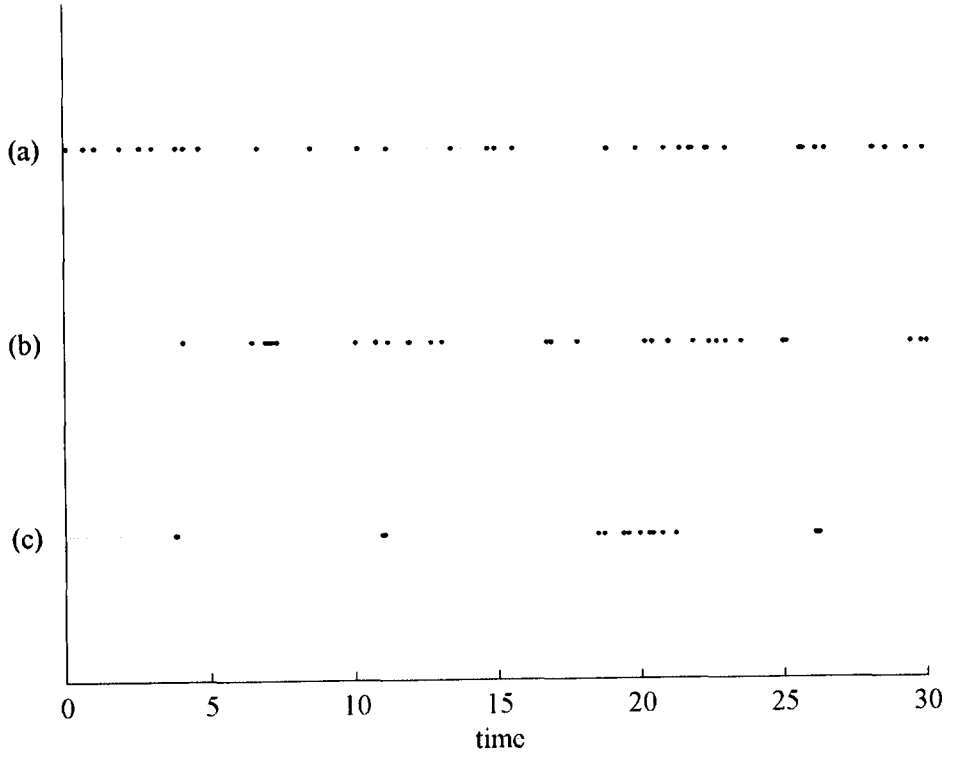


Figure 4-2 Realizations of three point processes of unit rate: (a) homogenous Poisson process, $\lambda = 1$; (b) 2-MMPP, $\lambda = [0.5, 3]$, $\mu = [0.5, 2]$; (c) IPP $\lambda = [0, 5]$, $\mu = [0.5, 2]$ (particular case of 2-MMPP with one intensity rate zero).

4.2 Poisson process occupancy model

4.2.1 Model formulation and assumptions

Let us consider a study with S sampling sites where surveys have been carried out along transects, recording the location of each detection of the species. As in section 2.1.2, here the assumption is that sites are closed to changes in species occupancy within the sampling season, and that each site has a probability ψ of being occupied. We also assume that detections along each transect at occupied sites can be considered independent and so can be modelled as a Poisson process with intensity λ, where λ represents the average number of detections per unit length. Note that, at this stage, we consider that occupancy probability ψ and detection intensity λ are constant.

The likelihood function for such a model is constructed as that of a series of independent Poisson processes but allowing for zero-inflation to account for unoccupied sites, thus resulting in a zero-inflated Poisson model. The contribution to the likelihood for a site in which the species was detected at least once is

$$\psi \prod_{j=1}^{R_i} \left\{ \lambda \exp(-\lambda l_{ij_1}) \dots \lambda \exp\left(-\lambda l_{ij_{d_{ij}}}\right) \exp\left(-\lambda l_{ij_{d_{ij}+1}}\right) \right\} = \psi \lambda^{d_i} e^{-\lambda L_i},$$

where R_i is the number of independent transects in site i, d_{ij} is the number of detections along transect j in site i and $l_{ij_1} \dots l_{ij_{d_{ij}+1}}$ are the inter-detection distances (Figure 4-3), with l_{ij_1} defined as the distance to first detection from the beginning of the transect and $l_{ij_{d_{ij}+1}}$ the distance from the last detection until the end of the transect; d_i and L_i represent the total number of detections and the total length surveyed in

site i respectively, so that $d_i = \sum_{j=1}^{R_i} d_{ij}$ and $L_i = \sum_{j=1}^{R_i} \sum_{k=1}^{d_{ij}+1} l_{ij_k}$. The likelihood contribution from a site with no detections of the species is

$$(1 - \psi) + \psi\, e^{-\lambda L_i},$$

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

$$L(\psi, \lambda) = \prod_{i=1}^{S} \{\psi \lambda^{d_i} e^{-\lambda L_i} + (1 - \psi) \mathrm{I}(d_i = 0)\}, \tag{4.9}$$

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

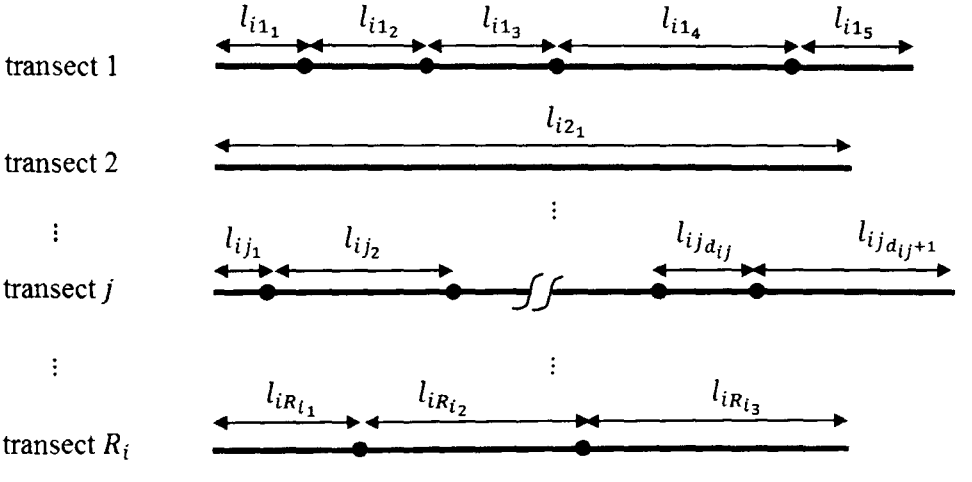


Figure 4-3 Notation used for the inter-detection distances from site i.

To construct the likelihood in (4.9), we interpret the detection data as a series of inter-detection distances, as in (4.2), rather than directly as a count of detections, as in (4.1). The latter would have introduced additional factors $L_i^{d_i}/d_i!$, but we note that these do not involve the model parameters and therefore do not affect their estimation. Looking at the data in this manner is necessary to make sure that the likelihood function is comparable to that corresponding to more general models based on inhomogeneous Poisson processes (models with covariates in section 4.2.6 or the MMPP model in section 4.3), and for which data can no longer be summarized as a count.

4.2.2 Relationship to the Bernoulli process occupancy model

The basic occupancy model discussed in chapter 2 models species detections at an occupied site as coming from K independent Bernoulli trials, each with probability of success p, the detection probability at each survey replicate. Such a random process, can in fact be interpreted as a point process along a discretized axis, and consequently is sometimes referred to as a Bernoulli process (e.g. Kingman 1993 p. 21). Here we adopt this terminology to distinguish this model from the model based on a (continuous) Poisson process.

The binomial distribution is a good approximation to the Poisson distribution when the number of trials is large and the probability of success at each trial is small (Haight 1967, p. 15). Therefore the Bernoulli process occupancy model will be a good approximation to the Poisson process occupancy model if the continuous detection data are discretized using sufficiently small intervals, that is, if the transect is cut into sufficiently small segments (Figure 4-4). The corresponding parameter p would be $1 -$

$\exp(-\lambda L/K)$, i.e. the probability of having at least one detection in a segment of length L/K given a Poisson process with detection rate λ.

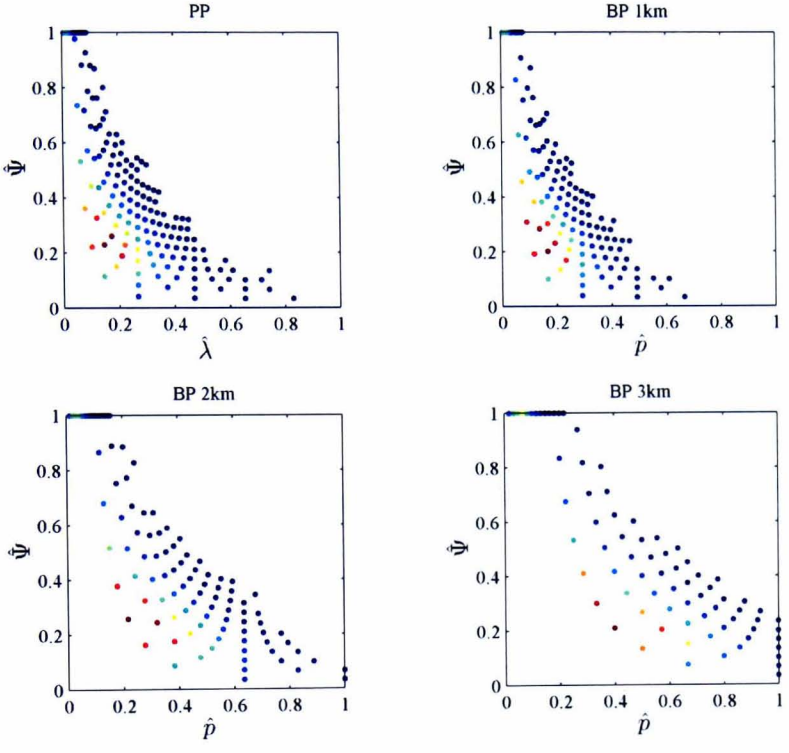


Figure 4-4 Comparison of the distribution of MLEs for the Poisson process occupancy model (PP) and the Bernoulli process occupancy model (BP) at different segment lengths when data are generated according to a Poisson process, for $\lambda = 0.25$ detections/km, $\psi = 0.25$, $S = 30$ sites and $L = 6$ km. Plots show part of the distribution that contains 0.999 probability, with no point that is excluded having higher probability than any of the points displayed (based on 10^4 simulated data sets).

4.2.3 MLEs and estimator properties

In order to explore the general properties of the maximum-likelihood estimators for the Poisson process occupancy model, let us assume that the total transect length surveyed within each site is constant, L, which is equivalent to the assumption of standard survey design made in section 2.1.3. The likelihood in (4.9) can now be written in a compact form as follows

$$L(\psi, \lambda) = \psi^{S_d} \lambda^{d_T} e^{-S_d \lambda L} (1 - \psi\lambda^*)^{S - S_d}, \tag{4.10}$$

where $d_T = \sum_{i=1}^{S} d_i$ is the total number of detections in the survey, S_d is the number of sites where the species was detected at least once, and $\lambda^* = 1 - e^{-\lambda L}$ is the probability of not missing the species at an occupied site (denoted λ^* here for consistency with the p^* notation in chapter 2). Note that (S_d, d_T) is a sufficient statistic with respect to this model.

As done in section 2.1.3 for the Bernoulli process occupancy model, the likelihood in (4.10) can be rewritten using a reparameterisation of the type suggested by Morgan, Revell & Freeman (2007) for simplifying the likelihood of site occupancy models. Setting $\theta = \psi\lambda^*$, the probability that a site is occupied and the species is detected there, leads to

$$L(\theta, \lambda) = \{\theta^{S_d}(1 - \theta)^{S - S_d}\} \left\{ \lambda^{d_T} \left(\frac{1 - \lambda^*}{\lambda^*} \right)^{S_d} \right\}. \tag{4.11}$$

From this factorization into two parts, each one involving only one of the parameters (θ or λ), the expressions that the MLEs must fulfil can be easily derived by differentiating and setting equal to zero each of the parts. The MLE for $\hat{\theta}$ is the proportion of cells where the species was detected, $\hat{\theta} = S_d/S$, and therefore the estimator of occupancy $\hat{\psi}$ satisfies

$$\hat{\psi} = \frac{S_d}{S\hat{\lambda}^*}. \tag{4.12}$$

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

$$\frac{\hat{\lambda}}{\hat{\lambda}^*} = \frac{d_T}{S_d L}. \tag{4.13}$$

There is an evident parallelism between these two expressions and the equivalent ones for the Bernoulli process occupancy model in (2.3). It can be seen that, as $\hat{\lambda}^*$ decreases, that is, as the estimated probability of detecting the species at an occupied site when a transect of length L is surveyed decreases, the estimate of occupancy ($\hat{\psi}$) increases and the estimate of detection rate ($\hat{\lambda}$) decreases relative to the naïve estimates obtained assuming that the species is always detected at occupied sites: $\hat{\psi}_{naive} = S_d/S$, $\hat{\lambda}_{naive} = d_T/(S_d L)$. Similarly to the Bernoulli model, the MLE expressions in (4.12) and (4.13) do not always hold and, depending on the observed detection data, the actual MLEs may be on the boundary $\hat{\psi} = 1$, as we show below.

Let us denote by $h(\lambda)$ the part of the likelihood (4.11) involving λ

$$h(\lambda) = \lambda^{d_T} \left(\frac{1-\lambda^*}{\lambda^*}\right)^{S_d}. \tag{4.14}$$

Equation (4.13), which is satisfied by those points for which the first derivative of $h(\lambda)$ is zero, can be conveniently written as an explicit expression for λ using the Lambert W function (Corless *et al.* 2005), which is defined as the inverse of xe^x. This function, commonly implemented in software packages (e.g. function *LambertW* in MATLAB), is multivalued and has two real branches (Figure 4-5). It is known that the solution to an equation of the form $a^x = x + b$ is $x = -b - W(-a^{-b}\log a)/\log a$. Therefore, in our case setting $a = e^{d_T/S_d}$, $b = 1$ and $x = -\lambda L S_d/d_T$, we have that λ satisfies

$$\lambda = \frac{1}{L}\left\{\frac{d_T}{S_d} + W\left(-\frac{d_T}{S_d}e^{-\frac{d_T}{S_d}}\right)\right\}. \tag{4.15}$$

Note that here 'hats' are removed from the notation, to recognize that not all the solutions to (4.15) lead to MLEs.

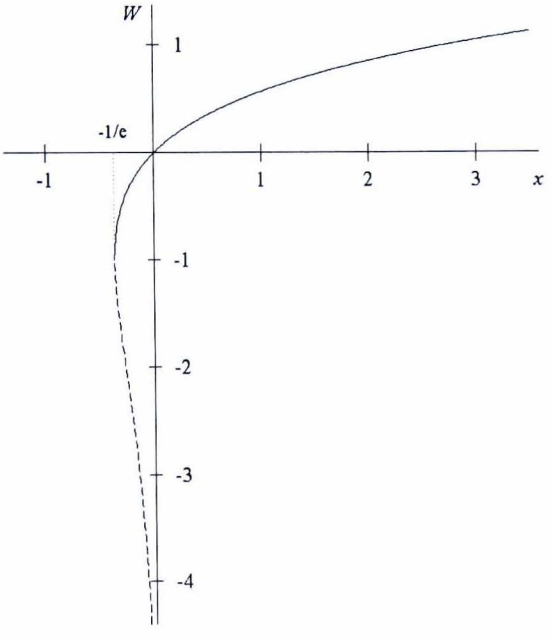


Figure 4-5 The two real branches of the Lambert W function $W(x)$. The solid line indicates the upper branch with $W(x) \geq -1$, usually called the principal branch $W_0(x)$. The dashed line represents the lower branch with $W(x) \leq -1$, usually denoted $W_{-1}(x)$.

When $d_T = S_d$, in (4.15) both branches of the Lambert W function lead to the same solution ($\lambda = 0$) given that $W_0(-1/e) = W_{-1}(-1/e) = -1$. Since the solution is unique it must correspond to a maximum in the function $h(\lambda)$. When $d_T \neq S_d$, there are two possible real solutions to (4.15), one corresponding to each real branch of the Lambert W function. The solution given by the lower branch, W_{-1}, is $\lambda = 0$, since

$W_{-1}(xe^x) = x$. This solution corresponds to a minimum in $h(\lambda)$ given that, if $d_T \neq S_d$,

$$\lim_{\lambda \to 0} h(\lambda) = \lim_{\lambda \to 0} \frac{\lambda^{d_T}}{(\lambda^*)^{S_d}} = C \cdot \lim_{\lambda \to 0} \frac{\lambda^{d_T - S_d}}{(\lambda^*)^0} = 0\,, \tag{4.16}$$

where C is a constant. Therefore, the solution given by the principal branch corresponds to a maximum in $h(\lambda)$. Considering this and provided the MLEs are not on the boundary, we can write that the MLE for λ satisfies

$$\hat{\lambda} = \frac{1}{L}\left\{\frac{d_T}{S_d} + W_0\left(-\frac{d_T}{S_d} e^{-\frac{d_T}{S_d}}\right)\right\}, \tag{4.17}$$

where $W_0(.)$ is the principal branch of the Lambert W function.

According to (4.12), $\hat{\psi}$ would take values larger than unity if $\lambda^* < S_d/S$, or equivalently, considering that $\lambda^* = \lambda S_d L/d_T$, if $\lambda < d_T/(SL)$. This implies that the MLE expressions in (4.12), (4.13) and (4.15) hold when the observed detection history fulfils the condition

$$\log\left(\frac{S - S_d}{S}\right) \geq -\left\{\frac{d_T}{S_d} + W_0\left(-\frac{d_T}{S_d} e^{-\frac{d_T}{S_d}}\right)\right\}. \tag{4.18}$$

If a detection history does not satisfy (4.18) then the MLEs are

$$\hat{\psi} = 1, \qquad \hat{\lambda} = \frac{d_T}{SL}. \tag{4.19}$$

In summary, the MLEs of the Poisson process occupancy model are such that:

(i) if $d_T = S_d$ (i.e. the species is detected at most once at any site with detections) there is only one solution for (4.13), the trivial solution $\lambda = 0$, which is a maximum of (4.14) and leads to a boundary occupancy estimate, $\hat{\psi} = 1$, and $\hat{\lambda} = d_T/(SL)$.

(ii) if $d_T/S_d \to \infty$ (i.e. there is a very large number of detections of the species at sites where it was detected) there are two solutions for (4.13), the trivial solution $\lambda = 0$, which is a minimum of (4.14), and a second one, which is the maximum, for $\lambda \to \infty$. In this case the occupancy estimate coincides with the naive estimate, $\hat{\psi} = S_d/S$.

(iii) if $\log\left(\frac{S-S_d}{S}\right) < -\left\{\frac{d_T}{S_d} + W_0\left(-\frac{d_T}{S_d}e^{-\frac{d_T}{S_d}}\right)\right\}$ there are two solutions for (4.13), the trivial solution $\lambda = 0$, which is a minimum of (4.14), and a second one, which is the maximum and leads to a boundary occupancy estimate, $\hat{\psi} = 1$, and $\hat{\lambda} = d_T/(SL)$.

(iv) if $\log\left(\frac{S-S_d}{S}\right) \geq -\left\{\frac{d_T}{S_d} + W_0\left(-\frac{d_T}{S_d}e^{-\frac{d_T}{S_d}}\right)\right\}$ there are two solutions for (4.13), the trivial solution $\lambda = 0$, which is a minimum of (4.14), and a second one given by (4.15), which is the maximum and leads to an occupancy estimate which lies within the probability boundaries, determined as in (4.12).

For the Poisson process occupancy model, the first derivative of the log-likelihood function, the scores vector, is

$$\begin{aligned}\mathbf{G} &= \left[\frac{\partial \mathcal{L}}{\partial \psi} \quad \frac{\partial \mathcal{L}}{\partial \lambda}\right] \\ &= \left[\frac{S_d - S\psi\lambda^*}{\psi(1-\psi\lambda^*)} \quad \frac{d_T - S_d\lambda L}{\lambda} - \frac{(S-S_d)\psi L(1-\lambda^*)}{1-\psi\lambda^*}\right].\end{aligned} \tag{4.20}$$

The elements of the observed information matrix $\mathbf{O}$ are

$$\begin{aligned}\mathbf{O}[1,1] &= -\frac{\partial^2 \mathcal{L}}{\partial \psi^2} = \frac{S_d}{\psi^2} + \frac{(S-S_d)\lambda^{*2}}{(1-\psi\lambda^*)^2} \\ \mathbf{O}[1,2] &= -\frac{\partial^2 \mathcal{L}}{\partial \psi \partial \lambda} = \frac{(S-S_d)L(1-\lambda^*)}{(1-\psi\lambda^*)^2} \\ \mathbf{O}[2,2] &= -\frac{\partial^2 \mathcal{L}}{\partial \lambda^2} = \frac{d_T}{\lambda^2} + \frac{(S-S_d)\psi L^2(1-\lambda^*)(\psi-1)}{(1-\psi\lambda^*)^2}.\end{aligned} \tag{4.21}$$

Since the expectations for the data are

$$\begin{aligned}\mathbb{E}[S_d] &= S\psi\lambda^* \\ \mathbb{E}[d_T] &= S\psi\lambda L,\end{aligned}$$

the expected information matrix $\mathbf{I} = \mathbb{E}[\mathbf{O}]$ has elements

$$\begin{aligned}\mathbf{I}[1,1] &= \frac{S\lambda^*}{\psi(1-\psi\lambda^*)} \\ \mathbf{I}[1,2] &= \frac{SL(1-\lambda^*)}{(1-\psi\lambda^*)} \\ \mathbf{I}[2,2] &= S\psi L\left\{\frac{1}{\lambda} + \frac{L(1-\lambda^*)(\psi-1)}{(1-\psi\lambda^*)}\right\},\end{aligned} \tag{4.22}$$

which in this case, given (4.12) and (4.13), are the same as those of the observed information matrix evaluated at the MLEs. From (4.22) the asymptotic variance-covariance matrix can be derived as $\boldsymbol{\Sigma} = \mathbf{I}^{-1}$, which leads to

$$\begin{aligned}
\boldsymbol{\Sigma}[1,1] &= \mathrm{var}(\hat{\psi}) = \frac{\psi}{S}\left\{(1-\psi) + \frac{1-\lambda^*}{\lambda^* - (1-\lambda^*)\lambda L}\right\} \\
\boldsymbol{\Sigma}[1,2] &= \mathrm{cov}(\hat{\psi}, \hat{\lambda}) = \frac{-\lambda}{S}\left\{\frac{1-\lambda^*}{\lambda^* - (1-\lambda^*)\lambda L}\right\} \\
\boldsymbol{\Sigma}[2,2] &= \mathrm{var}(\hat{\lambda}) = \frac{\lambda}{\psi SL}\left\{\frac{\lambda^*}{\lambda^* - (1-\lambda^*)\lambda L}\right\}.
\end{aligned} \tag{4.23}$$

Looking at (4.23) it can be seen that, as λ^* approaches unity, that is, as the probability of missing the species at occupied sites approaches zero,

(i) the variance of the occupancy estimator $\hat{\psi}$ tends to the variance dictated by the binomial proportion $\psi(1-\psi)/S$ and decreases as the number of sites increases;

(ii) the variance of the intensity parameter estimator $\hat{\lambda}$ tends to $\lambda/(\psi SL)$ and decreases as the total effort (SL) increases regardless of whether it is spent on surveying more sites or longer transects within each site;

(iii) the covariance approaches zero.

4.2.4 Design recommendations

For the purpose of deriving general survey design recommendations, let us continue assuming a survey design in which the same transect length L is surveyed in all S sites. Suppose now that a study can employ a fixed amount of surveying effort ($E = SL$) and

that we wish to maximise the precision of the estimator of occupancy, $\hat{\psi}$. Survey design recommendations for this scenario can be derived looking at the expression of the asymptotic variance of $\hat{\psi}$ in (4.23). Table 4-1 shows the optimal survey design, assuming that survey costs per unit length surveyed are constant and consequently assessing optimality with respect to total survey effort E. As discussed in previous chapter, in some scenarios other survey designs or cost functions might be more appropriate and the same exploration could be reproduced incorporating these.

Table 4-1 Mean number of detections at occupied sites (λL) to minimize the variance of the occupancy estimator in the Poisson process occupancy model for different levels of occupancy (ψ). The corresponding probability of detecting the species at occupied sites (λ^*) is also shown.

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

The optimal length to survey per site is determined by the parameter values (ψ and λ), irrespective of the total effort allocated to the survey. The probability of occupancy (ψ) determines the mean number of detections at occupied sites (λL) that maximises the precision of $\hat{\psi}$, from which the optimal length can be derived. The number of sites immediately follows by considering the total effort E. For instance, if $\psi = 0.4$, $\lambda = 3$ and $E = 40$, then the optimal design would be to survey $L = 1.74/3 = 0.58$ units of length in $S = E/L = 40/0.58 \approx 69$ sampling sites. Table 4-1 suggests that for rare species (i.e. low ψ) it is best to sample more sites (at the expense of increasing the probability of false absence), while for commoner species (i.e. higher ψ) it is best to allocate the effort so that fewer sites are surveyed more intensively. This general pat-

tern is in line with that observed for the Bernoulli process occupancy model in section 3.1. Note that the optimal length is the same if the study is designed to minimize the total surveying effort for a given precision of the occupancy estimator.

Note also that, when λL is small, the zero-inflated Poisson distribution with zero inflation $1 - \psi$ and rate λL is approximately a Bernoulli distribution with probability parameter $\psi\lambda L$. In such circumstances ψ and λ become non-identifiable. This is equivalent to not having replication in the basic Bernoulli process occupancy model.

4.2.5 Performance simulations

We conducted a simulation study to evaluate the performance of the Poisson process occupancy model compared to the Bernoulli process occupancy model for detection data collected under a continuous sampling protocol. For each sampling site the occupancy status was determined as the outcome of a Bernoulli trial with probability ψ. Detection data for occupied sites were generated following a homogenous Poisson process. We ran 10,000 simulations of a study design where the total survey effort available was 600 km of transect. We first set the length surveyed per site to 6 km, therefore resulting in 100 sampled sites. Occupancy was set to $\psi = 0.25$, 0.5 or 0.75. The detection rate was $\lambda = 0.1$, 0.2 or 0.3 detections/km, which give probabilities λ^* of detecting the species at an occupied site of 0.45, 0.70 and 0.83, respectively. For the Bernoulli process occupancy model the detection data were discretized based on three segment lengths (1, 2 and 3 km), assigning a success ('1') to those segments in which there was at least one detection. We then reran the simulations with the optimal per-site survey length for each scenario (i.e. $\lambda L = 1.57$, 1.88 and 2.54, respectively).

The simulation results (Table 4-2a) show that, at this sample size (600 km), the occupancy estimator has in general little bias. However, for rare and elusive species (low occupancy and low detection rate), the occupancy estimator is biased and its MSE is larger than that predicted by the asymptotic approximation. For instance, for $\psi = 0.25$, $\lambda = 0.1$ and $L = 6$, the asymptotic MSE, which corresponds to the asymptotic variance in (4.23) given that the estimator is asymptotically unbiased, is 0.013 while the actual MSE is 0.055.

The results also show that, as expected, when the detection process is a continuous process, the discretization of the data produces an occupancy estimator with larger bias and variance, especially when transects are divided into a few large segments. The estimator is more prone to estimates at the boundary of the parameter space (e.g. the proportion of estimates $\hat{\psi} = 1$ obtained was 0.052 in the Poisson process model, and 0.078, 0.118 and 0.188 in the Bernoulli process model with 1, 2 and 3 km segment lengths, when $\psi = 0.25$, $\lambda = 0.1$ and $L = 6$). This confirms that it is important to avoid discretizing continuous data if it is not really necessary.

