Bayesian hierarchical methods for species distribution modelling under imperfect detection

Jafet M. Belmont Osuna

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

December 23, 2021

Abstract

Monitoring the distribution of wildlife populations has become essential for the understanding of how species are affected by environmental changes and to provide adequate management plans and effective strategies for the conservation of biodiversity. The growing concern about biodiversity loss has led to a rapid development of sampling methods and data collection schemes that enables data of the distributions for multiple species to be obtained at different temporal and spatial scales. Nowadays, biodiversity conservation involves monitoring programs that target multiple species within a community where individual species responses vary widely. This high variability makes the task of identifying the ecological processes that drive species distributions challenging and complex. This complexity has led to the development of a wide range of species distribution models that allow the identification of the most important areas for biodiversity conservation. However, describing such processes is no easy task due to the sources of uncertainty that occur at different spatial and temporal scales and that are induced by imperfect detectability. Thus, modern methods in statistics are increasingly being used to analyse the distribution and abundance of wildlife populations while accounting for the multiple sources of error associated with both, the ecological process of interest and the data collection process.

The present work extends some of the well-established species distribution modelling techniques that address imperfect detection and propose new methods to describe different attributes of biological communities (e.g. species rarity) and their relationship with the environment. Computer simulations were used to assess models performance. Then, the proposed methods were applied to a data set of Odonata occurrence records in water bodies across the UK that were partially observed due to imperfect detection. The data for this research were provided by the Hydroscape project (www.hydroscapeblog.wordpress.com), a project that aim to determine how different connectivity metrics interact with environmental stressors to affect species diversity in UK freshwaters.

Chapter 1 gives a background of the ecological concepts and statistical principles that are commonly used in species distribution modelling, introduces the questions of interest and the aims of this research. It also shows an overview of the data and presents an exploratory analysis that will be necessary to take into account for the analysis in subsequent chapters.

Chapter 2 reviews some of the existing models that have been developed to investigate species distributions under imperfect detection and apply such methods to the Odonata case study, discusses the importance of accounting for imperfect species detection through a simulation study, and compares different software and approaches that have been developed to fit such models.

Chapter 3 proposes a new method to quantify species rarity in a community when species are detected imperfectly. Then, a two-step modeling framework is proposed as an approach that enables for a complex hierarchical model to be analyzed in different stages to provide a pragmatic computationally efficient method for choosing the most relevant predictors affecting an ecological response of interest while propagating the uncertainty associated with the estimation of this quantity on a second analysis.

In chapter 4, a new method that accounts for non-linear relationships between species distribution and environmental conditions is developed. A simulation study is presented to assess the model performance under different scenarios and different methodological consideration practices are discussed.

Chapter 5 provides an application of the model developed in chapter 4 to investigate Odonata species distribution temporal patterns and discusses how these results can be used for biodiversity conservation and management.

Finally, chapter 6 summarizes the main outcomes of this research and discusses the methodological innovations and challenges of the proposed methods with a final discussion on possible future work.

Contents

List of Tables

List of Figures

Acknowledgements

This work would not have been possible if it were not for all the amazing people that supported me during all the good and not-so-good times of this PhD.

First and foremost, I would like to thank my supervisors Claire Miller and Marian Scott for taking me as their student, I could not have asked for more caring, kinder and wiser supervisors. I am truly grateful for their support, patience, and guidance that encouraged me to keep moving forward even in the most difficult times. I cannot thank you enough for all the opportunities you gave me, learning from you has been the most enjoyable and I hope one day I will be as an amazing teacher and statistician as you both are. Secondly, I would like to thank to CONACyT (Consejo Nacional de Ciencia y Tecnologí a - Scholarship 494334) and NERC (NE/N005740/1) for their financial support, without which it would have been impossible for me to undertake my doctoral studies in Glasgow. I would also like to thank the Hydroscape team, especially Dr. Craig Wilkie, Dr. Stephen Brooks, and Dr. Tom August, for their valuable insights and comments and for supplying and helping me to understand the data.

Thanks to the School of Mathematics and Statistics at the University of Glasgow for providing me with the necessary academic bases and support without which I would not have been able to carry out this work. I will always be grateful to the staff of this department for all the opportunities they gave me and for the amazing support during my doctoral studies. My fellow PhD students, especially Dimitra and Yoana who were always happy to help and with who I shared lots of laughs and joy. All my friends in Glasgow, you have made this journey a life-changing experience. Special thanks to all my friends in the Mexican Society, you made Glasgow feel like home even in the coldest winter.

To my friends in Mexico and second family, Héctor, Panch, Rudo, Paris, and Fernando for always being there when I've needed them and for supporting me all these years no matter the distance. I would like to thank my family, my mum, my aunt, and my grandma for raising me and always encouraging me to pursue my dreams. I would not be here if it were not for you, you have taught me not to give up and always look ahead. During these difficult times, you have always guided me and helped me to stay on the right path, I will keep working hard to make you proud and will always be grateful for all your love and support. Finally, I would like to give a special thanks to my beloved Nathy for always believing in me and supporting me. You have been by my side in every step of the way and always had my back no matter the distance or how hard things could get. You have lent me your hand and lighten up my path even if we were miles apart. Paraphrasing Roger de Bussy-Rabutin, "distance is to love what wind is to fire; it extinguishes the small, it inflames the great". Thanks a lot for your patience and love, I cannot imagine achieving this without you.

"It is that range of biodiversity that we must care for - the whole thing - rather than just one or two stars." Sir David Attenborough.

In beloved memory of Sherlock.

Declaration

I, Jafet Belmont, hereby declare that this thesis entitled Bayesian hierarchical methods for species distribution modelling under imperfect detection and the work presented in it are entirely my own except as otherwise explicitly stated and referenced in the text. I confirm that this work has not been previously submitted for any degree or professional qualification and was done while a candidate for a research degree at this University. Some of the material in Chapter 3 was published in Proceedings of the 35th International Workshop on Statistical Modelling (2020), with the title " Hierarchical Species Distribution Modelling Across High Dimensional Nested Spatial Scales". Further results in chapter 3 were presented as a poster at the Royal Statistical Society conference in 2019. A manuscript based upon this material has been submitted to Diversity and Distributions and is currently under review. The material in Chapter 4 and 5 was presented as a talk at the 8th Channel Network Conference (2021). Some of the material in Chapter 2 was presented as a talk at the XII Latin American Botanical Congress (LABC) in 2018.

Chapter 1

Introduction

1.1 Overview

Understanding species distribution is essential to investigate biological diversity and to develop efficient strategies and management plans for ecological conservation (Pollock et al., 2017). Species distributions are affected by different geographic factors and environmental drivers that occur at varying temporal and spatial scales. Environmental drivers are those complex natural or human-induced phenomena that ultimately causes direct or indirect effects on biodiversity (e.g. change-climate change, population and economic growth, land-use change) through the different pressures they generate (e.g. hydrology alteration, habitats fragmentation, pollution),Oesterwind et al. (2016). Thus, methods that describe and predict how species distributions will change in space and time due to environmental changes have become an important tool for the conservation and management of natural areas (Elith and Leathwick, 2009). In consequence, an important effort has been made to produce metrics that relate species distribution and the physical environment to have a better understanding of biodiversity changes (Outhwaite et al., 2020).

However, addressing such complex spatial and temporal occurrence patterns can be difficult for both the collection and analysis of data, especially on large scale studies where data accessibility and quality depend on the target species distributions (inconspicuous and rare species will require a large amount of effort to be sampled) and the spatial scale of the study (the sampling effort involved in large spatial scales studies usually result in highly costly sampling schemes) (Elith et al., 2010; Rushing et al., 2019).

With the growing concern over the effects that environmental changes have on species diversity, monitoring the distribution and abundance of wildlife populations has become increasingly relevant for conservation studies (Devarajan et al., 2020). Thus, new technologies and information systems have been rapidly developed to facilitate access to species distribution data for an increasingly larger collection of species across different spatio-temporal scales (Elith and Leathwick, 2009). These data collections however, are prone to different sources of errors, from which the observational error, i.e. how data are collected, has been of major interest for scientists and statisticians and has been addressed by different methods through the last decade (Kellner and Swihart, 2014). Particularly, hierarchical models have been used to model species distribution when the species true presence at a site is masked by the capability of detecting it during a survey. However, incorporating this source of uncertainty can be computationally

expensive and little research to produce efficient algorithms that capture species distributions' spatial-temporal dynamics has been done (Clark and Altwegg, 2019).

Estimating multiple species occurrences in a community can be challenging due to the variation in the individual species responses, specifically for rare species since their limited distribution range makes the task of detecting them difficult (MacKenzie et al., 2017). These rare species are often of special interest to managers and conservationists as they represent the most threatened species by habitat fragmentation and connectivity loss (due to reduction in their habitat's area and populations isolation (Wan et al., 2018)). However, very few methods have been developed to identify rare species and assess how different environmental drivers shape their distributions to determine areas with relevance for their conservation (Leroy et al., 2013). In addition, whereas most of the methods to describe, map, and predict the distributions of multiple species have been applied mainly to terrestrial organisms (Kellner and Swihart, 2014), applications in freshwater environments have usually been over looked (Elith and Leathwick, 2009; Devarajan et al., 2020) despite the well-known importance of aquatic environments of maintaining ecosystem balance by connecting different hydrological units in which matter, energy, and organisms flow (Adrian et al., 2009; Pringle and Triska, 2000).

Monitoring the health of aquatic ecosystems requires understanding how freshwater connections are established between different waterbodies and how altering them will impact ecosystems integrity (Pringle, 2006; Crooks and Sanjayan, 2006). Thus, to manage and monitor freshwater effectively, there is an urgent need to better understand how local environmental conditions and processes on a national scale (for example, the dynamics of colonization and extinction in a collection of different biological communities) produce the observed species distribution patterns that can often be confounded due to species imperfect detection (Devarajan et al., 2020).

Unfortunately, monitoring freshwater diversity can be difficult not only because of the imperfect detection of species but because data distributions of certain biological groups such as invertebrates are hard to collect, requiring a high amount of effort and experience (van Strien et al., 2010). However, recent citizen science data projects where volunteers collect data from different biological groups enable species occurrence records to be obtained at large spatial and temporal scales that would be by any other means difficult and costly (Aceves-Bueno et al., 2017).

Within the different biological groups suitable for these citizen science projects, Dragonflies and Damselflies (Odonata) have proven to be an attractive group for the general public because adult species traits are appealing and can be identified relatively easily (encouraging a larger number of people to collect data about its distribution) van Strien et al. (2013). Additionally, Odonata species have proven to be useful to assess and monitor freshwater ecosystem quality because of their sensitivity to environmental changes at their breeding sites and surrounding areas (Golfieri et al., 2016). However, occurrence data for these species are observed incompletely due to imperfect detection (Outhwaite et al., 2020). In consequence, statistical methods that consider species imperfect detection need to be developed to quantify species distributions in a community characterized by complex occurrence patterns driven by population dynamics that vary across both space and time to drivers and pressures.

This research will explore and develop different statistical methods to quantify species distribution when species are detected imperfectly. This encompasses reviewing and comparing recent approaches in order to develop novel methodologies, metrics, and visualization tools to describe a different aspect of biological communities that can be used by ecologist and conservationists to characterize those communities on a national scale or to identify focal species that can be threatened by several factors such as environmental changes. Such methods will be applied to an Odonata case study to contribute towards the underrepresented field of freshwater ecology and small invertebrates species distribution modelling.

1.2 Aims and objectives

The main aims and objectives of this research are to:

- Explore and compare the computational efficiency of different state of the art statistical methods to provide valuable insights regarding species distribution modelling under an imperfect detection scenario.
- Develop statistical methods to identify and quantify species rarity in a community while accounting for imperfect detection.
- Develop a modelling framework approach to elucidate how stressors and connectivity interact to influence species distributions. Specifically:
  - Explore dimensionality reduction approaches
  - Develop statistical linkages that incorporate different sources of uncertainty.
- Develop spatio-temporal models to describe non-linear relationships between species occurrences under imperfect detection and environmental conditions. This will be assessed by:
  - Creating a simulation study to validate model results and to assess the model extensions and limitations.
  - Application to a real-life case study and appropriate selection of metrics to visualizing and assess model performance.
- Provide a detailed assessment of the contribution of this research to the state of the art in ecological modelling by discussing the following:
  - Practices and considerations in study designs based on the simulation work undergone in this research.
  - Model considerations and limitations and how these impact the collection of new data or which approaches involving existing datasets can be taken.
  - Identify areas of improvement from which new research fields can be established.

The remainder of this chapter will introduce the ecological principles required for understanding the role that environment has in shaping species distribution in aquatic ecosystems. Then, the key statistical concepts from which current methodological approaches to address species distribution modelling are based will be discussed. Finally, an initial exploratory analysis and data processing of the Odonata case study data set will be presented.

1.3 Background

1.3.1 Species distribution and environmental connectivity

Ecosystem connectivity has become a major study subject for conservationists and ecologists across the world, as it represents a potential mechanism to mitigate the impact that anthropogenic activities and other environmental stressors (i.e. drivers that lead to pressures which have a negative impact on ecosystems. e.g land-use change, overexploitation of natural resources and climate change) have on species diversity (Crooks and Sanjayan, 2006).

Connectivity refers to the landscape physical structures that either limit or facilitates organisms movement as well as the species behavioural strategies to overcome these physical features that constrain their distribution (Taylor et al., 2006). Connectivity is characterized by organisms movement across different spatio-temporal scales. For example, movement driven by foraging, dispersal and migration occur in very different spatial and temporal scales (whilst for any given species foraging may take place several times a day on a regular basis over the year, dispersal and migration occur in larger time intervals that show strong seasonality patterns with a longer range of movement) (Jeltsch et al., 2013).

Species biodiversity is influenced greatly by the landscape connectivity as it affects both population dynamics and communities assemblages by modifying species distributions patterns and their interactions with other landscape drivers such as climate change and land-use change (LaPoint et al., 2015). The study of how these drivers interact with the ecosystems connectivity to shape species distributions and biodiversity has not been fully understood yet, and little research is available (De Chazal and Rounsevell, 2009).

Aquatic ecosystems are well known for being characterized by their highly connected networks of waterbodies in which organisms, energy, materials and genetic resources move within and between hydrological units (Pringle, 2001). This helps to maintain ecological integrity of ecosystems. However, many human activities can alter hydrological connectivity at different spatial and temporal scales (e.g. Dams construction, flow regulation, water diversion, groundwater extraction), thereby affecting and damaging ecosystem functionality by disseminating human-derived nutrients, toxic wastes and facilitating the dispersion of exotic and invasive species (Pringle, 2003). Freshwater connectedness alteration has been increased at a worrying rate over the past decades, contributing to loss in aquatic biodiversity and ecosystem integrity on a global scale (Rosenberg et al., 2000). Thus, there is a timely need to understand how ecological responses of freshwaters to multiple stressors are affected by connectivity. Ecological con-

sequences of hydrological connectivity alteration can occur and affect species at different levels. For example, disruption of water body networks leads to a reduced gene flow which decreases aquatic organisms' genetic variation and thus, making populations more susceptible to local extinctions (Pringle, 1997; Winston et al., 1991).

Urban development such as dam construction, causes these genetic and species-specific effects that influence local biodiversity and can cause major economic losses when economically important species are affected Pringle (2006). For instance, Pringle and Triska (2000) report (1) extirpation of migratory species, (2) population isolation, (3) local extinction,(4) abundance reduction and (5) increase in exotic species, as some of the effects that hydrological connectivity disruption have on the fish and mollusc species diversity in a regional scale due to dam construction.

Hydrological connectivity alteration can also affect multiple species within ecological communities in various ways. For example, some North American mussel species numbers have declined dramatically in the past 20 years due to the removal of their host fish species by the construction of dams (Liittschwager, 1994). In some cases however, reducing hydrological connectivity has been proposed as a strategy to protect local endangered species from exotic and invasive species that displace local population from their habitats (Pringle, 2006). The effect that connectivity and land-use change have on species distributions can escalate to an ecosystem- and landscape-level. For instance, Freeman et al. (2003) explore how freshwater mussels' decline due to dams leads to lowered system productivity, nutrient retention and benthic stream stability. Increasing groundwater exploitation for agricultural purposes may also have large effects on aquatic ecosystems as they originate changes in drainage networks and the distribution of biota. For example, the Cuatro Cienegas basin in the northern part of Mexico is a unique system of springs, streams and pools that shelters over 70 endemic species and other microbial communities. This network has been threatened by the impact of agricultural development over the last 10 years resulting in temperature imbalance and water and biodiversity loss (Contreras-Balderas, 1984; Souza et al., 2006).

In summary, it is critical to understand how connectivity and stressors shape species distributions. Developing methodologies that describe how aquatic ecosystem connectivity and alterations to this property influence ecological patterns on regional and global scales is crucial to mitigate or even avoid environmental crises. The relationship between species distributions and the physical environment has been explored widely for different terrestrial, freshwater and marine biological groups. However, it is only relatively recently that freshwater and marine applications have gained an increasing attention from governmental and non-governmental entities (Elith and Leathwick, 2009). Thus, the challenge of estimating and predicting how species distributions and abundances will change due to environmental changes has led to the development of different species distribution models (SDMs) that explore the spatial and temporal patterns in species distribution (Hefley and Hooten, 2016). These models have become an important tool to explore in ecological studies with a wide range of application such as conservation management plans for threatened species and monitoring and predicting invasive species distribution ranges (Koshkina et al., 2017).

Often the only data available for species distribution modelling are those derived from partial observational processes due to imperfect detection of a species (MacKenzie et al., 2002; Koshkina et al., 2017). This has led to the development of different SDMs that account for inconspicuous species that are not always detectable in a survey (Elith and Leathwick, 2009). Thus, the following section describes how species distribution can be expressed from an statistical point of view and how it is related to other biodiversity metrics. At the end of this chapter, a case study for a dragonflies species distribution data set consisting of partial observation due to imperfect detection will be described.

1.3.2 Distribution, abundance, and species richness as a stochastic spatial point process

Biodiversity is described by three major attributes (Begon et al., 1986): Species distribution - the presence or absence of a species at a location, species richness - the different number of distinct species that occur at a site and Species abundance - the total number of individual member of each species that occur. These three attributes can be seen as derived quantities of a spatial point process (Kéry and Royle, 2015).
A spatial point process (SPP) is a stochastic process $Z(\mathbf{s}) : \mathbf{s} \in D$ where the random field $Z(\mathbf{s})$ is evaluated at location $\mathbf{s}$. The spatial domain D is a random set in $\mathbb{R}^2$, i.e. $D \subset \mathbb{R}^2$. Thus, the number of locations (and therefore the observations) occur at random.

Distribution and abundance are an aggregation of a SPP over some area A. Let $A \subset D \subset \mathbb{R}^2$. In this case, $Z(A)$ denotes the number of points in A. If D is bounded and $Z(A)$ is finite $\forall\ A \subset D$, then $\{Z(A) : A \subset D\}$ (Cressie and Wikle, 2015).

The expected number of points in A is given by the *first-order intensity* function evaluated at a centred region $\mathbf{s}$ with area $|d\mathbf{s}|$

$$\lambda_Z(\mathbf{s}) = lim_{|d\mathbf{s}|\to 0}\frac{\mathbb{E}[Z(d\mathbf{s})]}{|d\mathbf{s}|}. \tag{1.1}$$

Thus, the expectation $Z(A)$ is given by:

$$\begin{aligned}\mu_Z(A) &= \mathbb{E}[Z(A)] \\ &= \int_A \lambda_Z(\mathbf{s})ds,\end{aligned} \tag{1.2}$$

which measures the local contribution to the expected number of points at any location $\mathbf{s} \in D$. A common SPP $Z(\cdot)$ is the *homogeneous Poisson point process* in which for a given subset A with area $|A|$:

$$Z(A) \sim \text{Poisson}(\lambda_Z|A|), \quad A \subset D, \tag{1.3}$$

Moreover, for any subset $A_1, A_2 \subset D$, if $A_1 \cap A_2 = \emptyset$ then, $Z(A_1) \perp\!\!\!\perp Z(A_2)$.