The simulations corresponding to a study design based on the optimal per-site survey length (Table 4-2b) illustrate how the estimator properties improve in this case. With the same total survey effort, the occupancy estimator is less biased and more precise (e.g. for $\psi = 0.25$, $\lambda = 0.1$ the MSE of $\hat{\psi}$ decreased from 0.055 to 0.016). The improvement is especially noticeable for scenarios with low detection rate, as there was more discrepancy between the optimal survey length (15.7, 18.8 and 25.1 km respectively) and the 6 km initially used in the simulations. The loss in performance due to the discretization was less evident when working with an optimal survey length.

Table 4-2 Performance of the occupancy estimator in the Poisson process (PP) and Bernoulli process (BP) occupancy models when data are generated according to a Poisson process with detection rate λ, occupancy is ψ, the total survey effort available is 600 km and the survey design is based on either (a) $L = 6$ km or (b) the optimal L. Three segment lengths (1, 2 and 3 km) are tested for the BP model. Mean and mean square error (in square brackets) of the occupancy estimator $\hat{\psi}$ based on 10,000 simulations are shown. In (b) '----' is used to indicate that one case could not be evaluated as the total length was too short for the discretized data to contain more than one replicate.

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

4.2.6 Introducing covariates

The model structure in (4.9) may readily be expanded to allow the probability of occupancy (ψ) and/or the detection rate (λ) to depend upon site characteristics, such as habitat type or level of human disturbance, using a generalized linear model framework with a logit link function for the vector of site occupancies, $\boldsymbol{\psi}$, and a log link function for the vector of detection rates, $\boldsymbol{\lambda}$, so that

$$\begin{aligned} \boldsymbol{\psi} &= \frac{1}{1+\exp(-\boldsymbol{C\beta})}, \\ \boldsymbol{\lambda} &= \exp(\boldsymbol{D\alpha}), \end{aligned} \tag{4.24}$$

where $\boldsymbol{C}$ and $\boldsymbol{D}$ are matrices with site covariate information and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the vectors of model parameters.

The model can also be extended to accommodate within-site variation in the detection rate, by describing the detection process as a non-homogenous Poisson process. This results in the following construction for the likelihood function

$$\begin{aligned} L(\psi,\lambda) = \prod_{i=1}^{S} \Bigg[\psi \left\{ \prod_{j=1}^{d_i} \lambda(x_{ij}) \right\} \exp\left(- \int_{I_{L_i}} \lambda(x)dx \right) \\ + (1-\psi)\mathrm{I}(d_i = 0) \Bigg], \end{aligned} \tag{4.25}$$

where x_{ij} are the locations of the detections and I_{L_i} is the interval of length L_i surveyed. Obviously, since the locations of the detections are needed, the data can no longer be summarized by the number of detections at each site, d_i, as in (4.9).

The detection rate can now be expressed as a function of covariates that vary along the transect via a log link function

$$\lambda(x) = \exp(\alpha_0 + \alpha_1 D_1(x) + \alpha_2 D_2(x) + \cdots).$$

Most commonly in practice there will not be an explicit function describing the variation of the covariates along the transect, but rather discrete covariate values corresponding to transect sections, so the integration in (4.25) reduces to a finite summation as follows

$$L(\psi, \lambda) = \prod_{i=1}^{S} \left[\psi \left\{ \prod_{j=1}^{d_i} \lambda(x_{ij}) \right\} \exp\left(- \sum_{k=1}^{N_i} \lambda_{ik} L_{ik} \right) + (1 - \psi) \mathrm{I}(d_i = 0) \right] \quad (4.26)$$

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

4.3 2-MMPP occupancy model

4.3.1 Model formulation and assumptions

Considering the same sampling protocol as described in section 4.2.1, let us now suppose that the detections of the species along the transects exhibit some degree of clustering and thus cannot be considered independent. Here we propose to model the detections as a two-state Markov-modulated Poisson process (2-MMPP) with parameters $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$ and $\boldsymbol{\mu} = [\mu_{12}, \mu_{21}]$. The likelihood for such an occupancy model can be written as

$$L(\psi, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \prod_{i=1}^{S} \left\{ \psi \prod_{j=1}^{R_i} M_{ij} + (1-\psi)\mathrm{I}(d_i = 0) \right\}, \tag{4.27}$$

where M_{ij} is the expression for the contribution to the likelihood of the data from the j-th transect in site i, described as a 2-MMPP. Once again the model has a zero-inflation term, to reflect the fact that the species is absent from sites with probability $1-\psi$. Note that the assumption here is that when more than one transect is surveyed in a site, these can be considered independent and so their contributions to the likelihood can be simply multiplied.

From (4.7) we have that if the detection data along the transect is described as a n-MMPP, then M_{ij} can be written as

$$M_{ij} = \boldsymbol{\pi} \exp(\boldsymbol{C}l_{ij_1}) \boldsymbol{\Lambda} \ldots \exp\left(\boldsymbol{C}l_{ij_{d_{ij}}}\right) \boldsymbol{\Lambda} \exp\left(\boldsymbol{C}l_{ij_{d_{ij}+1}}\right) \boldsymbol{e}, \tag{4.28}$$

where $\boldsymbol{C} = \boldsymbol{Q} - \boldsymbol{\Lambda}$ and $l_{ij_1} \ldots l_{ij_{d_{ij}+1}}$ are the inter-detection distances with l_{ij_1} defined as the distance to the first detection from the beginning of the transect and $l_{ij_{d_{ij}+1}}$ defined as the distance from the last detection until the end of the transect (Figure 4-3). In the case of a 2-MMPP $\boldsymbol{e}^T = [1, 1]$, the generator matrix $\boldsymbol{Q}$ of the underlying Markov process is

$$\boldsymbol{Q} = \begin{bmatrix} -\mu_{12} & \mu_{12} \\ \mu_{21} & -\mu_{21} \end{bmatrix},$$

and the rate matrix $\boldsymbol{\Lambda}$ is

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Recall that $\boldsymbol{\pi}$ is the initial probability vector of the MMPP, that is, the probability of being in each of the two detection rate states (λ_1 or λ_2) at the beginning of the transect. If the start of the transect is chosen randomly, an appropriate specification for $\boldsymbol{\pi}$ is the equilibrium distribution of the underlying Markov process which fulfils (4.4) and which for a 2-MMPP is given by

$$\boldsymbol{\pi}' = [\, \pi_1' \quad \pi_2'] = \left[\frac{\mu_{21}}{\mu_{12} + \mu_{21}} \quad \frac{\mu_{12}}{\mu_{12} + \mu_{21}} \right]. \tag{4.29}$$

If instead the transect would be started at a point of detection, an appropriate specification for $\boldsymbol{\pi}$ is given by (4.5), which for a 2-MMPP is

$$\boldsymbol{\pi}^* = \left[\frac{\lambda_1 \pi_1'}{\lambda_1 \pi_1' + \lambda_2 \pi_2'} \quad \frac{\lambda_2 \pi_2'}{\lambda_1 \pi_1' + \lambda_2 \pi_2'} \right], \tag{4.30}$$

where π_1' and π_2' are given by (4.29). Although in our application transects are not started at a point of detection, (4.30) is relevant for the assessment of model fit, as we will see in section 4.4.3.

For the 2-MMPP, the matrix exponential $\exp(\boldsymbol{C}l)$ can be written in closed form (Rydén 1994) as

$$\begin{aligned} \exp[\boldsymbol{C}l] = D^{-1} \Bigg\{ & e^{-\theta_2 l} \begin{bmatrix} S_2 - \theta_2 & \mu_{12} \\ \mu_{21} & S_1 - \theta_2 \end{bmatrix} \\ & - e^{-\theta_1 l} \begin{bmatrix} S_2 - \theta_1 & \mu_{12} \\ \mu_{21} & S_1 - \theta_1 \end{bmatrix} \Bigg\}, \end{aligned} \tag{4.31}$$

where

$$\begin{aligned} S_1 &= \lambda_1 + \mu_{12}, \\ S_2 &= \lambda_2 + \mu_{21}, \\ K &= \lambda_1 \lambda_2 + \lambda_1 \mu_{21} + \mu_{12} \lambda_2, \\ \theta_1 &= (S_1 + S_2 + D)/2, \\ \theta_2 &= (S_1 + S_2 - D)/2, \\ D &= \sqrt{(S_1 + S_2)^2 - 4K}. \end{aligned}$$

These calculations can be computationally faster than using a general implementation of the matrix exponential. In a performance comparison with the function *expm* in MATLAB (version 7.12.0) we found around a six-fold difference in computing times.

For instance, the matrix exponential computation for $\mu = [0.1, 0.02]$, $\lambda = [2, 0.5]$ and $l = 20$ took an average of 0.0168 ms when using (4.31) compared to 0.1020 ms when using *expm*. This is especially relevant when fitting the model as the optimization process involves many evaluations of the likelihood function and, therefore, of matrix exponentials.

4.3.2 Relationship to the 2-MMBP occupancy model

The counterpart to the 2-MMPP occupancy model for discretized data would be a two-state Markov-modulated Bernoulli process (2-MMBP) occupancy model (Figure 4-6). In such a model, detections at occupied sites are described as coming from Bernoulli trials with two possible probabilities of success (detection probabilities p_1 and p_2). Which of the two probabilities of success is effective at each survey replicate is governed by a Markov chain (i.e. a discrete-time Markov process), with transition probabilities q_{12} (from the state with detection probability p_1 to the state with detection probability p_2) and q_{21} (from the state with detection probability p_2 to the state with detection probability p_1). We can expect the discrete 2-MMBP occupancy model to be a good approximation to the continuous 2-MMPP occupancy model if the continuous detection data are discretized using sufficiently small intervals, with $p_i \approx 1 - \exp(-\lambda_i L/K)$ and $q_{ij} \approx 1 - \exp(-\mu_{ij} L/K)$, where L/K is the length of transect segments.

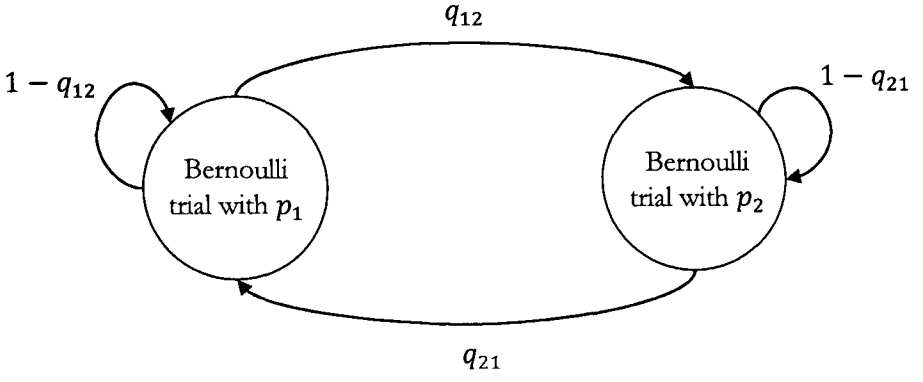


Figure 4-6 Detection process at occupied sites modelled as a 2-MMBP

In fact, the two models discussed in section 2.2.2, proposed by Hines *et al.* (2010) to generalize the Bernoulli process occupancy model and account for dependence between consecutive spatial replicates, are particular cases of a 2-MMBP occupancy model where p_2 is set to zero. Detections are therefore modelled as an interrupted Bernoulli process (IBP). In their 'trap response model', p_1 is also fixed, set to 1. The detection process at occupied sites in the 'Markov process for segment occupancy model', described by the Hidden Markov model shown in Figure 2-8, corresponds to a 2-MMBP with parameters $q_{12} = 1 - \theta', q_{21} = \theta, p_1 = p, p_2 = 0$. The detection process at occupied sites in the 'trap response model', described by the Markov chain shown in Figure 2-9, corresponds to a 2-MMBP with parameters $q_{12} = 1 - p', q_{21} = p, p_1 = 1, p_2 = 0$.

4.3.3 Maximum-likelihood estimation

In our study we estimated the parameters of the 2-MMPP occupancy model via maximization of the likelihood in (4.27). When this approach is followed one problem may arise. Since (4.28) involves matrix multiplications it is not possible to take logarithms

within the likelihood calculations corresponding to each transect, as is customary. The consequence is that, during the computation, the likelihood is more prone to take extreme values which are outside the range of the computer's floating point implementation, thus leading to optimization problems, as pointed out by various authors (Meier-Hellstern 1987; Rydén 1994; Skaug 2006). To prevent this, customized floating-point code needs to be written. The impossibility of taking logarithms becomes an issue when dealing with realizations of the MMPP consisting of many occurrences, that is, in our type of application if long transects were surveyed for a species producing many detections per unit length (Figure 4-7).

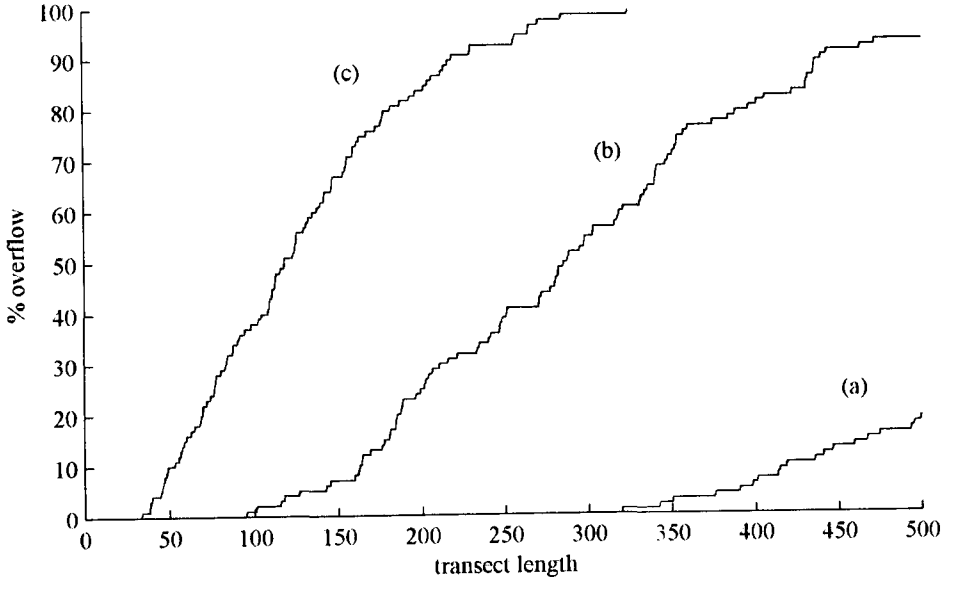


Figure 4-7 Percentage of simulations of a 2-MMPP transect that caused an arithmetic overflow when the likelihood was evaluated at the true parameter values. Scenario simulated: $\mu = [1/2, 1/15]$ and $\lambda = [\lambda_1, 0]$, with (a) $\lambda_1 = 10$, (b) $\lambda_1 = 15$ and (c) $\lambda_1 = 25$; 100 simulations.

Operations in MATLAB (version 7), the programming environment used in this thesis, are carried out in double-precision arithmetic conforming to the IEEE standard 754, so the largest finite floating-point number that can be represented is 1.7977e+308 and the smallest positive floating point number represented is 2.2251e-308. Although this representation was largely sufficient for the analysis of the Sumatran tiger data set in section 4.4, a simplified version of a customized floating-point code (in fact an adaptive scaling of the inbuilt floating point representation) was implemented for the simulation study to ensure that no numerical problems would be encountered. In the likelihood computation, for each inter-detection distance processed, we used two numbers to represent each of the elements in the 2×2 matrices obtained from calculating $\exp(\boldsymbol{C}l_i)\,\boldsymbol{\Lambda}$. A floating-point number (a) was used to represent the significant digits of the quantity, which was scaled as necessary to be kept within a reasonable range. The second number (b) was a counter representing the amount of scaling applied. Each element was represented by $a \cdot s^b$, with $s = 10^{50}$ and b chosen to ensure that $a \in [1/s, s]$. The product of the 2×2 matrices was carried out using this numerical representation. Once all the inter-detection distances in the transect were processed, the final quantity was transformed to a single floating point number on the logarithmic scale, $\log(a) + 50b\log(10)$, and this way the computation of the log-likelihood contribution for the transect was completed.

Provided that the numerical issues are satisfactory resolved, maximum-likelihood based model fitting via the Nelder-Mead simplex method has been shown to perform well compared to other approaches for MMPPs (Rydén 1994). Within the maximum-likelihood framework, another way to avoid the problem of numerical under/overflow would be to use an expectation-maximization (EM) algorithm. This kind of algorithm

works on the complete likelihood, which is a scalar, and therefore does not pose problems in terms of taking logarithms. For the case of standard MMPP models (without zero-inflation) at least two types of EM algorithm have been evaluated (Rydén 1994; Rydén 1996). In comparison to the optimization of the marginal likelihood via the Nelder-Mead simplex method, Rydén (1996) shows that the EM algorithm converges in a smaller number of iterations. However, he also points out that, since each of the EM algorithm iterations is more complicated than a likelihood evaluation, the actual implementation of the methods needs to be considered on a case by case basis to assess how they compare in terms of real-time performance.

4.3.4 Simulation study: PP and 2-MMPP model performance under clustering

We used simulations to investigate the performance of the two occupancy models presented when the detections along the transect are clustered. We generated data based on a 2-MMPP with one state without detections and another state, in which less time is spent, with detections. The probability of occupancy was set to $\psi = 0.25$, 0.5 or 0.75 and the sampling design had 100 sites with one transect of length $L = 20$ km surveyed per site. The 2-MMPP parameters used were detection rates $\boldsymbol{\lambda} = [0\ 5]$ and state switching rates $\boldsymbol{\mu} = [1/15\ 1/10]$, $\boldsymbol{\mu} = [1/15\ 1/2]$, $\boldsymbol{\mu} = [1/15\ 1]$, $\boldsymbol{\mu} = [1/15\ 1/0.5]$ and $\boldsymbol{\mu} = [1/15\ 1/0.2]$. With these parameter values and sampling design the probabilities of detecting the species at an occupied site, $1 - \boldsymbol{\pi} \exp(\boldsymbol{CL})\boldsymbol{e}$, were 0.83, 0.73, 0.68, 0.62 and 0.49 respectively.

The simulation results (Table 4-3) show that the occupancy estimator in the Poisson process occupancy model is negatively biased in the presence of clustering in the detection data. An informal interpretation for this negative bias is straightforward: given the average observed detection rate, the Poisson process model 'expects' that the sites without detections are less likely to be occupied than they really are; the model does not consider that the detections come in clusters so that, at occupied sites, relatively long stretches without detections leading to the species being missed are possible. As expected, the occupancy estimates obtained with the 2-MMPP model are much better, with only a very slight positive bias in the simulated scenarios.

Table 4-3 Performance of the occupancy estimator in the Poisson process (PP) and 2-state Markov-modulated Poisson process (2-MMPP) occupancy models when data are generated according to a 2-MMPP with detection rates $\boldsymbol{\lambda} = [0\ 5]$ and switching rates (a) $\boldsymbol{\mu} = [1/15\ 1/10]$, (b) $\boldsymbol{\mu} = [1/15\ 1/2]$, (c) $\boldsymbol{\mu} = [1/15\ 1/1]$, (d) $\boldsymbol{\mu} = [1/15\ 1/0.5]$ and (e) $\boldsymbol{\mu} = [1/15\ 1/0.2]$. The simulated sampling design consists of surveying 100 sites, with one 20 km transect surveyed per site. Mean and mean square error (in square brackets) of the occupancy estimator $\hat{\psi}$ based on 500 simulations are shown. The probability of detecting the species at occupied sites is also indicated.

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

4.3.5 Limiting case: PP mixture model

In a limiting situation with $\mu_{12} \to 0$ and $\mu_{21} \to 0$, the 2-MMPP tends to a mixture of two Poisson processes. Therefore, a model based on such a mixture will be appropriate in scenarios in which the surveyed transects are short with respect to the switching be-

tween detection rate states. Under this model the detections occur at two different rates λ_1 and λ_2 and detections along each transect take place at one of these two rates, with transitions between detection rates not allowed within individual transects. The detection process within each transect can therefore be described as a homogenous Poisson process of rate either λ_1 or λ_2, and there is a probability of the transect being in each of the two detection rates states (probability π_1 for state λ_1). The likelihood for this model is thus

$$L(\psi, \boldsymbol{\lambda}, \pi_1) = \prod_{i=1}^{S} \left[\psi \prod_{j=1}^{R_i} \left\{ \pi_1 \lambda_1^{d_{ij}} e^{-\lambda_1 L_{ij}} + (1-\pi_1) \lambda_2^{d_{ij}} e^{-\lambda_2 L_{ij}} \right\} + (1-\psi) \mathrm{I}(d_i = 0) \right], \tag{4.32}$$

where L_{ij} is the total length surveyed in the j-th transect in site i and d_{ij} is the total number of detections in that transect.

4.3.6 Identifiability

In a 2-MMPP, clustering arises as the two intensity parameters λ_1 and λ_2 differ from each other. When the two intensities are the same, the 2-MMPP reduces to a homogeneous Poisson process and the parameters μ_{12} and μ_{21} become unidentifiable. In the 2-MMPP occupancy model, for any given value of ψ, $\tilde{\psi}$ say, the likelihood function takes exactly the same value for all combinations of parameters satisfying:

(i) $\lambda_1 = \lambda_2 = \tilde{\lambda}$ regardless of the values of μ_{12} and μ_{21};

(ii) $\lambda_1 = \tilde{\lambda}$ when $\mu_{12} \ll \mu_{21}$, regardless of λ_2, as in effect this represents a case with only one state;

(iii) $\psi = \tilde{\psi}(1 + \tilde{\mu}_{12}/\tilde{\mu}_{21})$, $\lambda_1 = \tilde{\lambda}$, $\lambda_2 = 0$, $\mu_{12} \to 0$, $\mu_{21} \to 0$, if only one transect is surveyed per site. The model has two alternative explanations for sites without detections (unoccupied or occupied in state 2), so ψ and the ratio of switching parameters μ_{12}/μ_{21} are not separately identifiable.

The same identifiability issue arises for species that occupy sampling sites partially (i.e. $\lambda_1 = \tilde{\lambda}$, $\lambda_2 = 0$), if transects are short and without replication within sites.

4.3.7 Introducing covariates

As in the Poisson process occupancy model, information on site covariates can be incorporated easily in the 2-MMPP model using a logit link function for occupancy probability ψ and a log link function for the detection rates $\boldsymbol{\lambda}$ (4.24). Within-site detection covariates can also be incorporated in a similar way as in (4.26). Let us consider a case in which transect ij consists of two sections (A and B) with different detection rate parameter vectors, $\boldsymbol{\lambda_A}$ and $\boldsymbol{\lambda_B}$, and suppose that there were two detections in section A and none in section B. The expression for M_{ij} can then be written as

$$M_{ij} = \boldsymbol{\pi_A} \exp(\boldsymbol{C_A} l_{ijA1})\, \boldsymbol{\Lambda_A} \exp(\boldsymbol{C_A} l_{ijA2}) \boldsymbol{\Lambda_A} \exp(\boldsymbol{C_A} l_{ijA3}) \exp(\boldsymbol{C_B} l_{ijB1})\, \boldsymbol{e}, \quad (4.33)$$

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

4.4 Analysis of the Kerinci tiger data set

We illustrate the application of the models proposed in this chapter with an analysis of the Sumatran tiger data set from Kerinci Seblat National Park described in section 1.3.3, which consists of the location of footprint detections along transects and details on the transect routes followed in the survey.

4.4.1 Methods

We fitted the Poisson process model and the 2-MMPP model, as well as two variations of the latter: a particular case with one of the detection rates set to zero and the limiting Poisson process mixture model. Data were processed for the analysis by measuring the distances between detection points along the transect, as well as the distance from the start of the transect to the first detection and from the last detection until the end. In some of the sites more than one transect were surveyed. For the analysis, these transects were assumed to be statistically independent, so that their contributions to the likelihood could be multiplied, as in (4.27). In the surveys, transects were not started at a point of detection. We considered that the starting point was a random location in the landscape and consequently used (4.29) as an initial probability vector in the likelihood function.

4.4.2 Model selection and parameter estimates

The results from fitting the homogenous Poisson process occupancy model and the 2-MMPP occupancy model (Table 4-4) indicate that the latter model fits the Kerinci tiger data substantially better. Its AIC value was almost 60 units lower despite the penalty due to having three additional parameters. Fitting an interrupted Poisson process for the detection process (i.e. 2-MMPP with λ_2 fixed to zero) was also better than the

homogenous Poisson process but considerably worse than the general 2-MMPP. This suggests that, although rare, some detections take place outside areas of high detection rate.

Table 4-4 Parameter estimates with standard errors and AIC values for the occupancy models with a continuous detection process fitted to the Kerinci tiger data. Note - PP: homogenous Poisson process, 2-MMPP: 2-state Markov-modulated Poisson process, IPP: Interrupted Poisson process and PP Mixture: mixture of two homogenous Poisson process. The unit of $\hat{\lambda}_i$ is km^{-1} and the unit of $\hat{\rho}_{ij} = 1/\hat{\mu}_{ij}$ is km.

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

The estimate of occupancy under the 2-MMPP model is higher ($\hat{\psi} = 0.96$) than under the homogenous Poisson process model ($\hat{\psi} = 0.82$), which concurs with the fact that disregarding the dependence between detections causes a negative bias in the occupancy estimator. For this parameter the symmetric 95% confidence interval derived from the point estimate extended beyond unity (0.835-1.090). We also derived an interval based on the profile log-likelihood, that is, the log-likelihood maximised over all parameters other than ψ, with respect to ψ. The 95% confidence interval based on the profile log-likelihood contains all the values of ψ for which the profile log-likelihood

lies within $\chi^2_{1;0.05}/2 = 1.92$ of the maximum (Figure 4-8). In our case we obtained (0.831-1.000).

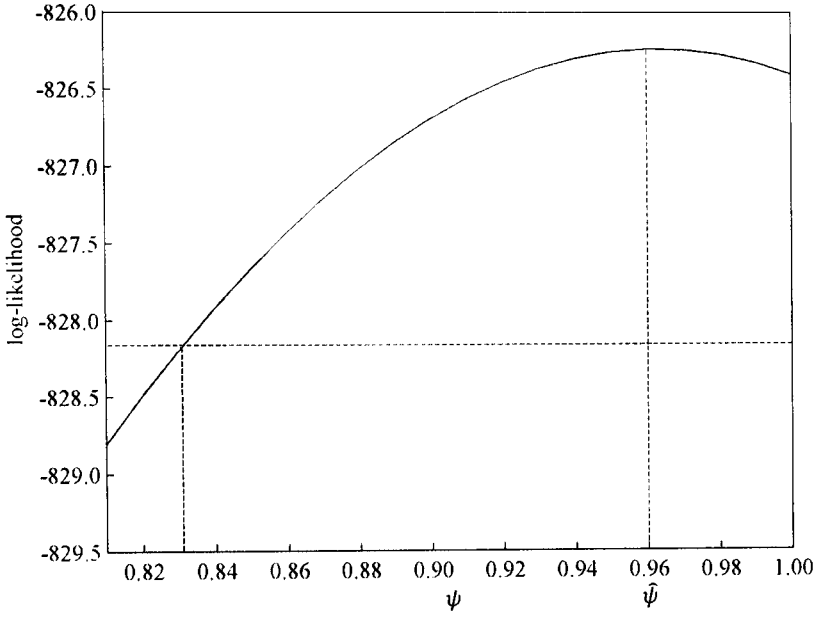


Figure 4-8 Profile log-likelihood for the occupancy parameter in the 2-MMPP tiger occupancy model. A 95%-confidence interval for the estimate of occupancy is derived as all occupancy values corresponding to profile log-likelihood values less than 1.92 units below the maximum. This threshold is shown as a dashed horizontal line.

The estimates associated with the detection process in occupied cells suggest that the rate of encounter of tiger footprints is ten times higher in some areas compared to others and that the average distance travelled in areas with high encounter rate ($\rho_{12} = 1/\mu_{12}$) is half the average distance travelled in areas with low encounter rates ($\rho_{21} = 1/\mu_{21}$). We chose to use a parameterization for the embedded Markov process in terms of expected holding times (ρ) instead of switching rates (μ) as this provides a more straightforward interpretation of our results. The point estimates of the ρ parameters are larger than we expected initially; however, their standard errors indicate poor precision. It is known that, in general, the switching rate parameters are more difficult

to estimate than the event (detection) intensities (Rydén 1996), which is in fact not surprising as the states themselves are not directly observable.