Hence, this type of process is often referred to as a completely spatially random process (CRS), since the first order intensity function is constant for any $\mathbf{s} \in D$. This can be derived from equation (1.1) as follows:

$$\begin{aligned}\lambda_Z(\mathbf{s}) &= lim_{|d\mathbf{s}|\to 0}\frac{\mathbb{E}[Z(d\mathbf{s})]}{|d\mathbf{s}|} = lim_{|d\mathbf{s}|\to 0}\frac{\lambda_Z|d\mathbf{s}|}{|d\mathbf{s}|} \\ &= \lambda_Z. \end{aligned} \tag{1.4}$$

A homogeneous Poisson point process simulation using the `sim.fn()` function from the `AHMbook` package in `R` (Kery et al., 2017) is illustrated in Fig. 1.1. A total of 114 observations (individuals) are simulated in a quadrat of length 10 divided by cells of length 1. A constant intensity function across space λ_Z gives the mean number of points per unit area [1].

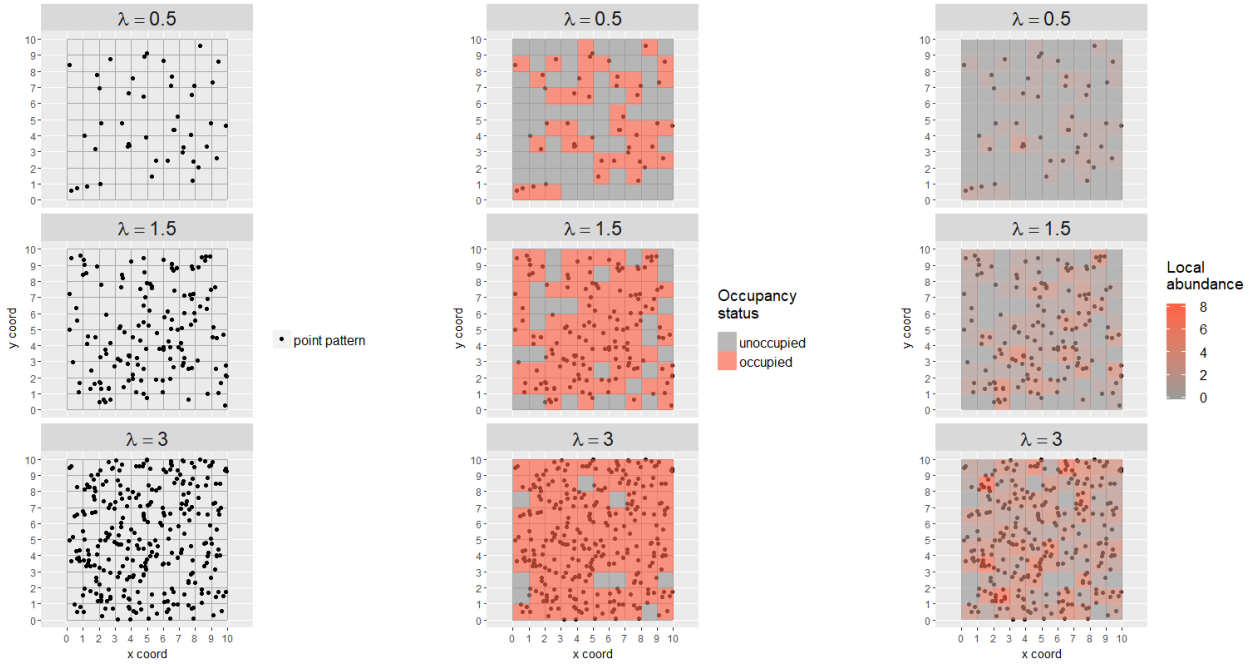


Figure 1.1: Homogeneus Poisson point process simulation following Hefley and Hooten (2016) over a 10 x 10 grid with a cell size of 1 unit with different intensities λ_Z. The left panels shows the point pattern of the simulated observations, the right indicates the local abundance (i.e. number of counts per cell) and the middle shows the occupancy (gray cells indicate unoccupied sites).

[1] Relationship between occupancy, distribution and point pattern based on a homogeneous Poisson point process simulation App available at `https://shiny.maths-stats.gla.ac.uk/2259971b/PoissProcSim/`

Figure 1.1 exemplifies the relationship between a point pattern, abundance and distribution (species distribution from now on will be referred as occupancy or true occurrences of a species). First, the point pattern can define the abundance N of a given species by the discretization of the spatial domain over a grid of quadratic cells of length m. The cell size, often denoted as grain, determines the spatial scale at which species occurrences are recorded.

The occupancy y_j is then determined depending on the presence/absence of the species on the j-th cell or site:

$$y_j = \begin{cases} 1 \text{ if } N > 0, \\ 0 \text{ otherwise.} \end{cases}$$

Adding up all of the i different species that occur at a site defines the sites' local species richness, i.e. $\sum_i y_{ij}$ for the jth site. This quantity is often referred to as α diversity (Kéry and Royle, 2015).

1.3.2.1 Information exchange and scale dependency

Richness, abundance and occupancy are scale dependent processes, in which the grain (cell size) and the intensity determine the conclusions that can be drawn about these quantities. Figures 1.2 and 1.3 portray the relationship between occupancy and abundance for different intensities and grain sizes.

A small grain size with low intensity yields to a very small proportion of occupied cells with just a few or even a single observation per occupied cell. At these values, the mean abundance and occupancy are proportional to each other (a straight positive line can be fitted for these quantities) (Fig. 1.2).
As the grain and intensity increases, (i.e. the expected number of points becomes larger on each cell) the slope for the relationship between occupancy and abundance decreases up to the point where all sites become occupied (Fig. 1.3). At this stage, the slope for the association becomes close to zero (Fig. 1.2) and the occupancy becomes a non informative measure for either the underlying point process or the abundances.

Thus, point patterns, abundance and occupancy have a unidirectional exchangeable information structure (Kéry and Royle, 2015), i.e. occupancy can be inferred from abundance which in turn can be defined by the point pattern, but the same information exchange is not always possible the other way around due to information loss. An in-depth discussion of this point within the context of the models proposed in this research is provided in chapter 6.

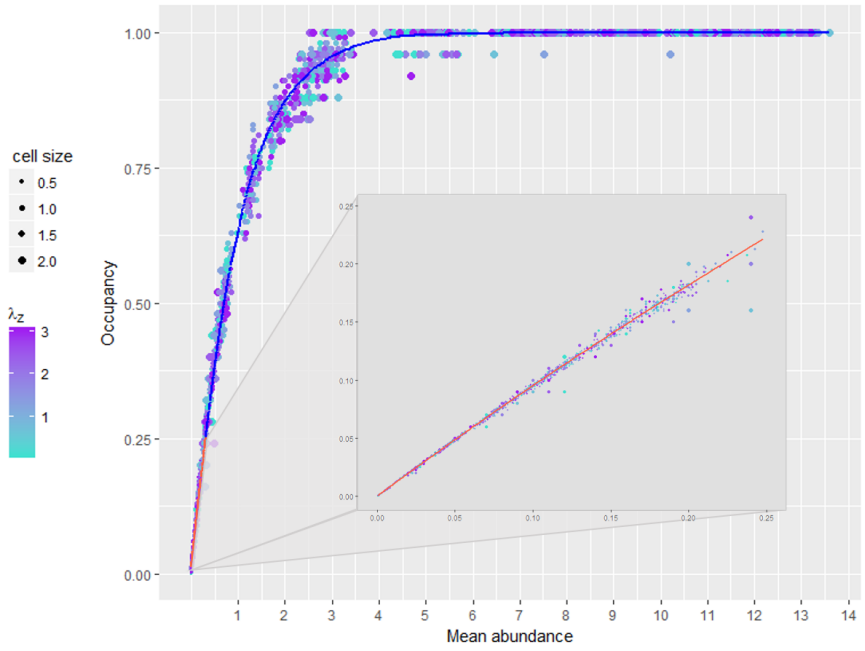


Figure 1.2: Effect of intensity and cell size on the occupancy and abundance of an homogeneous Poisson point process. The line is a smoothing spline (d.f = 5) (red line is the smoothing where occupancy values are < 0.25)

.

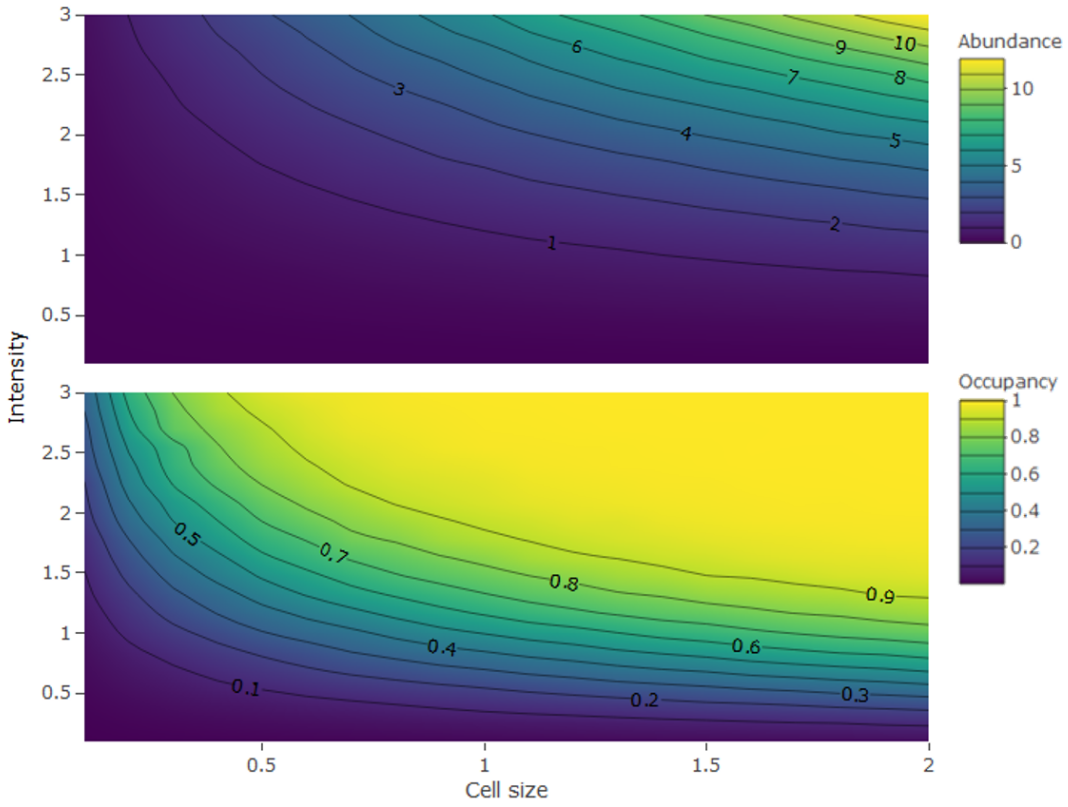


Figure 1.3: Relationship between occupancy and mean abundance (N) with varying intensity (λ_z) values and different cell sizes.

1.3.2.2 A spatio-temporal model

Very often, ecological data are collected over time to infer species distributions' spatial and temporal patterns where each individual of a species is usually represented as a point in space in a given time. As illustrated in section 1.3.2, an underlying model that can be used to describe different types of data (presence-only data, abundance-counts data and presence-absence data) is the spatial point process (Hefley and Hooten, 2016). Thus, a general framework that describes the patterns created by points in time and space is the spatio-temporal point process. In fact, an extension of the homogeneous Poisson point process that can be used to describe the location and time an individual was observed is the space-time Poisson point process given by the log-likelihood function in equation (1.5) (Cressie and Wikle, 2015; Vere-Jones, 1988).

$$l(\lambda, \mathbf{U}) = \sum_{j}^{n} \log\lambda(\mathbf{s}_j, t_j) - \int_D \int_0^T \lambda(\mathbf{s}, t) dtds - \log(n!), \tag{1.5}$$

where the matrix $\mathbf{U}$ denotes the two dimensional spatio-temporal point process up to time t and the intensity function determines the rate at which points occur over the spatial domain D at the time interval $[0, T]$. Moreover, let $A \subset D$ be an area of of interest within the spatial domain at a time interval $[t_1, t_2]$ such that $0 \leq t_1 < t_2 \leq T$. Then, the expected number of points (u) is given by:

$$u \sim \text{Poisson}(\bar{\lambda}),$$

Where $\bar{\lambda}$ is the integrated intensity function derived from Eqn. (1.5):

$$\bar{\lambda} = \int_A \int_{t_1}^{t_2} \lambda(\mathbf{s}, t) dtds. \tag{1.6}$$

Therefore, discretization of a point processes over the time interval $[0, T]$ (e.g by recording the number of species observed every 15 minutes within an hour) also enables information exchange between occupancy and abundance quantities (Hefley and Hooten, 2016) as previously illustrated in Fig. (1.1).

One of the main statistical challenges is to determine how the point pattern, abundance and occupancy of a population changes under the effect of different covariates such as habitat or landscape features. Exploring the influence that factors and their interaction have on the occurrence and distribution are the target of many ecological studies.

To incorporate the effect that different covariates may have on the ecological responses, the intensity function can be defined as:

$$\log(\lambda(\mathbf{s},t)) = X(\mathbf{s},t)'\beta + \varepsilon. \tag{1.7}$$

The term ε represent the random error associated with the ecological process of interest, $X(\mathbf{s},t)$ are the set of time-varying covariates at each location $\mathbf{s}$ associated with a set of unknown parameters β. However, in most of the ecological studies, the estimation process is also affected by an observational error related to how data were collected (e.g. while spatial or temporal processes that occur at smaller scales than the study grain inflate the ecological process error, measuring error incorporate uncertainty into the observation process) (Cressie et al., 2009). Thus, in order for statistical inferences made about the ecological process of interest to be valid, a more flexible class of models that incorporate different sources of uncertainty is needed. Formulating space-time Poisson point process model within a hierarchical modeling framework, enables components that affect both the quantity of interest (i.e. species presence, occurrence or abundance derived from the underlying space-time intensity function) and the observational process to be incorporated (Hefley and Hooten, 2016). The next section provides an introduction to Hierarchical models and how these different components can be incorporated into a spatial-temporal modelling framework .

1.3.3 Hierarchical models in Ecology

Hierarchical models (HM) have proved to be a powerful tool for analysing complex scenarios with drivers that act on species distributions and abundances at different spatial and temporal scales by simultaneously incorporating different sources of uncertainty within each level or hierarchy (Wikle, 2003). For example, species abundances or distributions may be influenced at a first level by the species traits, environmental factors or topographic features but also, at a second level, there may be a temporal or spatial variation involved, each level has its own source of error due to processes that occur in two different scales.
The drivers that are directly involved with the mechanisms that produce the observations define the observation model. The variables that define the variation in the ecological processes of interest (e.g. temporal or spatial variation) constitute the process or state model (Royle and Dorazio, 2008). A hierarchical structure is presented in Figure 1.4 where the first level unit is portrayed by the individual observations (organisms) that are nested in the second unit (4 different sites) with n different replications or surveys.

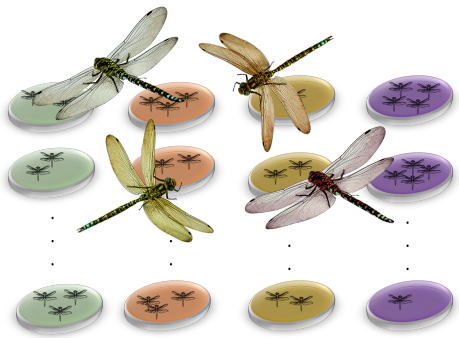


Figure 1.4: Graphical representation of hierarchical space-state structured data: number of dragonflies observed in 4 sites represented by different colours. Each site is visited n times. The bigger dragonflies represent the mean abundance for each site over n time points.

Thus, the observations $\mathbf{y} = \mathbf{y}_1, \ldots, \mathbf{y}_J$ can be written for any J number of cells or sites as follows:

$$\begin{aligned} y_1 &= (y_{11}, \ldots, y_{n_1 1}) \\ y_2 &= (y_{12}, \ldots, y_{n_2 2}) \\ &\vdots \\ y_j &= (y_{1j}, \ldots, y_{n_j j}) \\ &\vdots \\ y_J &= (y_{1J}, \ldots, y_{n_J J}), \end{aligned} \tag{1.8}$$

where, $i = 1, \ldots, n_j$ identifies the replicates (or surveys) of each site $j = 1, \ldots, J$. The parameters of interest are $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_J\}$ the site index allow for all the surveys of the same site to share the same parameter.

The full hierarchical structure of the data can be specified by assuming that the parameters $\theta_1, \ldots, \theta_J$ arise from a common distribution $p(\theta_j|\boldsymbol{\Psi})$ characterized by the hyperparameters $\boldsymbol{\Psi} = \Psi_1, \ldots, \Psi_K$. The structure of this hierarchy is shown in Figure 1.5.

The model shown on Fig. 1.5 can be assessed at two different levels, one at the different θ_j's and the other described by $\boldsymbol{\Psi}$. By allowing θ_j's to come from the same distribution, the exchange of information between θ_j and θ_r (for $r \neq j$) is possible (it would not be if the model was just described by a single θ or independent θ_j's).

All the θ_j's share the same distribution characterized by the hyperparameters $\boldsymbol{\Psi}$, which are at the same time characterized by a prior distribution. Hence, the elements of $\boldsymbol{\theta}$ are conditionally independent given $\boldsymbol{\Psi}$, and the marginal prior distributions of $\boldsymbol{\theta}$ can be decomposed as the product of their conditionals given the hyperparameters $\boldsymbol{\Psi}$ and their prior $p(\boldsymbol{\Psi})$:

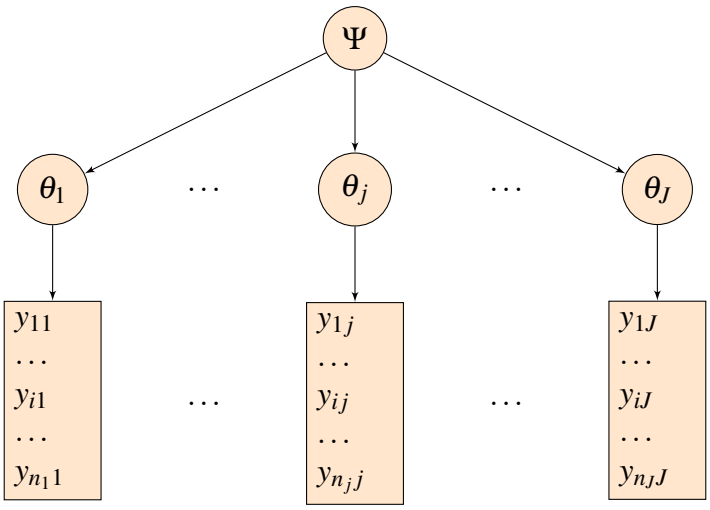


Figure 1.5: Acyclic diagram for a two level hierarchical structure characterized by hyperparameters $\boldsymbol{\Psi}$.

$$p(\theta) = \int p(\theta|\Psi)p(\Psi)d\Psi. \tag{1.9}$$

Now extending it to L different hierarchy levels:

$$p(\theta) = \int p(\theta|\Psi_1)p(\Psi_1|\Psi_2)\dots p(\Psi_{L-1}|\Psi_L)p(\Psi_L)d\Psi_1\dots d\Psi_L$$

Potentially, many hierarchies could be included. However, the more levels are included, the more complex becomes the interpretation of the parameters (Blangiardo et al., 2013).