A reparameterisation of the model using the ratio and sum of the ρ parameters ($R = \rho_{12}/\rho_{21}$ and $A = \rho_{12} + \rho_{21}$) gave estimates and standard errors $\hat{R} = 0.5$ (0.15) and $\hat{A} = 365$ (628) showing that the estimate of the ratio is relatively precise and that most of the uncertainty lies in the magnitude of the parameters. This is interesting, as it suggests that the tiger data are informative in terms of the probability of being in each of the two states but carry little information on the actual rate at which transitions between states take place. The transects walked in this survey were short compared to the rate at which state transitions seem to take place, which prevents the accurate estimation of these parameters.

Our results suggest that, due to transect length, transitions are unlikely to occur within transects so we could expect that the model based on a mixture of two Poisson processes in section 4.3.5 would describe our detection data well. Indeed, fitting this model provided a very similar likelihood value (826.4 vs. 826.2) and practically the same estimates of occupancy and detection rates as under the 2-MMPP model. The estimate of the probability of being in the state with high detection rate $\hat{\pi}_1 = 0.35$ (0.064) also matches the corresponding estimate based on the 2-MMPP model $\hat{\pi}_1 = \hat{R}/(\hat{R} + 1) = 0.33$ (0.067).

4.4.3 Model diagnostics

We explored two aspects of goodness-of-fit for the models fitted by assessing how well they describe

(i) the inter-detection distances and

(ii) the distances from the beginning of each transect until the first detection.

We did this by comparing the survivor function of these two random variables according to the fitted models with the empirical survivor function obtained directly from the recorded data. The survivor function, also known as survival function or reliability function depending on the application, reflects the probability that a characteristic of a system will 'survive' beyond a specified time. For a continuous random variable X, with $f(x)$ its probability density function and $F(x)$ its cumulative distribution function, the survivor function is

$$S(x) = P(X > x) = \int_x^\infty f(u)\,du = 1 - F(x). \tag{4.34}$$

Obviously, the survivor function is monotone decreasing and unity at time zero (unless there is a non-negligible probability that the system will 'fail' immediately, which is not our case here as the inter-detection distances are strictly larger than zero).

The survivor function for the inter-detection distances in the Poisson process model is given by

$$S_{PP}(l) = e^{-\lambda l}, \tag{4.35}$$

and in the 2-MMPP model is

$$S_{2-MMPP}(l) = \boldsymbol{\pi}^{*} \exp(\boldsymbol{C}l)\, \boldsymbol{e}. \tag{4.36}$$

Since we are assessing the distance between detections it implies that the interval starts at a point of detection, therefore in (4.36) we use $\boldsymbol{\pi}^{*}$, given in (4.30), for the initial probability vector.

The survivor function for the distance until first detection in the Poisson process model is given by

$$S_{PP}(l) = \psi e^{-\lambda l} + (1 - \psi), \tag{4.37}$$

and for the 2-MMPP model is

$$S_{2-MMPP}(l) = \psi\, \boldsymbol{\pi}' \exp(\boldsymbol{C}l)\, \boldsymbol{e} + (1 - \psi). \tag{4.38}$$

These expressions account for the fact that we are dealing with zero-inflated models. The survivor function is a mixture of two functions. With probability ψ, the function is that given by a Poisson process or a 2-MMPP. With probability $1 - \psi$, the site will not be occupied by the species and therefore the probability that the distance to first detection in these sites is larger than a given quantity is always one. Note that in (4.38) we now use $\boldsymbol{\pi}'$ for the initial probability vector, as we are considering that the transects start at a random point.

As an empirical survivor function we used the Kaplan-Meier curve (Kaplan & Meier 1958), which takes account of the right-censoring in the data, due to transects that

ended before detecting the next (or any) tiger footprint. The Kaplan-Meier estimator $\hat{s}(l)$ is calculated as the product

$$\hat{S}(l) = \prod_{l_i \leq l} \left(1 - \frac{d_i}{n_i}\right), \tag{4.39}$$

where $\{l_i\}$ are the observed distances to detection (first or next), n_i represents the number of intervals longer than l_i (including those that ended without detection) and d_i is the number of detections made at a distance l_i.

The comparison of the two fitted survivor functions with the corresponding empirical survivor functions confirms that the 2-MMPP occupancy model has a better fit than the Poisson process occupancy model (Figure 4-9).

As an additional assessment of fit we fitted a three-state MMPP occupancy model. The maximum-likelihood estimates obtained were: $\hat{\psi} = 0.96$, $\hat{\lambda}_1 = 0.99$, $\hat{\lambda}_2 = 0.23$, $\hat{\lambda}_3 = 0.03$, $\hat{\rho}_{12} = 2681$, $\hat{\rho}_{13} \approx 0$, $\hat{\rho}_{21} = 125$, $\hat{\rho}_{23} = 4667$, $\hat{\rho}_{31} = 3093$ and $\hat{\rho}_{32} = 243$. Looking at these estimates it is evident that this 3-MMPP model in effect collapses to the 2-MMPP model in Table 4-4, therefore indicating that a model structure with two states provides good fit for these data in this regard.

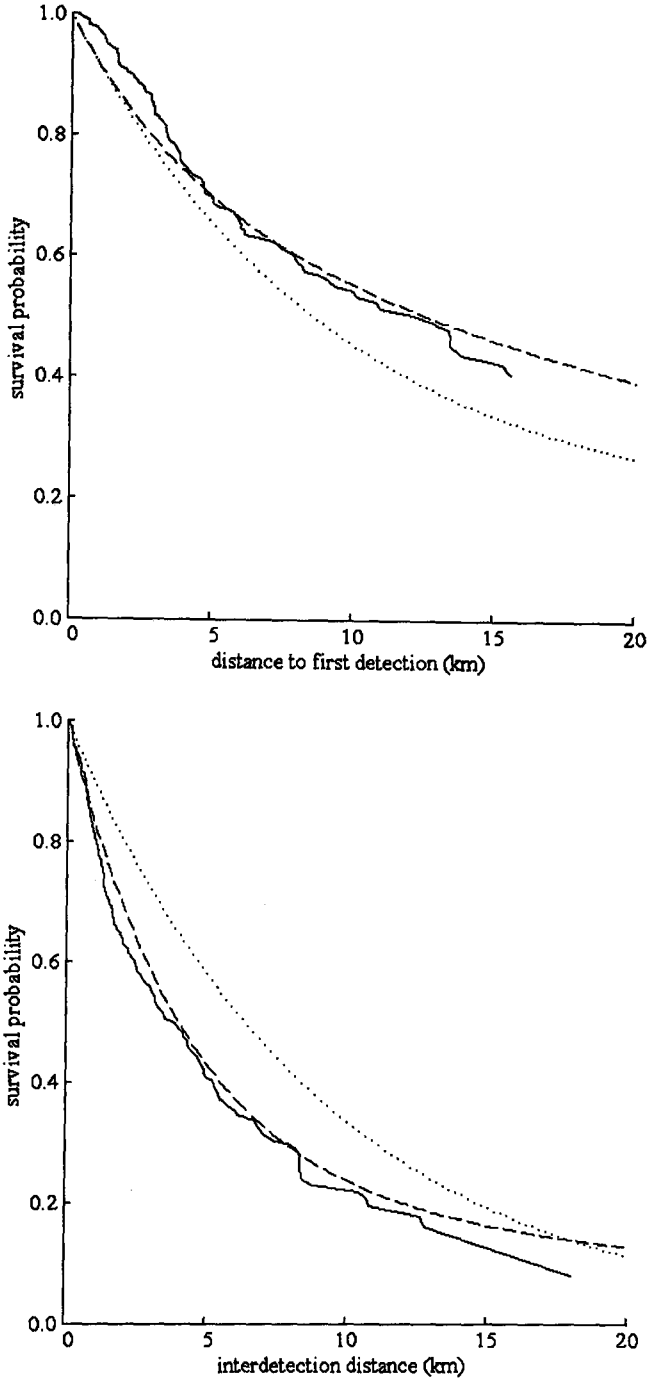


Figure 4-9. Empirical and fitted survivor functions for 'distance to first detection' (top) and 'inter-detection distances' (bottom) for the Kerinci Sumatran tiger data. The empirical survivor function (solid line) is the Kaplan-Meier plot. The survivor functions for the fitted 2-MMPP occupancy model (dashed line) and the Poisson process occupancy model (dotted line) are derived using parameter estimates in Table 4-4.

4.5 Discussion

In this chapter we describe the detection process as a continuous process when modelling occupancy from detection data collected in transect surveys. In particular we develop occupancy models that use a Poisson process to describe detections that can be assumed independent and alternatively a 2-MMPP for cases with clustered detections. We found that the 2-MMPP occupancy model was more appropriate than a model based on a homogeneous Poisson process for the Kerinci-Seblat tiger detection data. We also propose an occupancy model that describes species detections as a mixture of two Poisson processes, which can be seen as a limiting case of the 2-MMPP occupancy model when $\mu_{12} + \mu_{21}$ approaches zero (i.e. when switching between detection states is extremely unlikely within a transect). This model is simpler to fit and provided an adequate description of our tiger data. However, in general, the 2-MMPP model will be more appropriate for modelling data of this type as it accounts for potential state transitions within transects, which is relevant if transects are long compared to the scale of clustering in detections of the species.

When discussing the use of the discrete occupancy model to analyze data collected along transects, Hines *et al.* (2010) raise the question of whether there is an optimal transect segment length with respect to estimator properties (in terms of precision or MSE). Here we argue that in general it is more appropriate to model the detection process along transects as a continuous point process, as this provides a more natural description for this kind of data. In practice this is equivalent to dividing the transects into infinitesimal segments. Pooling detections from larger transect segments may result in a poorer occupancy estimator, less precise and more prone to boundary esti-

mates, which is not surprising as data are lost in the discretization process. The use of relatively long transect segments can however help mitigate the potential lack of independence between adjacent segments, when not explicitly accounted for in the model. An approximation to the continuous model can of course also be implemented by using a short segment length in the discrete models.

The use of point processes for the description of the detection process in occupancy modelling opens the door to a number of different model developments, including the extension of the models described to account for abundance-induced heterogeneity, which we investigate next in chapter 5. It is also interesting to note the parallelism of the occupancy models described here with models used in other applications. In survival analysis for instance the target is to model the time to an event, such as the onset of a medical condition or the failure of an electrical system (Cox & Oakes 1984). Survival data tend to be right-censored as experiments often end before all susceptible individuals develop the condition, in the same way as surveys may end before detecting the species of interest at occupied sites, causing the so-called 'false absences'. Furthermore, there are models that account for individuals that are 'immune' to the condition of interest. These individuals, referred to as 'long-term survivors', introduce zero-inflation in the same way that non-occupied sites do in species occupancy data. Survival analysis, counting processes and, in general, point process theory therefore provide an opportunity for bringing new ideas into species occupancy modelling, and vice versa.

5 EXTENSIONS FOR ABUNDANCE-INDUCED HETEROGENEITY

In the context of occupancy models based on discrete sampling protocols, it has been shown that heterogeneity in detection probability can induce bias in the occupancy estimator (MacKenzie *et al.* 2006; Royle 2006), as discussed in section 2.2.2. In particular, differences in abundance between sites can sometimes be a significant source of heterogeneity and therefore of bias in occupancy (Dorazio 2007). To deal with this situation, Royle & Nichols (2003) propose a mixture model that directly links the latent local abundance with the species detectability at the site (section 2.2.4). Provided that the model assumptions are satisfied, this model yields better occupancy estimates in the presence of abundance-induced heterogeneity in detectability than the model that assumes constant detection probability, as well as providing an estimate of local abundance.

In this chapter we extend the models introduced in chapter 4 to account for abundance-induced heterogeneity in the detection process. Once again we consider a sampling scenario in which species detection surveys are carried out along one or more transects at a number of sampling sites, recording the location of each detection. The new models assume that the species detection process at each site can be well described as the

result of the superposition of n_i identical point processes, where n_i is the number of individuals present at site i. Each point process describes the detections corresponding to one individual and the assumption is that all individuals present at the site are equally detectable from the transect. As n_i is unknown, it is modelled as a random variable with some probability distribution and inference is made on the marginal likelihood, which is a discrete mixture over all the possible abundance values, as in the model of Royle & Nichols (2003).

We start the chapter by describing in section 5.1 the extension of the Poisson process model presented in section 4.2 to account for abundance-induced heterogeneity. This model assumes independence within the detections of each individual. However, as we showed in chapter 3, clustering in species detections along transects can induce bias in the estimator of site occupancy. Therefore we can also expect bias to be induced in the abundance distribution estimation if there is clustering within the detections of individuals and this is not accounted for. To deal with this scenario, in section 5.2 we propose a model for the detection data which accounts simultaneously for both abundance-induced heterogeneity and clustering in individuals' detections. The performance of the models proposed in this chapter is explored via simulations in section 5.3, including the impact that unmodelled detection clustering can have in the estimator of abundance. Finally, in section 5.4 we illustrate the application of these new models by fitting them to the Kerinci Seblat tiger data set, comparing the results with those that were obtained from fitting the occupancy models in the previous chapter.

To deal with the problem of accounting for both spatial clustering and abundance-induced heterogeneity in detection data collected along transects (i.e. continuous pro-

tocols), Hines *et al.* (2010) suggest a 2-step ad hoc approach: to use their discrete clustering model to explore how the dependence among adjacent replicates decreases as data are collapsed using larger segments and then, using the data resulting from a segment length chosen so that clustering can be considered unimportant, to carry out the actual analysis with the standard Royle–Nichols model. The work in this chapter provides an alternative solution to this problem based on a description of the detection process that allows us to account for both aspects simultaneously. This model is, as far as we are aware, the first that allows species-detection data from continuous sampling protocols to be analysed by explicitly accounting for both clustering and abundance-induced heterogeneity in the detection process. By providing a description of the detection process that explicitly incorporates both aspects, the model allows not only the estimation of abundance but also of the parameters associated with the clustering pattern. The work in this chapter has been published in Guillera-Arroita *et al.* (2012).

5.1 Model for abundance-induced heterogeneity

5.1.1 Model formulation and assumptions

Let us first consider a case where the successive detections of each individual can be considered independent of one another and so can be modelled as a homogenous Poisson process with intensity γ, where γ is the average number of detections per individual over a unit length. The detection process for the species at site i, modelled as the superposition of n_i independent identical Poisson processes, results in a Poisson process with rate γn_i. Under this model the detection data can be summarized by the total number of detections at each site, d_i, and the likelihood is

$$L(\boldsymbol{\theta}, \gamma) = \prod_{i=1}^{S} \left[\sum_{n_i=0}^{\infty} \{(\gamma n_i)^{d_i} \exp(-\gamma n_i L_i) \Pr(n_i|\boldsymbol{\theta})\} \right] \tag{5.1}$$

where L_i is the total length surveyed in site i, S is the total number of sampling sites and $\Pr(n_i|\boldsymbol{\theta})$ denotes the probability mass function for the distribution of abundances described by the parameter vector $\boldsymbol{\theta}$. Species occupancy ψ would therefore be derived here as $1 - \Pr(n_i = 0|\boldsymbol{\theta})$.

As in the likelihood for the Poisson process occupancy model in (4.9), the likelihood in (5.1) lacks the factors $L_i^{d_i}/d_i!$ that would result from considering the detection data as a Poisson count rather than as a series of independent exponentially distributed inter-detection distances. As a consequence, the magnitude of the likelihood is comparable to that from the clustering models which is necessary for the purpose of model selection.

As discussed in connection with the Royle-Nichols model in section 2.2.4, site abundance can be described by parametric or non-parametric distributions. In the non-parametric approach the number of parameters required increases with the number of support points in the abundance mixture and so it may become impracticable when working with certain species; however this method provides flexibility and can be useful for instance when working with species for which abundance is low at the spatial scale of the survey. Within the parametric approach a first candidate is the Poisson distribution, which assumes that individuals occur completely at random. In section 5.1.3 we study this model in further detail, and derive some study design guidelines based on asymptotic approximations.

While in (5.1) we have assumed constant parameters, the model can be expanded to incorporate covariates following a generalized linear model approach. Under the parametric approach, the abundance distribution can be readily allowed to depend upon site characteristics, for instance via a log link function on the density parameter when a Poisson distribution is used. Covariates can be incorporated in the same manner to allow the detection rate to vary with respect to site characteristics, while within-site variation in the detection rates can also be accommodated as explained in section 4.2.6.

5.1.2 Relationship to other models

The model proposed here can be interpreted as the continuous counterpart of the Royle-Nichols model described in section 2.2.4, which relies on a discrete sampling protocol that records the detection/non-detection of the species in each replicate (records consist of 0s and 1s), and models the detections at each site as a series of inde-

pendent Bernoulli trials. However, in effect, our model is closer to the N-mixture abundance model for repeated counts (Royle 2004b) described in section 2.2.5 but with an important difference. The model for repeated counts, based on discrete replicates, describes the number of detections in each replicate (i.e. the counts) as binomially distributed. This implies that an already-detected individual cannot be detected again in the same replicate. Our Poisson process based model arises if the repeated counts are instead described using a Poisson distribution, i.e. one can detect in the same replicate an individual that has already been detected. This difference has important implications as, if the binomial model is used when the absence of repeated detections of individuals cannot be guaranteed within a replicate, then the estimation of abundance can be significantly inflated. This may be relevant for some surveys based on direct observations (e.g. camera-trap surveys without individual identifications, bird point counts) and is often crucial when modelling indirect observation data (e.g. pug-marks), as each individual can leave more than one sign. Therefore, apart from its utility to model detection data collected along a transect or period of time, the model described here can also be useful when modelling repeated counts obtained from (discrete) sampling protocols based on separate survey visits.

Note also that, if the mixing distribution in (5.1) is a non-parametric distribution with mass only at $n_i = 0$ and $n_i = m$, the model reduces to the Poisson process occupancy model (section 4.2) with likelihood

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

where $\gamma m = \lambda$ and the abundance probability distribution is $\Pr(0) = 1 - \psi$ and $\Pr(m) = \psi$.

5.1.3 Poisson mixture model: design recommendations and performance

Under the parametric approach a natural first candidate for the distribution describing site abundance is the Poisson distribution, which provides an appropriate description when the individuals are distributed completely at random. In this case the likelihood for the model in (5.1) becomes

$$L(\delta, \gamma) = \prod_{i=1}^{S} \left[\sum_{n_i=0}^{\infty} \left\{ (\gamma n_i)^{d_i} \exp(-\gamma n_i L_i) \frac{e^{-\delta} \delta^{n_i}}{n_i!} \right\} \right], \tag{5.2}$$

where δ is the average site abundance for the species.

In this model the resulting distribution describing the number of detections of the species in site i, d_i, is a Poisson mixture of Poisson distributions,

$$d_i \sim \text{Poisson}(\gamma n_i L_i) \text{ where } n_i \sim \text{Poisson}(\delta),$$

which gives rise to a Neyman type A distribution with parameters $(\gamma L_i, \delta)$ (Neyman 1939; Douglas 1980, pp. 153-258; Johnson, Kemp & Kotz 2005, pp. 403-410). There are two standard expressions for the probability mass function in a Neyman type A distribution. The first, which follows directly from considering the above mixture, is

$$\begin{aligned}\Pr(d_i) = P_{d_i} &= \sum_{n_i=0}^{\infty} \left\{ \frac{(\gamma n_i L_i)^{d_i} \exp(-\gamma n_i L_i)}{d_i!} \frac{e^{-\delta} \delta^{n_i}}{n_i!} \right\} \\ &= \frac{(\gamma L_i)^{d_i} e^{-\delta}}{d_i!} \sum_{n_i=0}^{\infty} \left\{ \frac{{n_i}^{d_i} (\delta e^{-\gamma L_i})^{n_i}}{n_i!} \right\}.\end{aligned} \tag{5.3}$$

It can be shown that (5.3) can also be written as

$$P_{d_i} = \frac{(\gamma L_i)^{d_i} \exp(-\delta + \delta e^{-\gamma L_i})}{d_i!} \sum_{k=0}^{d_i} \mathcal{S}(d_i, k) (\delta e^{-\gamma L_i})^k , \tag{5.4}$$

where $\mathcal{S}(d_i, k)$ are Stirling numbers of the second kind, an expression first given by Cernuschi & Castagnetto (1946). The Stirling numbers of the second kind (Douglas 1980, p. 471) compute the number of ways of partitioning a set of d_i elements into k nonempty sets and are calculated as

$$\mathcal{S}(d_i, k) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^{d_i} . \tag{5.5}$$

The expression for the probability mass function in (5.4) involves a finite summation, limited to d_i, while (5.3) involves an infinite summation in n_i. Although in practice in the computation of probabilities the infinite summation would be truncated to the support points with non-negligible probability, we can expect (5.4) to be better computationally for low values of d_i and high δ. In a comparison of the time it takes to compute a probability with the two expressions we found that, in our implementation, (5.4) was considerably better than (5.3) for $d_i < 5$, with a saving factor for the cases we evaluated of up to 66 times when $d_i = 0$ (Table 5-1).

Table 5-1 Comparison of probability computation times between the two forms of the probability mass function in a Neyman type A distribution with parameters $(\gamma L, \delta)$. Numbers represent the ratio between the computation time of expression (5.3) and expression (5.4) and were obtained based on average times from 100,000 calculations. In the computation of (5.3) the summation on n_i was truncated to those values outside the 0.001 probability tails in the Poisson with parameter δ.

		d_i					
		0	1	2	5	10	20
	0.5	0.031	0.154	0.260	0.728	1.671	4.981
	1.0	0.026	0.141	0.247	0.615	1.478	4.737
δ	2.0	0.022	0.123	0.204	0.582	1.354	4.385
	5.0	0.018	0.087	0.176	0.458	1.067	3.293
	10.0	0.015	0.074	0.143	0.304	0.808	2.167

This reduction in computing times can be valuable when obtaining maximum-likelihood estimates, as this process involves many probability evaluations. An alternative form for the likelihood function in (5.2) based on the probability mass function expression in (5.4), and removing as before the factors $L_i^{d_i}/d_i!$, is

$$L(\delta, \gamma) = \prod_{i=1}^{S} \left\{ \gamma^{d_i} \exp(-\delta + \delta e^{-\gamma L_i}) \sum_{k=0}^{d_i} S(d_i, k)(\delta e^{-\gamma L_i})^k \right\}. \qquad (5.6)$$

Note that this expression is based on the assumption that the Poisson processes describing individuals' detections are homogenous. If detection rate changes within the site, the data can no longer be summarized by the number of detections in each site, d_i, and cannot be described as coming from a Neyman type A distribution.

In order to explore study design trade-offs and model performance, let us assume that the parameters δ and γ are constant and that we have a design in which the same length L is surveyed in each site. These assumptions are equivalent to those made in previous chapters when deriving general study design guidelines for other models. Under these conditions, the data for the Neyman type A model can be summarized by $\{F_x\}$, with F_x denoting the number of sites in which x detections were recorded. It can be shown than the MLEs of the parameters satisfy (Douglas 1980, p. 187)

$$\hat{\delta}\hat{\gamma}L = \frac{S_1}{S_0}, \tag{5.7}$$

$$\sum_{x=0}^{\infty} F_x \hat{\Pi}_x = S_1, \tag{5.8}$$

where

$$S_i = \sum_{x=0}^{\infty} x^i F_x \ , \quad i = 0, 1,$$

and

$$\Pi_x = (x+1)\frac{P_{x+1}}{P_x}, \tag{5.9}$$

with P_x the probability of drawing a count of x from a Neyman type A distribution with parameters δ and γL, given by (5.3) and (5.4); in (5.8) $\hat{\Pi}_x$ represents Π_x evaluated at the MLE's $\hat{\delta}$ and $\hat{\gamma}$. Note that (5.7) allows one of the parameters to be expressed as a function of the other, thereby reducing the problem of finding the MLEs to solving a single equation, (5.8).

The total effort available for the survey can be allocated in different ways: more sites with shorter transects or fewer sites with longer transects. In section 4.2.4 we identified the design that minimizes the variance of the occupancy estimator in the Poisson process occupancy model for different occupancy scenarios. Similarly here we can expect an optimum design in terms of the precision of the estimator of average site abundance $\hat{\delta}$, as increasing the number of sites S involves a trade-off: it provides more samples for estimating the abundance process but, at the same time, the detection data from shorter transects provide less information about the number of individuals at each site.

In order to explore study design trade-offs here we assume that the effort involved in surveying a unit length is constant, with no significant overheads when adding new sites, so that the cost of the survey can be evaluated in terms of the total survey effort $E = SL$. To evaluate study design performance we can look at the first-order asymptotic expressions for the variance-covariance matrix of the MLEs, which for a Neyman type A distribution with parameters δ and γL is given by (Shenton 1949; Douglas 1980, pp. 190-191)

$$\Sigma(\hat{\delta}, \hat{\gamma}L) = \frac{1}{S}\frac{1}{(1+\gamma L)\phi - \delta(\gamma L)^2(\delta + \delta\gamma L + \gamma L)} \cdot \begin{pmatrix} \delta\gamma L\{\phi + \delta\gamma L - \delta(\gamma L)^2(1+\delta)\} & -\gamma L\{\delta(\gamma L)^2(1+\delta) - \phi\} \\ -\gamma L\{\delta(\gamma L)^2(1+\delta) - \phi\} & \frac{\gamma L}{\delta}\{\phi - (\delta\gamma L)^2\} \end{pmatrix}, \tag{5.10}$$

where

$$\phi = \mathbb{E}[\Pi_x^2] = \sum_{x=0}^{\infty} \{(x+1)^2 P_{x+1}^2 / P_x\}. \tag{5.11}$$

The variance-covariance matrix for $\hat{\delta}$ and $\hat{\gamma}$ is

$$\Sigma(\hat{\delta}, \hat{\gamma}) = \begin{pmatrix} \Sigma(\hat{\delta}, \hat{\gamma}L)_{1,1} & \frac{1}{L}\Sigma(\hat{\delta}, \hat{\gamma}L)_{1,2} \\ \frac{1}{L}\Sigma(\hat{\delta}, \hat{\gamma}L)_{1,2} & \frac{1}{L^2}\Sigma(\hat{\delta}, \hat{\gamma}L)_{2,2} \end{pmatrix}. \tag{5.12}$$

Let f, h and g be functions of δ and γL, then (5.12) can be written as

$$\Sigma(\hat{\delta}, \hat{\gamma}) = \begin{pmatrix} \frac{1}{\gamma E} f(\delta, \gamma L) & \frac{1}{E} h(\delta, \gamma L) \\ \frac{1}{E} h(\delta, \gamma L) & \frac{\gamma}{E} g(\delta, \gamma L) \end{pmatrix}, \tag{5.13}$$

where $E = SL$ is the total survey effort. As indicated by (5.13), for a particular scenario of δ and γ the optimal design is determined by δ and γL, as γ and E are only scaling the variance-covariance functions. The process of identifying an appropriate design requires an initial assumption for the values of the parameters γ and δ. Given the assumed value for δ, an optimum γL can be identified. The optimal survey length can then be derived from this based on the assumed γ. The number of sites to survey follows immediately from considering the total effort available, $E = SL$. Finally the precision of the estimators for the identified design can be assessed using (5.13).