In terms of a spatial-temporal Poisson point process, the hierarchical species distribution model hierachy following Cressie and Wikle (2015) and Hefley and Hooten (2016) notation is given by

$$g(\mathbf{Y}) \sim p(g(\mathbf{Y})|\mathbf{U},\theta) \tag{1.10}$$

$$\mathbf{U} \sim p(\mathbf{U}|\lambda(\mathbf{s},t)) \tag{1.11}$$

$$\lambda(\mathbf{s},t) \sim p(\lambda(\mathbf{s},t)|\psi), \tag{1.12}$$

where Eqn. (1.10) corresponds to the observational model for which $\mathbf{Y}$ observations are aggregated by a function $g(\cdot)$ (e.g. aggregated into binary or count data) given the parameters θ and the spatial-temporal pattern $\mathbf{U}$. The sampling model (Eqn.1.11) is defined by the true point pattern produced by the space-time Poisson process (Eqn. 1.5) given the intensity function $\lambda(\mathbf{s},t)$. Finally, the process model (Eqn. 1.12) is defined by the distribution of the intensity function given the set of parameters Ψ where different smoothing functions can be used to model the spatial or temporal variation (Hefley et al., 2017). In a lot of situations, the observed point-pattern $\mathbf{Y}$ (or aggregated through $g(\cdot)$ during the data collection process) are observed with error with respect to the true pattern $\mathbf{U}$. For example, species distributions data could be collected from biased sampling schemes that result in some areas not being sampled (e.g. due to accessibility or sampling costs), or because $\mathbf{Y}$ is the result of an imperfect detection of $\mathbf{U}$ (species'

detection can be difficult due to species traits, rarity or because detection varies with environmental conditions) (Hefley and Hooten, 2016).

Spatial-temporal HMs have been used for some time now to address different ecological processes (Holmes et al., 1994; Wikle and Hooten, 2010). However, it is not up until the last decade that their applications became increasingly frequent in the context of species distribution modelling, making researchers more aware about the importance of accounting for the species imperfect detection and sampling bias (Hefley and Hooten, 2016).

Inference and estimates of abundances and other biodiversity metric often overlook or fail to directly incorporate information regarding the processes that differentially affect the detection of species in a survey (Denes et al., 2015). Thus, models that account for imperfect detection have been proposed as a sensible way of quantifying and predicting species distributions (MacKenzie et al., 2002; MacKenzie and Hines, 2017). Particularly, in aquatic studies, modelling species distributions has become a challenging task due to detection of mobile species being problematic (Elith and Leathwick, 2009). Through the next section, a case study for dragonflies incorporated partial occurrences due to imperfect detection across freshwaters in the UK will be explored.

1.4 Case Study: Odonata distribution patterns in the UK

Odonata is a taxonomic order of small invertebrate insects conformed by dragonflies (Anisoptera) and damselflies (Zygoptera) (Corbet and Brooks, 2011). This group is characterized by having an aquatic larval stage, compound eyes and two pairs of wings with muscles attached to the thorax that allow for independent wing movements (Bomphrey et al., 2016). According to Smallshire and Swash (2018) there are over 6000 different species of dragonflies worldwide from which only ≈ 40 species with breeding populations have been identified in Great Britain and Ireland.

Since most of the species have an aquatic larval stage, adults are more likely to be seen near water bodies such as streams and rivers. In the UK the most common habitat these species occupy are lakes, ponds, rivers, streams, canals and ditches. However, some species can also be detected in bogs, flushes and even far away from water bodies (Smallshire and Swash, 2018). Odonates have proved to serve as important bioindicators to asses quality of water bodies and integrity of ecosystems because their early-stage development depends on high quality water. Thus, ecological studies of these species distributions are crucial for waterbody management and restoration plans (Golfieri et al., 2016).

In the UK, the four-year Hydroscape project compiled information of different biological groups' distributions to determine how anthropogenic stressors and connectivity interact to influence biodiversity and ecosystem function in freshwaters across Britain (`https://hydroscapeblog.wordpress.com/about/`). Within the biological groups of interest, odonates occurrence records from 2000-2016 were taken for 41 species. Data were compiled by the Hydroscape multidisciplinary group from the National Biodiversity Network (NBN), the Biological Records Centre (BRC) and from the British Dragonfly Society. Species occurrence records were obtained on a 1 km grid for the UK.

Along with the occurrences data set, connectivity (e.g. Lake area (ha, Lake altitude (m), Mean Depth (m), Volume (m3) , Perimeter (km), Distance to sea (km)) and stressors (% of agricultural and urban development) metrics were calculated at different spatial scales (buffers from 500 m up to 2 km) by the Centre of Ecology and Hydrology, Edinburgh. Also, species-specific level covariates that may affect the species detection probability were taken from the traits data set provided by Powney et al. (2014), which contains 43 different species traits (e.g. mean body size, flight duration, number of habitat each species occupy and family).

1.4.1 Species occurrence data processing

Odonata observed occurrence records are available on over 4000 1 km grid cells (sites) based on OSNG (Ordnance survey national grid) references, each of them defined by one or more individual water bodies. The total number of visits each site had across all years was calculated based on the dates when each species was recorded after removing any duplicated visits (i.e. sightings of two or more species on the same date at the same site were assumed to be recorded on a single visit) (Table 1.1).

Species observed absences were inferred for each of the J sites with no detections of the ith species at the kth visit. The occurrences data base was matched against the traits data set to assure both data sets had the same species i.e. species occurrences for which species level covariates were not available were removed along with species in the trait data set which did not have any occurrences at all.

Table 1.1: Odonata occurrences data structure example and number of visits per site after processing.

Grid cell ID (sites)	taxa	date
HU3272	*E. cyathigerum*	2008-06-01
HU3272	*E. cyathigerum*	2008-06-30
HU3282	*E. cyathigerum*	2001-07-24
HU3977	*E. cyathigerum*	2004-08-02
HU3977	*E. cyathigerum*	2001-07-24
HU4075	*L. quadrimaculata*	2004-08-01
⋮		

remove duplicated records →

Grid cell ID (sites)	visits
HU3272	2
HU3977	2
HY2102	10
HY2103	6
HY2310	2
HY2314	2
⋮	⋮

After matching both data sets, a total of 41 species were found to have both occurrence records and traits information. Two out of the 41 species were identified as invasive species (from discussion with British Dragonflies Society experts), *Chalcolestes viridis* and *Erythromma viridulum* (Fig. 1.6 and 1.7). *C. viridis* occurrences were first recorded in 2007. Then, the number of detections started to rise until 2013 when occurrences decreased for the following 3 years. *C. viridis* was detected once again in 2016 in several low-latitude sites (Fig.1.6). Unfortunately , the occurrence data for this species is very sparse from 2013 to 2015 and could indicate that either (1) the species was not detected due to time-varying sampling effort or (2) the local populations of this species underwent a local extinction process in which previously occupied sites became unoccupied until the further colonization in 2016 (an in-depth discussion of these two concepts will be provided in chapter 4).

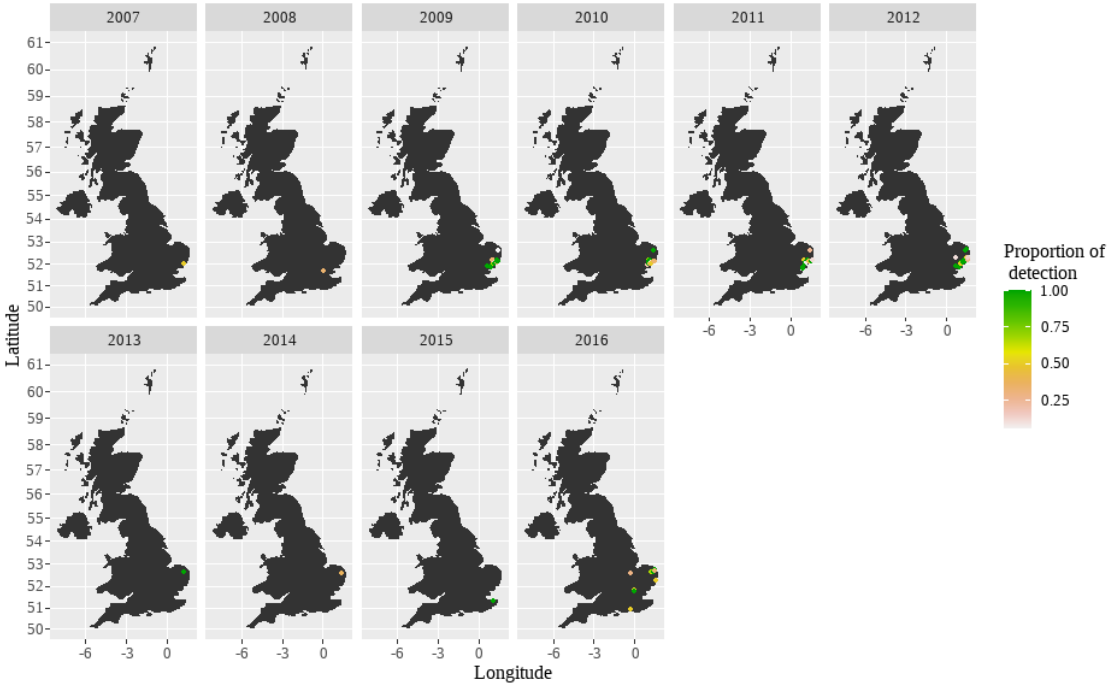


Figure 1.6: Observed proportion of occurrences for *C. viridis* through 2000-2016 at sites which had at least one detection of the species.

Unlike *C. viridis*, *E. viridulum* first occurrences were recorded in 2000 and began to rise in the subsequent years at lower latitudes.

These two species were removed from further analysis in the current chapter, since they provide evidence against the assumption of constant occupancy across visits that will be revisited in the next chapter. For this same reason, the analysis and methods presented in chapters 2 and 3 are also restricted to these 39 non invasive species. Then, the distributions for the complete species list (i.e. including the two-invasive species *E. viridulum* and *C. viridis*) will be revisited in chapters 4 and 5 under a temporal model that accounts for uneven time-varying sampling effort.

The occurrences data base with and without *E. viridulum* and *C. viridis* was matched against the site level covariates data set to ensure all sites in the occurrences data set had associated site level metrics.

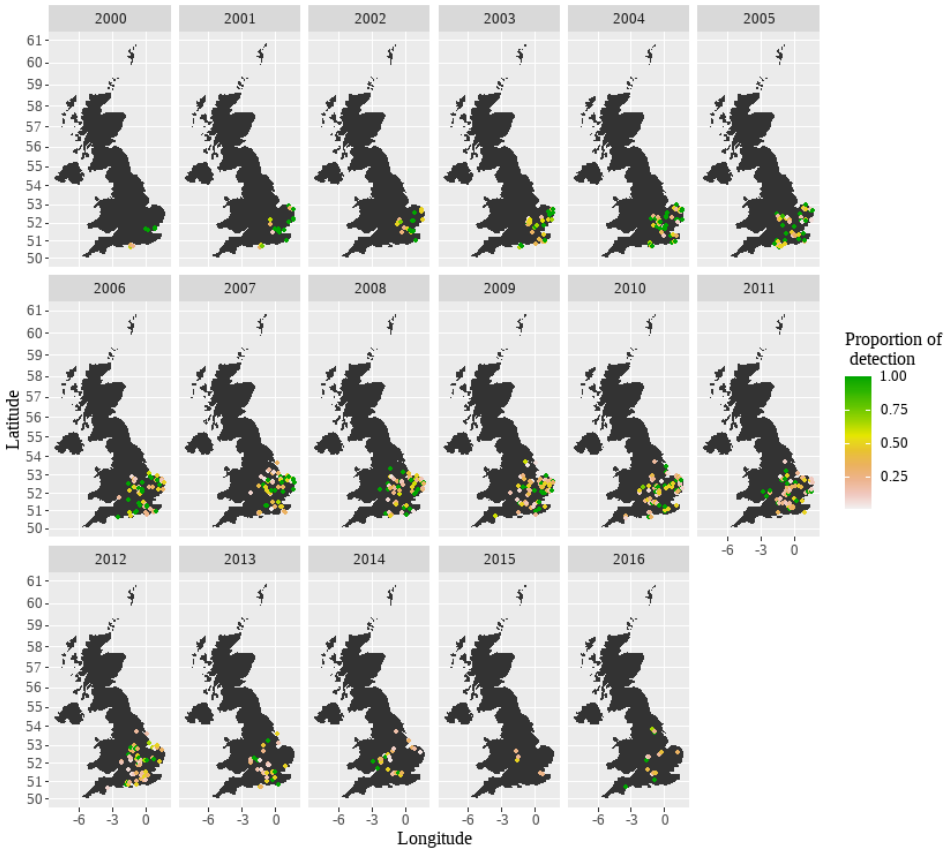


Figure 1.7: Observed proportion of occurrences for *E. viridulum* through 2000-2016 at sites which had at least one detection of the species.

1.4.2 Measures of connectivity and stressors

Stressors and connectivity metrics were calculated by the hydroscape team on 7 geographical scales to investigate the impact that different types of connectivity have on freshwater ecosystem in the UK (Table 1.2).

Table 1.2: Geographical scale buffers that defines the different types of connectivity.

Connectivity type	**Scale**
Hydrological connectivity	Catchment buffer
Landscape connectivity	500 m, 1 km, 1.5 km and 2 km buffer
Riparian/Lateral connectivity	100 m
Lake	water body level

Freshwater ecosystem connections can be established through proximity riparian zones that link the main course of a river with various waterbodies lying in the alluvial floodplain (Riparian/Lateral connectivity) (Amoros and Bornette, 2002). Freshwater can also be connected by the spatial proximity and barriers of a landscape (Landscape connectivity) (Fergus et al., 2017). Moreover, it has been reported that there is a scale-dependent effect of land use on freshwaters due to organisms' differential dispersal abilities (Pedersen et al., 2006; Steffan-Dewenter et al., 2002). Therefore, establishing increasing distances of buffer zones from each waterbody enables investigation of the scale effect of landscape connectivity on different organisms. Finally, hydrological units can be connected by a temporary or permanent flow of water between sites in which organisms material, energy and nutrients flow (hydrological connectivity). The hydroscape project focused on the upstream hydrological connectivity by accounting for metrics such as river length and counts, area or perimeter of lakes or ponds in the upstream catchment which are thought to affect the distribution of Odonata. Figure 1.8 portrays the different connectivity types developed by Hydroscape. Initially, the Hydroscape model foresaw freshwater connectedness through dispersal via the movement of humans or other animal vectors, but these metrics have been discarded as they do not affect Odonata distribution directly.

Anthropogenic stressors defined by the % agricultural and urban land use and connectivity metrics (e.g. perimeter, number of lakes, river length, number of obstacles, area, among others) are available for the different buffers capturing hydrological, landscape and riparian connectivity. A ratio between covariates that depend strongly on the catchment/buffer area was computed for all of the different geographical scales, i.e. rivers and canals length (m), number of obstacles, number of lakes and ponds, lengths of Strahler 1, 2, 3 and 4 and ponds and rivers perimeter (m). Lake level data was matched to the grid level data. Grid cells values were defined by the Hydroscape team based on the largest lake or catchment within the cell, for example, if the jth grid cell contains several lakes, then the values for that cell will be assigned based on the largest lake metrics unless its part from a greater catchment for which the largest catchment values will be used.

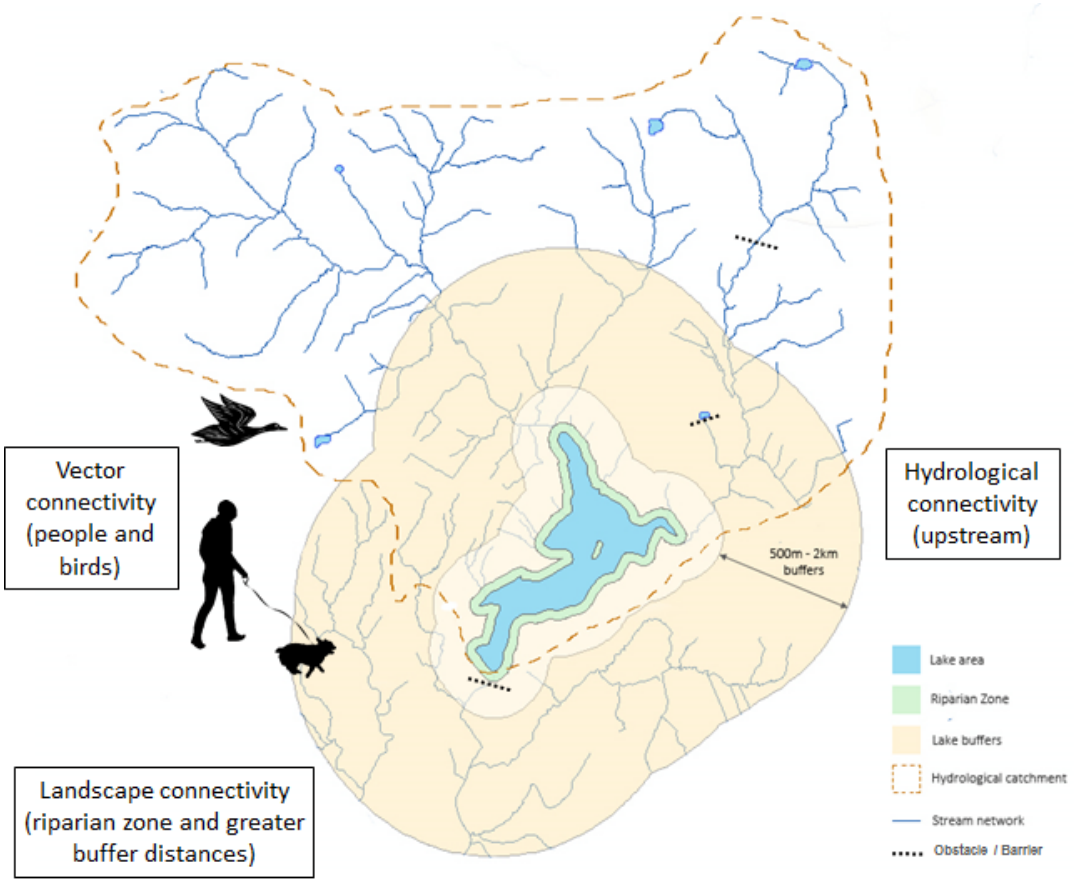


Figure 1.8: Connectivity of freshwaters via catchment hydrology, landscape and dispersal vectors model (image by permission of Hydroscape team).

1.4.3 Exploratory analysis

After the matching, there were 4953 sites coded according to national grid reference in which 39 non-invasive odonata species where detected from 2000-2016. The observed occurrences vary widely between species. While there are some widespread species (e.g. *Sympetrum striolatum*) that occur in more than 3500 sites, there are others with just a few detections (e.g. *Aeshna affinis* was observed in only 5 sites) (Fig.1.9).

The observed species richness on each site is presented on Figure 1.10. The observed local richness ranges between 2-7 species for most of the sites.