In practice, to compare the performance of designs under different scenarios of δ and γ, it is useful to consider the coefficient of variation (CV) of the estimators rather than their variances, as the CV takes into account the magnitude of δ and γ. Note that the

CV of both $\hat{\delta}$ and $\hat{\gamma}$ can be written as a function of δ and γL, $\tilde{f}$ and $\tilde{g}$ say, scaled by a factor $(\gamma E)^{-1/2}$:

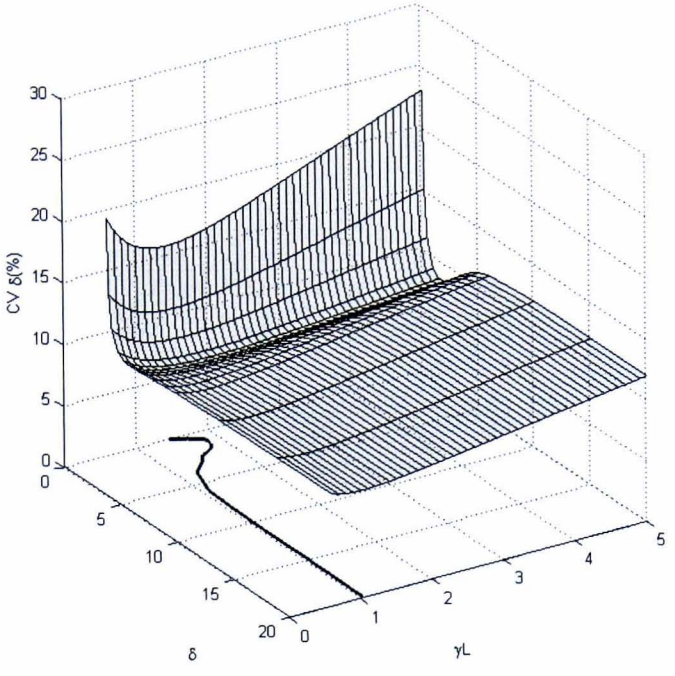
$$CV_{\hat{\delta}} = \frac{\sqrt{\text{var}(\delta)}}{\delta} = \frac{1}{\sqrt{\gamma E}} \tilde{f}(\delta, \gamma L),$$

$$CV_{\hat{\gamma}} = \frac{\sqrt{\text{var}(\gamma)}}{\gamma} = \frac{1}{\sqrt{\gamma E}} \tilde{g}(\delta, \gamma L).$$

Therefore the plots in Figure 5-1, which compare the CV for different combinations of δ and γL, are essentially independent of γ and E except for a scaling factor.

Regardless of whether the coefficient of variation or the variance is considered, the optimal design for a given scenario (Figure 5-2) is the same. When the average abundance δ is very small, the optimum design in terms of minimizing the variance of $\hat{\delta}$ is to use a transect length so that γL is about 1.5. As δ increases, the optimum γL increases to reach 1.8 when δ is approximately 1.4 and from this point on it decreases approaching unity (i.e. one detection on average per individual at each site). In terms of minimizing the variance of the individual detection intensity estimator, $\hat{\gamma}$, when δ is low the best strategy is to have long transects. However, as δ increases the optimum γL decreases, once again approaching unity.

(a)



(b)

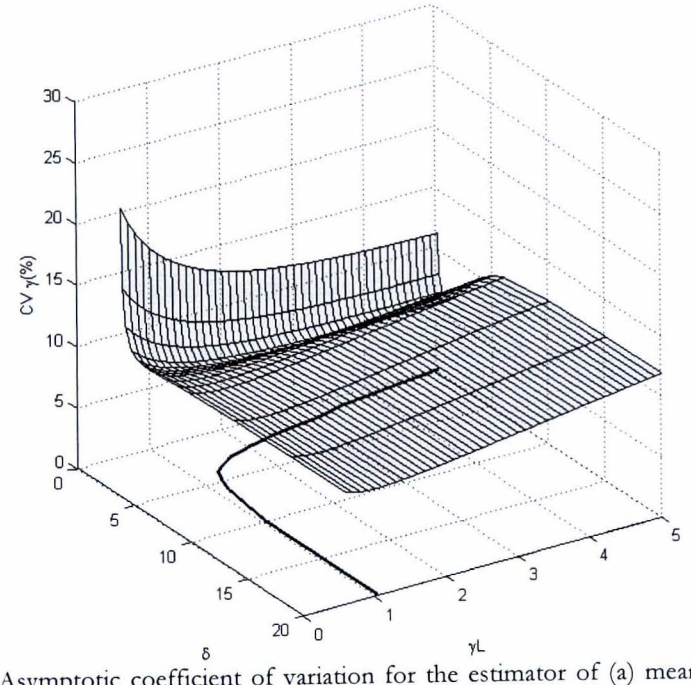


Figure 5-1 Asymptotic coefficient of variation for the estimator of (a) mean abundance $\hat{\delta}$ and (b) mean individual detection rate $\hat{\gamma}$, as a function of δ and γL. The black line shows the value of γL that minimizes the coefficient of variation for each value of δ. Plots were produced based on $\gamma = 1$ and $E = 1000$. For other values of γ and E the resulting surfaces will be a scaled version of the ones shown here.

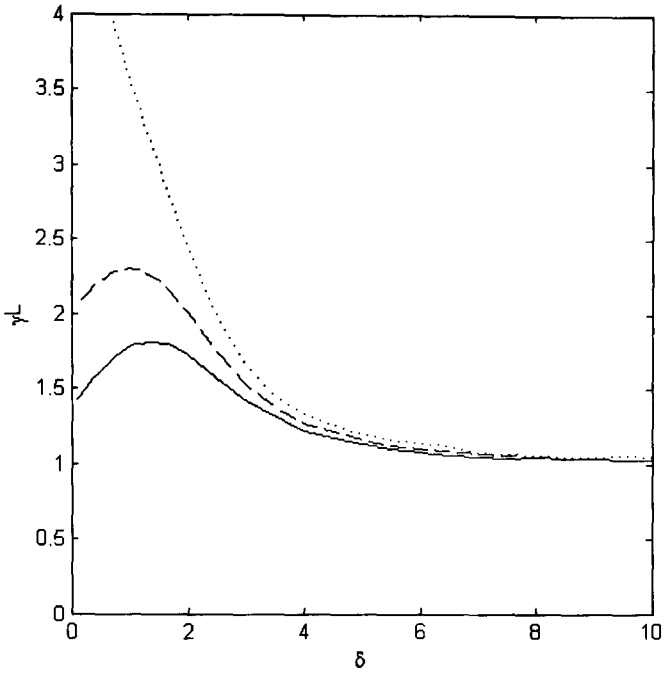


Figure 5-2 Average number of individual detections per site γL to minimize the asymptotic variance of the abundance estimator $\hat{\delta}$ (solid line), the asymptotic variance of the individual detection rate estimator $\hat{\gamma}$ (dotted line) and the sum of the asymptotic variances of $\hat{\delta}$ and $\hat{\gamma}$ (dashed line), as a function of average abundance δ.

For small δ the probability mass function for a Neyman type A distribution can be approximated by

$$\begin{aligned} P_0 &\approx (1-\delta) + \delta e^{-\gamma L}, \\ P_x &\approx \delta \frac{e^{-\gamma L}(\gamma L)^x}{x!}, \quad x = 1, 2, \ldots, \end{aligned} \tag{5.14}$$

which corresponds to a zero-inflated Poisson distribution with zero-inflation parameter $1 - \delta$ and rate γL (Martin & Katti 1962; Douglas 1980, p. 166; Johnson, Kemp & Kotz 2005, p. 406). We derived design recommendations for such a model in section 4.2.4, where the optimal survey length was identified using the asymptotic variance of the

occupancy estimator $\hat{\psi}$ (i.e. 1 – zero-inflation) as a design criterion. As expected, the optimal transect length values identified here to minimize the variance of $\hat{\delta}$ for small δ match those derived for the zero-inflated Poisson model for small ψ (Figure 5-3).

A second limiting form worth mentioning here is that, if γL is small, the Neyman type A distribution is approximately a Poisson distribution with expected value $\delta\gamma L$. In such circumstances δ and γ become non-identifiable, that is, transects are not long enough to estimate the system process (i.e. abundance) and the detection process separately, as also happens with the lack of replication in the standard Bernoulli process occupancy model or with short transects in the Poisson process occupancy model.

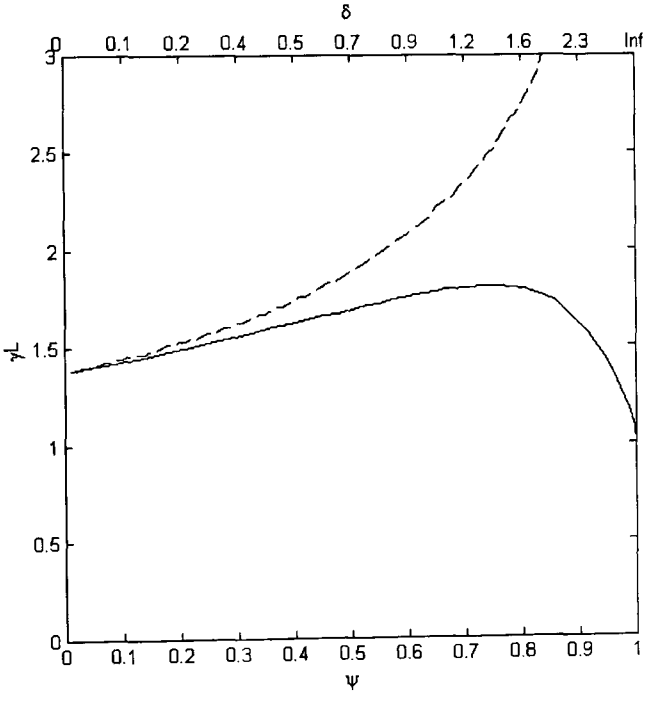


Figure 5-3 Comparison of the optimal design (average number of individual detections per site γL) for the Neyman type A model (solid) and the zero-inflated Poisson model (dashed), as a function of occupancy ψ. In the Neyman type A model the design is that which minimizes the asymptotic variance of the abundance estimator $\hat{\delta}$, and in the zero-inflated Poisson model that which minimizes the variance of the occupancy estimator $\hat{\psi}$. In the top axis the corresponding value of average abundance δ in the Neyman type A model is shown.

While the recommendations based on asymptotic properties of the estimators provide a useful design guideline, it is important once again to bear in mind that these approximations may not be appropriate when working with a small sample size and that it is always advisable to verify the performance of the chosen design via simulations. Shenton and Bowman (1967) explore the asymptotics for the Neyman type A distribution and observe that, even for samples of size 100 (i.e. 100 sites in our case), the true distribution of the estimators can have bias and higher variance than that predicted by the first-order asymptotic approximation. They note that this departure is more relevant for $\hat{\delta}$ than for $\hat{\gamma}$, and that it is greater when δ is large in comparison to γL, particularly when γL is small.

Figure 5-4 provides an illustration of how the actual distribution of the estimators compares with the bivariate normal distribution predicted by first-order asymptotics for two levels of total effort (E = 100 and 1000), two levels of mean abundance (δ = 0.5 and 5), γ = 0.5 and an optimal design based on minimizing the variance of the abundance estimator (γL = 1.6 and 1.1, respectively). While the departure is quite noticeable for the lower effort level (which resulted in 31 and 45 sites), the asymptotic approximation is good in the higher-effort case (which resulted in 312 and 454 sites). The departure is more noticeable in the scenarios with $\delta = 5$, which is larger than γL, therefore in agreement with the observation made by Shenton and Bowman (1967). Figure 5-5 shows how the properties of the estimators worsen for the case of $\delta = 5$ when the design departs from the optimal and longer transects are used (γL = 5.5 instead of 1.1).

Apart from having larger variance, smaller data sets are also more prone to extreme estimates. Although not shown in Figure 5-4, in our simulations the case of low effort ($E = 100$) produced some extreme estimates (e.g. $\hat{\delta} > 10^7$ in 16 out of 10,000 simulated data sets when $\delta = 0.5$). In fact, as we demonstrate now, if the data set is such that there is at most one detection per site, i.e. $\max(d_i) = 1$, then $\hat{\gamma} = 0$ and $\hat{\delta} = \infty$. In this case $F_x = 0$ for $x > 2$ and therefore equation (5.8) becomes

$$F_0\hat{\Pi}_0 + F_1\hat{\Pi}_1 = F_1, \tag{5.15}$$

which, given that now $S_0 = F_0 + F_1$ and $S_1 = F_1$, is equivalent to

$$(S_0 - S_1)\hat{\Pi}_0 + S_1\hat{\Pi}_1 = S_1\,. \tag{5.16}$$

Considering in (5.16) that

$$\begin{aligned}\hat{\Pi}_0 &= \hat{\gamma}L\hat{\delta}e^{-\hat{\gamma}L},\\ \hat{\Pi}_1 &= \frac{1}{2}\hat{\gamma}L\left(1 + \hat{\delta}e^{-\hat{\gamma}L}\right),\end{aligned} \tag{5.17}$$

and that, according to (5.7), $S_1 = S_0\hat{\delta}\hat{\gamma}L$, we arrive at the following expression for $\hat{\gamma}$ in this scenario

$$e^{-\hat{\gamma}L} + \hat{\gamma}L = 1, \tag{5.18}$$

which leads to $\hat{\gamma} = 0$ and consequently to $\hat{\delta} \to \infty$. Similarly we can expect that, if the frequency of $d_i > 1$ is very low, the model will produce extreme estimates, as observed in our simulations.

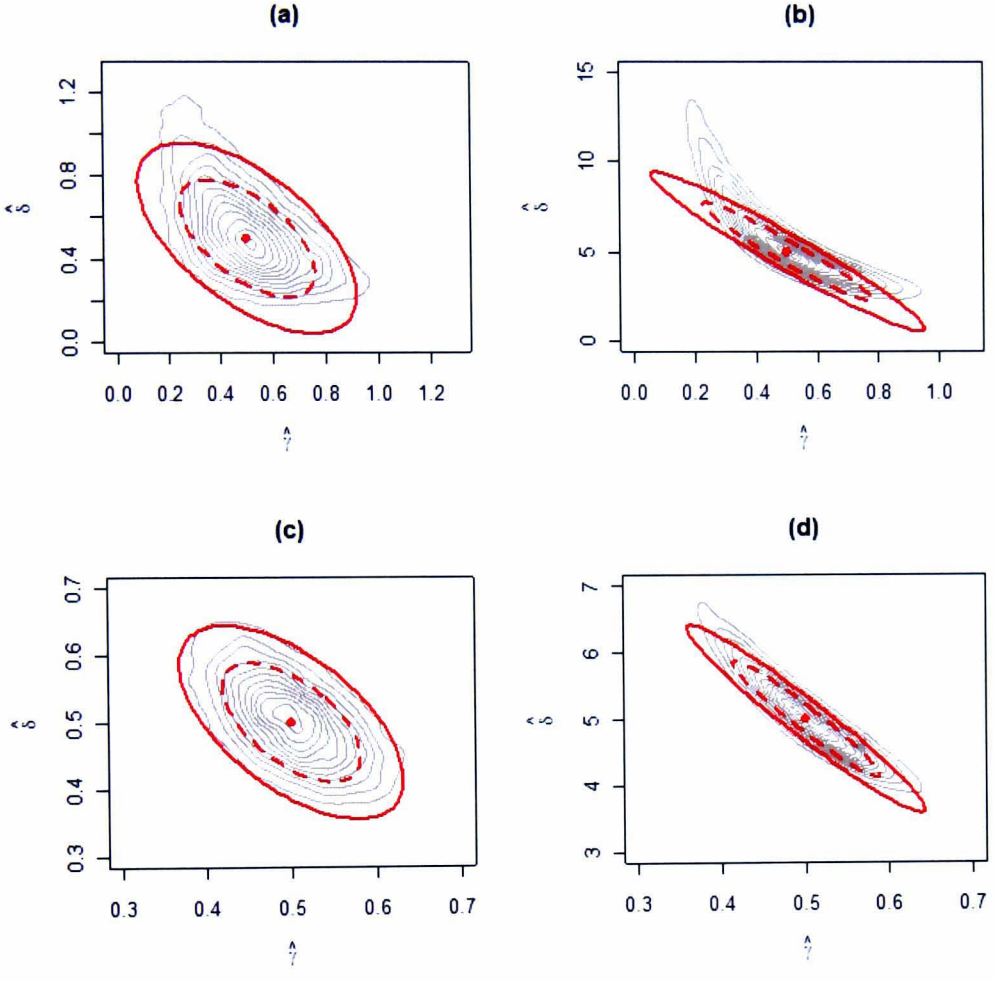


Figure 5-4 True and asymptotic distribution for the MLEs of the Neyman type A model for a case with total effort $E = 100$ and parameter values $\gamma = 0.5$ and (a) $\delta = 0.5$, (b) $\delta = 5.0$. Study design is chosen to minimize the asymptotic variance of $\hat{\delta}$, i.e. (a) $\gamma L = 1.6$ and (b) $\gamma L = 1.1$. The true distribution is based on the analysis of 10,000 simulated data sets and represented as a kernel density estimate contour plot. The asymptotic distribution is shown as two ellipses defined by the part of the bivariate normal distribution contained within one (dashed line) and two (solid line) standard deviations. In (c) and (d) the total effort was increased to $E = 1000$ (note the difference in the scale of the axes).

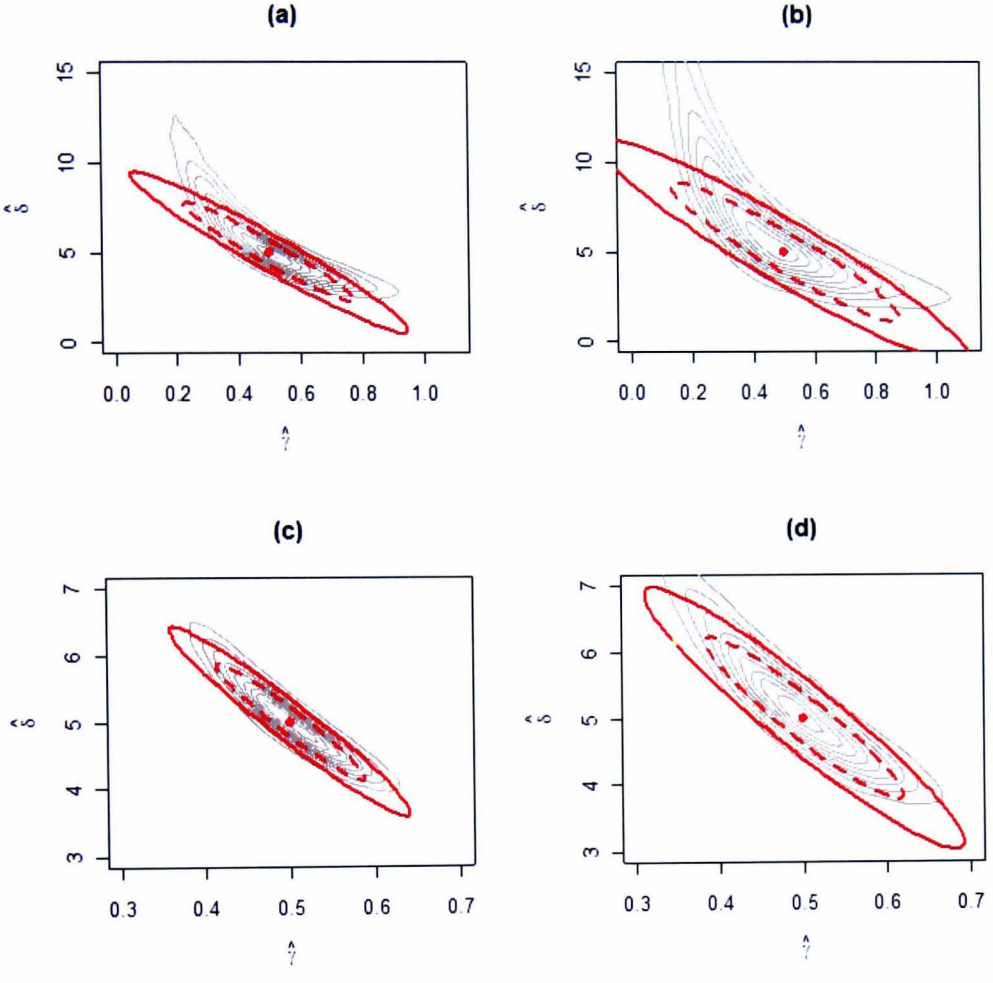


Figure 5-5 True and asymptotic distribution for the MLEs of the Neyman type A model for a case with total effort $E = 100$ and parameter values $\gamma = 0.5$ and $\delta = 5.0$, for a design (a) optimal to minimize the asymptotic variance of $\hat{\delta}$, i.e. $\gamma L = 1.1$, and (b) a non optimal design $\gamma L = 5.5$. The true distribution is based on the analysis of 10,000 simulated data sets and represented as a kernel density estimate contour plot. The asymptotic distribution is shown as two ellipses defined by the part of the bivariate normal distribution contained within one (dashed line) and two (solid line) standard deviations. In (c) and (d) the total effort was increased to $E = 1000$ (note the difference in the scale of the axes).

5.2 Model for abundance-induced heterogeneity and clustering

5.2.1 Model formulation and assumptions

We consider now a scenario in which the detections of individuals exhibit some degree of clustering and therefore cannot be considered independent. One possible way to account for this is to model the detection process for each *individual* as a two-state Markov-modulated Poisson process (2-MMPP) with detection intensity parameters $\boldsymbol{\gamma} = [\gamma_1 \quad \gamma_2]$ and state switching rate parameters $\boldsymbol{\mu} = [\mu_{12} \quad \mu_{21}]$. Assuming that detections among individuals are independent and the detection process has the same characteristics for all individuals, the resulting detection process for the species at site i can be modelled as the superposition of n_i identical 2-MMPPs.

The likelihood for this model is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\mu}) = \prod_{i=1}^{S} \left[\sum_{n_i=0}^{\infty} \left\{ \left(\prod_{j=1}^{R_i} M_{ij|n_i} \right) \Pr(n_i) \right\} \right], \tag{5.19}$$

where R_i is the number of independent transects surveyed in site i and $M_{ij|n_i}$ is the likelihood contribution of transect j in site i, given that detections along it are described by the superposition of n_i identical 2-MMPPs. For this model, the detection data can no longer be summarized by the number of detections at each site as in (5.1), and the distances between consecutive detections are required.

To construct $M_{ij|n_i}$ we make use of a key result: the superposition of MMPPs is a MMPP. The generator matrix and the rate matrix for the composite MMPP resulting

from the superposition of n MMPPs are calculated from the individual generator matrices $\boldsymbol{Q}_i$ and rate matrices $\boldsymbol{\Gamma}_i$ as follows (Fischer & Meier-Hellstern 1993)

$$\boldsymbol{Q}_C = \boldsymbol{Q}_1 \oplus \boldsymbol{Q}_2 \oplus \ldots \oplus \boldsymbol{Q}_n,$$
$$\boldsymbol{\Gamma}_C = \boldsymbol{\Gamma}_1 \oplus \boldsymbol{\Gamma}_2 \oplus \ldots \oplus \boldsymbol{\Gamma}_n,$$

where $\oplus$ represents the Kronecker sum, defined as in Appendix A.5. The dimension of $\boldsymbol{Q}_C$ and $\boldsymbol{\Gamma}_C$ is $k \times k$, where $k = \prod_{i=1}^{n} k_i$ and $k_i \times k_i$ is the dimension of $\boldsymbol{Q}_i$ and $\boldsymbol{\Gamma}_i$.

For a composite MMPP resulting from the superposition of n identical 2-MMPPs, such as in our case, the calculations simplify significantly and lead to a $(n+1)$-MMPP. Given n 2-MMPPs with infinitesimal generator matrix $\boldsymbol{Q}$

$$\boldsymbol{Q} = \begin{bmatrix} -\mu_{12} & \mu_{12} \\ \mu_{21} & -\mu_{21} \end{bmatrix},$$

and rate matrix $\boldsymbol{\Gamma}$,

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix},$$

the $(n+1) \times (n+1)$ generator matrix $\boldsymbol{Q}_C$ for the composite MMPP is

$$\begin{aligned}
\boldsymbol{Q}_C[i,i] &= -i\mu_{21} - (n-i)\mu_{12}, && \text{for } 0 \le i \le n \\
\boldsymbol{Q}_C[i,i-1] &= i\mu_{21}, && \text{for } 1 \le i \le n \\
\boldsymbol{Q}_C[i,i+1] &= (n-i)\mu_{12}, && \text{for } 0 \le i \le n-1 \\
\boldsymbol{Q}_C[i,j] &= 0, && \text{otherwise},
\end{aligned}$$

and the $(n+1) \times (n+1)$ rate matrix $\boldsymbol{\Gamma_C}$ is

$$\boldsymbol{\Gamma_C} = \text{diag}\{i\gamma_2 + (n-i)\gamma_1\}, \qquad 0 \le i \le n\,.$$

The equilibrium distribution of the underlying Markov process $\boldsymbol{\pi'_C}$ is

$$\boldsymbol{\pi'_C}[i] = \binom{n}{i} \frac{\mu_{21}^{n-i}\mu_{12}^{i}}{(\mu_{12}+\mu_{21})^n}\,, \quad \text{for } 0 \le i \le n\,.$$

For instance, the superposition of three 2-MMPPs would lead to

$$\boldsymbol{Q_C} = \begin{bmatrix} -3\mu_{12} & 3\mu_{12} & 0 & 0 \\ \mu_{21} & -\mu_{21}-2\mu_{12} & 2\mu_{12} & 0 \\ 0 & 2\mu_{21} & -2\mu_{21}-\mu_{12} & \mu_{12} \\ 0 & 0 & 3\mu_{21} & -3\mu_{21} \end{bmatrix},$$

$$\boldsymbol{\Gamma_C} = \begin{bmatrix} 3\gamma_1 & 0 & 0 & 0 \\ 0 & \gamma_2+2\gamma_1 & 0 & 0 \\ 0 & 0 & 2\gamma_2+\gamma_1 & 0 \\ 0 & 0 & 0 & 3\gamma_2 \end{bmatrix},$$

and

$$\boldsymbol{\pi'_C} = [\mu_{21}^3 \quad 3\mu_{21}^2\mu_{12} \quad 3\mu_{21}\mu_{12}^2 \quad \mu_{12}^3]/(\mu_{12}+\mu_{21})^3.$$

Once $\boldsymbol{Q_C}$, $\boldsymbol{\Gamma_C}$ and the initial probability vector of the MMPP ($\boldsymbol{\pi'_C}$ or $\boldsymbol{\pi^*_C}$) are computed, $M_{ij|n_i}$ is constructed as in (4.28).

5.2.2 *'Synchronized' clustering in individuals detections*

In section 5.2.1 we have assumed that the processes describing the detections of individuals are independent (Figure 5-6a). Let us consider here a second case in which the 2-MMPPs describing individuals' detections are 'synchronized', such that at a given point all the individuals are in the same detection rate state (high or low) but, within this, detections are still independent (Figure 5-6b). This can be a useful model for scenarios in which the clustering in individuals' detections arises due to the difference in substrate conditions (e.g. when some patches are better than others for capturing footprints), or if the individuals in the site move closely together as a group. The likelihood for this model is the one given in (5.19), now constructing $M_{ij|n_i}$ considering that the superposition of n_i aligned 2-MMPPs with generator matrix $\boldsymbol{Q}$ and rate matrix $\boldsymbol{\Gamma}$ is a 2-MMPP with $\boldsymbol{Q_C} = \boldsymbol{Q}$ and $\boldsymbol{\Gamma_C} = n_i\boldsymbol{\Gamma}$.

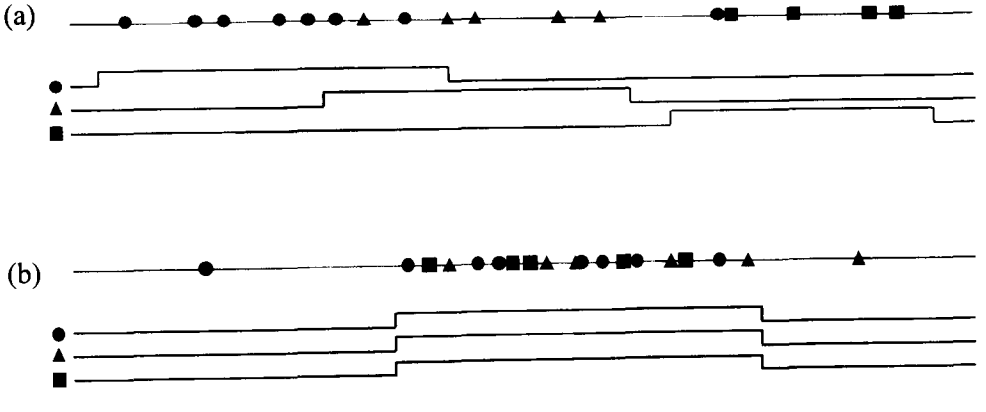


Figure 5-6 Hypothetical clustered detections of three individuals along a transect and corresponding realizations of the underlying Markov process. In (a) the clustering pattern is independent among individuals while in (b) the clustering pattern is 'synchronised'. Detections from each individual are represented with different symbols here only for explanatory purposes. Actual data would not contain information on individual identities.