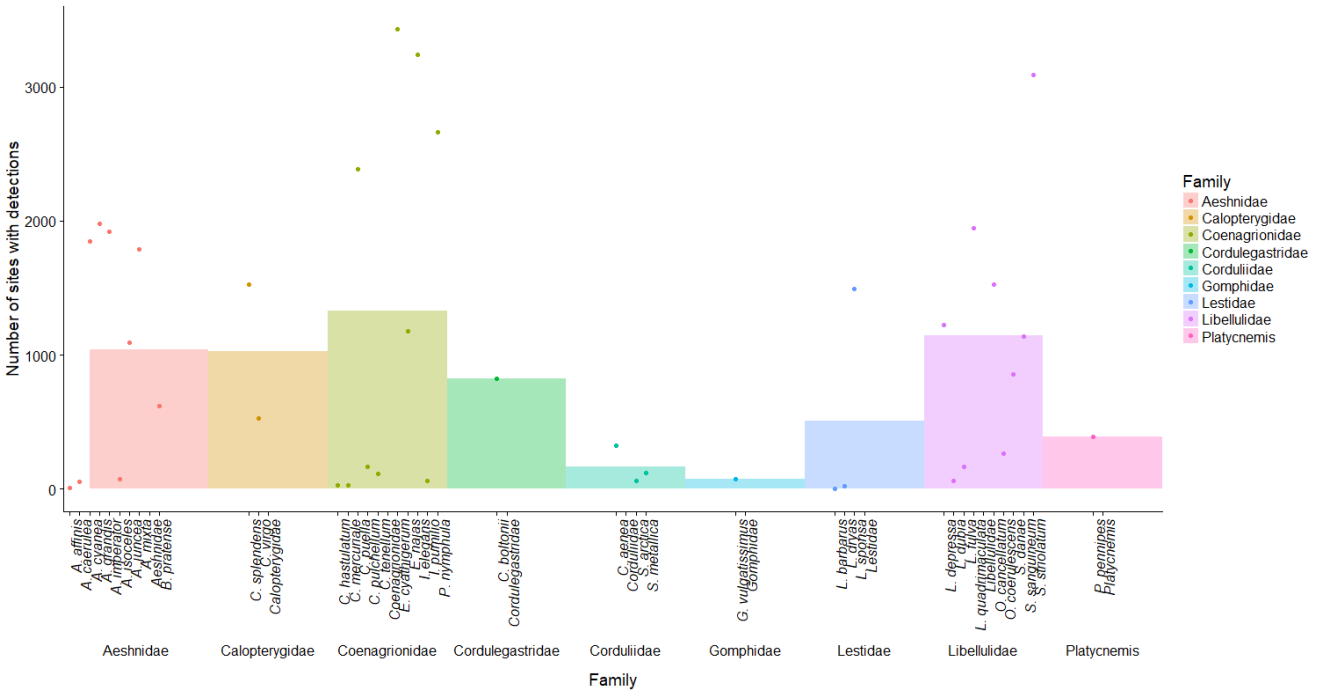


Figure 1.9: Number of sites with detections for each species grouped by family. The graph bars show the mean number of occupied sites for each family.

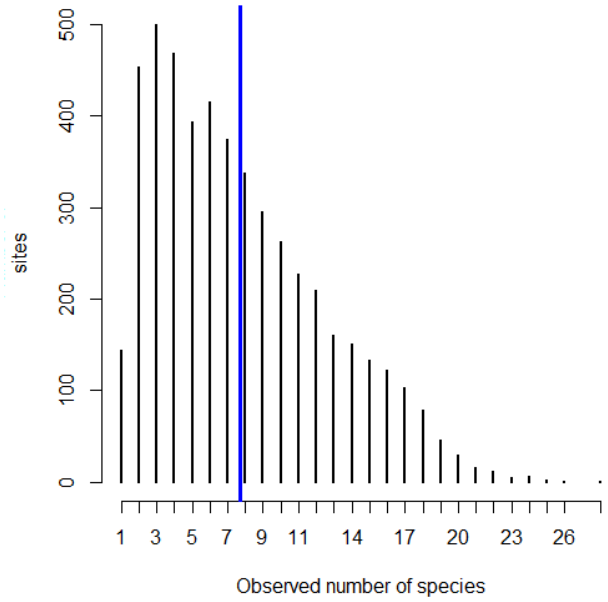


Figure 1.10: Observed number of species per site. The blue line indicates the species richness mean.

In terms of species-specific covariates, the proportion of detections given the number of visits is greater for smaller species, thus smaller species seems more likely to be detected (Fig. 1.11 left). On the other hand, the longer the flight duration, the more likely species detection will be (Fig. 1.11 right).

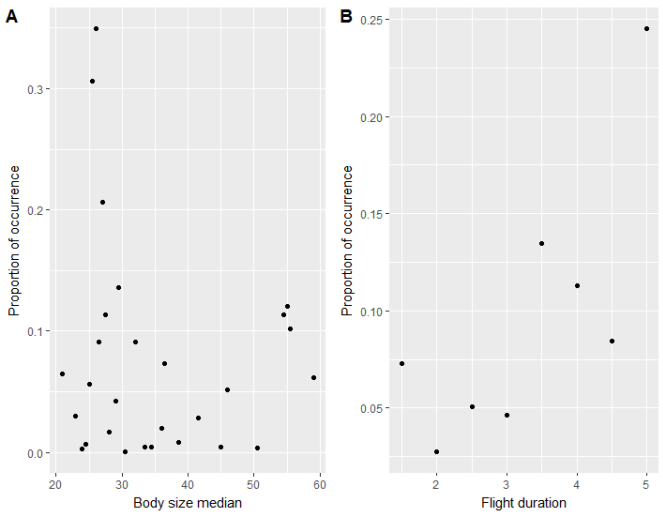


Figure 1.11: Relationship between the proportion of occurrence mean body size (left) and flight duration (right).

To investigate a potential interaction between mean body size and flight duration, species medium body size was discretized into "small", "medium" and "large" and flight duration was categorized into "short", "medium" and "long" flight periods based on 0.25 and 0.75 quantiles. Figure 1.12 shows how small species with longer flight duration are then more likely to occur, whereas large species with short duration have almost no occurrences.

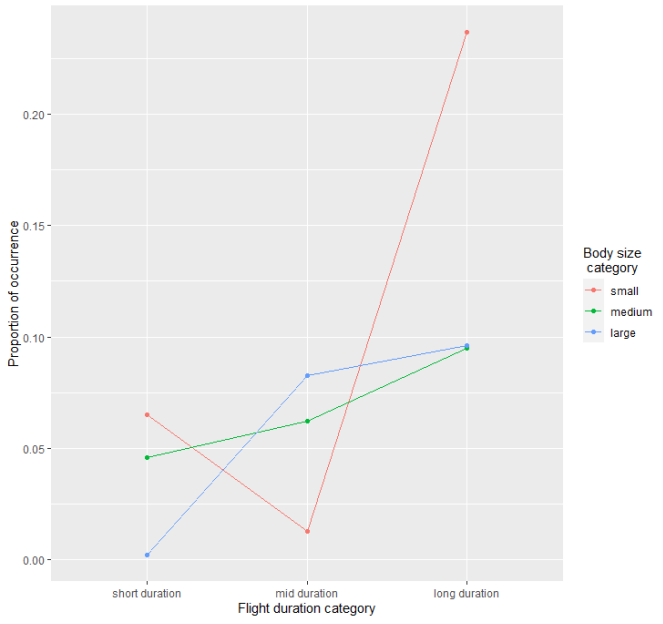


Figure 1.12: Interaction plot for proportion of detections for size and flight duration categories

.

The number of different habitats (distribution proxy) species occupy, may have an important effect on the occurrences (Figure 1.13) .The more wide spread species are, the more likely they are to be detected.

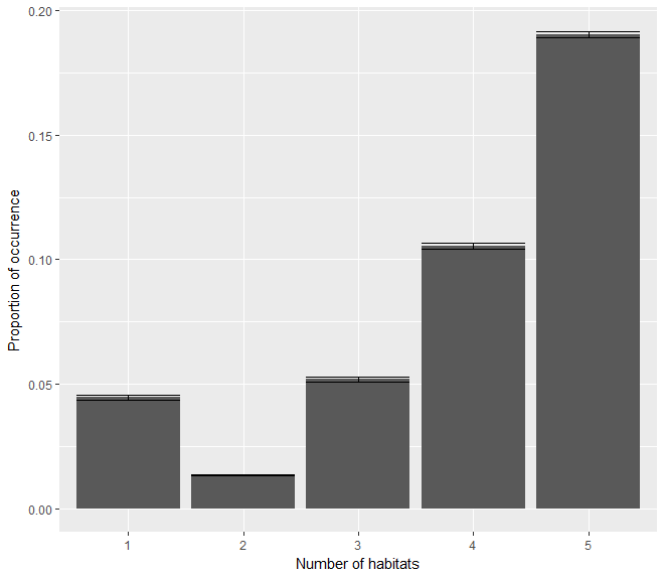


Figure 1.13: Number of different habitat occupied by species.

Smaller species tend to occur in a higher number of habitats (Fig. 1.14 left), this could explain the negative trend on the detectability based on species size. Moreover, the occurrences for long flight duration seems to be higher regardless of the number of different habitats occupied by species.

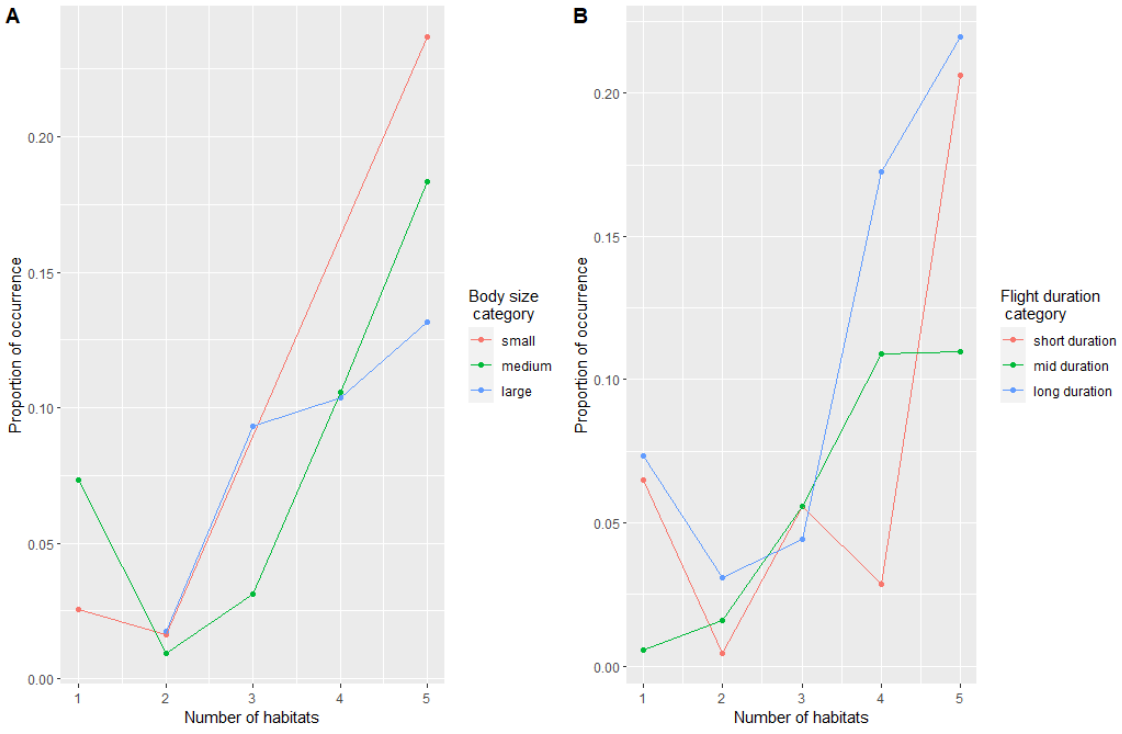


Figure 1.14: Interaction plot for proportion of detections for number of habitat with size (left) and flight duration (right) categories respectively.

Each species has been given an empirical distribution status by Powney et al. (2014) for how common or rare they are. Fig. 1.15 shows the comparison of the empirical species distribution status against the species sizes. Very wide spread species have a smaller body size. Thus, smaller species are more widespread than larger ones and therefore, more likely to be detected (Fig. 1.15).

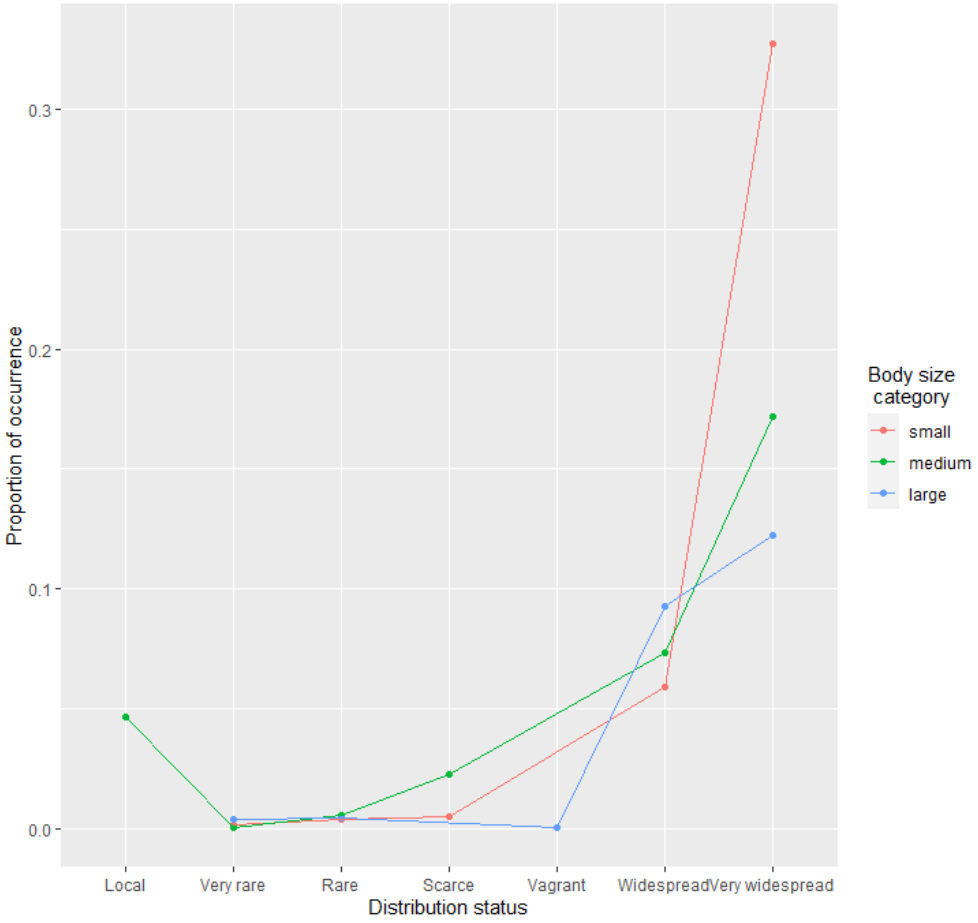


Figure 1.15: Interaction plot between the empirical species distribution and body size category.

1.5 Summary

Freshwater connectivity allows for the exchange of species, material and nutrients between hydrological units thereby affecting species diversity. These connections can be established at different spatial and temporal scales depending on species-specific dispersal abilities and they can also be altered by anthropogenic stress, affecting thus, ecosystems integrity. How these mechanisms operate to modify species distributions patterns can be complex and represent different theoretical and statistical challenges.

When modelling species distributions, the sampling bias or observational process errors mask the effect of the ecological process of interest, with imperfect detection being one of the main sources of uncertainty. To address this, hierarchical models (HMs) have proven to be a powerful tool to incorporate different sources of uncertainty. Within HMs a special class of species distribution models (SDMs) that

analyze species distributions by taking into account detectability bias will be explored in the following chapter. Different occurrence-based models and fitting strategies will be compared in chapter 2 to develop a suitable approach to be applied within the Odonata case study to evaluate species distribution patterns in waterbodies across the UK while accounting for detection bias. The performance and computational efficiency of current approaches will be compared with computer simulations. This information will be used in the following chapters to develop novel statistical methods to describe the relationship between species distribution and environmental drivers.

Specifically, chapter 3 will explore the connectivity metrics discussed in this chapter to determine how multiple stressors interact with measures of ecosystem connectivity to affect Odonata biodiversity. Specifically, rare species as they represent the most vulnerable species to environmental changes. The aim is to determine how Odonata rare species distributions are affected by differing measures of landscape connectivity and to identify potential environmental stressors that limit their distribution. To do so, methods will be developed to assess species rarity in Odonata communities across the UK to identify relevant areas for biodiversity conservation and then evaluate potential drivers that may affect the composition of rare species at a site level. The central hypothesis is that the connectivity effect will depend on the degree of anthropogenic stress, shifting from having a positive effect on diversity in areas with low degree of stress and negative in high stress environments.

In chapter 4, a flexible model that accommodates temporal population dynamics will be developed to estimate occupancy temporal changes through space while incorporating environmental site-level non-linear effects. A simulation study will be used to assess the model performance and to gain information about modelling considerations and limitations. The proposed flexible dynamic model will be applied to the Odonata case study in chapter 5 to describe the Odonata species temporal distribution patterns in UK freshwaters, including the invasive species discussed in this chapter. These models will investigate different terms that are expected to occur in real biological populations and communities such as population colonization/extinction dynamics changes over time while exploring different modelling structures that incorporate uneven sampling effort, a common issues found species distributions data from citizen science projects. Additionally, different biodiversity metrics (e.g. Species richness) derived from these methods will be investigated to detect Odonata species occupancy trends over time which could help to evaluate potential risks or threats experienced by some members of the whole biological community.

Finally, chapter 6 will provide a detailed assessment of the contribution of this research and raise caveats involving future research while addressing the statistical and ecological challenges discussed throughout this work. A thorough discussion of the topics discussed in this chapter along with several methodological considerations of the proposed methods in this research will be provided. This encompasses caveats and practices of the proposed methods that, in conjunction with current methodological innovations and approaches that are being developed, represent a promising area of future research for ecological studies where imperfect detection is an issue.

Chapter 2

Modelling species detection uncertainty

Taking into account the uncertainty in ecological studies is not an easy task because of the different sources of error that occur at different spatial and temporal scales (Cressie et al., 2009). Thus, several statistical methods have been developed for quantifying and predicting the effect that different environmental factors will have on species distributions through time and space (Guisan and Thuiller, 2005; Schroeder, 2008; Elith and Leathwick, 2009). However, many data collection processes are prone to errors associated with the observational process of detecting the species. This is the case for the dragonflies and damselflies communities reviewed in the last chapter. The data available for this group are partially observed occurrence records due to imperfect detection that makes the task of identifying the ecological processes that drive the true species occupancy patterns difficult.

Over the last decade an important effort has been made to model species occurrence under imperfect detection (MacKenzie et al., 2017). For instance, in a meta-analysis by Kellner and Swihart (2014) of over 500 ecological articles in 5 different decades (from 1970 to 2011) of different taxonomic groups, just 23% of the studies incorporated or discussed species imperfect detection. Within this small amount of articles that accounted for detectability, the yearly mean percent of studies accounting for imperfect detection showed an increasing trend over time (up to a 35% increase between 2001 and 2011) but vary widely between taxonomic groups, with plants and invertebrates being the groups for which imperfect detection is less likely to be addressed. More recently, Devarajan et al. (2020) also reported that while the number of studies addressing imperfect detection for multiple species has been increasing over the last decade, studies involving species imperfect detection of small invertebrates are still under-represented compared to other taxa such as birds, fish, and mammals. And thus, this chapter will investigate a special class of models to estimate species distributions while addressing detection bias. These methods will be analyzed through computer simulations and will then be applied to the Odonata case study, a taxonomic group that has been systematically overlooked in studies involving detection bias (Termaat et al., 2019).

2.1 Detection bias

In ecological studies, the distribution of a species is defined by the presence or absence of a species in a particular site. Whilst presence can be confirmed by the detection of a species, non-detections may be the result of either the species being undetected during the surveying or being truly absent at a site. Thus, detectability refers to the fact that a species may be present at a site but undetected. These "false absences" occur frequently for those rare species which are often a point of interest for conservationists and ecologists (MacKenzie et al., 2017).

Detectability can be formulated as follows: Let $N \sim (\lambda_Z|A|)$ be the total number of individuals (i.e. the abundance) of a particular species that are detected independently in a sample with a probability of θ. Then, the number of individuals that are actually observed is given by

$$n|N \sim \text{Binomial}(N, \theta).$$

Thus, the expected number of observed individuals is $\mathbb{E}(n|N) = \mathbb{E}(N) \times \theta$ such that $\mathbb{E}(N) = \lambda|A|$ is the *first-order intensity* function of the homogeneous Poisson point process discussed in the previous chapter. When $\theta < 1$ then bias in sample is introduced, i.e. $\mathbb{E}(n) < N$.

If several surveys are carried out and detection is independent and constant between surveys, then the probability of an individual appearing in the sample is a function of each individual sampling probability of detection p. For example, if sampling is conducted k times, the probability of an individual being detected at least once during a survey is:

$$\theta = 1 - (1-p)^k.$$

The more surveys or sampling occasions, the smaller the detection bias will be. Spatial sampling units also have an important role in detectability, thereby affecting inference about occurrence, abundance or any other biodiversity measures. Denote $n(\mathbf{s})$ to be the number of individuals that occupy a location $\mathbf{s}$. Then, the total variance can be written as the sum of two components:

$$\begin{aligned}
\text{Var}[n(\mathbf{s})|N(\mathbf{s})] &= \mathbb{E}\{\text{Var}[n(\mathbf{s})|N(\mathbf{s})]\} + \text{Var}\{\mathbb{E}[n(\mathbf{s})|N(\mathbf{s})]\} \\
&= \mathbb{E}\{\theta(1-\theta)N(\mathbf{s})\} + \text{Var}\{N(\mathbf{s})\theta\} \\
&= \theta(1-\theta)\mathbb{E}\{N(\mathbf{s})\} + \theta^2\text{Var}\{N(\mathbf{s})\}. \qquad (2.1)
\end{aligned}$$

The first component ($\mathbb{E}\{\text{Var}[n(\mathbf{s})|N(\mathbf{s})]\}$) is the variance associated with the observational process. It is, a spatial average of the observational variance where $N(\mathbf{s})$ is an inhomogeneous Poisson process describing the total number of individuals occurring at each location. The second component ($\text{Var}\{N(\mathbf{s})\}$) is the process variance, i.e. the variance across replicated spatial sample units. As detection probability increases, i.e. $\theta \to 1$, then the proportion of variance explained by the spatial variance becomes larger and the observational variance decreases.

When θ is known, occupancy probabilities can be estimated using maximum likelihood estimators (MLEs) (MacKenzie et al., 2017). Assuming a random sample of M different sites, each of which have the same probability ψ of becoming occupied by a given species, then the number of sites that are occupied by the species is given by

$$x \sim \text{Binomial}(M, \psi),$$

such that $\mathbb{E}(x) = M\psi$ and $\text{Var}(x) = M\psi(1-\psi)$. Thus, the occupancy probability estimator is $\hat{\psi} = x/M$. However, if the species detection probability is less than 1 across different k sampling occasions the expected number of sites at which a species is detected is given by

$$x_d \sim \text{Binomial}(M, \psi \times \theta),$$

such that $\mathbb{E}(x_d) = M\psi\theta$ with variance $\text{Var}(x_d) = M\psi\theta(1-\psi\theta)$. Where the probability of the species being present and detected is given by $\psi\theta$. Thus, if the detection probability is known, MLE for the proportions are

$$\hat{\psi} = \frac{x_d}{M\theta}, \tag{2.2}$$

with variance given by:

$$\begin{aligned} \text{Var}(\hat{\psi}) &= \frac{\psi(1-\psi\theta)}{M\theta} = \frac{\psi - \psi^2\theta}{M\theta} = \frac{\psi\theta - \psi^2\theta + \psi - \psi\theta}{M\theta} \\ &= \frac{\psi\theta(1-\psi)}{M\theta} + \frac{\psi(1-\theta)}{M\theta} = \frac{\psi(1-\psi)}{M} + \frac{\psi(1-\theta)}{M\theta}. \end{aligned} \tag{2.3}$$

Note that the variance again consists of two components. The first one corresponding to the uncertainty around the true value ψ and the second one which arises from having imperfect detection and for estimating the number of sites that were occupied in the sample. In summary, detection bias introduces a variance component that is essentially a source of error that masks the ecological process of interest. Moreover, both occupancy and detection probabilities are very unlikely to be known prior to conducting the study. Thus, an additional source of error is induced due to estimating these parameters jointly from the collected data. These sources of errors (process, observation and estimation variances) fit in quite naturally within the structure of hierarchical models, which provides a useful model-based framework to accommodate these different variance components. A special class of HMs, named occupancy models, will be introduced in the next section to address this.

2.2 The Occupancy model for species distributions

Hierarchical models are widely used in ecology for modelling species distribution or abundance. Within hierarchical models, occupancy models predict species occurrence or distribution by taking into account detectability bias (MacKenzie et al., 2002). Suppose the presence or absence of a species is registered at M different sites which are surveyed on K occasions. Let y_j be the outcome of the species being observed ($y_j = 1$) or not ($y_j = 0$) in a particular site j. The occupancy models deal with the fact that a species can go undetected even if it is present on the site. Thus, imperfect detection should be taken into account in order for any inferences made about the ecological process of interest to be valid. Species abundance or distribution patterns are commonly the main ecological process of interest (state process). However, the mechanisms that operate on the observation process (i.e. variables that affect directly the species detectability) represent an additional source of variation that masks the effects of the state variables (Denes et al., 2015).

Detectability bias is defined by two type of errors, the *false absence errors* which relates to the probability of not detecting a species given it is present, and the *false positive errors* which is the probability of detecting the species given it is not present (Royle and Dorazio, 2008). The latter, can be a result of double counting or misclassified individuals in a survey (commonly seen in capture-recapture designs). However, this is a less studied subject and very few researchers have tackled this problem (Royle, 2006). Therefore, in order to account for imperfect detection (assuming no false-positive errors), occupancy models specify the true presence/absence state of a species with the latent variable $z_j \in [0,1]$ for each site. This model is often referred to as a two state model because of the two possible outcomes the occupancy random variable may have.

The observation process is conditioned on the state variable in the observation model:

$$y_j | z_j \sim \text{Bernoulli}(z_j p), \tag{2.4}$$

where $p = \Pr(y_j = 1 | z_j = 1)$ is the detection probability given the occupancy status of the species being present ($z_j = 1$). Probability of detection is constant across sites unless covariates $g = \{g_{1j}, g_{2j}, \ldots\}$ that may influence the detection probability are specified on the logit scale, i.e. $\text{logit}(p_j) = \alpha_0 + \alpha_1 g_{1j} + \alpha_2 g_{2j} \ldots$.

Then, the state model is defined as follows:

$$z_j \sim \text{Bernoulli}(\psi). \tag{2.5}$$

Where $\psi = Pr(z_j = 1)$ is the occupancy probability. Similarly to the observational model, different site-level covariates that may affect the occupancy probability can be specified as $\text{logit}(\psi_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots$.

Assuming no false-positive errors implies that $\Pr(y_j = 1|z_j = 0) = 0$, i.e. if a species is absent on a site, then it can no longer be detected. Thus, marginalization of the observational process yields:

$$\begin{aligned}\Pr(y_j) &= \sum_{z_j} \Pr(y_j|z_j)\Pr(z_j) \\ &= \Pr(y_j = 1|z_j = 1)Pr(z_j = 1) + \Pr(y_j = 1|z_j = 0)Pr(z_j = 0) \\ &= p\psi. \end{aligned} \tag{2.6}$$

Hence, parameters for the state and observational model parameters (p, ψ) are indistinguishable unless different observations or replicates for each site j are made, resulting in $\mathbf{y}_j = y_{j1}, \ldots, y_{jK}$ observations for K sampling occasions or visits.

Assuming the occupancy status does not change over the K replicate surveys, the observational model can then be seen as an aggregated process over K visits (Eqn. 2.7). Thus, the individuals have a probability p of being detected given the occupancy status.

$$y_j \begin{cases} \sim \text{Binom}(K, z_j p), & \text{if } z_j = 1 \\ = 0, & \text{if } z_j = 0. \end{cases} \tag{2.7}$$

The occupancy model assumes that, (1) in the absence of site-level covariates the occupancy probability is constant across sites (i.e. the underlying spatial point process has a constant intensity function across space), (2) the detection probability given the species is present is the same for all sites, (3) the probability of detection between surveys is independent, i.e. $p_k \perp\!\!\!\perp p_l \quad \forall \quad k \neq l$, (4) the detection histories (i.e. the detection record across the K surveys for the site j) are independent at each site and (5) there are no false-positive errors. Occupancy models can be seen as a two stage logistic regression that takes into account detection bias (MacKenzie et al., 2017). Figure 2.1 shows the fit of a two state occupancy model and an ordinal logistic regression that ignores detectability bias, to a simulated data set with different detection and occupancy probabilities at 200 different sites with three replicates each and the effect of a covariate X on the state process [2]. Note that the fitted logistic regression that ignores the detectability bias underestimates the true occupancy relationship when the detection probability is low.

Occupancy models can be analysed either by classical or Bayesian methods. Bayesian inference has become a popular choice for analysing hierarchical models due to the conditional probability structures that allow MCMC methods to retain the latent variables (Royle and Dorazio, 2008; Wikle, 2003; Kéry and Royle, 2015). In contrast, classical methods rely on modelling the marginal likelihood for each latent variable. Different ways for implementing likelihood-approach inference have become available through different platforms (Denes et al., 2015; MacKenzie and Hines, 2017; Fiske and Chandler, 2011).

[2]Further simulation App available at `https://shiny.maths-stats.gla.ac.uk/2259971b/LogisticvsOccupancy`

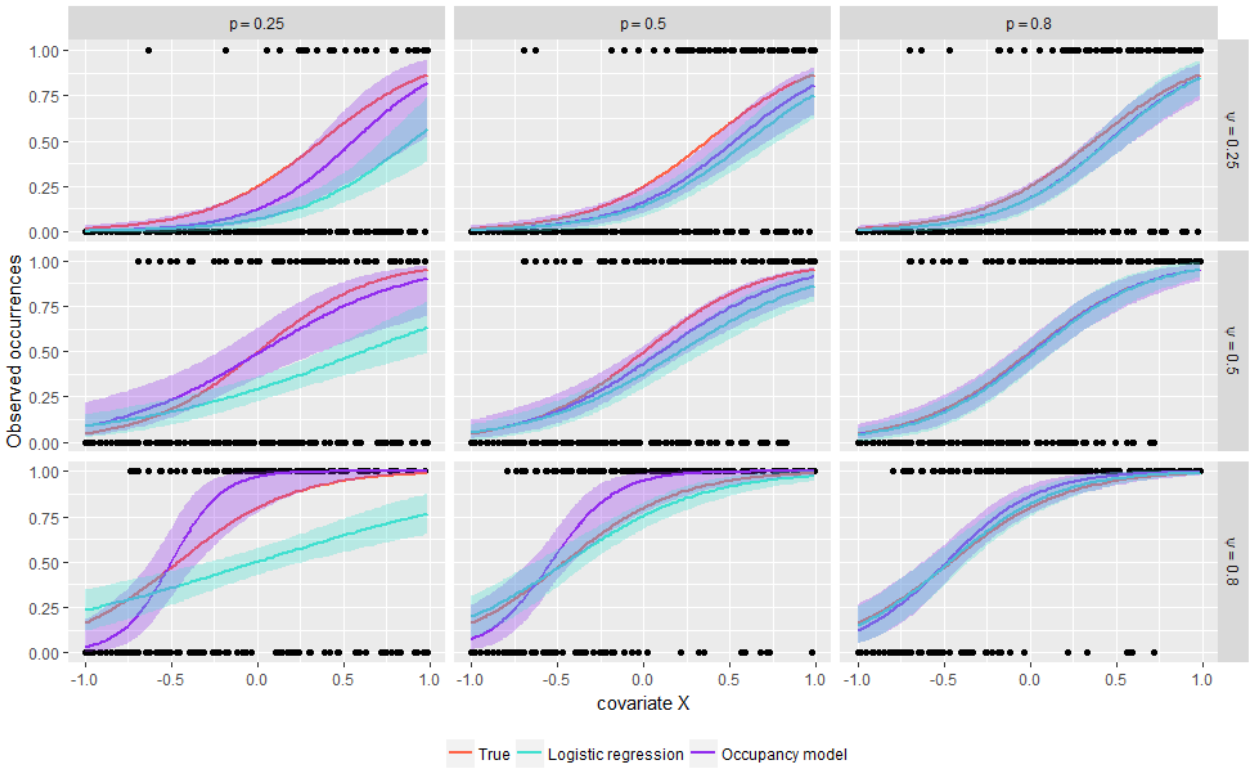


Figure 2.1: Comparison between the true occupancy and covariate relationship (red) and the estimates from the occupancy model (purple) and a simple logistic regression (turquoise); circles represent the observed occurrences of each site across all surveys (i.e. the maximum number of occurrences ever detected).

2.2.1 Classical inference approach to occupancy models

Analysing species occurrences from a classical point of view is an optimization procedure based on the likelihood of the detection histories $\mathbf{y}_j = y_{j1}, \ldots, y_{jK}$ in a set of M discrete sites in space (MacKenzie et al., 2002):

$$\begin{aligned} L(\psi, p|\mathbf{y}_1, \ldots, \mathbf{y}_M) &= \left[\psi^J \prod_k^K p_k^{J_k}(1-p_k)^{J-J_k}\right]\left[\psi \prod_k^K (1-p_k)(1-\psi)\right]^{M-J} \quad \text{assuming } p_k = p \;\forall\, k \\ &= \left[\psi^J p^{\Sigma_k^K J_k}(1-p)^{KJ-\Sigma_k^K J_k}\right]\left[\psi(1-p)^K + (1-\psi)\right]^{M-J}. \end{aligned} \tag{2.8}$$

In the first part of the likelihood (2.8), J are the number of sites where the species was detected on at least one sampling occasion and J_k is the number of sites where the species was detected during visit k. Thus, and assuming a constant detection probability across visits, the target species must be present at those J sites with a probability ψ and either being detected in J_k sites with a probability p or not detected in $J - J_k$ sites with a probability $(1-p)$. The second part of the likelihood is formulated for those $M - J$ sites where species is not detected during any of the K visits (note that the number of visits can be set to vary between sites, i.e. K_j).

Under the assumption of constant detection probabilities, the likelihood in equation (2.8) describes the number of detections at each site as a zero-inflated binomial random variable (Hall, 2000). Thus, classical inference for both occupancy and detection probabilities is based on the marginalization of the latent variable z_j. For the two-state occupancy model, because y is a discrete random variable, the marginal probability is given by averaging over the conditional on $z_j \in [0,1]$:

$$\Pr(y) = \text{Binomial}(y|K, z=1)\Pr(z=1) + \text{Binomial}(y|K, z=0)\Pr(z=0), \tag{2.9}$$

where $\Pr(z=1) = \psi$ denotes the probability of a site being occupied and $\Pr(z=0) = (1-\psi)$ the probability that a site in unoccupied and therefore, the species is absent.

Assuming no false positive errors the expression $\Pr(y|K, z=0)\Pr(z=0)$ from Eqn. (2.9) can only be evaluated at the mass point $y=0$ (i.e. the probability of not observing a species given it is not present). Thus, the likelihood becomes the product shown in equation (2.10).

$$L(\psi, p|y_1, \ldots, y_M) = \prod_j^M \text{Binomial}(y|K, p)\psi + I(y=0)(1-\psi), \tag{2.10}$$

which is a zero-inflated binomial distribution where the indicator variables takes a value of 1 if $y_j = 0$ and zero otherwise. When survey-level covariates are present, detection probabilities may vary by visit, making the individual detection history defined as:

$$\Pr(\mathbf{y}_j|z_j=1) = \prod_k p_{jk}^{y_{jk}}(1-p_{jk})^{1-y_{jk}}. \tag{2.11}$$

The effect on any survey-level covariate g_{jk} can be specified on the logit scale. Hence, the probability of detecting the species at site j during visit k can be expressed as:

$$\text{logit}(p_{jk}) = \alpha_0 + \alpha_1 g_{jk1} + \ldots. \tag{2.12}$$

Then, to account for non-detection record histories a zero-inflated multinomial can be specified, following Royle and Nichols (2003) this is:

$$\Pr(\mathbf{y}) = \prod_j \Pr(\mathbf{y}_j|p_{j1}, \ldots, p_{jK})\psi_j + I(\mathbf{y}_j=0)(1-\psi_j). \tag{2.13}$$

This allow for the detection probability to vary by the k-th replicate of the j-th site, while the occurrence probability can be determined by site-level covariates:

$$\text{logit}(\psi_j) = \beta_0 + \beta_1 x_{j1} + \ldots. \tag{2.14}$$

Estimates for both ψ and p can be obtained by maximizing the log likelihood (Eqn. 2.8) for each parameter (MacKenzie et al., 2005). However, when occupancy and detection probabilities depend on

covariates, estimators can no longer be expressed in closed-form and need to be estimated through numerical methods. Moreover, computing variance for these estimators involves asymptotic theory or implementing non-parametric bootstraping methods which can be computationally demanding (MacKenzie et al., 2002).

Another important caveat pointed out by Welsh et al. (2013) is the fact that estimating equations based on the score functions, obtained through the differentiation of the log-likelihood with respect to each parameter, can yield solutions for ψ or p that are close to zero and one. Solutions at these boundaries are problematic as they produce estimates with very large variability and in consequence, unstable fitted probabilities when data are very sparse (i.e. species have very high or low occurrences) (Welsh et al., 2013).

2.2.2 Bayesian inference approach to occupancy models

The occupancy model which allows for imperfect detection has a natural hierarchical formulation for Bayesian methods because of the simple conditional structure of the model (Royle and Dorazio, 2008). The two state occupancy model assuming observations were recorded at M sites on K occasions with a constant probability of detection p across replicates is given by:

- $z_j \sim \text{Bernoulli}(\psi)$
- $y_{j\cdot} \sim \text{Binomial}(M, z_j p)$,

where each observation y_{jk} is grouped for each site (i.e. $y_{j\cdot} = \sum_k^K y_{jk}$). The joint distribution given the hierarchical structure of the model is given by:

$$f(\mathbf{y}, \mathbf{z}, p, \psi) = \left\{ \prod_j^M \prod_k^K [y_{jk}|z_j, p][z_j|\psi] \right\} [\psi].$$

To generate samples from the posterior, the full conditionals of each parameter should be specified. In this case, besides the conditional distributions on p and ψ, wherever $y_j = 0$ a realization of z_j is obtained from the conditional $z_j|y_j = 0$ (Royle and Dorazio, 2008).