5.2.3 Relationship to other models

If the mixing distribution in (5.19) is a non-parametric distribution with mass at $n_i = 0$ and $n_i = 1$, the model reduces to the 2-MMPP occupancy model (section 4.3) with likelihood

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

where $\Pr(0) = 1 - \psi$, $\Pr(1) = \psi$ and $M_{ij} = M_{ij|1}$.

To our knowledge, no discrete counterpart of the model in (5.19) has been proposed for the analysis of species detection data. However such a model could be devised by considering that the detections of each individual are produced by a two-state Markov-modulated Bernoulli process (2-MMBP) and that, at the species level, detections are described as the superposition of such processes.

As mentioned in the introduction, to deal with clustering and abundance-induced heterogeneity in the detection process, Hines *et al.* (2010) propose a two-step ad hoc approach based on discrete models. This method tries to avoid the effect of clustering rather than modelling it explicitly as in our model.

5.3 Performance simulation study

We used simulations to explore the performance of the models, looking at the three following aspects:

(i) first we explored how unmodelled abundance-induced heterogeneity in the detection process can induce bias in the occupancy estimator (section 5.3.1),

(ii) second we explored the impact that unmodelled clustering can have in the abundance estimators (section 5.3.2), and

(iii) third, we explored the performance of the MMPP abundance model for different sample sizes (section 5.3.3).

For the simulations we generated scenarios in which the true site abundance was such that the probabilities of having 0 to 3 individuals at a sampling site were 0.05, 0.5, 0.3 and 0.15 respectively, a distribution that is plausible ecologically for our motivating tiger example. To ensure in the analysis that the estimated probabilities of the non-parametric abundance distribution summed to one, we used for the unconstrained optimization a multinomial logit transformation

$$\theta_j = \frac{e^{\phi_j}}{S} \text{ for } j = 0 \ldots N-1, \qquad \theta_N = \frac{1}{S},$$

where $S = 1 + e^{\phi_0} + e^{\phi_1} + \cdots + e^{\phi_{N-1}}$, N is the number of support points in the non-parametric distribution, $\boldsymbol{\theta}$ are the corresponding probabilities and $\boldsymbol{\phi}$ are the $N-1$ unconstrained parameters used in the optimization process.

5.3.1 Impact of abundance-induced heterogeneity in occupancy estimation

To assess the impact that abundance-induced heterogeneity in the detection process has on the estimation of occupancy, we simulated scenarios with the abundance distribution specified above, which implies a probability of site occupancy of 0.95, assuming independence within the detections of each individual, that is, modelling the individual detection process as a Poisson process with rate γ. The scenarios consisted of $S = 100$ sites surveyed for $L = 30$ unit length, and increasing individual detection rates γ, from 0.01 to 0.20. Data were analyzed with the Poisson process occupancy model and with the Poisson process abundance model assuming the correct abundance structure (i.e. a non-parametric distribution with 4 support points in 0-3).

The simulation results illustrate how abundance-induced heterogeneity in the detection process can induce bias in the estimator of occupancy (Table 5-2). If the abundance structure was disregarded the occupancy estimator was more negatively biased than when modelled. As expected, as the probability of detecting an individual at a site increased, the estimator bias tended to zero in all methods.

Table 5-2 Impact of unmodelled abundance-induced heterogeneity in the occupancy estimator. The table displays the mean of the occupancy estimator obtained from 1000 simulations through three methods: naïve estimate, Poisson process occupancy model (PP) and Poisson process abundance model (PPab). The true abundance distribution was such that the probability of occupancy was $\psi = 0.95$. The probability of detecting an individual at a site, $\gamma^* = 1 - e^{\gamma L}$, is also shown.

	γ					
	0.01	0.02	0.05	0.10	0.15	0.20
γ^*	0.26	0.45	0.77	0.95	0.99	1.00
Naïve $\hat{\psi}$	0.097	0.237	0.595	0.846	0.919	0.943
PP $\hat{\psi}$	0.780	0.865	0.892	0.933	0.945	0.951
PPab $\hat{\psi}$	0.814	0.920	0.943	0.952	0.949	0.951

5.3.2 Impact of unmodelled detection clustering in the abundance estimators

To assess the impact that unmodelled detection clustering can have on the estimation of abundance we simulated scenarios with the abundance distribution specified above, and generated individual detection data according to a 2-MMPP, with the processes being independent among individuals. We explored four detection scenarios with equal average individual detection rates and increasing levels of clustering:

(i) Case nc: $\gamma = 1/4$ (i.e. no clustering),

(ii) Case c1: $\boldsymbol{\gamma} = [0.5, 0.125]$, $\boldsymbol{\rho} = [6, 12]$,

(iii) Case c2: $\boldsymbol{\gamma} = [1, 0.0625]$, $\boldsymbol{\rho} = [3, 12]$,

(iv) Case c3: $\boldsymbol{\gamma} = [5, 0.0125]$, $\boldsymbol{\rho} = [0.6, 12]$,

where $\boldsymbol{\rho} = 1/\boldsymbol{\mu}$ represent the average length spent in each state before switching to the other state.

We simulated a study design in which 100 sites were surveyed for 30 length units, and ran 100 simulations per scenario. Data were analyzed with the Poisson process abundance model, that is, assuming no clustering within the detections of individuals. We fitted the data assuming first a non-parametric abundance distribution with mass in categories 0-3, as used for data generation, and then assuming a distribution with mass in 0-5 to explore the impact of allowing for this extra flexibility. We used AIC to compare the fit of the models based on each of these two abundance distributions.

The simulation results show that the abundance distribution can be estimated relatively well with the given sampling effort when, as the model assumes, there is no clustering

in the detections (Table 5-3). However, as clustering increases, the estimators became biased. For instance the means for the abundance distribution estimators were [0.10, 0.51, 0.14, 0.25] in scenario (c3). The number of empty sites was overestimated, influenced by the long stretches without detections. The highest abundance category was also overestimated, which suggests that detection clusters may be interpreted by the model as the result of many individuals being present at the site. In fact, a model based on an abundance distribution with mass in 0-5 tended to provide a better fit for the clustered data. For instance, in all 100 simulations for scenario (c3), this model was selected as a better explanation for the data based on its AIC. However the model did not provide a satisfactory estimate of abundance. The distribution probabilities were biased, and the overall abundance was overestimated. Here as well the probability mass for the highest abundance category increased with the level of clustering.

Table 5-3 Impact of unmodelled detection clustering on the estimators of abundance. Mean and RMSE (in square brackets) of the estimators are shown. Results were obtained from 100 simulations of a design with 100 sites surveyed for 30 units of length, assuming a true site abundance distribution with probabilities $\boldsymbol{\theta}$ = [0.05, 0.5, 0.3, 0.15] for 0-3 individuals (average abundance N = 1.55). Four detection scenarios with equal average individual detection rate and increasing levels of clustering were tested: (nc) γ = 0.25 (no clustering), (c1) $\boldsymbol{\gamma}$ = [0.5, 0.125], $\boldsymbol{\rho}$ = [6, 12], (c2) $\boldsymbol{\gamma}$ = [1, 0.0625], $\boldsymbol{\rho}$ = [3, 12], and (c3) $\boldsymbol{\gamma}$ = [5, 0.0125], $\boldsymbol{\rho}$ = [0.6, 12]. The upper part of the table shows the results of fitting a model based on a non-parametric abundance distribution with four abundance categories. nAIC indicates the number of simulations in which this model produced an AIC smaller than that from a model with six abundance categories. The lower part of the table shows the results of the six-category model for the simulations in which it was selected as best model.

	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{N}$	nAIC
(nc)	0.05 [0.021]	0.50 [0.082]	0.29 [0.078]	0.16 [0.084]	--	--	1.56 [0.154]	93
(c1)	0.05 [0.022]	0.48 [0.082]	0.25 [0.091]	0.22 [0.092]	--	--	1.63 [0.148]	25
(c2)	0.06 [0.026]	0.51 [0.069]	0.18 [0.135]	0.24 [0.113]	--	--	1.60 [0.134]	1
(c3)	0.10 [0.059]	0.51 [0.059]	0.14 [0.172]	0.25 [0.120]	--	--	1.54 [0.106]	0
(nc)	0.06 [0.023]	0.31 [0.269]	0.40 [0.150]	0.12 [0.111]	0.08 [0.128]	0.03 [0.059]	1.95 [0.546]	
(c1)	0.05 [0.022]	0.28 [0.238]	0.29 [0.104]	0.22 [0.142]	0.05 [0.099]	0.11 [0.130]	2.26 [0.774]	
(c2)	0.06 [0.026]	0.36 [0.161]	0.21 [0.133]	0.22 [0.117]	0.02 [0.047]	0.13 [0.143]	2.17 [0.669]	
(c3)	0.10 [0.058]	0.38 [0.139]	0.15 [0.173]	0.20 [0.101]	0.000 [0.028]	0.16 [0.168]	2.11 [0.598]	

5.3.3 Performance of the abundance model with clustering (MMPP)

To assess the performance of the MMPP abundance model we simulated data with the same abundance distribution and clustering scenarios as in the previous section. Data were fitted assuming a non-parametric abundance distribution with mass in categories 0-3, using different starting values (10 sets) to reduce the chances of hitting a local maximum. We considered first a design with $S = 100$, $L = 30$, and then increased the sampling effort by either adding more sites, increasing survey length or both. Since fitting this model is more demanding in terms of computational time, we only ran 20 simulations for each scenario. In our implementation, the duration of each simulation varied from less than an hour to several hours. For comparison, data were also analysed assuming independence in the detections (i.e. Poisson process).

The simulation results show that the model can estimate all the parameters and that the estimators are unbiased given large, yet realistic, data sets (Table 5-4). The precision of the abundance estimators was poor for the initial survey design considered. As expected, increasing the amount of sampling effort improved the quality of the estimators. In the scenario with strong clustering, (c3), it was more effective to increase the length of the survey rather than the number of sampling sites: doubling the survey length gave better results than a 5-fold increase in sampling sites. The initial survey length used was relatively short considering the average interval spent in the low-detection state, in which the rate of detections was very low (0.0125). Individuals were therefore likely to remain undetected at surveyed sites. Increasing the survey length provided data that reflected better the number of individuals present at each surveyed site. In this case this was more critical than increasing the number of surveyed sites to obtain a better estimation of the abundance distribution.

Table 5-4 Estimator mean and RMSE (in square brackets) for the MMPP abundance model under three clustering scenarios: (c1) $\boldsymbol{\gamma}$ = [0.5, 0.125], $\boldsymbol{\rho}$ = [6, 12], (c2) $\boldsymbol{\gamma}$ = [1, 0.0625], $\boldsymbol{\rho}$ = [3, 12], and (c3) $\boldsymbol{\gamma}$ = [5, 0.0125], $\boldsymbol{\rho}$ = [0.6, 12]. Based on a true site abundance distribution with probabilities $\boldsymbol{\theta}$ = [0.05, 0.5, 0.3, 0.15] for 0-3 individuals (average abundance N = 1.55). Results were obtained from 20 simulations for four study designs in which S sites are surveyed for L units of length.

		$S = 100$ $L = 30$	$S = 100$ $L = 60$	$S = 200$ $L = 30$	$S = 500$ $L = 30$	$S = 500$ $L = 60$
(c1)	$\hat{\theta}_0$	0.05 [0.017]	0.05 [0.021]	0.05 [0.013]	0.05 [0.010]	0.05 [0.009]
	$\hat{\theta}_1$	0.44 [0.169]	0.51 [0.063]	0.54 [0.089]	0.52 [0.045]	0.50 [0.038]
	$\hat{\theta}_2$	0.30 [0.155]	0.29 [0.097]	0.25 [0.081]	0.30 [0.051]	0.30 [0.037]
	$\hat{\theta}_3$	0.21 [0.162]	0.15 [0.092]	0.16 [0.080]	0.13 [0.061]	0.15 [0.031]
	$\hat{\gamma}_1$	0.56 [0.152]	0.51 [0.077]	0.52 [0.127]	0.52 [0.049]	0.50 [0.023]
	$\hat{\gamma}_2$	0.11 [0.052]	0.12 [0.027]	0.13 [0.033]	0.13 [0.015]	0.12 [0.010]
	$\hat{\rho}_1$	5.50 [2.533]	7.85 [5.539]	6.03 [1.983]	6.88 [1.529]	6.28 [0.988]
	$\hat{\rho}_2$	15.3 [11.08]	14.6 [5.906]	12.5 [5.980]	14.5 [3.701]	12.2 [1.741]
	$\hat{N}$	1.67 [0.290]	1.54 [0.152]	1.52 [0.152]	1.51 [0.095]	1.55 [0.064]
(c2)	$\hat{\theta}_0$	0.05 [0.027]	0.05 [0.024]	0.06 [0.019]	0.05 [0.011]	0.05 [0.011]
	$\hat{\theta}_1$	0.50 [0.201]	0.53 [0.091]	0.51 [0.104]	0.46 [0.083]	0.50 [0.042]
	$\hat{\theta}_2$	0.29 [0.147]	0.26 [0.103]	0.26 [0.078]	0.27 [0.069]	0.30 [0.038]
	$\hat{\theta}_3$	0.15 [0.162]	0.15 [0.087]	0.16 [0.132]	0.21 [0.102]	0.15 [0.032]
	$\hat{\gamma}_1$	1.00 [0.081]	1.00 [0.066]	0.99 [0.043]	0.97 [0.040]	1.00 [0.024]
	$\hat{\gamma}_2$	0.06 [0.018]	0.06 [0.010]	0.06 [0.009]	0.06 [0.009]	0.06 [0.005]
	$\hat{\rho}_1$	3.15 [0.648]	2.97 [0.280]	3.08 [0.343]	2.94 [0.178]	3.00 [0.130]
	$\hat{\rho}_2$	12.2 [3.239]	11.5 [1.517]	12.2 [2.460]	12.2 [1.475]	12.0 [0.764]
	$\hat{N}$	1.55 [0.329]	1.50 [0.141]	1.54 [0.220]	1.65 [0.179]	1.54 [0.066]
(c3)	$\hat{\theta}_0$	0.05 [0.034]	0.05 [0.024]	0.05 [0.036]	0.05 [0.021]	0.05 [0.009]
	$\hat{\theta}_1$	0.55 [0.215]	0.54 [0.118]	0.53 [0.167]	0.52 [0.122]	0.52 [0.047]
	$\hat{\theta}_2$	0.17 [0.220]	0.25 [0.117]	0.23 [0.174]	0.22 [0.128]	0.27 [0.055]
	$\hat{\theta}_3$	0.23 [0.186]	0.16 [0.114]	0.19 [0.137]	0.21 [0.136]	0.16 [0.051]
	$\hat{\gamma}_1$	4.98 [0.277]	5.02 [0.146]	4.97 [0.172]	5.02 [0.114]	5.00 [0.073]
	$\hat{\gamma}_2$	0.013 [0.005]	0.013 [0.003]	0.012 [0.003]	0.012 [0.002]	0.013 [0.001]
	$\hat{\rho}_1$	0.59 [0.046]	0.61 [0.043]	0.61 [0.034]	0.60 [0.022]	0.61 [0.014]
	$\hat{\rho}_2$	12.2 [2.979]	11.8 [1.282]	11.9 [1.673]	12.6 [2.041]	12.1 [0.605]
	$\hat{N}$	1.58 [0.354]	1.53 [0.215]	1.56 [0.274]	1.60 [0.251]	1.55 [0.081]

Analyzing the data assuming independence in the detections provided a much poorer model fit, with large differences in terms of AIC for all simulations. Although biased, the abundance estimators obtained from this model were more precise (Figure 5-7). A similar phenomenon was observed by Morgan & Ridout (2008) on closed population capture–recapture models for estimating population size while accounting for heterogeneity in capture probability. Their simulations showed that, when the true model was a beta-binomial, fitting a binomial model produced a very precise but biased estimator of abundance. The beta-binomial model performed well in terms of bias but provided much poorer precision.

Despite being more precise, the coverage properties of the estimator from the model that assumes independence were in general poorer driven by the bias (Table 5-5). In fact, as illustrated in the previous section, an analysis based on the assumption of independence would actually favour models based on abundance distributions with more support points, which provide estimators that are both biased and imprecise.

Table 5-5 Comparison of the Poisson process and 2-MMPP abundance models in terms of coverage of 95% confidence intervals based on 20 simulations for the abundance distribution estimators under clustering scenario $\boldsymbol{\gamma} = [1, 0.0625]$, $\boldsymbol{\rho} = [3, 12]$, and true site abundance distribution $\boldsymbol{\theta} = [0.05, 0.5, 0.3, 0.15]$.

	No clustering model					Clustering model				
S, L	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{N}$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{N}$
100, 30	0.80	0.85	0.65	0.65	0.75	0.85	0.80	0.95	0.70	0.80
100, 60	0.90	0.90	0.80	0.90	0.90	0.90	0.95	1.00	0.90	0.95
200, 30	0.90	1.00	0.40	0.55	0.95	0.95	0.95	1.00	0.90	0.95
500, 30	0.85	1.00	0.05	0.00	0.65	0.95	0.90	1.00	1.00	0.95
500, 60	0.95	0.55	0.90	0.20	0.50	0.95	0.95	0.95	0.95	0.95

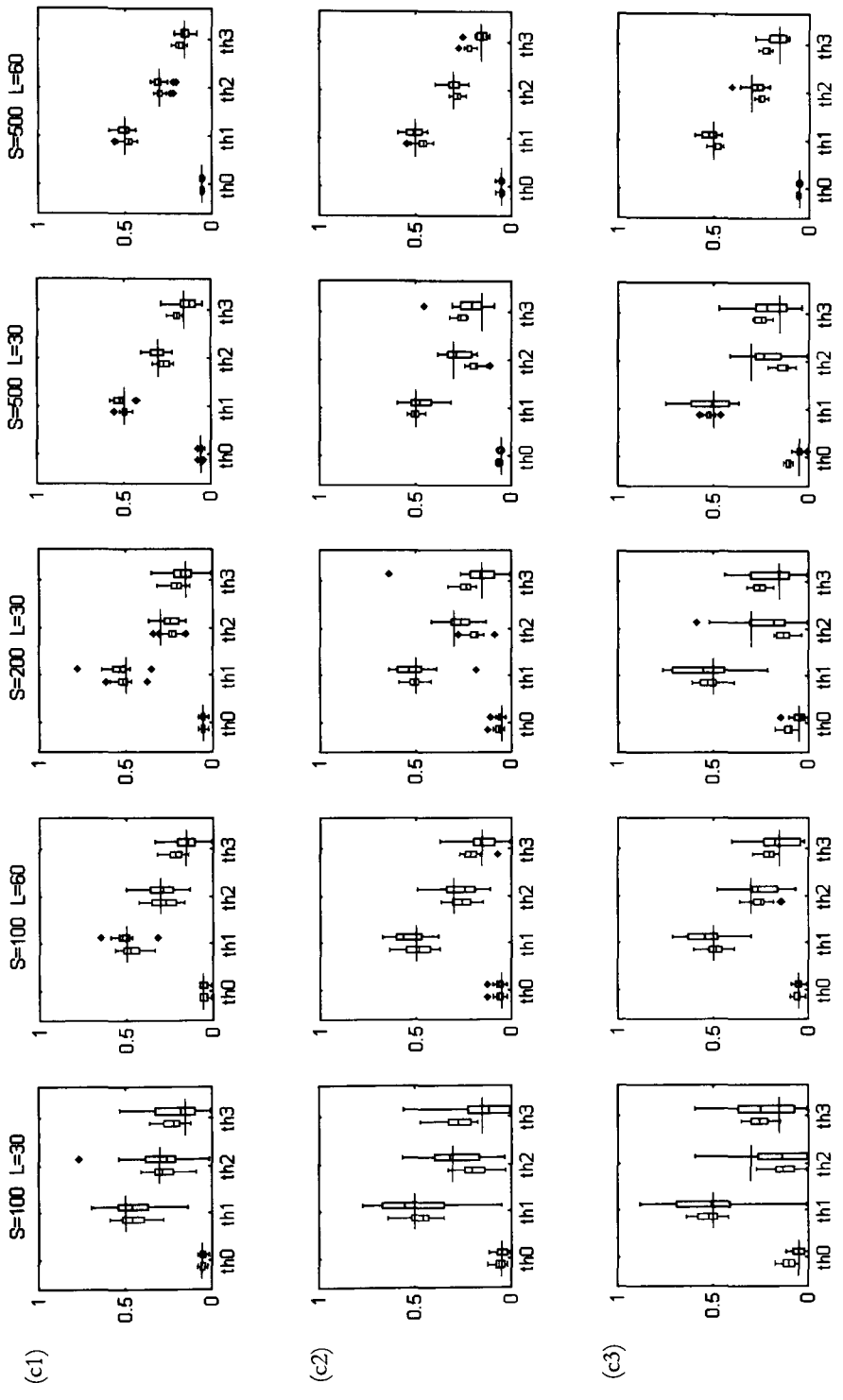


Figure 5-7 (landscape) Box-plot of the abundance estimates $\hat{\theta}_0 - \hat{\theta}_3$ (denoted as th0 – th3) for the simulated scenarios obtained from the Poisson process abundance model (left of each pair, in blue) and from the 2-MMPP abundance model (right of each pair, in black). Red horizontal lines indicate the true value of the parameters.

5.4 Analysis of the Kerinci tiger data set

Here we illustrate the application of the new models proposed in this chapter with the analysis of the Sumatran tiger data set from Kerinci Seblat National Park. These data were analyzed in chapter 3 assuming no abundance-induced heterogeneity in the detection process. In this section we present the results of an analysis relaxing this assumption.

5.4.1 Methods

First we used the abundance model that assumes independence and describes the detection process of individuals as a Poisson process (section 5.1). Second, we analyzed the data with the model structure that incorporates independent clustering in the detection process of individuals (section 5.2.1). Finally, we fit the data to the model that allows for clustering in individuals' detections but assumes that this clustering is 'synchronized' (section 5.2.2).

To describe the abundance process we considered non-parametric distributions with support points from 0 to a maximum of 5 individuals. Since tigers are territorial, one might expect underdispersion in the abundance distribution with respect to a Poisson distribution. Using a non-parametric distribution provides the flexibility to account for this. The scale of the survey (i.e. sampling site size) was chosen based on the size of an adult male tiger territory, so it is reasonable to expect a low number of individuals per site and therefore a density of more than five individuals per site to be highly unlikely.

5.4.2 Results

Looking in isolation at the results from models that assume no clustering in detections would suggest that abundance-induced heterogeneity in the detection process is very relevant for this data set (Table 5-6). The model that assumes no abundance-induced heterogeneity (labelled 'NP1a' in the table) was over 30 AIC units worse than the best model in this subset (NP3a). By allowing more abundance categories, the abundance estimates tended to give considerable weight to the highest abundance category (e.g. $\hat{\theta}_4 = 0.26$ in NP4a or $\hat{\theta}_5 = 0.23$ in NP5a).

Table 5-6 Parameter estimates and ΔAIC for the analysis of Kerinci tiger data with abundance models that describe the detection process of individuals as a Poisson process with intensity γ, i.e. independence among detections. Model labels 'NPXa' indicate that the abundance distribution is non-parametric with support points 0 to X and corresponding probabilities $\boldsymbol{\theta}$. ΔAIC is calculated both with respect to the models in the table set and to all the models fitted in this section (i.e. Table 5-7 and Table 5-8). Model NP1 in this set is the Poisson process occupancy model (PP) fitted in Table 4-4.

Model	ΔAIC		Abundance $\hat{\boldsymbol{\theta}}$						Detection
	set	all	0	1	2	3	4	5	$\hat{\gamma}$
NP1a	30.9	62.3	.18	.82	--	--	--	--	0.11
NP2a	9.2	40.6	.16	.55	.29	--	--	--	0.08
NP3a	0.0	31.4	.13	.59	.00	.28	--	--	0.06
NP4a	0.2	31.6	.10	.51	.12	.00	.26	--	0.05
NP5a	1.7	33.1	.08	.43	.18	.08	.00	.23	0.04

However, models that allow for clustering in individuals' detections had substantially higher support that those that assume independent detections (Table 5-7 and Table 5-8). Modelling clustering provided an improvement of about 30 AIC units in model fit, which indicates is a relevant feature in this data set. In all of the fitted models with

clustering, estimates imply that tiger detections were about ten times more frequent in some areas compared to others. As observed when fitting the 2-MMPP model in section 4.4, the switching rate estimates were very small and imprecise. The data were however informative about the probability of being in each of the two detection states. In the tables we report the estimated probability $\hat{\pi}_1'$ of an individual being in the high detection rate state.

Unlike when fitting models that assume independent detections, allowing for abundance-induced heterogeneity in the detection process provided modest improvement in model fit when detection clustering was modelled. This suggests that the structure for modelling abundance-induced heterogeneity in the models that assumed independent detections in Table 5-6 was, at least partially, capturing the clustering in the detection process rather than actual variations in site abundance. Under the assumption of independence, the best fitting models (NP3a and NP4a) produced relatively large estimates for the highest support point in the abundance distribution. This effect was also observed in our simulation study, when exploring the effect of unmodelled clustering in the estimation of abundance.

If the clustering pattern among individuals' detections was assumed 'non-synchronized' (Table 5-7) the best model in the subset was the one with three support points in the abundance distribution (NP2b). In this case the difference in AIC units between the best model (NP2b) and the model with two support points (NP1b) was just 2.7. Adding one extra abundance category resulted in a model with similar fit. Note that models NP4b and NP5b are in effect the same as NP3b. In the computation of the AIC we used the actual number of parameters in each model to draw attention to

the simpler model. Another approach for models that can be collapsed would be to use the number of parameters in the reduced model, as recommended by Burnham & Anderson (2002, pp. 342-343) when discussing the use of AIC in finite mixture models.

Table 5-7 Parameter estimates and ΔAIC for the analysis of Kerinci tiger data with abundance models that describe the detection process of individuals as a 2-MMPP with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$, i.e. clustering in detections, and the detection processes of individuals are independent. The estimate $\hat{\pi}_1' = \hat{\mu}_{21}/(\hat{\mu}_{12} + \hat{\mu}_{21})$ represents the probability of being in the high detection rate state for each individual. Model labels 'NPXb' indicate that the abundance distribution is non-parametric with support points 0 to X and corresponding probabilities $\boldsymbol{\theta}$. ΔAIC is calculated both with respect to the models in the table set and to all the models fitted in this section (i.e. Table 5-6 and Table 5-8). Model NP1b in this set is the 2-MMPP occupancy model fitted in Table 4-4.

Model	ΔAIC		Abundance $\hat{\boldsymbol{\theta}}$						Detection	
	set	all	0	1	2	3	4	5	$\hat{\pi}_1'$	$\hat{\boldsymbol{\gamma}}$
NP1b	2.7	2.7	.04	.96	--	--	--	--	0.33	0.23, 0.03
NP2b	0.0	0.0	.00	.59	.41	--	--	--	0.27	0.19, 0.02
NP3b	0.5	0.5	.00	.63	.03	.35	--	--	0.24	0.17, 0.02
NP4b	2.5	2.5	.00	.63	.03	.35	.00	--	0.24	0.17, 0.02
NP5b	4.5	4.5	.00	.63	.03	.35	.00	.00	0.24	0.17, 0.02

Analyzing the data under the assumption that clustering in detections was 'synchronized' among individuals did not provide an improvement in terms of model support (Table 5-8). The best model in this set (NP2c) was 2.7 AIC units better than the model that assumes no abundance-induced heterogeneity (NP1c), and had very similar fit to the model that assumes independent clustering (0.4 AIC units compared to model NP2b in Table 5-7).

Table 5-8 Parameter estimates and ΔAIC for the analysis of Kerinci tiger data with abundance models that describe the detection process of individuals as a 2-MMPP with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$, i.e. clustering in detections, and the detection processes of individuals are 'synchronized'. The estimate $\hat{\pi}_1' = \hat{\mu}_{21}/(\hat{\mu}_{12} + \hat{\mu}_{21})$ represents the probability of being in the high detection rate state for each individual. Model labels 'NPXc' indicate that the abundance distribution is non-parametric with support points 0 to X and corresponding probabilities $\boldsymbol{\theta}$. ΔAIC is calculated both with respect to the models in the table set and to all the models fitted in this section (i.e. Table 5-6 and Table 5-7). Model NP1c in this set is the 2-MMPP occupancy model fitted in Table 4-4 (and the same as NP1b in Table 5-7).

Model	ΔAIC		Abundance $\hat{\boldsymbol{\theta}}$						Detection	
	set	all	0	1	2	3	4	5	$\hat{\pi}_1'$	$\hat{\gamma}$
NP1c	2.3	2.7	.04	.96	--	--	--	--	0.33	0.23, 0.03
NP2c	0.0	0.4	.02	.65	.33	--	--	--	0.39	0.15, 0.02
NP3c	1.5	1.9	.00	.52	.26	.22	--	--	0.44	0.11, 0.01
NP4c	3.1	3.5	.00	.26	.47	.00	.27	--	0.45	0.08, 0.01
NP5c	5.1	5.5	.00	.26	.47	.00	.27	.00	0.45	0.08, 0.01

The best fitting models in Table 5-7 and Table 5-8 both provided a similar estimated abundance distribution, suggesting that tigers are absent from very few sites, that in about two thirds of the sites there is one individual and that in the remaining third, there are two. It is however important to note that the estimates obtained were fairly imprecise, with estimated standard errors $\mathrm{SE}(\hat{\boldsymbol{\theta}}) = [0.00^*, 0.27, 0.27]$ for NP2b and $\mathrm{SE}(\hat{\boldsymbol{\theta}}) = [0.06, 0.21, 0.20]$ for NP2c (*boundary estimate). This is actually not very surprising considering that the sample size was not particularly large (89 sampling sites) compared to the sample sizes required to obtain precise estimates according to our simulation study. The lack of precision was also reflected by the fact that the models suffered from local maxima and needed to be fitted several times from different starting values. When fitting some of the models (in particular the more complicated

ones) alternative solutions provided by these local maxima could receive very similar support. For instance in model NP4b in Table 5-7 the value of the log-likelihood function was only slightly lower than the maximum we found (~ 0.2 units) at points that correspond to quite different abundance distributions such as $\boldsymbol{\theta}$ = [0.00, 0.50, 0.20, 0.25, 0.04] and $\boldsymbol{\theta}$ = [0.00, 0.34, 0.39, 0.00, 0.28]. On the other hand, the estimates of the detection rate parameters and $\hat{\pi}_1'$ were considerably more precise. In particular for the best fitting models in Table 5-7 and Table 5-8 the corresponding standard errors were $\mathrm{SE}(\hat{\pi}_1', \hat{\gamma}_1, \hat{\gamma}_2)$ = [0.076, 0.033, 0.007] for NP2b and $\mathrm{SE}(\hat{\pi}_1', \hat{\gamma}_1, \hat{\gamma}_2)$ = [0.080, 0.031, 0.007] for NP2c. The abundance estimates obtained from the models without clustering were more precise that when clustering was accounted for (e.g. $\mathrm{SE}(\widehat{\boldsymbol{\theta}})$ = [0.05, 0.09, 0.00, 0.08] for NP3a in Table 5-6), which is in line with the results of our simulations.

5.5 Discussion

In this chapter we have extended the models in chapter 4 to account for abundance-induced heterogeneity in the detection process. We have shown how disregarding abundance-induced heterogeneity can induce bias in the occupancy estimator. We have also demonstrated that modelling clustering in the detection process can be relevant when estimating site abundance. We have proposed a model to deal with the problem of accounting for both clustering and abundance-induced heterogeneity, considering two model variants that reflect different sources of clustering giving rise to either independent or 'synchronized' clustering patterns. To date this problem had been addressed by following a 2-step ad hoc method, rather than by explicitly modelling both aspects in the detection process.

One limitation of the model proposed is that it is relatively demanding in terms of computing time, owing to the matrix operations involved in the likelihood. For example, fitting the independent clustering model with five abundance support points to the Sumatran tiger data set took in our implementation around 10 minutes (on a 2.40 GHz computer) compared to the few seconds required by the Poisson process abundance model that assumes independent detections. As computing time increases with the number of support points in the abundance distribution, this becomes more relevant when dealing with abundant species.

Another issue is that, in the presence of clustering and abundance-induced heterogeneity, sample size requirements increase as the detection process becomes more complicated to describe. However our simulations indicated that precise estimates can be obtained with sample sizes that, although large, can still be achievable in some ecological

studies. The simulation results also reveal that, while unmodelled clustering induces bias in the abundance estimators, these estimators are more precise than those from the clustering model. Depending on the sample size, disregarding clustering may result in better estimators in terms of RMSE, but with poorer coverage properties.

The complexity of the model implies that, unless enough data are available, the likelihood can be rather flat with various local maxima, which can be of similar magnitude while providing different explanations for the data. This means that the outcome of the analysis can be sensitive to the point from which the optimization routine is initialized and that, therefore, it is a good approach to explore the likelihood function by trying out different starting values. In our tiger analysis we encountered this problem: the standard errors of the estimates were quite large and, as we allowed for more support points in the non-parametric abundance distribution, alternative explanations fitted the data equally well. This suggests that the model had too much flexibility for the amount of information in the data.

Despite its limitations, our analysis of the tiger data set provides a nice illustration of how accounting for clustering can change the conclusions regarding the underlying abundance distribution. When we analyzed the data with the model structure that assumes independent detections, there was strong support for models that allowed for higher abundance categories in the abundance mixture distribution; however, there was much less evidence for differences in site abundance when clustering was accounted for. Our results reflect how other sources of heterogeneity in the detection process can creep into the estimated abundance distribution if not incorporated in the model.

Finally, when applying the models proposed in this chapter, it is important to remember that the appropriateness of interpreting the estimates obtained as abundance estimates is contingent on how well the model assumptions are met. First of all, the models assume that differences in site abundance are reflected as heterogeneity in the detection process and that this is the only source of such heterogeneity. Unmodelled heterogeneity coming from other sources would be interpreted by the model as abundance-induced, which can cause bias in the estimators. For instance, we have seen how unmodelled clustering in the detection process can induce bias in the abundance estimators. An appropriate description of the detection process is therefore critical for a reliable estimation of abundance. If some measurable factors are thought to affect detection rates appreciably, these can and should be incorporated into the models as covariates. Second, the models assume that all the individuals present at the site are detectable over the whole site, i.e. at any point in the site, the detection process is the superposition of n_i individual detection processes. Whether this assumption is satisfied depends on the characteristics of the species and the survey, including the choice of site size (we discuss the tiger case in chapter 6). Third, all individuals are assumed to exhibit similar movement patterns, so that their detections are well described by identical detection processes. While for most species this may be a reasonable approximation, for some there may be marked differences among groups, such as males vs. females. In such cases, heterogeneity could potentially be addressed by modelling the detection process as a mixture of non-identical point processes, although we suspect that this increased flexibility may also imply more difficulties for model fitting.

6 CONCLUSIONS

In this thesis we have explored two aspects of occupancy modelling. We have first dealt with several issues related to the design of occupancy studies. Carefully addressing study design is essential to ensure that meaningful results are obtained in the most efficient way. This is important for any statistical study, but it is particularly crucial in ecology and conservation where resources are often fairly limited. Being a relatively recently developed modelling framework, the design of occupancy studies is currently an area of interest and development (e.g. Efford & Dawson 2012; Pacifici, Dorazio & Conroy 2012), and our work contributes in this direction.

The rest of the thesis focussed on developing and assessing occupancy models for species detection data collected along transects. The interest in such data was prompted by the analysis of the Sumatran tiger data set, which was collected following such a survey protocol. Transects had previously been used as the basis for collecting detection data in surveys targeted to study species occupancy but the analysis of such data had however been carried out by modelling the detection process as a discrete process after pooling detection from transect segments of a given length (e.g. Hines *et al.* 2010). The analysis of this kind of data had previously raised two questions

(i) How to discretize the detection data in the best way?

(ii) How to model the data accounting both for potential clustering and abundance-induced heterogeneity in the detection process?

In this thesis we addressed these two issues. The key to our work is to model the detection process along the transect as a continuous point process on the real half-line. Point processes are attractive in this context as they provide a natural framework for the description of the arrival of events (the detections) along a line (the transect). Such description of the detection process avoids the need to pool data, an approach which may ultimately result in poorer estimators if large transect segments are used, as information is lost in the discretization process. Regarding the first question above, we believe there is no such a thing as an optimal segment length, neither in fact, an intrinsic benefit of discretizing continuous data. However we do acknowledge that such an approach can be convenient from the implementation point of view (e.g. it allows the use of existing discrete models) and that it can for example help to mitigate the effects of potential clustering in the detections when this aspect is not explicitly modelled.

The point process models that we explored for species detection data are summarized by Figure 6-1. We started from a simple model that assumes independent detections taking place at a rate not influenced by differences in site abundance, and built up more complex models that take account of clustering and/or abundance-induced heterogeneity in the detection process. The model that accounts for both clustering and abundance-induced heterogeneity is the most general model among those studied, with the others being particular or limiting cases of this one.

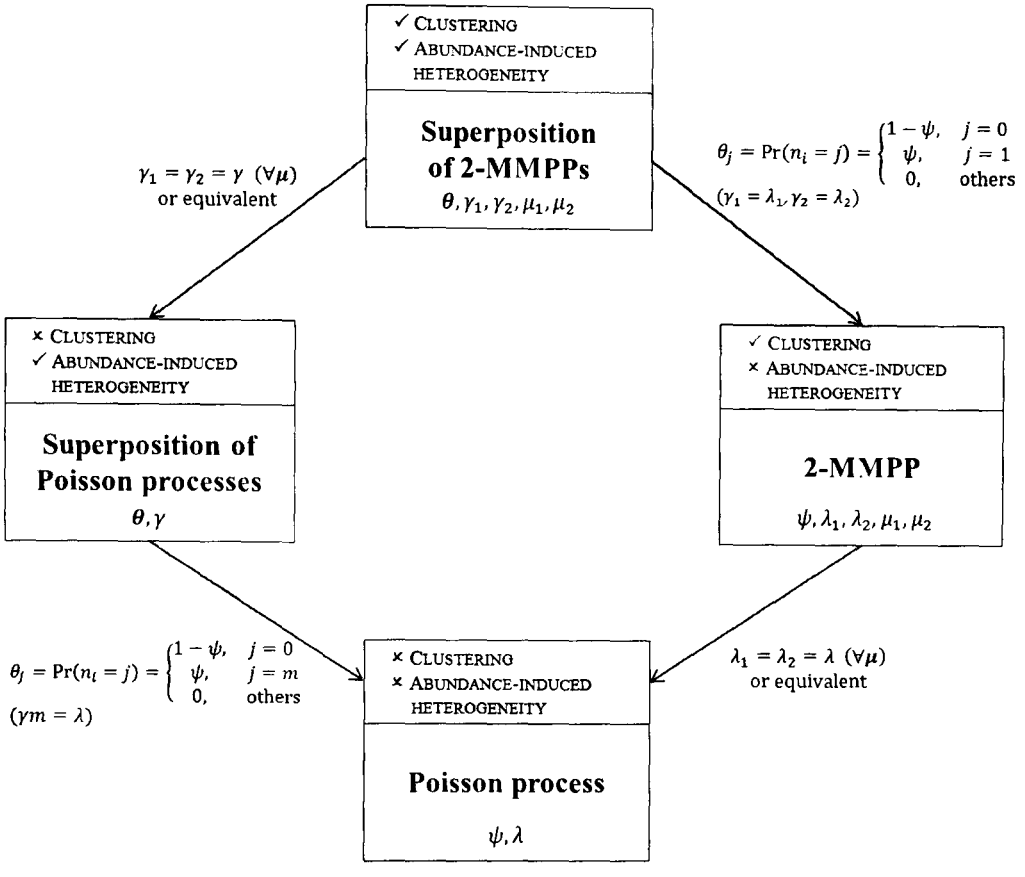


Figure 6-1 Summary diagram of the models developed in this thesis. The diagram reflects how simpler models result as particular (or limiting) cases of the most complicated model. Parameters are as defined in the corresponding chapters.

Although in this thesis the models have been explored with the transect survey protocol as a motivating example, they are also potentially applicable to surveys that collect detection data by monitoring each sampling site over a continuous interval of time. Camera-trap surveys provide an obvious example of such a design. To date this kind of data has been used to estimate occupancy by collapsing trapping times in intervals of a given duration (e.g. 2 weeks, Linkie *et al.* 2007) and then using discrete models.

Another example is point count surveys, often targeted to avian species, which involve collecting detection data over an interval of time at a number of sampling stations. The interval of time is usually relatively short (e.g. 10 minutes) and independent repeated surveys are carried out at each sampling station. In the model structure presented in this paper, these replicated surveys would be analogous to the independent transects surveyed within each sampling site.

We have shown that modelling clustering in the detection process can be relevant when estimating site occupancy and abundance, as the lack of independence can induce bias in the estimators. In our models, the description of clustered detections, either for species or for individuals, is based on 2-MMPPs. This kind of point process provides a simple description for clustered arrivals. Although other models are possible for describing varying detection rates (e.g. having more states in the underlying hidden process or even allowing detection rates to vary continuously), we believe that 2-MMPPs will often provide a sufficiently flexible description for this kind of data.

An advantage of using point processes to model detections along transects is in the interpretation of the parameters. It can often be easier and more natural to communicate that 'the detection rate of a species is X detections per kilometre surveyed', rather than 'the probability of detecting the species at least once in a segment of a given length Y is Z'. Equally, the parameters related to the switching between states, which reflect the clustering pattern, are more readily interpreted in terms of average length in each state, rather than as switching probabilities which are dependent on an arbitrary segment-length definition.

A different matter is the ecological interpretation of the observed clustering patterns, regardless of whether a discrete or continuous model is used for the analysis. Various mechanisms can give rise to clustering in species detections and the models we have proposed in this thesis can be readily used to model different scenarios. For instance, in sign surveys along transects, clustering may be due to individuals intermittently following the paths used as transects or due to patches of different substrate conditions. In camera-trap surveys, clustering can be expected if the movement patterns of individuals are such that they remain for a while in the area around the camera before moving to other parts of their territory. In transect surveys, clustering can also be induced if the species only covers the sampling sites partially, as detections are only possible on the transect sections overlapping with the portion of the site that is occupied. For instance this can be the case for surveys based on observations of species that exhibit clumped distributions (e.g. due to patchy resources or limited dispersal ability of offspring) or for sign surveys of territorial species if the home-range size is smaller than the sampling site. In these scenarios of partial occupation, modelling the clustering pattern provides an estimate of the actual area occupied by the species at sites where it is present ($\hat{\pi}_1'$).

In the case of the tiger data set, our analysis revealed patent clustering in the footprint detections, however the ecological interpretation of the pattern observed is not clear cut. On the one hand tigers are territorial and the size of the sampling sites was chosen to be comparable to a male tiger large home range, so we could expect the clustering to be reflecting the fact that sampling sites overlap partially with tiger territories. On the other hand it could be argued that the clustering may rather be a consequence of individuals following the transects (often along ridges) intermittently, leaving clusters

of footprints within their territories. Of course, in practice the observed clustering may be due to a combination of these two processes. Looking at the estimates of the 2-MMPP occupancy model in section 4.4 we see that the spatial scale of the clustering pattern picked up by the analysis is large: the rate of detections in the high detection state implies that on average one footprint is detected every 4.3 km; the estimates of the switching rate parameters suggest that whole transects are embedded in each detection-rate state. This would be compatible with a scenario in which the clustering pattern is largely dominated by the overlap of territories and sampling sites. Tigers are wide-ranging animals and individuals move long distances across their territories every day. Tiger footprints usually remain detectable for a few days, depending on the substrate and weather conditions. We can therefore speculate that a large part of the home range of an individual tiger would contain footprints to be detected, at least along the features that the animals follow as routes to move across. However, unfortunately we cannot tell from the data the extent to which the areas of high footprint detection rate coincide with the territories as this depends on the tiger movement patterns and the persistence of the footprints in the landscape.

In fact the tiger case becomes even more complicated if we consider that, while male tigers keep relatively exclusive territories, female tigers hold smaller territories that overlap with male territories and which can have some degree of overlapping among them (Karanth & Nichols 2002, p. 11). This complex territorial structure means that parameter estimates from the models, and particularly those from the abundance models, should be interpreted with care. While for the surveys of some species the model assumptions will more closely match the reality and thus a relatively direct interpretation of the estimates will be possible, we recommend caution with the literal reading of

the abundance estimates as individual numbers in the tiger scenario and would suggest limiting the interpretation to a coarse indication of site abundance.

Regarding the models that account for abundance-induced heterogeneity, we started by presenting a basic model that assumes independent detections. As we discussed, apart from its utility as a model for species-detection data collected along transects, this model can also be useful as an alternative to the binomial N-mixture model of Royle (2004) for the analysis of spatially-replicated repeated count data, allowing for the possibility of detecting each individual more than once per replicate. This is relevant when individuals cannot be identified within survey visits, which can happen in surveys based on direct observations (e.g. camera-trap surveys without individual identifications or bird point counts) and it is crucial when modelling detection data from indirect observations, as individuals can leave more than one sign. In fact, the Poisson–Poisson mixture model in particular had been previously mentioned as a potentially useful model development for encounter-rate data (Stanley & Royle 2005, in Discussion; Royle & Dorazio 2008, p. 413).

We then explored the more general model that accounts for both abundance-induced heterogeneity and clustering in the detection process. Our approach represents an attractive alternative to the previously proposed two-step ad hoc solution (Hines *et al.* 2010), as it provides a description that explicitly incorporates both aspects, and therefore allows the estimation of not only abundance but also of the parameters associated with the clustering pattern. This is, as far as we know, the first model that accounts for both clustering within the detections of individuals and abundance-induced heterogeneity in the species-detection process. Martin *et al.* (2011) studied the effect of corre-

lated behaviour in the binomial N-mixture model used to estimate abundance from repeated counts. Although dealing with an essentially different problem (i.e. dependence among individuals instead of within detections from each individual), they also found that lack of independence in the detections can lead to a poor estimation of abundance.

Our model tackles the problem posed by question (ii), however it is important to realize the limitations that we face when dealing with such a scenario. We have illustrated how the lack of identifiability between alternative explanations and the poor precision of the estimators can indeed be problematic, unless a relatively large data set is available. It is not surprising that disentangling the clustering and abundance processes from detection data of unmarked individuals is in fact challenging. We have also discussed how other unmodelled sources of heterogeneity can be 'captured' within the estimation of abundance.

It is unfortunately not only within a given model structure that inferential problems arise. In connection with this, it is worth noting that other models might be devised to account for heterogeneity in the species detection rate. For instance, a negative-binomial model could be used, which would arise if the rate in the Poisson detection process is allowed to vary among sites according to a gamma distribution. Finite mixtures could be used to characterize a system in which sites can belong to a finite number of classes with distinct species-detection rates, as is done in capture–recapture to model heterogeneous recapture probabilities (Pledger 2000). A crucial problem is that verifying whether abundance is the source of heterogeneity in the detection process may be difficult, or even impossible. As discussed in the context of occupancy models for discrete sampling protocols, different descriptions for heterogeneity in detection

probability may fit the detection/non-detection data equally well and yet produce different estimates of occupancy (Royle 2006). This kind of identifiability problem can also be expected when modelling data collected along transects and adds to our previous discussion on the need to address other sources of heterogeneity in detection rate to obtain reliable estimates of abundance. This should not only be relegated to the development of advanced models with sophisticated descriptions of the detection process but should be dealt with at the early stages of the survey by carefully addressing sampling design to minimize unwanted sources of heterogeneity.

We have seen that there are some limitations on the robustness and precision of the abundance estimates when abundance is estimated solely from detections of unmarked individuals. In this connection, we would like to highlight that, if the real focus of the study is to obtain a precise estimation of population abundance, then other survey techniques specifically developed to estimate abundance may be more suitable. However these methods tend to be more resource-intensive which can limit their application to large geographical scales. An interesting strategy for such scenarios relies on the development of approaches that integrate large-scale low cost surveys with more targeted resource-intensive sampling methods (e.g. Conroy *et al.* 2008).

Apart from the issues discussed above, there are a couple of directions in which our work could be further refined. First of all, our analysis disregards the fact that the actual transects in the surveys were not straight lines but somewhat wiggly. This simplification is necessary to apply the theory of one-dimensional point processes. However, in reality detections take place on a surface, rather than along a true one-dimensional line. Ignoring the spatial nature of the data may result in detections that are spatially

close appearing distant along the transect, which might affect the conclusions of this type of study. Moving to a more spatially-explicit model is an avenue for further research. A possible associated difficulty is the fact that we are dealing with detections *on* the transect, given that the window of observation is virtually of zero width, which might complicate the fitting of spatial point process models.

Secondly, our clustering models are based on MMPPs, which assume that the time spent in each of the detection states is exponentially distributed. Depending on the type of clustering that is to be captured, the state holding times might be more suitably described by a different distribution. For instance, if clustering is induced by the survey routes cutting through individual territories, perhaps less dispersed holding times would be more appropriate. In this case, rather than describing the detections as a Markov-modulated Poisson process, a Poisson process modulated by a semi-Markov process would be used. What conceptually seems a small modification in the model, in practice implies a considerable increase in complexity, as the relatively simple form of the MMPP likelihood relies on the assumption of exponential holding times. Using simulation-based inferential methods can therefore be a useful approach in this case. An alternative is to use MMPPs to construct a semi-MMPP using the 'method of stages'. In a preliminary literature search we found no examples of model fitting based on a semi-Markov-modulated Poisson process. The closest piece of work was an application to modelling neural spike bursts (Tokdar *et al.* 2010), in which such a point process was used but imposing, as a simplification, some particular restrictions in terms of when states were allowed to switch. Further investigation of such processes would therefore be useful, not only from the point of view of modelling species detections, but also as a tool of interest in other applications.

CITED LITERATURE

Abdelbasit, K.M. & Plackett, R.L. (1983) Experimental design for binary data. *Journal of the American Statistical Association,* **78,** 90-98.

Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* (eds B.N. Petrov & F. Csaki), pp. 267-281. Akademiai Kiado, Budapest.

Anderson, D.R. (2008) *Model Based Inference in the Life Sciences.* Springer, New York.

Atkinson, A.C. & Donev, A.N. (1992) *Optimum Experimental Designs.* Clarendon Press, Oxford.

Azuma, D.L., Baldwin, J.A. & Noon, B.R. (1990) Estimating the occupancy of spotted owl habitat areas by sampling and adjusting for bias. *General Technical Report GTR-PSW-124.* USDA Forest Service, Berkeley.

Bailey, L.L., Hines, J.E., Nichols, J.D. & MacKenzie, D.I. (2007) Sampling design trade-offs in occupancy studies with imperfect detection: examples and software. *Ecological Applications,* **17,** 281-290.

Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O.U., Swartz, B., Quental, T.B., Marshall, C., McGuire, J.L., Lindsey, E.L., Maguire, K.C., Mersey, B. & Ferrer,

E.A. (2011) Has the Earth's sixth mass extinction already arrived? *Nature,* **471,** 51-57.

Best, L.B. & Petersen, K.L. (1982) Effects of stage of the breeding cycle on sage sparrow detectability. *The Auk,* **99,** 788.

Bled, F., Royle, J.A. & Cam, E. (2011) Hierarchical modeling of an invasive spread: the Eurasian collared-dove *Streptopelia decaocto* in the United States. *Ecological Applications,* **21,** 290-302.

Boos, D.D. (1992) On generalized score tests. *The American Statistician,* **46,** 327-333.

Borchers, D.L., Buckland, S.T. & Zucchini, W. (2003) *Estimating Animal Abundance: Closed Populations*. Springer, New York.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J., Borchers, D.L. & Thomas, L. (2008) *Advanced Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, Oxford.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2001) *Introduction to Distance Sampling*. Oxford University Press, Oxford.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* 2nd edn. Springer-Verlag, New York.

Casagrande, J.T., Pike, M.C. & Smith, P.G. (1978) An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics,* **34,** 483-486.

Catchpole, E.A. & Morgan, B.J.T. (1996) Model selection in ring-recovery models using score tests. *Biometrics,* **52,** 664-672.

Cernuschi, F. & Castagnetto, L. (1946) Chains of rare events. *The Annals of Mathematical Statistics,* **17,** 53-61.

Chaloner, K. & Verdinelli, I. (1995) Bayesian experimental design: a review. *Statistical Science,* **10,** 273-304.

Chen, G., Kéry, M., Zhang, J. & Ma, K. (2009) Factors affecting detection probability in plant distribution studies. *Journal of Ecology,* **97,** 1383-1389.

Cochran, W.G. & Cox, G.M. (1957) *Experimental Designs,* 2nd edn. Wiley & Sons, Oxford.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences,* 2nd edn. Lawrence Erlbaum Associates, Hilldale.

Conroy, M.J., Runge, J.P., Barker, R.J., Schofield, M.R. & Fonnesbeck, C.J. (2008) Efficient estimation of abundance for patchily distributed populations via two-phase, adaptive sampling. *Ecology,* **89,** 3362-3370.

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

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

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

Danielsen, F., Burgess, N.D. & Balmfort, A. (2005) Monitoring matters: examining the potential of locally-based approaches. *Biodiversity and Conservation,* **14,** 2507–2542.

Davison, A.C. (2003) *Statistical Models.* Cambridge University Press, Cambridge.

Daykin, C.D., Pentikäinen, T. & Pesonen, M. (1994) *Practical Risk Theory for Actuaries.* Chapman and Hall, London.

Dinerstein, E., Loucks, C., Wikramanayake, E., Ginsberg, J., Sanderson, E., Seidensticker, J., Forrest, J., Bryja, G., Heydlauff, A., Klenzendorf, S., Leimgruber, P., Mills, J., O'Brien, T.G., Shrestha, M., Simons, R. & Songer, M. (2007) The fate of wild tigers. *Bioscience,* **57,** 508-514.

Dobson, A.J. & Gebski, V.J. (1986) Sample sizes for comparing two independent proportions using the continuity-corrected arcsine transformation. *Journal of the Royal Statistical Society, Series D,* **35,** 51-53.

Donner, A. (1984) Approaches to sample size estimation in the design of clinical trials - A review. *Statistics in Medicine,* **3,** 199-214.

Dorazio, R.M. (2007) On the choice of statistical models for estimating occurrence and extinction from animal surveys. *Ecology,* **88,** 2773-2782.

Dorazio, R.M., Kéry, M., Royle, J.A. & Plattner, M. (2010) Models for inference in dynamic metacommunity systems. *Ecology,* **91,** 2466-2475.

Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association,* **100,** 389-398.

Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology,* **87,** 842-854.

Douglas, J.B. (1980) *Analysis with Standard Contagious Distributions.* International Cooperative Publishing House, Fairland.

Efford, M.G. & Dawson, D.K. (2012) Occupancy in continuous habitat. *Ecosphere,* **3,** Article 32.

Field, S.A., Tyre, A.J. & Possingham, H.P. (2005) Optimizing allocation of monitoring effort under economic and observational constraints. *Journal of Wildlife Management,* **69,** 473-482.

Fischer, W. & Meier-Hellstern, K. (1993) The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation,* **18,** 149-171.

Fiske, I. & Chandler, R. (2011) Unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software,* **43,** 1-23.