Since the parameters p and ψ are probabilities, usually a $p \sim \text{Uniform}(0,1)$ and $\psi \sim \text{Uniform}(0,1)$ are specified as priors. Note that $z_j = 0 \rightarrow y_j = 0$, so the conditionals are required only for the case where $z_j = 1$. Due to conjugacy, the conditional posteriors for p and ψ are:

- $\Pr(p|\psi, y, z) = \text{Beta}(a_p, b_p)$ with $a_p = 1 + \sum_j^J z_j y_{j\cdot}$ and $b_p = 1 + M\sum_j z_j - \sum_j y_{j\cdot} z_j$
- $\Pr(\psi|z) = \text{Beta}(a_\psi, b_\psi)$ with $a_\psi = 1 + \sum z_i$ and $b_\psi = 1 + M - \sum_j z_j$.

Note that whenever $y_j > 0$, $z_j = 1$. Thus, for the case when $y_j = 0$, the conditional on the site being occupied ($z_j = 1$) is given by:

$$\begin{aligned}\Pr(z_j = 1|y_j = 0, p, \psi) &= \frac{\Pr(y_j = 0|z_j = 1)\Pr(z_j = 1)}{\Pr(y_j = 0)} \\ &= \frac{(1-p)^K \psi}{(1-p)^K \psi + (1-\psi)}. \end{aligned} \tag{2.15}$$

The probability of not detecting the species given it is present is given by $p(y_j|z_j = 1) \sim p^{y_j}(1-p)^{K-y_j}$, $\Pr(z_j = 1)$ is the occupancy probability ψ, and $\Pr(y = 0)$ is the likelihood from equation (2.13) evaluated at $y_j = 0$. Having all the full conditionals enable us to use any MCMC algorithm (e.g. Gibbs sampler) to estimate the posteriors of each parameter (see Appendix A.1 for the implementation of a Metropolis (MH)/Gibbs sampler algorithm to estimate occupancy model parameter under space-varying occupancy probabilities and constant detection). To assess the algorithms convergence to the posterior distribution, conventional convergence monitoring tools can be used by running several Markov chains on different starting values (Plummer et al., 2006). For instance, if chains have converged, then trace plots will show the parameter values mixing without a trend and with constant variance around the parameter space at each MCMC iteration (examples are presented in the appendix and discussed in the following section). Also, autocorrelation plots can be used to measure the correlation between the draws of the Markov chain. The greater the autocorrelation is, the larger the number of samples required to achieve the desired level of precision for MCMC estimates. Finally, Gelman-Rubin diagnostics (Gelman et al., 1992) can be used to assess whether the distribution of the chain does not change over time by comparing the within-chain and between-chain variance of multiple chains. The mean of the variances of each chain W is given by:

$$W = \frac{1}{C}\sum_{j}^{C} s_j^2, \text{ such that } s_j^2 = \frac{1}{L-1}\sum_{i}^{L}(\theta_{ij} - \bar{\theta}_j)^2, \tag{2.16}$$

where C is the number of independent chains, L is the number of samples in each chain, s_j^2 is the variance of the jth chain for the parameter θ_{ij}. Then, the variance of the chain means is given by:

$$B = \frac{L}{C-1}\sum_{j}^{C}(\bar{\theta}_j - \bar{\theta})^2 \text{ where } \bar{\theta} = \frac{1}{C}\sum_{j}^{C}\bar{\theta}_j. \tag{2.17}$$

The variance of θ is computed as a weighted average of W and B, i.e. $\widehat{\text{Var}}(\theta) = (1 - \frac{1}{L})W + \frac{1}{L}B$. Then, the Gelman-Rubin potential scale reduction factor statistic is computed as $\hat{R} = \left(\widehat{\text{Var}}(\theta)/W\right)^{-\frac{1}{2}}$.

Typically, an $\hat{R}$ value < 1.1 or close to one indicates that the chains have converged and that sufficient posterior samples have been obtained.

2.2.3 Multiple covariates and model selection

Very often, applications of occupancy models rely on assessing the effect that different covariates may have or not on the occupancy and detection probabilities. Hence, the occupancy and detection probabilities can be defined as linear functions of x_u different site-level covariates and possible g_v survey-specific effects as follows:

State/Occupancy Model

- $z_j \sim \text{Bernoulli}(\psi_j)$
- $\text{logit}(\psi_j) = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_u x_{ju}$

Observational/Detection Model

- $y_{jk} \sim \text{Bernoulli}(z_j p_{jk})$
- $\text{logit}\ (p_{jk}) = \alpha_0 + \alpha_1 x_{j1} + \ldots + \alpha_u x_{ju} + \alpha_{u+1} g_{jk1} + \ldots + \alpha_{u+v} g_{jkv}$

Different combinations of covariates yield a wide range of candidate models. How to choose the best model that fits the data is not trivial and thus, different methods have been proposed to choose the "best" model (Kéry and Royle, 2015).

When using likelihood-based methods, Akaike Information Criteria (AIC) has become a popular choice for model selection. AIC is defined in (Eqn. 2.18) as the negative log-likelihood of the model penalized by the number of parameters (r).

$$AIC = 2r - 2\text{log}L(\hat{\theta}_{MLE}|\mathbf{y}), \tag{2.18}$$

where $\hat{\theta}_{MLE}$ is the maximum likelihood estimates of the different parameters in the model. Thus, for a given set of candidate models, the model with the minimum AIC value is preferred. Moreover, AIC weights can be produced for the pair exponential differences between the best model AIC (AIC_{min}) and another model candidate (AIC_m) (Eqn. 2.19).

$$w_m = \frac{\exp(-(1/2)\Delta_m)}{\sum_m \exp(-(1/2)\Delta_m)}, \tag{2.19}$$

where $\Delta_m = AIC_m - AIC_{min}$. These weights can be used to model average estimates of different models rather than relying on the results of one single model (Burnham and Anderson, 2003).
Adjustment for small sample sizes on the AIC gives the sample criteria shown in equation (2.20).

$$AICc = \frac{2r(r+1)}{n-r-1} - 2\text{log}L(\hat{\theta}_{MLE}|\mathbf{y}), \tag{2.20}$$

where n is the sample size. In a HM context, the sample size can be determined by the number of sites or by the number of replicates depending on which structure of the model (occupancy or detection) is of most interest.

Deviance information criteria (DIC) is an analogous version of the AIC for Bayesian analysis (Spiegelhalter et al., 2002)

$$DIC = \text{Dev}(\hat{\theta}) + 2p_D. \tag{2.21}$$

DIC is defined by the posterior mean of the deviance ($\overline{\text{Dev}}(\theta)$) penalized by a model complexity measure $p_D = \overline{\text{Dev}}(\theta) - \text{Dev}(\hat{\theta})$, where :

$$\begin{aligned} \overline{\text{Dev}}(\theta) &= \frac{1}{S}\sum_s -2\log(\mathbf{y}|\theta^{(s)}) \quad \text{for } s = 1,\ldots,S \text{ MCMC samples,} \\ \text{Dev}(\hat{\theta}) &= -2\log(\mathbf{y}|\hat{\theta}). \end{aligned} \tag{2.22}$$

The complexity measure p_D is the difference between the posterior mean deviance and the deviance evaluated at the posterior mean ($\hat{\theta}$) of the model parameters. However, as mentioned by Lunn et al. (2012) and Kéry and Royle (2015), evaluating the deviance at the posterior mean of categorical data (such as the occupancy state) is not meaningful. Alternatively, deviance could be computed by marginalizing out the z_j latent parameters by using the likelihood in (2.10). However, the Watanabe-Akaike information criteria (WAIC; Watanabe and Opper (2010)) was proposed as a fully Bayesian information criteria that is better suited for hierarchical models as it does not depend on a single point-estimate and rather uses the whole posterior distribution for its calculation. The WAIC is computed as follows:

$$WAIC = \sum_j^M \log\left(\frac{1}{S}\sum_s p(y_j|\theta^{(s)})\right) - \sum_j Var\left(\log(p(y_j|\theta^{(1)}),\ldots,\log(p(y_j|\theta^{(S)})\right), \tag{2.23}$$

where the first part of equation (2.23) corresponds to the log-pointwise-predictive-density (lppd) which is an average of the model fit across all the posterior draws. The second part of this equation is the variance of the log-likelihood across MCMC samples, i.e. the sum of the posterior variance of the log predictive density representing the number of effective parameters. Unfortunately, computing WAIC can be computationally intensive when monitoring a large number of parameters (e.g. when investigating multiple species occurrences) Hobbs and Hooten (2015); Tenan et al. (2014). Hence, model selection by introducing an indicator variable that imposes a prior distribution to the model itself, has become more relevant within hierarchical models framework. This approach was first introduced by Kuo and Mallick (1998), the idea is to define a latent indicator variable w_u for a variable u in the model such that:

$$w_k = \begin{cases} 1 & \text{if covariate } k \text{ coefficient} \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.24}$$

Moreover, it is assumed that $w_k \overset{iid}{\sim} \text{Bernoulli}(\pi_k = 0.5)$, where π_k is the prior for each model. For instance, the state model with multiple covariates shown above would be re-parametrized as:

$$\text{logit}\,(\psi_j) = \beta_0 + w_1\beta_1 x_{j1} + \ldots + w_u\beta_u x_{ju}.$$

The null model occurs when $w_1 = \ldots = w_u = 0$. The closer w_u is to 1, the stronger is the evidence that the uth variable should remain in the model. The posterior frequencies for $w_1, \ldots, w_u$ can be processed using MCMC to produce the posterior probabilities for each of the 2^u models.
This method allows averaged model predictions to be produced since the MCMC outcome contains samples from all possible models defined by different combinations of the indicator variables (Kéry and Royle, 2015).

2.2.4 Multispecies occupancy model

The models discussed so far allow the occupancy patterns for a single species to be estimated. However, making inference about multiple species and understanding how their distributions are shaped by different ecological processes is often of main interest in ecological studies (Dorazio and Royle, 2005). These scenarios can be easily accommodated within hierarchical Bayesian modelling. Let z_{ij} be a latent variable for the true presence/absence of species i at the jth site. Then, the observational process given the true state is given by the number of times species i was detected at site j over k different occasions or visits (i.e. y_{ijk}). These can be seen as an aggregated process over the K visits, i.e. $\sum_{k=1}^{K} y_{ijk} = y_{ij\cdot}$.

The model can then be written as follows:

$$\begin{aligned} z_{ij} &\sim \text{Bernoulli}(\psi_{ij}) \\ y_{ij\cdot}|z_{ij}, K_j &\sim \text{Binomial}(K_j, p_{ij}z_{ij}), \end{aligned} \tag{2.25}$$

where ψ_{ij} and p_{ij} are the occupancy and detection probabilities respectively. Thus, the joint distribution is given by:

$$p(y_{ij\cdot}, z_{ij}|p_{ij}\psi_{ij}) = \left[\binom{K_j}{y_{ij\cdot}} p_{ij}^{y_{ij\cdot}}(1-p_{ij})^{K_j - y_{ij\cdot}}\right]^{z_{ij}} \times \psi_{ij}^{z_{ij}}(1-\psi_{ij})^{1-z_{ij}}. \tag{2.26}$$

The marginal distribution is then given by:

$$\begin{aligned} p(y_{ij\cdot}|p_{ij}\psi_{ij}) &= \sum_z \Pr(y_{ij\cdot}|z_{ij})\Pr(z_{ij}) \\ &= \text{Binomial}(K_j, y_{ij\cdot}|z_{ij}=1)\Pr(z_{ij}=1) + \text{Binomial}(K_j, y_{ij\cdot}|z_{ij}=0)\Pr(z_{ij}=0) \\ &= \binom{K_j}{y_{ij\cdot}} p_{ij}^{y_{ij\cdot}}(1-p_{ij})^{K_j - y_{ij\cdot}}\psi_{ij} + \mathbb{I}(y_{ij\cdot}=0)(1-\psi_{ij}). \end{aligned} \tag{2.27}$$

The first term of equation (2.27) evaluates point-mass at $y_{ij} = 1$ when species are present, i.e. $z_{ij} = 1$. The second term is an indicator function when the species is absent. Therefore, species i is absent at the jth site with a probability of $1 - \psi_{ij}$, or it is present but undetected with a probability of $(1 - p_{ij})^{K_j - y_{ij\cdot}} \psi_{ij}$.

Site-level covariates and species-specific effects can be modeled as linear functions of the ocupancy and detection probabilities. Thus, the effect model can then be written by following Dorazio et al. (2006) and Dorazio et al. (2011) as:

$$\begin{aligned} \phi_{ij} &= \text{logit}(\psi_{ij}) = b_{i0} + b_{i1} x_{1j} + \ldots + b_{iu} x_{uj} \\ \eta_{ij} &= \text{logit}(p_{ij}) = a_{i0} + a_{i1} g_{1j} + \ldots + g_{ih} x_{hj}. \end{aligned} \tag{2.28}$$

The latter formulation allows for occupancy and detection probabilities of each species to vary with the different u and g site level covariates. The parameter b_{i0} and a_{i0} correspond to the species-specific logit-scale baseline probability of occupancy and detection respectively. Then, b_{iu} and a_{ih} denote the effect of the hth and uth site level covariates on the occupancy and detection probabilities.

The marginal density of the observed number of detections can be written in terms of logit-scale parameters as follows:

$$\begin{aligned} p(y_{ij\cdot}|p_{ij}\psi_{ij}) &= \frac{exp(\phi_{ij})}{1+exp(\phi_{ij})} \binom{K_j}{y_{ij\cdot}} \frac{exp(\eta_{ij} y_{ij\cdot})}{\left[1+exp(\eta_{ij})\right]^{y_{ij\cdot}}} \left[\frac{1}{1+exp(\eta_{ij})}\right]^{K_j - y_{ij\cdot}} + \mathbb{I}(y_{ij\cdot} = 0) \left[\frac{1}{1+exp(\phi_{ij})}\right] \\ &= \frac{exp(\phi_{ij})}{1+exp(\phi_{ij})} \binom{K_j}{y_{ij\cdot}} \frac{exp(\eta_{ij} y_{ij\cdot})}{\left[1+exp(\eta_{ij})\right]^{y_{ij\cdot}}} \frac{\left[1+exp(\eta_{ij})\right]^{y_{ij\cdot}}}{\left[1+exp(\eta_{ij})\right]^{K_j}} + \mathbb{I}(y_{ij\cdot} = 0) \left[\frac{1}{1+exp(\phi_{ij})}\right] \\ &= \frac{exp(\phi_{ij})}{1+exp(\phi_{ij})} \binom{K_j}{y_{ij\cdot}} \frac{exp(\eta_{ij} y_{ij\cdot})}{\left[1+exp(\eta_{ij})\right]^{K_j}} + \mathbb{I}(y_{ij\cdot} = 0) \left[\frac{1}{1+exp(\phi_{ij})}\right]. \end{aligned} \tag{2.29}$$

Estimating occupancy and detection probabilities independently for each species however, has some drawbacks. For example, the number of parameters estimated by this model increases with the size of the community ($2n$ parameters where n is the number of species within the community). Also, Kéry and Royle (2015) reported that the uncertainty associated with the estimates can be remarkably high for rare and elusive species that occur in low abundances and with low detectability. Thus, instead of estimating occupancy and detection parameters for each species independently, the parameters can be assumed to arise from the same prior distribution by adding one more level into the hierarchical model so species are treated as a random effect. Species heterogeneity in occupancy and detection probabilities is modeled by this new hierarchy representing the parameters' average value among all the species in a community. Thus, species-specific parameters are shrunk towards the community mean enabling information between species to be exchanged and allowing for rare and elusive species parameters to be estimated that would not otherwise be possible due to their low occurrences (Dorazio et al., 2011). By taking this approach and assuming all species are related to each other by being part of the same biological community, parameters of species with scarce occurrence records can be estimated. This is essential to provide an accurate description of the diversity in the system.

Occurrence and detection parameters can be drawn from a multivariate normal distribution as follows:

$$\begin{pmatrix}\mathbf{b}_i\\ \mathbf{a}_i\end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix}\beta\\ \alpha\end{pmatrix}, \Sigma\right), \tag{2.30}$$

where $\mathbf{b}_i = (b_{i0}, b_{i1}, \ldots, b_{iu})$ and $\mathbf{a}_i = (a_{i0}, a_{i1}, \ldots, a_{iv})$ denotes a vector of logit-scale occurrence and detection species-specific parameters. The community mean hyper parameters are β and α from which individual species parameters are drawn. The diagonal of Σ holds the variance for each parameter whereas the off-diagonal elements of Σ are all zero except for $\sigma_{b_{i0},a_{i0}} > 0$ to allow for correlated intercepts. Correlation between occupancy and detection accounts for the fact that species with high abundance will be more widespread and thus, more likely to be detected (Royle and Dorazio, 2008).

If site level covariates are not available, site-level effects can be assumed to be constant as reported by Dorazio and Royle (2005) and Dorazio et al. (2006), i.e. $[\phi_i, \eta_i|\Sigma] \sim \text{Normal}([\mu_\psi, \mu_p], \Sigma_{2\times 2})$, where $[\mu_\psi, \mu_p]$ are the logit-scale community baseline occupancy and detection hyperparameters respectively. Σ is a 2×2 matrix with variances $(\sigma^2_\psi, \sigma^2_p)$ and covariance $\sigma_{\psi p}$.

By setting $\sigma_{\psi p} = 0$, Kéry and Royle (2008) specified an uncorrelated intercept model. So species specific effects are drawn from two independent normal distributions, i.e.

$$\begin{aligned}\text{logit}(\psi_i) = \phi_i &\sim \text{Normal}(\mu_\psi, \sigma^2_\psi)\\ \text{logit}(p_i) = \eta_i &\sim \text{Normal}(\mu_p, \sigma^2_p).\end{aligned} \tag{2.31}$$

This model assumes that the only source of variation in detection and occurrence probabilities is the species identity. Thus, $[\phi_i, \eta_i]$ are species specific random effects and $[\mu_\psi, \mu_p]$ are the mean logit scale occupancy and detection among all species. Thus, species-specific effects can be defined as linear functions of either the occupancy or the detection hyperparameters e.g. $\mu_{\psi_i} = \delta_0 + \delta_1 v_{i1} + \ldots \delta_p v_{ip}$ where v_{ip} are species-specific covariates.

In addition to the single species two-state occupancy model assumptions, the multispecies occupancy model also assumes that (1) species occurrences are independent to each other and (2) all inference made is restricted to the list of species that were detected at least once at any given site.

Multiple species occupancy models allow for several important ecological metrics to be derived. For instance, when dealing with multiple species the total number of species estimated to be present at a site j define the site-specific species richness (also called α diversity), i.e. $\alpha_j = \sum_i z_{\cdot j}$, which is an important ecological metric used to describe a site's biodiversity (Begon et al., 1986). Pairwise difference in species compositions between two sites r and s allow for a Jaccard Similarity index to be computed for a reference site or species (Kéry and Royle, 2015):

$$Jaccard_{r,s} = \frac{\sum_i z_{\cdot r} z_{\cdot s}}{\sum_i z_{\cdot r} + \sum_i z_{\cdot s} - \sum_i z_{\cdot r} z_{\cdot s}}. \tag{2.32}$$

Another quantity of interest that can be derived from either the single species or multispecies occupancy models is the occupancy rate $\psi^{(fs)}$ which is a point estimate for the total number of sites in the sample that were occupied expressed as a proportion out of the total number of sites.