Fleishman, E., Mac Nally, R., Fay, J.P. & Murphy, D.D. (2002) Modeling and predicting species occurrence using broad-scale environmental variables: An example with butterflies of the Great Basin. *Conservation Biology,* **15,** 1674-1685.

Fleiss, J.L. (1973) *Statistical Methods for Rates and Proportions.* Wiley & Sons, New York.

Fleiss, J.L., Tytun, A. & Ury, H.K. (1980) A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics,* **36,** 343-346.

Geissler, P.H. & Fuller, M.R. (1986) Estimation of the proportion of area occupied by an animal species. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* pp. 533-538. American Statistical Association, Alexandria.

Gordon, I. & Watson, R. (1996) The myth of continuity-corrected sample size formulae. *Biometrics,* **52,** 71-76.

Gu, W. & Swihart, R.H. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation,* **116,** 195-203.

Guillera-Arroita, G. (2011) Impact of sampling with replacement in occupancy studies with spatial replication. *Methods in Ecology and Evolution,* **2,** 401-406.

Guillera-Arroita, G., Morgan, B.J.T., Ridout, M.S. & Linkie, M. (2011) Species occupancy modeling for detection data collected along a transect. *Journal of Agricultural, Biological, and Environmental Statistics,* **16,** 301-317.

Guillera-Arroita, G., Ridout, M.S. & Morgan, B.J.T. (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution,* **1,** 131-139.

Guillera-Arroita, G., Ridout, M.S., Morgan, B.J.T. & Linkie, M. (2012) Models for species-detection data collected along transects in the presence of abundance-induced heterogeneity and clustering in the detection process. *Methods in Ecology and Evolution,* **3,** 358-367.

Haight, F.A. (1967) *Handbook of the Poisson Distribution.* Wiley & Sons, New York.

Hall, D.B. (2000) Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics,* **56,** 1030-1039.

Hanski, I. (1999) *Metapopulation Ecology.* Oxford University Press, Oxford.

Hardin-Waddle, J., Dorazio, R.M., Walls, S.C., Rice, K.G., Beauchamp, J., Schuman, M.J. & Mazzotti, F.J. (2010) A new parameterization for estimating co-occurrence of interacting species. *Ecological Applications,* **20,** 1467-1475.

Hauck, W.W. & Donner, A. (1977) Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association,* **72,** 851-853.

Heffes, H. & Lucantoni, D.M. (1986) A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications,* **4,** 856-868.

Hines, J.E. (2006) PRESENCE2 - Software to estimate patch occupancy and related parameters. USGS-PWRC www.mbr-pwrc.usgs.gov/software/presence.html.

Hines, J.E., Nichols, J.D., Royle, J.A., MacKenzie, D.I., Gopalaswamy, A.M., Samba Kumar, N. & Karanth, K.U. (2010) Tigers on trails: Occupancy modeling for cluster sampling. *Ecological Applications,* **20,** 1456–1466.

Hurvich, C.M. & Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika,* **76,** 297-307.

IUCN (2001) *Red List Categories and Criteria: Version 3.1.* IUCN Species Survival Commission, Gland, Switzerland.

Johnson, D.H. (1996) Point process models of single-neuron discharges. *Journal of Computational Neuroscience,* **3,** 275-299.

Johnson, N.L., Kemp, A.W. & Kotz, S. (2005) *Univariate Discrete Distributions,* 3rd edn. Wiley & Sons, Hoboken.

Joseph, L.N., Elkin, C., Martin, T.G. & Possingham, H.P. (2009) Modeling abundance using N-mixture models: the importance of considering ecological mechanisms. *Ecological Applications,* **19,** 631-642.

Kalish, L.A. (1990) Efficient design for estimation of median lethal dose and quantal dose-response curves. *Biometrics,* **46,** 737-748.

Kaplan, E.L. & Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* **53,** 457–481.

Karanth, K.U. (1995) Estimating tiger *Panthera tigris* populations from camera-trap data using capture-recapture models. *Biological Conservation,* **71,** 333-338.

Karanth, K.U. & Nichols, J.D. (1998) Estimation of tiger densities in India using photographic captures and recaptures. *Ecology,* **79,** 2852-2862.

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

Karanth, K.U., Nichols, J.D., Seidenstricker, J., Dinerstein, E., Smith, J.L.D., McDougal, C., Johnsingh, A.J.T., Chundawat, R.S. & Thapar, V. (2003) Science deficiency in conservation practice: the monitoring of tiger populations in India. *Animal Conservation,* **6,** 141-146.

Kendall, W.L. & White, G.C. (2009) A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *Journal of Applied Ecology,* **46,** 1182-1188.

Kéry, M. (2011) Towards the modelling of true species distributions. *Journal of Biogeography,* **38,** 617-618.

Kéry, M., Dorazio, R.M., Soldaat, L., Strien, A.v., Zuiderwijk, A. & Royle, J.A. (2009a) Trend estimation in populations with imperfect detection. *Journal of Applied Ecology,* **46,** 1163-1172.

Kéry, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography,* **37,** 1851-1862.

Kéry, M., Royle, J.A., Plattner, M. & Dorazio, R.M. (2009b) Species richness and occupancy estimation in communities subject to temporary emigration. *Ecology,* **90,** 1279-1290.

Kéry, M., Royle, J.A. & Schmid, H. (2005) Modeling avian abundance from replicated counts using binomial mixture models. *Ecological Applications,* **15,** 1450-1461.

Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Hafliger, G. & Zbinden, N. (2010) Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology,* **24,** 1388-1397.

Kéry, M. & Schaub, M. (2012) *Bayesian Population Analysis using WinBUGS – A Hierarchical Perspective*. Academic Press, Waltham.

Kingman, J.F.C. (1993) *Poisson Processes*. Oxford University Press, Oxford.

Laake, J. & Rextad, E. (2008) RMark – an alternative approach to building linear models in MARK. *Program MARK: A Gentle Introduction - Appendix C* (eds E. Cooch & G.C. White). www.phidot.org/software/mark/docs/book/.

Lee, I.W.C. & Fapojuwo, A.O. (2005) Stochastic processes for computer network traffic modeling. *Computer Communications,* **29,** 1-23.

Legg, C.J. & Nagy, L. (2006) Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management,* **78,** 194-199.

Link, W.A. (2003) Nonidentifiability of population size from capture-recapture data with heterogenerous detection probabilities. *Biometrics,* **59,** 1123-1130.

Linkie, M., Chapron, G., Martyr, D.J., Holden, J. & Leader-Williams, N. (2006) Assessing the viability of tiger subpopulations in a fragmented landscape. *Journal of Applied Ecology,* **43,** 576-586.

Linkie, M., Dinata, Y., Nugroho, A. & Haidir, I.A. (2007) Estimating occupancy of a data deficient mammalian species living in tropical rainforests: Sun bears in the Kerinci Seblat region, Sumatra. *Biological Conservation,* **137,** 20-27.

Linkie, M., Guillera-Arroita, G., Smith, J. & Rayan, D.M. (2010) Monitoring tigers with confidence. *Integrative Zoology,* **5,** 342-350.

Linkie, M., Martyr, D.J., Holden, J., Yanuar, A., Hartana, A.T., Sugardjito, J. & Leader-Williams, N. (2003) Habitat destruction and poaching threaten the Sumatran tiger in Kerinci Seblat National Park, Sumatra. *Oryx,* **37,** 41-48.

Linkie, M., Wibisono, H.T., Martyr, D.J. & Sunarto, S. (2008) *Panthera tigris ssp. sumatrae*. In: IUCN 2011. IUCN Red List of Threatened Species. Version 2011.2. www.iucnredlist.org.

Lu, S. (2012) Markov modulated Poisson process associated with state-dependent marks and its applications to the deep earthquakes. *Annals of the Institute of Statistical Mathematics,* **64,** 87-106.

Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000) WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing,* **10,** 325-337.

MacKenzie, D.I. (2005) What are the issues with presence absence data for wildlife managers? *Journal of Wildlife Management,* **69,** 849-860.

MacKenzie, D.I. (2006) Modeling the probability of resource use: The effect of, and dealing with, detecting a species imperfectly. *Journal of Wildlife Management,* **70,** 367-374.

MacKenzie, D.I. & Bailey, L.L. (2004) Assessing the fit of site-occupancy models. *Journal of Agricultural, Biological, and Environmental Statistics,* **9,** 300-318.

MacKenzie, D.I., Bailey, L.L., Hines, J.E. & Nichols, J.D. (2011) An integrated model of habitat and species occurrence dynamics. *Methods in Ecology and Evolution,* **2,** 612-622.

MacKenzie, D.I., Bailey, L.L. & Nichols, J.D. (2004) Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology,* **73,** 546–555.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology,* **84,** 2200-2207.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology,* **83,** 2248-2255.

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

MacKenzie, D.I., Nichols, J.D., Seamans, M.E. & Gutiérrez, R.J. (2009) Modeling species occurrence dynamics with multiple states and imperfect detection. *Ecology,* **90,** 823-835.

MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology,* **42,** 1105-1114.

Martin, D.C. & Katti, S.K. (1962) Approximations to the Neyman type A distribution for practical problems. *Biometrics,* **18,** 354-364.

Martin, J., Royle, J.A., MacKenzie, D.I., Edwards, H.H., Kéry, M. & Gardner, B. (2011) Accounting for non-independent detection when estimating abundance of

organisms with a Bayesian approach. *Methods in Ecology and Evolution,* **2,** 595-601.

McCarthy, M.A. & Possingham, H.P. (2007) Active adaptive management for conservation. *Conservation Biology,* **21,** 956-963.

McCarthy, T., Murray, K., Sharma, K. & Johansson, O. (2010) Preliminary results of a long-term study of snow leopards in South Gobi, Mongolia. *IUCN Cat News,* **53.**

McClintock, B.T., Bailey, L.L., Pollock, K.H. & Simons, T.R. (2010) Experimental investigation of observation error in anuran call surveys. *Journal of Wildlife Management,* **74,** 1882-1893.

Meier-Hellstern, K.S. (1987) A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European Journal of Operational Research,* **29,** 370-377.

Millennium Ecosystem Assessment (2005) *Ecosystems and Human Well-being: Synthesis.* Island Press, Washington.

Miller, D.A., Nichols, J.D., McClintock, B.T., Campbell Grant, E.H., Bailey, L.L. & Weir, L.A. (2011) Improving occupancy estimation when two types of observation error occur: non-detection and species misidentification. *Ecology,* **92,** 1422-1428.

Ministry of Forestry of Indonesia (2010) The National Tiger Recovery Program: The Road to the Tiger Summit. The Ministry of Forestry of Indonesia and the Global Tiger Initiative National Consultation, Jakarta.

Mordecai, R.S., Mattsson, B.J., Tzilkowski, C.J. & Cooper, R.J. (2011) Addressing challenges when studying mobile or episodic species: hierarchical Bayes estimation of occupancy and use. *Journal of Applied Ecology,* **48,** 56-66.

Moreno, M. & Lele, S.R. (2010) Improved estimation of site occupancy using penalized likelihood. *Ecology,* **91,** 341-346.

Morgan, B.J.T. (2008) *Applied Stochastic Modelling,* 2nd edn. Chapman and Hall, London.

Morgan, B.J.T., Palmer, K.J. & Ridout, M.S. (2007) Negative score test statistic. *The American Statistician,* **61,** 285-288.

Morgan, B.J.T., Revell, D.J. & Freeman, S.N. (2007) A note on simplifying likelihoods for site occupancy models. *Biometrics,* **63,** 618–621.

Morgan, B.J.T. & Ridout, M.S. (2008) A new mixture model for capture heterogeneity. *Applied Statistics,* **57,** 433-446.

Mortelliti, A. & Boitani, L. (2008) Inferring red squirrel (*Sciurus vulgaris*) absence with hair tubes surveys: a sampling protocol. *European Journal of Wildlife Research,* **54,** 353–356.

Muscariello, L., Mellia, M., Meo, M., Ajmone Marsan, M. & Lo Cigno, R. (2005) Markov models of internet traffic and a new hierarchical MMPP model. *Computer Communications,* **28,** 1835-1851.

Nelder, J.A. & Mead, R. (1965) A simplex method for function minimization. *The Computer Journal,* **7,** 308-313.

Neyman, J. (1939) On a new class of "contagious" distributions applicable in entomology and bacteriology. *The Annals of Mathematical Statistics,* **10,** 35-57.

Nichols, J.D., Bailey, L.L., O'Connell Jr., A.F., Talancy, N.W., Campbell Grant, E.H., Gilbert, A.T., Annand, E.M., Husband, T.P. & Hines, J.E. (2008) Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology,* **45,** 1321–1329.

Nichols, J.D., Hines, J.E., MacKenzie, D.I., Seamans, M.E. & Gutiérrez, R.J. (2007) Occupancy estimation and modeling with multiple states and state uncertainty. *Ecology,* **88,** 1395-1400.

Nichols, J.D. & Karanth, K.U. (2002) Statistical concepts: assessing spatial distributions. *Monitoring Tigers and Their Prey: A Manual for Wildlife Managers, Researchers, and Conservationists* (eds K.U. Karanth & J.D. Nichols), pp. 29-38. Centre for Wildlife Studies, Bangalore.

Otis, D.L., Burnham, K.P., White, G.C. & Anderson, D.R. (1978) Statistical inference from capture data on closed animal populations. *Wildlife Monographs,* **62,** 1-135.

Pacifici, K., Dorazio, R.M. & Conroy, M.J. (2012) A two-phase sampling design for increasing detections of rare species in occupancy surveys. *Methods in Ecology and Evolution,* doi: 10.1111/j.2041-1210X.2012.00201.x.

Pledger, S. (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics,* **56,** 434-442.

Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Vienna.

Possingham, H.P., Andelman, S.J., Noon, B.R., Trombulak, S. & Pulliam, H.R. (2001) Making Smart Conservation Decisions. *Conservation biology: research priorities for the next decade* (eds M.E. Soule & G.H. Orians), pp. 225–244. Island Press, Washington.

Rabus, B., Eineder, M., Roth, A. & Bamler, R. (2003) The shuttle radar topography mission - a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing,* **57,** 241-262.

Ramesh, N.I. (1995) Statistical analysis on Markov-modulated Poisson processes. *Environmetrics,* **6,** 165-179.

Reunanen, P., Nikula, A., Mönkkönen, M., Hurme, E. & Nivale, V. (2002) Predicting occupancy for the Siberian flying squirrel in old-growth forest patches. *Ecological Applications,* **12,** 1188-1198.

Richmond, O.M.W., Hines, J.E. & Beissinger, S.R. (2010) Two-species occupancy models: a new parameterization applied to co-occurrence of secretive rails. *Ecological Applications,* **20,** 2036-2046.

Ridout, M.S. (1995) Three-Stage Designs for Seed Testing Experiments. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* **44,** 153-162.

Ridout, M.S. & Besbeas, P. (2004) An empirical model for underdispersed count data. *Statistical Modelling,* **4,** 77-89.

Royle, J.A. (2004a) Modeling abundance index data from anuran calling surveys. *Conservation Biology,* **18,** 1378-1385.

Royle, J.A. (2004b) N-mixture models for estimating population size from spatially replicated counts. *Biometrics,* **60,** 108-115.

Royle, J.A. (2006) Site occupancy models with heterogeneous detection probabilities. *Biometrics,* **62,** 97-102.

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology.* Academic Press, Amsterdam.

Royle, J.A. & Link, W.A. (2005) A general class of multinomial mixture models for anuran calling survey data. *Ecology,* **86,** 2505-2512.

Royle, J.A. & Link, W.A. (2006) Generalised site occupancy models allowing for false positive and false negative errors. *Ecology,* **87,** 835-841.

Royle, J.A. & Nichols, J.D. (2003) Estimating abundance from repeated presence-absence data or point counts. *Ecology,* **84,** 777-790.

Rydén, T. (1994) Parameter estimation for Markov modulated Poisson processes. *Communications in Statistics - Stochastic Models,* **10,** 795-829.

Rydén, T. (1996) An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis,* **21,** 431-447.

Salafsky, N., Margoluis, R. & Redford, K.H. (2001) Adaptive management: a tool for conservation practitioners. Biodiversity Support Program, Washington.

Sanderson, E., Forrest, J., Loucks, C., Ginsberg, J., Dinerstein, E., Seidensticker, J., Leimgruber, P., Songer, M., Heydlauff, A., O'Brien, T., Bryja, G., Klenzendorf, S. & Wikramanayake, E. (2006) Setting priorities for the conservation and recovery of wild tigers: 2005–2015: A technical assessment. Wildlife Conservation Society, World Wildlife Fund, Smithsonian, and Save the Tiger Fund, Washington.

Sewell, D., Beebee, T.J.C. & Griffiths, R.A. (2010) Optimising biodiversity assessments by volunteers: The application of occupancy modelling to large-scale amphibian surveys. *Biological Conservation,* **143,** 2102-2110.

Sharma, S., Jhala, Y. & Sawarkar, V.B. (2005) Identification of individual tigers (*Panthera tigris*) from their pugmarks. *Journal of Zoology,* **267,** 9-18.

Shenton, L.R. (1949) On the efficiency of the method of moments and Neyman's type A distribution. *Biometrika,* **36,** 450-454.

Shenton, L.R. & Bowman, K.O. (1967) Remaks on large sample estimators for some discrete distributions. *Technometrics,* **9,** 587-598.

Skaug, H.J. (2006) Markov modulated Poisson processes for clustered line transect data. *Environmental and Ecological Statistics,* **13,** 199-211.

Stanley, T.R. & Royle, J.A. (2005) Estimating site occupancy and abundance using indirect detection indices. *Journal of Wildlife Management,* **69,** 874-883.

Takada, H., Sumita, U. & Takahashi, K. (2011) Pricing collateralized debt obligations with Markov-modulated Poisson processes. *Quantitative Finance,* **11,** 1761-1771.

Thompson, W.A. (1988) *Point Process Models with Applications to Safety and Reliability*. Chapman and Hall, London.

Tokdar, S., Xi, P., Kelly, R. & Kass, R. (2010) Detection of bursts in extracellular spike trains using hidden semi-Markov point process models. *Journal of Computational Neuroscience,* **29,** 203-212.

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications,* **13,** 1790-1801.

Ury, H.K. & Fleiss, J.L. (1980) On Approximate Sample Sizes for Comparing Two Independent Proportions with the Use of Yates' Correction. *Biometrics,* **36,** 347-351.

Uryu, Y., Purastuti, E., Laumonier, Y., Sunarto, S., Setiabudi, Budiman, A., Yulianto, K., Sudibyo, A., Hadian, O., Kosasih, D.A. & Stüwe, M. (2010) Sumatra's

forests, their wildlife and the climate. Windows in time: 1985, 1990, 2000 and 2009. WWF Indonesia, Jakarta.

Vaeth, M. (1985) On the use of Wald's test in exponential families. *International Statistical Review,* **53,** 199-214.

Vere-Jones, D. (1970) Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society, Series B,* **32,** 1-62.

Vorburger, M. & Munoz, B. (2006) Simple power calculations: how do we know we are doing them the right way? *Proceedings of the Survey Research Methods Section of the American Statistical Association,* pp. 3809-3812.

Walston, J., Robinson, J.G., Bennett, E.L., Breitenmoser, U., da Fonseca, G.A.B., Goodrich, J., Gumal, M., Hunter, L., Johnson, A., Karanth, K.U., Leader-Williams, N., MacKinnon, K., Miquelle, D., Pattanavibool, A., Poole, C., Rabinowitz, A., Smith, J.L.D., Stokes, E.J., Stuart, S.N., Vongkhamheng, C. & Wibisono, H. (2010) Bringing the tiger back from the brink-The six percent solution. *PLoS Biology,* **8,** e1000485.

Walters, D.E. (1979) In defence of the arcsine approximation. *Journal of the Royal Statistical Society, Series D,* **28,** 219-222.

Wenger, S.J. & Freeman, M.C. (2008) Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology,* **89,** 2953-2959.

White, G.C. & Burnham, K.P. (1999) Program MARK: Survival estimation from populations of marked animals. *Bird Study,* **46,** 120-139.

Wibisono, H.T., Linkie, M., Guillera-Arroita, G., Smith, J.A., Sunarto, Pusparini, W., Asriadi, Baroto, P., Brickle, N., Dinata, Y., Gemita, E., Gunaryadi, D., Haidir, I.A., Herwansyah, Karina, I., Kiswayadi, D., Kristiantono, D., Kurniawan, H.,

Lahoz-Monfort, J.J., Leader-Williams, N., Maddox, T., Martyr, D.J., Maryati, Nugroho, A., Parakkasi, K., Priatna, D., Ramadiyanta, E., Ramono, W.S., Reddy, G.V., Rood, E.J.J., Saputra, D.Y., Sarimudi, A., Salampessy, A., Septayuda, E., Suhartono, T., Sumantri, A., Susilo, Tanjung, I., Tarmizi, Yulianto, K., Yunus, M. & Zulfahmi (2011) Population status of a cryptic top predator: An island-wide assessment of tigers in Sumatran rainforests. *PLoS ONE,* **6,** e25931.

Wild, C.J. & Seber, G.A.F. (2000) *Chance Encounters: A First Course in Data Analysis and Inference*. Wiley & Sons, New York.

Williams, B.K., Nichols, J.D. & Conroy, M.J. (2002) *Analysis and Management of Animal Populations*. Academic Press, San Diego.

Yamaura, Y., Andrew Royle, J., Kuboi, K., Tada, T., Ikeno, S. & Makino, S.i. (2011) Modelling community dynamics based on species-level abundance models from detection/nondetection data. *Journal of Applied Ecology,* **48,** 67-75.

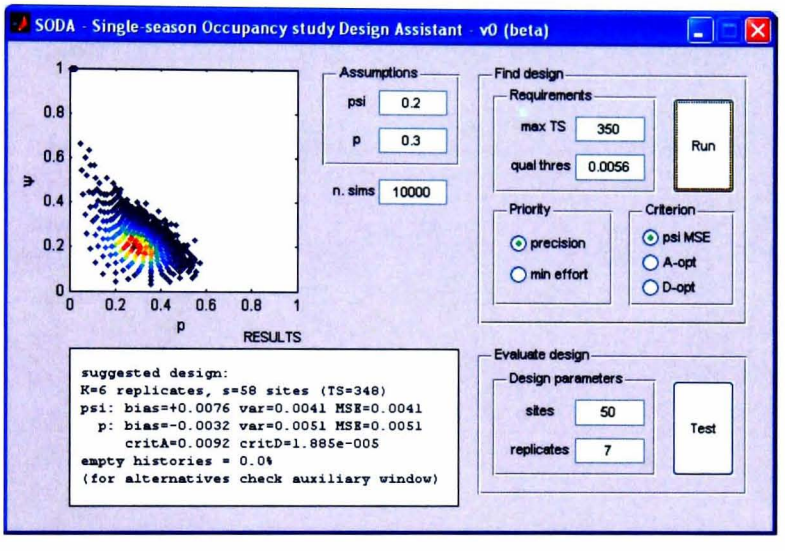
Yamaura, Y., Royle, J., Shimada, N., Asanuma, S., Sato, T., Taki, H. & Makino, S.i. (2012) Biodiversity of man-made open habitats in an underused country: a class of multispecies abundance models for count data. *Biodiversity and Conservation,* **21,** 1365-1380.

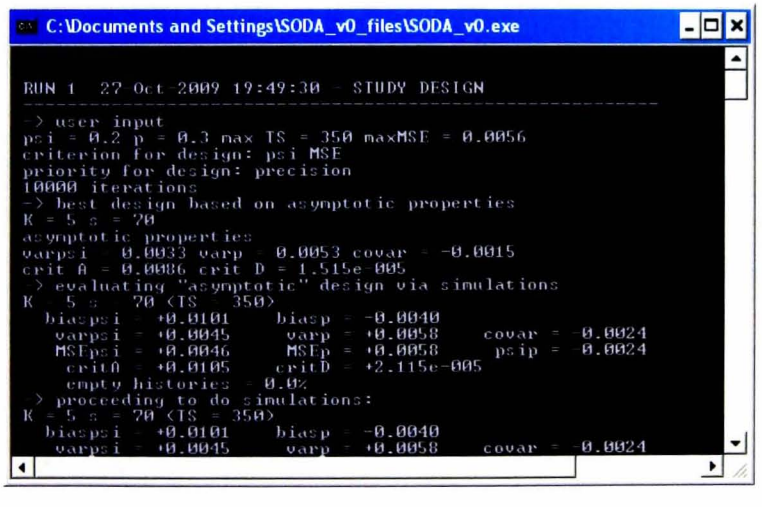
Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution,* **16,** 446-453.

APPENDICES

A.1 Software tool assistant for the design of occupancy studies (SODA)

SODA (Single-season Occupancy study Design Assistant) is a MATLAB-based stand-alone software tool that allows running an automated search for a suitable design given user-specified requirements for the basic single-season occupancy model, as well as testing the performance of specified designs. The tool and instructions on its use can be found at www.kent.ac.uk/ims/personal/msr. An R function for evaluating the performance of a given design is available at the same site.



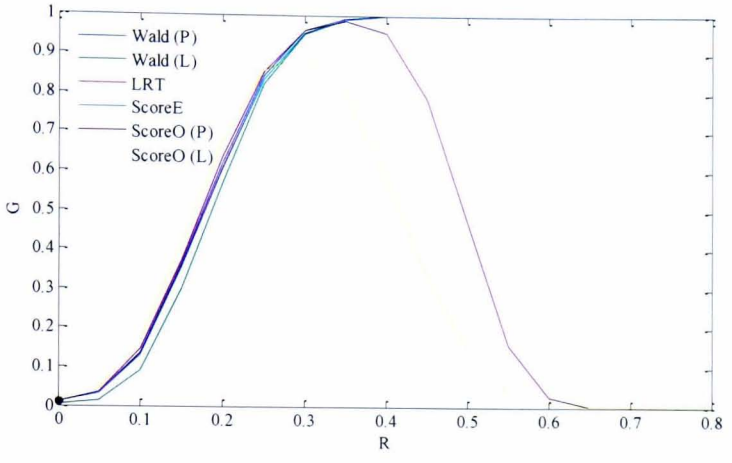


SODA results displayed in the main window and auxiliary window

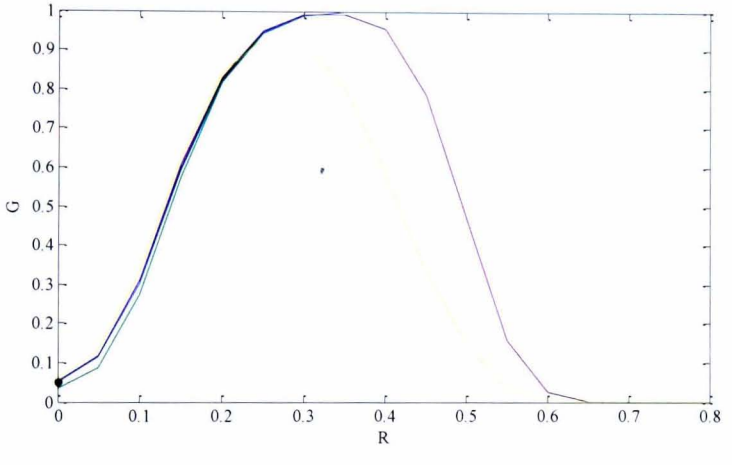
A.2 Additional simulation results for section 3.2.6

Case 3

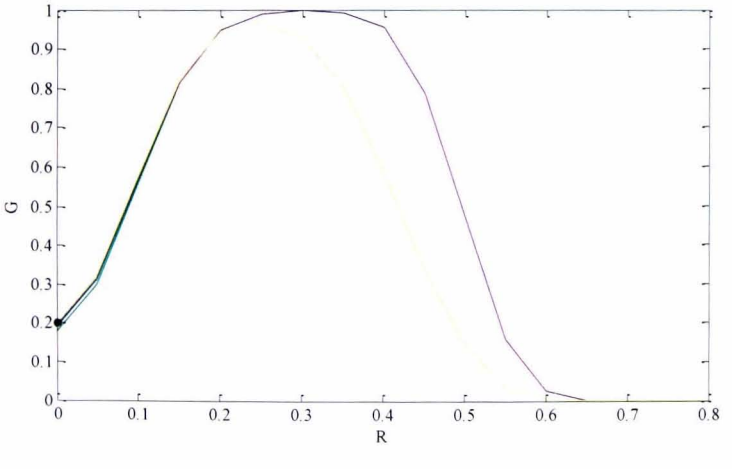
(a)



(b)

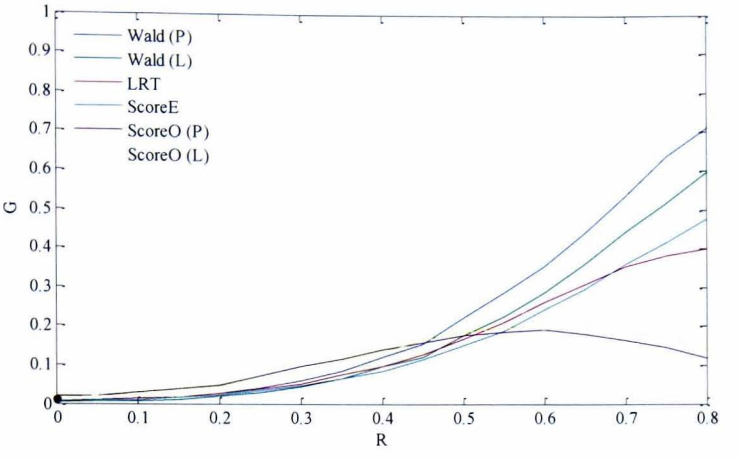


(c)

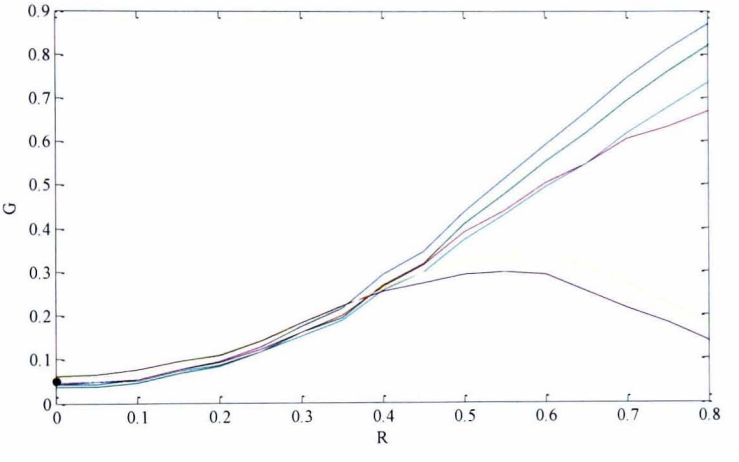


Case 4

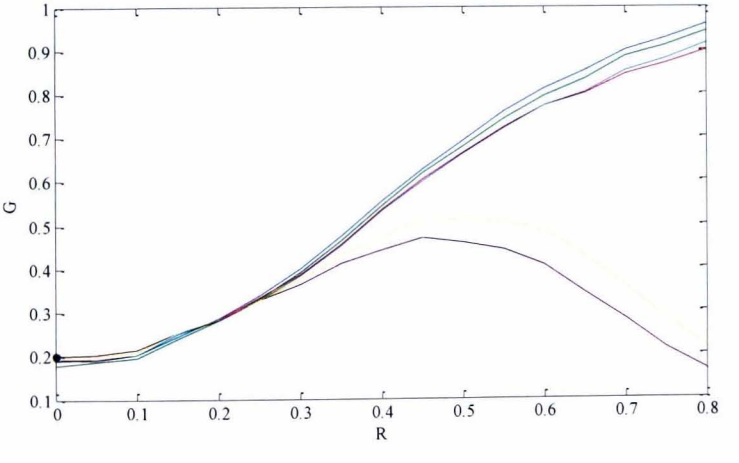
(a)



(b)

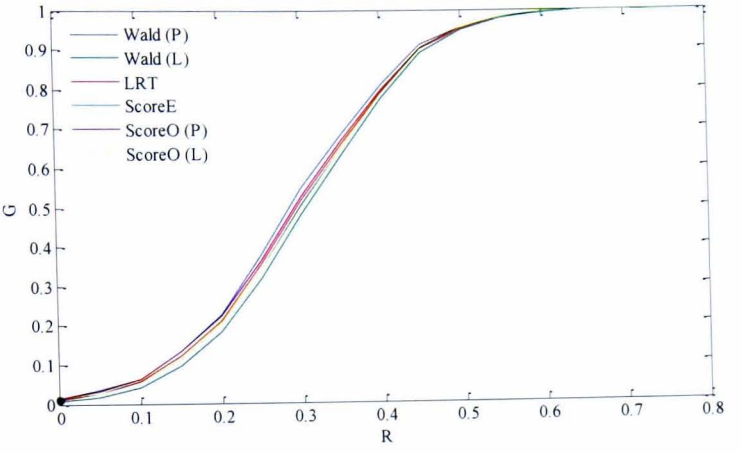


(c)

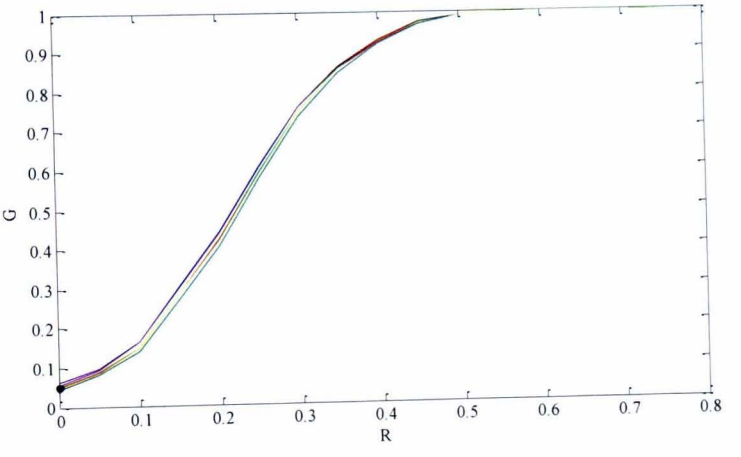


Case 5

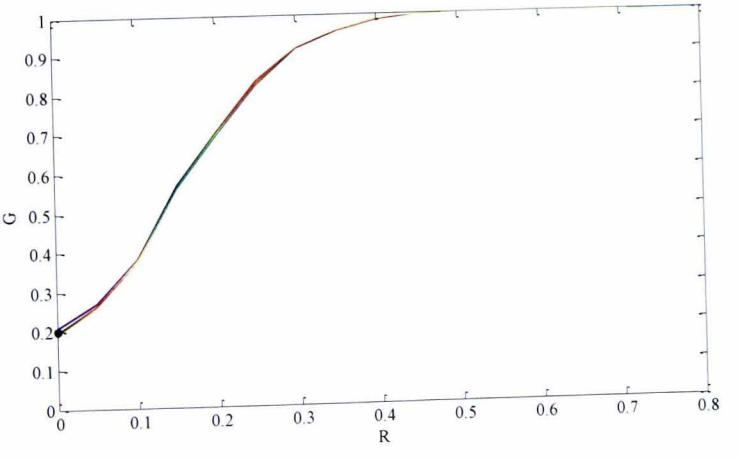
(a)



(b)

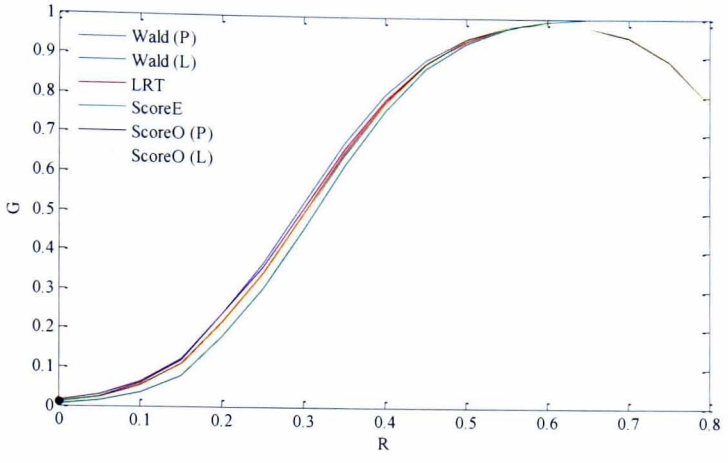


(c)

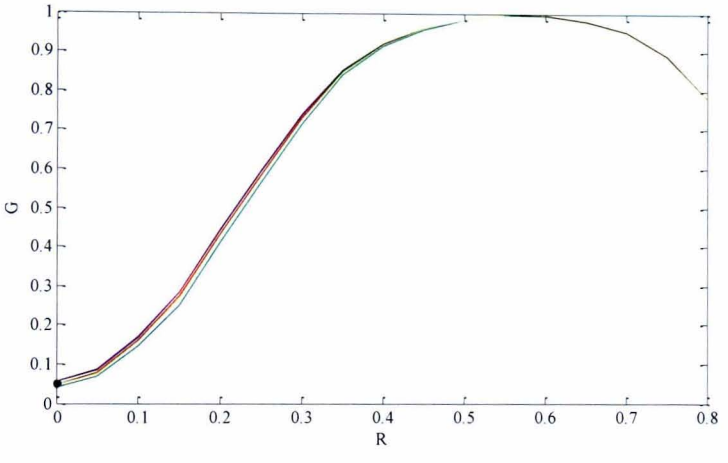


Case 6

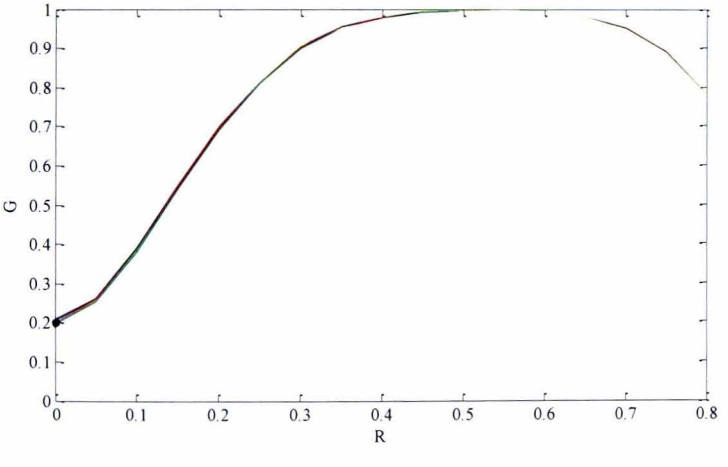
(a)



(b)



(c)



A.3 Mathematical equivalence of cases A2 and B1 in section 3.4

Case A2: Fixed proportion of occupied subunits at occupied sites (θ) and sampling with replacement (i.e. binomial distribution)

Case B1: Fixed probability of a subunit being occupied at occupied sites (θ) and sampling without replacement (i.e. hypergeometric distribution).

Let X be the number of replicate surveys at an occupied site that sample occupied subunits, N the number of subunits per site, K the number of replicates per site and θ the fixed proportion (Case A2) or the probability (Case B1) of occupied subunits at occupied sites. X is distributed as

$$X_{A2} \sim Bin(K, \theta) \text{ under case A2}$$

$$X_{B1} \sim HG\big(N, K, Bin(N, \theta)\big) \text{ under case B1}$$

The binomial mixture of hypergeometric distributions in B1 gives rise to the binomial distribution in A2 (Johnson *et al.* 2005, p. 377), as shown below

$$\Pr(X_{B_1} = x) = \sum_{i=0}^{N} \left[\frac{\binom{i}{x}\binom{N-i}{K-x}}{\binom{N}{K}} \right] \left[\binom{N}{i} \theta^i (1-\theta)^{N-i} \right]$$

$$= \binom{K}{x} \sum_{i=x}^{N-K+x} \binom{N-K}{i-x} \theta^i (1-\theta)^{N-i}$$

$$= \binom{K}{x} \theta^x (1-\theta)^{K-x} \sum_{i=0}^{N-K} \binom{N-K}{i} \theta^i (1-\theta)^{(N-K)-i}$$

$$= \binom{K}{x} \theta^x (1-\theta)^{K-x} = \Pr(X_{A_2} = x) \qquad x \in \{0 \dots K\}, K \leq N$$

A.4 MLEs for the two-season Markovian occupancy model

In section 2.2.6 we describe an occupancy model for multiple-season data with Markovian dependence in the occupancy status of the sites across seasons. For the particular case of two seasons it is evident that, under the assumption of perfect detection, the estimator of seasonal occupancy is the same regardless of whether independence or Markovian dependence is assumed (for season j, $\hat{\psi}_j = S_{d_j}/S$, where S_{d_j} is the number of sites out of S in which the species was detected in the season). Here we show that this also applies for the model in section 2.2.6, which accounts for imperfect detection.

For the two-season case, the contribution of site i to the likelihood function in (2.24) is

$$L_i(\psi_1, \gamma, \varepsilon, p_1, p_2) = \boldsymbol{\phi_0} \begin{bmatrix} p_1^{d_1}(1-p_1)^{K-d_1} & 0 \\ 0 & \mathrm{I}(d_1 = 0) \end{bmatrix} \boldsymbol{\phi} \begin{bmatrix} p_2^{d_2}(1-p_2)^{K-d_2} \\ \mathrm{I}(d_2 = 0) \end{bmatrix},$$

or what is the same

$$\begin{aligned} L_i = {} & \{(1-\varepsilon)\psi_1 p_1^{d_1}(1-p_1)^{K-d_1} + \gamma(1-\psi_1)\mathrm{I}(d_1 = 0)\} p_2^{d_2}(1-p_2)^{K-d_2} \\ & + \{\varepsilon\psi_1 p_1^{d_1}(1-p_1)^{K-d_1} + (1-\gamma)(1-\psi_1)\mathrm{I}(d_1 = 0)\}\mathrm{I}(d_2 = 0), \end{aligned} \tag{A5.1}$$

where K is the number of replicate visits per season, d_j is the number of detections at the site in season j, $j = 1, 2$, and $\mathrm{I}(.)$ denotes the indicator function which takes value one when the expression in brackets is true and zero otherwise. We can distinguish four types of site in the detection history:

- A: sites where the species is detected in both seasons ($d_1 > 0, d_2 > 0$),
- B: sites where the species is only detected in season 1 ($d_1 > 0, d_2 = 0$),

- C: sites where the species is only detected in season 2 ($d_1 = 0, d_2 > 0$),

- D: sites where the species is not detected in either season ($d_1 = 0, d_2 = 0$).

We define the following quantities, which provide a full data summary for this model,

- S_x: number of sites of type x, $x \in \{A, B, C, D\}$

- d_{T_j}: total number of detections of the species in season j, $j = 1, 2$.

For convenience, let us define as well

- d_{A1}: total number of detections of the species in season 1 in sites of type A,

- d_{B1}: total number of detections of the species in season 1 in sites of type B,

- d_{A2}: total number of detections of the species in season 2 in sites of type A,

- d_{C2}: total number of detections of the species in season 2 in sites of type C,

where $d_{T1} = d_{A1} + d_{B1}$, $d_{T2} = d_{A2} + d_{C2}$ and $S_A + S_B + S_C + S_D = S$.

The expression (A5.1) for each of the site types is

$$L_A = (1-\varepsilon)\psi_1 p_1^{d_1}(1-p_1)^{K-d_1} p_2^{d_2}(1-p_2)^{K-d_2},$$

$$L_B = (1-\varepsilon)\psi_1 p_1^{d_1}(1-p_1)^{K-d_1}(1-p_2)^{K} + \varepsilon\psi_1 p_1^{d_1}(1-p_1)^{K-d_1},$$

$$L_C = \{(1-\varepsilon)\psi_1(1-p_1)^{K} + \gamma(1-\psi_1)\} p_2^{d_2}(1-p_2)^{K-d_2},$$

$$L_D = \{(1-\varepsilon)\psi_1(1-p_1)^{K} + \gamma(1-\psi_1)\}(1-p_2)^{K} + \varepsilon\psi_1(1-p_1)^{K} + (1-\gamma)(1-\psi_1),$$

and the likelihood for the full history is

$$L = L_A^{S_A} \times L_B^{S_B} \times L_C^{S_C} \times L_D^{S_D}.$$

Since our interest is in the occupancy estimators, let us reparameterize the likelihood as a function of ψ_2. Considering that $\psi_2 = \psi_1(1-\varepsilon) + (1-\psi_1)\gamma$ and rearranging some terms we have

$$\begin{aligned} L = {} & p_1^{d_{T1}}(1-p_1)^{K(S_A+S_B)-d_{T1}} p_2^{d_{T2}}(1-p_2)^{K(S_A+S_C)-d_{T2}} \\ & \times \{(1-\varepsilon)\psi_1\}^{S_A} \\ & \times \{(1-\varepsilon)\psi_1(1-p_2^*) + \varepsilon\psi_1\}^{S_B} \\ & \times \{\psi_2 - \psi_1(1-\varepsilon)p_1^*\}^{S_C} \\ & \times \{1 - \psi_1 p_1^* - \psi_2 p_2^* + \psi_1(1-\varepsilon)p_1^* p_2^*\}^{S_D}, \end{aligned}$$

where $p_i^* = 1-(1-p_i)^K$ is the probability of detecting the species at least once in season i at a site, given that it is occupied. Using a reparameterization $\theta_1 = \psi_1 p_1^*$, $\theta_2 = \psi_2 p_2^*$ and $\delta = (1-\varepsilon)p_2^*$ allows the likelihood to be written as the product of two functions of one parameter and one involving three parameters

$$\begin{aligned} L &= F_1(p_1) \times F_2(p_2) \times F_3(\theta_1, \theta_2, \delta) \\ &= \frac{p_1^{d_{T1}}(1-p_1)^{K(S_A+S_B)-d_{T1}}}{(p_1^*)^{S_A+S_B}} \times \frac{p_2^{d_{T2}}(1-p_2)^{K(S_A+S_C)-d_{T2}}}{(p_2^*)^{S_A+S_C}} \\ &\times \theta_1^{S_A+S_B}\delta^{S_A}\{1-\delta\}^{S_B}\{\theta_2 - \theta_1\delta\}^{S_C}\{1-\theta_1-\theta_2+\theta_1\delta\}^{S_D}. \end{aligned}$$

This simplifies finding the MLEs, as now the problem involves maximizing lower dimensional functions separately. The MLEs for each of the two detection probability parameters, obtained by maximizing the one-parameter functions F_1 and F_2, satisfy

$$\frac{\hat{p}_1}{\hat{p}_1^*} = \frac{d_{T1}}{K(S_A + S_B)}, \qquad \frac{\hat{p}_2}{\hat{p}_2^*} = \frac{d_{T2}}{K(S_A + S_C)}.$$

To obtain the MLEs for the three remaining parameters $(\theta_1, \theta_2, \delta)$ we differentiate $f_3 = \log(F_3)$ with respect to each of the parameters and equate to zero which leads to the following system of equations

$$\frac{S_A + S_B}{\theta_1} - \frac{S_C\delta}{\theta_2 - \theta_1\delta} + \frac{S_D(\delta - 1)}{1 - \theta_1 - \theta_2 + \theta_1\delta} = 0$$

$$\frac{S_C}{\theta_2 - \theta_1\delta} - \frac{S_D}{1 - \theta_1 - \theta_2 + \theta_1\delta} = 0$$

$$\frac{S_A}{\delta} - \frac{S_B}{1 - \delta} - \frac{S_C\theta_1}{\theta_2 - \theta_1\delta} + \frac{S_D\theta_1}{1 - \theta_1 - \theta_2 + \theta_1\delta} = 0,$$

which can be easily solved to obtain the following expressions

$$\hat{\delta} = \frac{S_A}{S_A + S_B}, \qquad \hat{\theta}_1 = \frac{S_A + S_B}{S_A + S_B + S_C + S_D}, \qquad \hat{\theta}_2 = \frac{S_A + S_C}{S_A + S_B + S_C + S_D}.$$

Back-transforming we arrive to

$$\hat{\varepsilon} = 1 - \frac{S_A}{(S_A + S_B)\hat{p}_2^*},$$

$$\hat{\psi}_1 = \frac{S_A + S_B}{(S_A + S_B + S_C + S_D)\hat{p}_1^*},$$

$$\hat{\psi}_2 = \frac{S_A + S_C}{(S_A + S_B + S_C + S_D)\hat{p}_2^*}.$$

Let us write $S_{d1} = S_A + S_B$ and $S_{d2} = S_A + S_C$ for the number of sites in which the species was detected in the first season and second season respectively. Rewriting accordingly the MLE expressions for occupancy and detection probability we get

$$\hat{\psi}_1 = \frac{S_{d1}}{S\hat{p}_1^*}, \quad \hat{\psi}_2 = \frac{S_{d2}}{S\hat{p}_2^*}, \quad \frac{\hat{p}_1}{\hat{p}_1^*} = \frac{d_{T1}}{KS_{d1}}, \quad \frac{\hat{p}_2}{\hat{p}_2^*} = \frac{d_{T2}}{KS_{d2}},$$

which are the same expressions as those obtained under the assumption of independence, that is, when data are analyzed using two separate single-season models (2.3).

We show now that the asymptotic variance of the MLEs is also the same in both models. Under the $(\theta_1, \theta_2, \delta)$ parameterization the hessian matrix of the log-likelihood is

$$\boldsymbol{H} = \begin{pmatrix} \frac{\partial^2 f_1}{\partial p_1^2} & 0 & 0 \\ 0 & \frac{\partial^2 f_2}{\partial p_2^2} & 0 \\ 0 & 0 & \boldsymbol{H}_{f_3} \end{pmatrix}$$

where $f_j = \log(F_j)$ and

$$\boldsymbol{H}_{f_3} = \begin{pmatrix} \frac{\partial^2 f_3}{\partial \theta_1^2} & \frac{\partial^2 f_3}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 f_3}{\partial \theta_1 \partial \delta} \\ \frac{\partial^2 f_3}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 f_3}{\partial \theta_2^2} & \frac{\partial^2 f_3}{\partial \theta_2 \partial \delta} \\ \frac{\partial^2 f_3}{\partial \theta_1 \partial \delta} & \frac{\partial^2 f_3}{\partial \theta_2 \partial \delta} & \frac{\partial^2 f_3}{\partial \delta^2} \end{pmatrix}.$$

The variance for the detection parameters p_1 and p_2 is obtained as

$$\text{var}(p_j) = \left\{ -\mathbb{E}\left(\frac{\partial^2 f_j}{\partial p_j^2} \right) \right\}^{-1}, \qquad j = 1,2,$$

which, considering that $\mathbb{E}[d_{Tj}] = S\psi_j K p_j$, is

$$\mathrm{var}(p_j) = \frac{p_j(1-p_j)}{SK\psi_j}\left\{\frac{p_j^*}{p_j^* - Kp_j(1-p_j)^{K-1}}\right\}. \tag{A5.2}$$

The variance-covariance matrix for $(\theta_1, \theta_2, \delta)$ is obtained as

$$\Sigma_{\theta_1,\theta_2,\delta} = \left\{-\mathbb{E}(H_{f_3})\right\}^{-1},$$

which, considering $\mathbb{E}[S_A] = S\theta_1\delta$, $\mathbb{E}[S_B] = S\theta_1(1-\delta)$ and $\mathbb{E}[S_C] = S(\theta_2 - \theta_1\delta)$, is

$$\Sigma_{\theta_1,\theta_2,\delta} = \frac{1}{S}\begin{pmatrix} \theta_1(1-\theta_1) & \theta_1(\delta-\theta_2) & 0 \\ \theta_1(\delta-\theta_2) & \theta_2(1-\theta_2) & \delta(1-\delta) \\ 0 & \delta(1-\delta) & \delta(1-\delta)/\theta_1 \end{pmatrix}. \tag{A5.3}$$

From (A5.3) the variance of ψ_j can now be calculated as

$$\begin{aligned} \mathrm{var}(\psi_j) &= \left(\frac{\partial\psi_j}{\partial\theta_j}\right)^2 \mathrm{var}(\theta_j) + \left(\frac{\partial\psi_j}{\partial p_j}\right)^2 \mathrm{var}(p_j) \\ &= \frac{\psi_j}{S}\left\{(1-\psi_j) + \frac{1-p_j^*}{p_j^* - Kp_j(1-p_j)^{K-1}}\right\}, \end{aligned} \tag{A5.4}$$

and the covariance of $\psi_1$and ψ_2 as

$$\mathrm{cov}(\psi_1, \psi_2) = \frac{\partial\psi_1}{\partial\theta_1}\frac{\partial\psi_2}{\partial\theta_2}\mathrm{cov}(\theta_1, \theta_2) = \frac{\psi_1\{(1-\varepsilon) - \psi_2\}}{S}, \tag{A5.5}$$

which can be rewritten using other parameterizations as

$$\mathrm{cov}(\psi_1, \psi_2) = \frac{(\psi_2 - \gamma)(1-\psi_1)}{S} = \frac{\psi_1(1-\psi_1)}{S}(1-\varepsilon-\gamma). \tag{A5.6}$$

Note that both (A5.2) and (A5.4) are the same expressions as those obtained under the single-season model (2.13), and that therefore the occupancy estimators for the Markovian model for two seasons $\psi_1(.)\varepsilon(.)\gamma(.)p_1(.)p_2(.)$ have the same expression and asymptotic variance as those obtained assuming independence with a model $\psi_1(.)\psi_2(.)p_1(.)p_2(.)$. The covariance in (A5.5) and (A5.6) is zero under the assumption of independence, as then $1-\varepsilon=\gamma=\psi_2$.

A.5 The Kronecker sum

The Kronecker sum $\oplus$ of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined as

$$\boldsymbol{A} \oplus \boldsymbol{B} = (\boldsymbol{A} \otimes \boldsymbol{I_B}) + (\boldsymbol{I_A} \otimes \boldsymbol{B}),$$

where $\boldsymbol{I_A}$ and $\boldsymbol{I_B}$ are identity matrices of the same order as $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\otimes$ represents the Kronecker product defined as

$$\boldsymbol{C} \otimes \boldsymbol{D} = \begin{bmatrix} c_{11}\boldsymbol{D} & c_{12}\boldsymbol{D} & \dots & c_{1m}\boldsymbol{D} \\ \vdots & \vdots & & \vdots \\ c_{n1}\boldsymbol{D} & c_{n2}\boldsymbol{D} & \dots & c_{nm}\boldsymbol{D} \end{bmatrix}.$$

For example, the Kronecker sum of two matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$,

$$\boldsymbol{Q} = \begin{bmatrix} -\mu_{12} & \mu_{12} \\ \mu_{21} & -\mu_{21} \end{bmatrix}, \qquad \boldsymbol{R} = \begin{bmatrix} -r_{12} & r_{12} \\ r_{21} & -r_{21} \end{bmatrix},$$

which could represent the generator matrices of two different 2-MMPP processes, is

$$\boldsymbol{Q} \oplus \boldsymbol{R} = \begin{bmatrix} -\mu_{12} & 0 & \mu_{12} & 0 \\ 0 & -\mu_{12} & 0 & \mu_{12} \\ \mu_{21} & 0 & -\mu_{21} & 0 \\ 0 & \mu_{21} & 0 & -\mu_{21} \end{bmatrix} + \begin{bmatrix} -r_{12} & r_{12} & 0 & 0 \\ r_{21} & -r_{21} & 0 & 0 \\ 0 & 0 & -r_{12} & r_{12} \\ 0 & 0 & r_{21} & -r_{21} \end{bmatrix}$$

$$= \begin{bmatrix} -(\mu_{12} + r_{12}) & r_{12} & \mu_{12} & 0 \\ r_{21} & -(\mu_{12} + r_{21}) & 0 & \mu_{12} \\ \mu_{21} & 0 & -(\mu_{21} + r_{12}) & r_{12} \\ 0 & \mu_{21} & r_{21} & -(\mu_{21} + r_{21}) \end{bmatrix}.$$