There is a subtle yet clear distinction between the occurrence probability (ψ_i) and the occupancy rate ($\psi_i^{(fs)}$). The former, is a population parameter that indicates the probability of any site being occupied by each species. Whereas the latter, often referred in literature as finite-sample occupancy (Royle and Dorazio, 2008), is the proportion of sites occupied in the study sample given the observation process. For a given number of M sites, the finite sample occupancy is given by equation (2.33).

$$\psi_i^{(fs)} = M^{-1} \sum_j z_{ij}. \tag{2.33}$$

For any given species at a particular site j, the occupancy status z_{ij} is determined by the observation process, i.e. if $y_{ij} > 0$ then $z_{ij} = 1$. However, if $y_{ij} = 0$ then the conditional probability of the site being occupied given the observation status is determined by the expected occupancy probability, the detection probability, and number of surveys in each site (Eqn. 2.15). Thus, it is more likely for a species to occupy a site even though it was not detected, if it has high occupancy and low detection probabilities (i.e. it is a elusive but very common species) and the number of surveys for the j-the site is small. How the number of visits and number of sites affect the model performance is a recently studied subject which will be discussed in the next section (Kéry and Royle, 2015). Moreover, fitting these models using classical or Bayesian inference may be advantageous depending on the situation. Because of the conditional structure of the latent variables in the occupancy models, bayesian methods are usually preferred since Markov chain Monte Carlo (MCMC) algorithms allow retention of latent variables in the model (Clark and Altwegg, 2019). However, these methods can be computationally expensive and thus, classical methods based on the marginal likelihood are often used instead.

Nowadays, there are some platforms available to fit these classes of models. For instance, libraries `unmarked` and `hSDM` have been developed in R to fit single species occupancy models using classical or Bayesian approaches respectively (Fiske and Chandler, 2011; Vieilledent, 2019). For multispecies models, however, occupancy models have only been developed within a bayesian framework where `jags` has become a popular choice to implement MCMC methods in R (Kellner, 2017). In the following section a two stage single species occupancy model is compared using both classical and bayesian methods to provide a computational efficiency comparison.

2.3 Simulation study: Bayesian vs classical analysis comparison for a two state occupancy model

As pointed out in the previous section, both classical and Bayesian inference can be used to fit occupancy models. However, the performance of each approach will depend a great deal on the species distribution (i.e how common or rare the species are) and detectability (how elusive they are). Simulation analysis is an important tool for both statisticians and ecologists as they provide with a detailed description of the framework under which a statistical model is built. Simulation analysis enables comparison of a proposed model in different ways. For example, by calibrating the model parameters, i.e. how variation of certain parameters affects the estimates of other parameters, or by checking model parameters' identifiability and robustness of the estimators (Kéry and Royle, 2015). Simulations enable investigation of the model fit by comparing the estimates against the values from which the data were generated. Simulating different data sets can be used to explore the sampling error to assess the sample size needed to identify a certain effect with a given probability. Within the context of species distribution models, simulation studies can be used to produce guidelines that help researchers to take informed decisions in terms of sampling schemes and designs that can be applied on different biodiversity monitoring programs (Banner et al., 2019). In this section, populations with different detectability and occupancy parameters are simulated to evaluate the performance of Bayesian and classical methods. Observations with detectabilty bias were simulated using the `simOcc` function from `AHMbook` library (Kery et al., 2017) in `R` v.3.3.1 (R Core Team, 2016).

The function `occ()` from the `unmarked` library (Fiske and Chandler, 2011) was used for classical analysis of occupancy models. For Bayesian analysis the `jagsUI` library was used (Kellner, 2017).
For the first comparison, a simple occupancy model with no covariate effects (i.e. constant occupancy and detection across sites) is fitted. The data were simulated by drawing observations for J = 100 sites surveyed twice (K = 2) from the following model:

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\psi) \\ y_j &\sim \text{Binomial}(J, z_j p). \end{aligned} \tag{2.34}$$

For Bayesian analysis, a non-informative Uniform (0,1) prior is used for both ψ and p parameters (as suggested by Royle and Dorazio (2008)), 5000 iterations and a burnin period of 1000 are specified to run the Gibbs sampler [3]. The detectability and occupancy parameters were chosen to portray species over a wide range of occupancy and detection probabilities.

[3]Simulation App available on: `https://shiny.maths-stats.gla.ac.uk/2259971b/OccmodelSIM/`

The classical and Bayesian approach can produce similar estimators depending on the species occurrences and detectability. For example, Table 2.1 shows that Bayesian and classical estimates are close to each other with the exception of the cases where extremely high or low detection and occupancy probabilities are specified. For species with high or low probabilities of occurrence, classical methods estimate occupancy probabilities of 1 with very wide confidence intervals. In contrast, the uncertainty around Bayesian estimates is smaller for the occupancy probability. On the other hand, when species have a high probability of detection (i.e they are easy to spot), Bayesian and classical methods produce similar outcomes. However, for species with very low detection, Bayesian estimates are much closer to the true occupancy (although the credible intervals are wider) and MLE estimates are closer to the true detection value. In summary, Bayesian estimates generally perform better when species have widespread or very sparse distributions but still produce wide credible intervals (CRIs) when detection probability is low (i.e. when dealing with elusive species).

Table 2.1: Comparison between MLE and Bayesian estimates $\hat{\psi}$ and $\hat{p}$ for different detection and occupancy values.

True parameter	MLE	95% Confidence interval	Bayesian	95% Credible interval
$\psi = 0.96$	1	[3.41e-11 , 1]	0.954	[0.862 , 0.998]
p = 0.5	0.515	[0.446 , 0.584]	0.535	[0.459 , 0.616]
$\psi = 0.8$	0.81	[0.6 , 0.924]	0.818	[0.672 , 0.969]
p = 0.5	0.556	[0.431 , 0.673]	0.549	[0.437 , 0.662]
$\psi = 0.05$	0.03	[0.009 , 0.089]	0.042	[0.011 , 0.094]
p = 0.5	1	[8.54e-30 , 1]	0.835	[0.443 , 0.996]
$\psi = 0.5$	0.502	[0.324 , 0.679]	0.535	[0.364 , 0.797]
p = 0.5	0.468	[0.302 , 0.641]	0.452	[0.279 , 0.623]
$\psi = 0.5$	0.992	[1.44e-35,1]	0.382	[0.054 , 0.956]
p = 0.05	0.025	[0.008 , 0.074]	0.140	[0.019 , 0.487]
$\psi = 0.5$	0.482	[0.371 , 0.594]	0.491	[0.381 , 0.612]
p = 0.8	0.706	[0.57 , 0.813]	0.692	[0.556 , 0.808]
ψ= 0.5	0.5	[0.404 , 0.597]	0.501	[0.404 , 0.598]
p = 0.96	0.969	[0.907 , 0.990]	0.959	[0.911 , 0.988]

For a second comparison, the following occupancy model (Eqn. 2.35) with covariates effects was fitted:

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\psi_j) \\ \text{logit}(\psi_j) &= \beta_0 + \beta_1 x_{j1} \\ y_{jk} &\sim \text{Bernoulli}(z_j p_{jk}) \\ \text{logit}(p_{jk}) &= \alpha_0 + \alpha_1 g_{j1}, \end{aligned} \tag{2.35}$$

where the true occurrences z_j are drawn from a Bernoulli distribution with occupancy probability defined as a function of covariate x with $\beta_0 = 0$ and $\beta_1 = 3$ for the ecological process model (Fig. 2.2 (left)). The simulated observations (seed = 1234) y_{jk} for sites $j \in \{1,\ldots,100\}$ with $k \in \{1,\ldots,3\}$ replicates, are drawn from a Bernoulli distribution defined as a function of covariate g with $\alpha_0 = 0$ and $\alpha_1 = -3$ (Fig. 2.2 (right)). Parameter values were chosen to avoid generating very sparse data that cause unstable predicted probabilities of occurrences and detection as mentioned by Welsh et al. (2013). A total of 5000 iterations with a burnin period of 1000 were used for Bayesian analysis. [4]

Both Bayesian and classical estimates are close to each other (Fig. 2.2). However, the confidence region for the classical approach does not contain the true value for the occupancy probability across the whole range of covariate x. Moreover, when detection probability is below 0.2, confidence regions obtained through MLE are considerably wider than those obtained through Bayesian analysis (see appendix A.1.2 for simulation results, settings and graphical diagnostics).

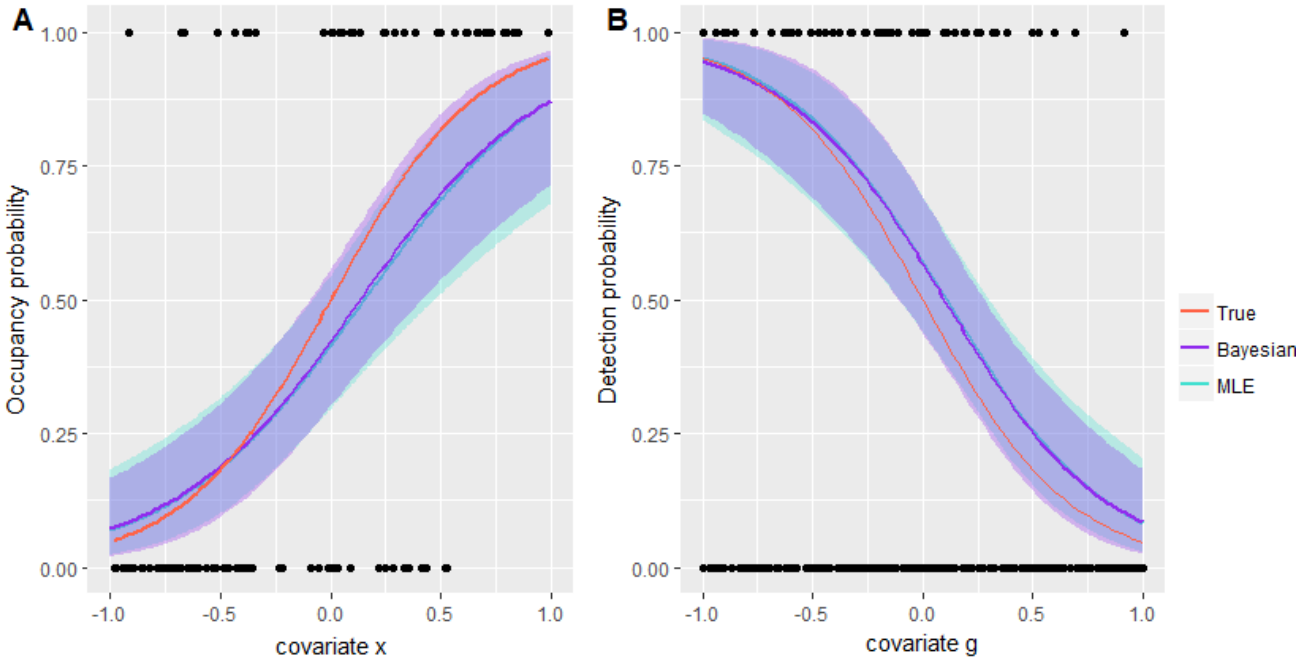


Figure 2.2: Estimated relationships between occupancy probability and simulated covariate x (left) and between detection probability and covariate g (right) from an occupancy model fit to the simulated data set with a baseline occupancy and detection probabilities of 0.5. Red lines represent the true relationship between covariates and the response with points showing the observed occurrences of each site across all surveys. Blue lines indicate maximum likelihood estimates and 95% CIs (turquoise shaded region). Purple lines indicate bayesian estimates with 95% Credible intervals (purple shaded region).

[4]Simulation App available on: `https://shiny.maths-stats.gla.ac.uk/2259971b/Occ_covariates`.

Figure 2.3 shows the finite-sample occupancy showing the estimated number of occupied sites in the sample. The true number of occupied sites in the sample was 65, from which 54 were the observed occupied sites (i.e. sites where there was at least one observation across all surveys). Parametric bootstrap confidence intervals were used to create a confidence region for the MLE. Both credible and confidence regions contain the true finite-sample occupancy, and the point estimates for both methods are very close to the true parameter value (Fig. 2.3).

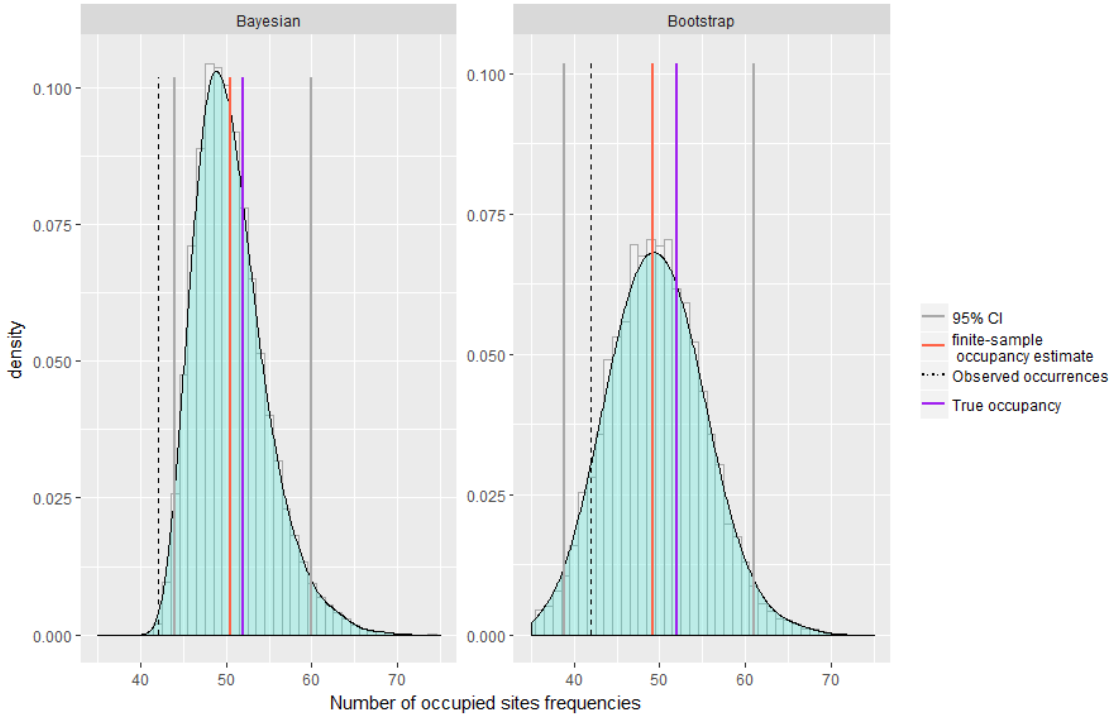


Figure 2.3: Bootstrap and posterior distributions for the finite-sample occupancy. The red lines indicates the 64 true number of occupied sites. Dashed lines indicates the 54 sites in which a species was observed. Blue lines shows the point estimate for each method and grey bands the confidence/credible intervals for the estimates.

Several simulation studies have explored experimental design of occupancy models (MacKenzie et al., 2002; Guillera-Arroita and Lahoz-Monfort, 2012; Ellis et al., 2015). For example, Kéry and Royle (2015) present a Bayesian and classical occupancy model comparison using different detectability values for simulated data sets with different number of sites and number of replicates (Fig. 2.4). Small detectability values ($p < 0.2$) tend to produce biased MLEs for cases with small number of sites and surveys/visits (e.g. visits < 5 and sites < 100). Therefore, uncertainty around the occupancy probability is higher in MLEs for low detectability values when the number of sites and surveys is small.

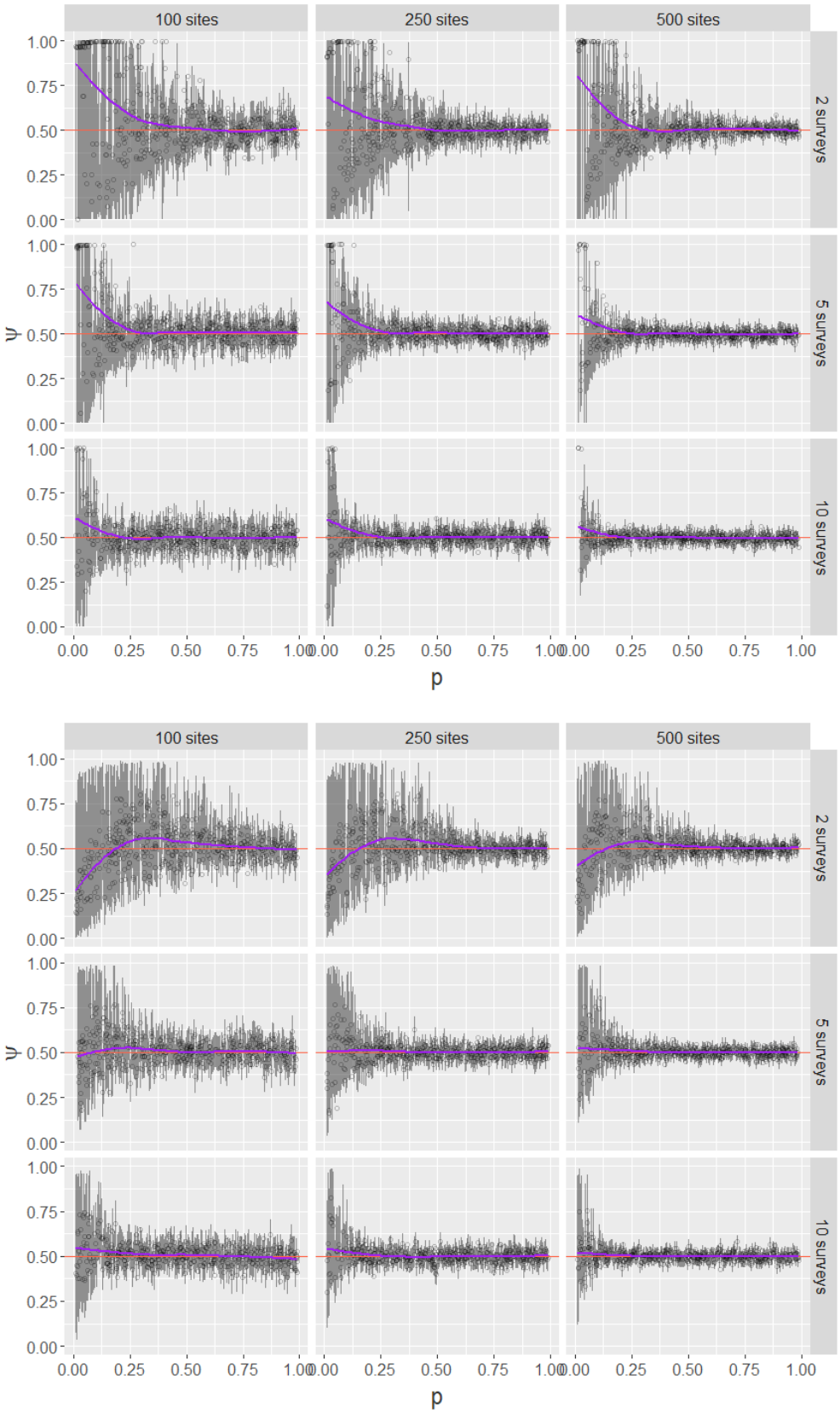


Figure 2.4: Simulation study results following Kéry and Royle (2015) for MLE (top) and Bayesian (bottom) estimates of the occupancy probability for 500 simulated data sets with different detection probabilities at three different number of sites and surveys. The red line shows the true occupancy probability (0.5) and the blue line is a spline smoother to show the average behaviour of the estimator for a given probability of detection.

To evaluate the predictive performance of the model with covariates effects (Eqn. 2.35), the root mean square predictive error (RMSPE) was computed according to Eqn. (2.36) for the occupancy and detection probabilities $\theta = [\psi, p]$ at each site $j = 1, \ldots, M$ on the k-th replicate/visit using 500 simulations.

$$\text{RMSPE} = \sqrt{\frac{1}{M}\sum_{1}^{n}(\theta - \hat{\theta})^2}. \tag{2.36}$$

When the detection probability is low, RMSPE for both state and observation models decreases as the number of sites increases. This means that sampling more sites rather than surveying the same sites multiple times improves the model for species that are hard to detect. On the other hand, when species are easy to spot, i.e. the detection probability is high, the RMSPE is smaller and decreases with higher number of surveys (the more replicates at each site, the better the predictive performance , Fig.2.5).

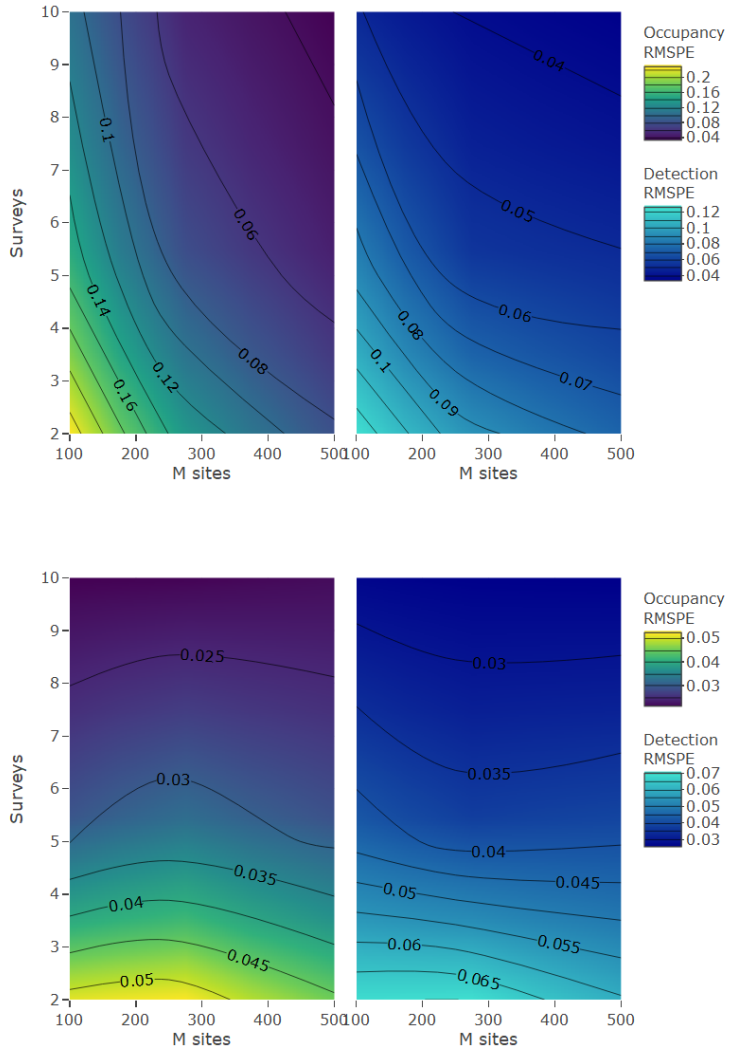


Figure 2.5: RMSPE for an occupancy model with a detection probability $p = 0.05$ (top) and $p = 0.95$ (bottom) and a constant occupancy probability $\psi = 0.5$ across a gradient of different sites and visits.

Unfortunately, Bayesian methods can be computationally expensive. Computational times of different algorithms to fit these models are shown in Table 2.2. The classical model proved to be the most computationally efficient method for fitting a two stage occupancy model, followed by Gibbs sampler written in C++ with `JAGS` being the most computationally expensive method. However, when BUGS-written models were compiled through `nimble` (which generates C++ code in R)(de Valpine et al., 2017), their computational time was equivalent to the metropolis-hastings algorithm written in C++. For very complex models for which the full conditional distributions are not always known, the metropolis-hastings algorithm written in C++ or in Nimble represent a computationally efficient alternative to classical methods (specially to fit occupancy models for multiple species).

Table 2.2: Computational times comparison between different algorithms for bayesian analysis of a two state Occupancy model with $\psi = 0.8$ and $p = 0.6$ for 100 sites and three surveys each.

Method	$\hat{\psi}$	$\hat{p}$	95 % CI	**System.Time (sec)**
Classical	0.843	0.557	ψ_{CI} [0.744, 0.942] p_{CI}[0.481. 0.634]	0.02
Gibbs Sampler in R	0.841	0.555	ψ_{CI} [0.741, 0.938] p_{CI}[0.478, 0.628]	0.62
Gibbs Sampler in C++	0.842	0.555	ψ_{CI} [0.740, 0.941] p_{CI}[0.478, 0.629]	0.16
JAGS	0.841	0.555	ψ_{CI} [0.740, 0.939] p_{CI}[0.478, 0.629]	5.71
R nimble compiler	0.842	0.555	ψ_{CI} [0.740, 0.946] p_{CI}[0.474, 0.628]	1.82
Metropolis in R	0.842	0.554	ψ_{CI} [0.737, 0.938] p_{CI}[0.479, 0.626]	3.81
Metropolis in C++	0.843	0.554	ψ_{CI} [0.738, 0.949] p_{CI}[0.473, 0.630]	1.53

2.4 Odonata multispecies occupancy model - estimating species occurrences while accounting for detection bias

The Odonata data set described in chapter 1 illustrates an example of a biological community integrated by a set of species with different responses and distribution ranges. However, records for Odonata are partially observed due to detection bias. Thus, the multispecies occupancy model reviewed in section 2.2.4 allows for the species occurrences within a community to be estimated when detection probability is less than one. A hierarchical formulation of a multispecies occupancy model is given by:

$$y_{ij}|z_{ij} \sim \text{Binomial}(K_j, p_i z_{ij}) \tag{2.37}$$

$$zij \sim \text{Bernoulli}(\psi_i), \tag{2.38}$$

where y_{ij} is the number of times species i was detected at site j across all K surveys, p_i is the detection probability for the ith species, z_{ij} is the latent variable for true species occurrences and ψ_i is the occupancy probability.

In this model, separate models are fitted to each species. This is particularly useful when the interest is to compare between species. Detection frequencies for each species are computed, i.e. $\sum y_{ij}$ for species $i = 1,\ldots,S$ at the sites $j = 1,\ldots,M$ with K_j visits such that $S = 39$ and $M = 4953$. The number of visits K_j ranged broadly between sites, from 2 to 477 visits across all years.

Non-informative Uniform (0,1) priors were used for both ψ and p parameters, 20000 iterations and a burnin period of 5000 and a thinning of 15 were specified to run the Gibbs sampler in the `nimble` compiler (approximate run time - 8 hours). The large number of parameters that need to be estimated due to the large collection of sites and species makes the MCMC computational time expensive. Thus, recent approximation methods to Bayesian inference such as the Integrated Laplace approximation (INLA) are currently being developed to investigate species distribution modelling under imperfect detection. An in-depth discussion of this point is provided in chapter 6. Moreover, fitting occupancy models to each individual species occurrence data separately, makes the parameter estimates of each species exclusively determined by the data for that species. Thus, species with very few observed presences will produce estimates with large variability as shown in Figure 2.6.

When occupancy is smaller than about 0.6 and detection smaller than about 0.015, estimates become very imprecise making some algorithm's convergence criteria dubious. For instance, Gelman-Rubin $\hat{R}$ statistic was found to be close to 1.1 for some of the rare and elusive species (see model diagnostics for *A. affinis* in the appendix A.1.3). Therefore, it is reasonable to fit a random effect for the species to improve these estimates due to information exchange (see section 2.2.4). Also, fitting a random effect model would enable abundance-induced detection effects to be modelled, i.e. include a correlation parameter to model the fact that widespread species are more likely to be detected. Figure 2.7, shows the importance of taking into account the detection bias since the observed number of occupied sites for each species differs widely from the estimated finite sample occupancy.

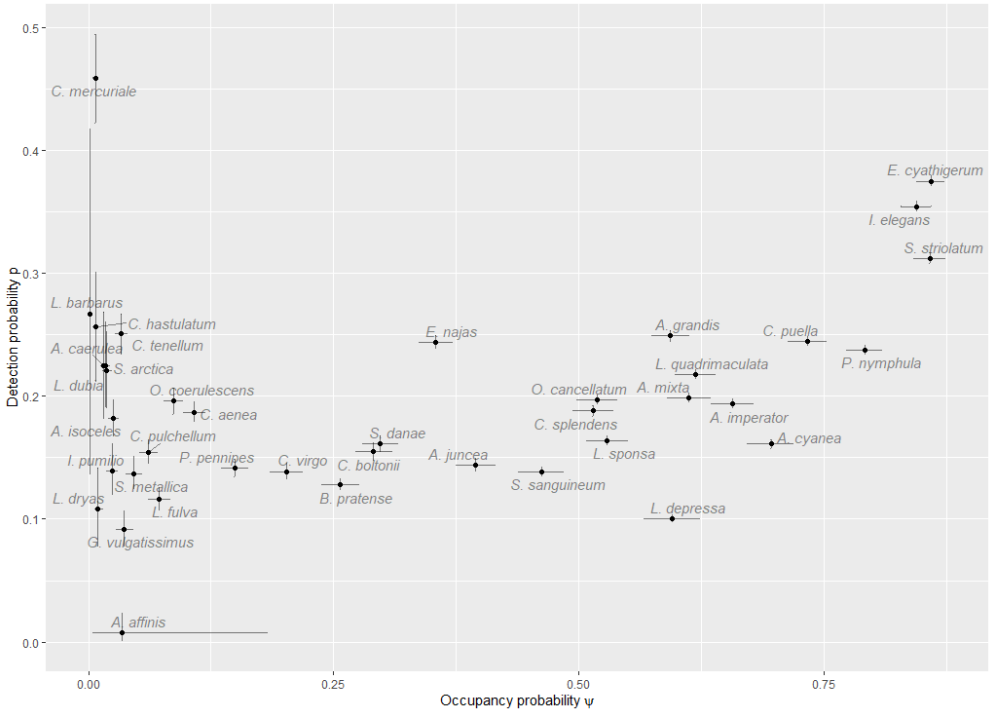


Figure 2.6: Estimated occupancy and detection probabilities. Error bars represent 95% Credible intervals.

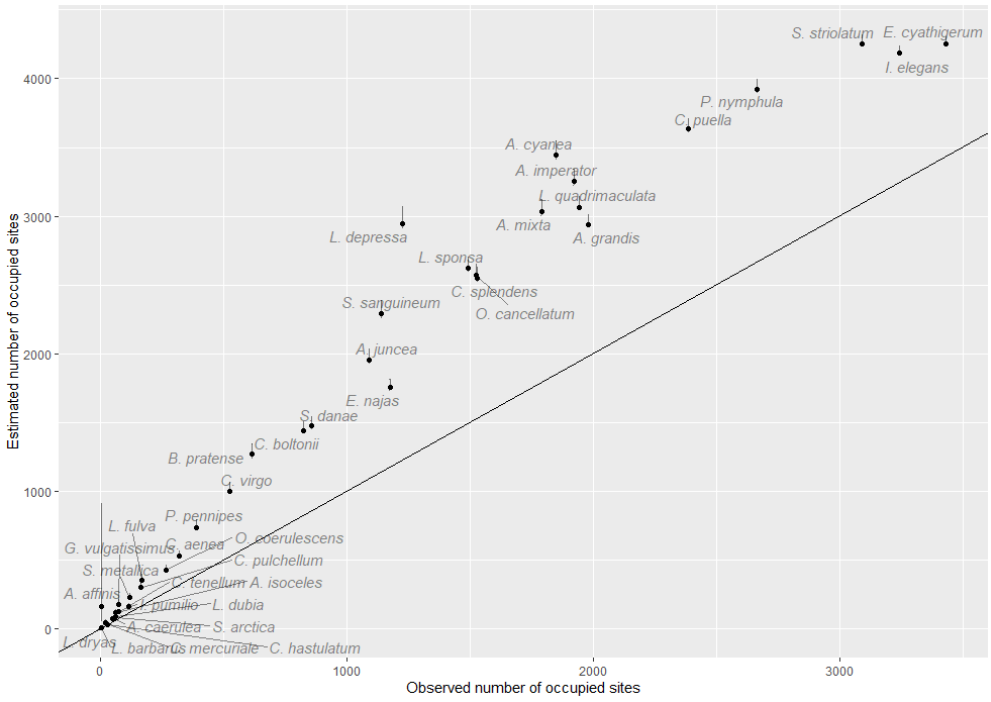


Figure 2.7: Estimated vs. observed finite sample occupancy (equality line represented in solid black).

2.4.1 Species random effect occupancy model - modelling abundance induced detections

Instead of fitting separate occupancy models for each species, a species can be treated as a random effect by assuming that detection and occupancy parameters arise from the same prior distributions. Let $u_i = \text{logit}(\psi_i)$ and $v_i = \text{logit}(p_i)$ the logit-scale species-specific baseline occupancy and detection such that:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} \mu_\psi \\ \mu_p \end{pmatrix}, \Sigma\right), \tag{2.39}$$

where $[\mu_\psi, \mu_p]$ denote the logit scale community mean site-level occupancy and detection respectively, with covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}.$$

The diagonal entries of $[\sigma_u^2, \sigma_v^2]$ hold the variability in species occupancy and detection respectively and the covariance $\sigma_{uv} = \rho\sigma_u\sigma_v$ captures the correlation between occupancy and detection, i.e. abundance-induced detection (the more widespread species are, the more likely they are to be detected). To fit this model, an inverse Wishart (IW) prior was used as a conjugate for the Covariance-Variance matrix as suggested by Kéry and Royle (2015), i.e. $\Sigma^{-1} \sim \text{Wishart}(\delta, \Omega_0)$ where δ corresponds to the degrees of freedom. Non-informative priors were specified for the rest of parameters. Specifically, Uniform (0,1) for probability scale site-level detection ($\text{logit}^{-1}(\mu_\psi)$) and occupancy ($\text{logit}^{-1}(\mu_p)$) were used. A total of 20000 iterations were run with a burnin period of 5000 and a thinning of 15 to run the Gibbs in `nimble` (approximate run time - 6.7 hours).

Table 2.3 shows the posterior means and 95% CRIs for the different parameters in the random effects occupancy model. Site level occupancy and detection ($\psi = \text{logit}^{-1}(\mu_\psi); p = \text{logit}^{-1}(\mu_p)$) are around 20% for all communities. Moreover, the association between species occupancy and detection is low ($\rho = 0.25$, Appendix figure 12).

Table 2.3: Summary for posterior estimates for the multispecies random effect occupancy model

	Mean	95% CRI
ψ	0.183	[0.100, 0.316]
p	0.184	[0.158, 0.214]
ρ	0.252	[-0.049 , 0.534]
σ_u	5.416	[3.458 , 8.580]
σ_v	0.340	[0.206 , 0.558]
σ_{uv}	0.346	[-0.064 , 0.883]

Figure 2.8 shows the estimated species-level occupancy and detection probabilities along with the estimated total number of sites each species occupy. Species detection probabilities are estimated to be below 50% for all species, highlighting the importance to take into account the detection bias for this system. The variability of the species-specific occupancy and detection probabilities is lower than the variability of the estimates produced by fitting separate occupancy models to each species. Also, convergence criteria was improved for even the most rare and elusive species (e.g. Gelman-Rubin $\hat{R} \approx 1$ for *A. affinis* occupancy and detection parameters). This can also be observed in the traceplots mixing and autocorrelation plots shown in appendix A.1.3. The finite sample occupancy (i.e. the number of sites in sample where each individual species occur) indicates that almost half of the species are estimated to be present in more than 50% of the sites. Note that the credible intervals show that the uncertainty for detection probabilities becomes larger as the occupancy probability decreases as there is less information from which the detection parameters are estimated resulting in greater uncertainty around these estimates.

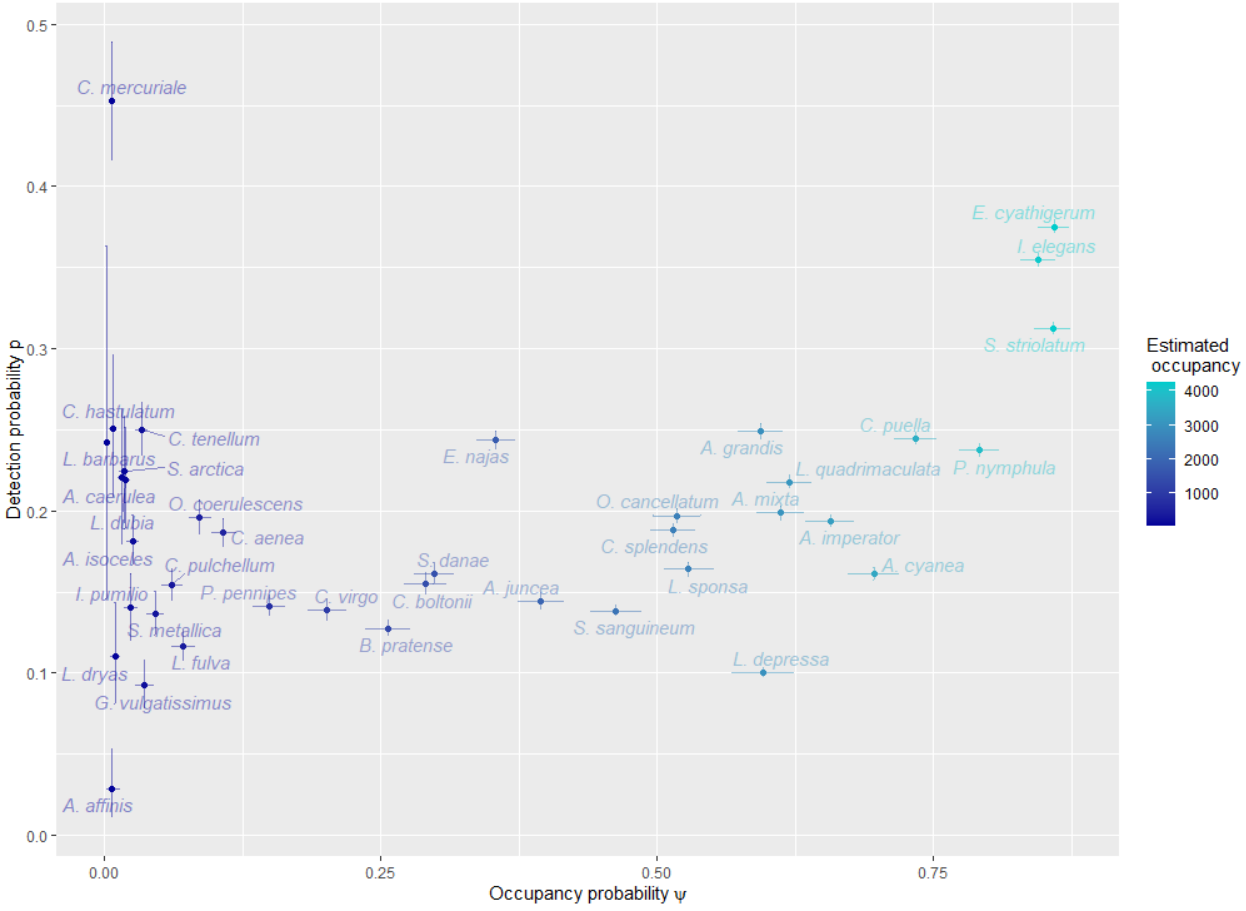


Figure 2.8: Estimated number of occupied sites, detection and occupancy probability for each species.

Figure 2.9 (left) shows the estimated local richness (α diversity) at each site which seems to be greater at lower latitudes. This pattern depends on each individual species distribution range, as environmental factors occurring up in the north may limit the distribution of several species occurring in the south. Species richness was calculated with a reasonably small error as the standard deviations (SD) in Figure 2.9 (right) suggest (SD < 2 for most sites).

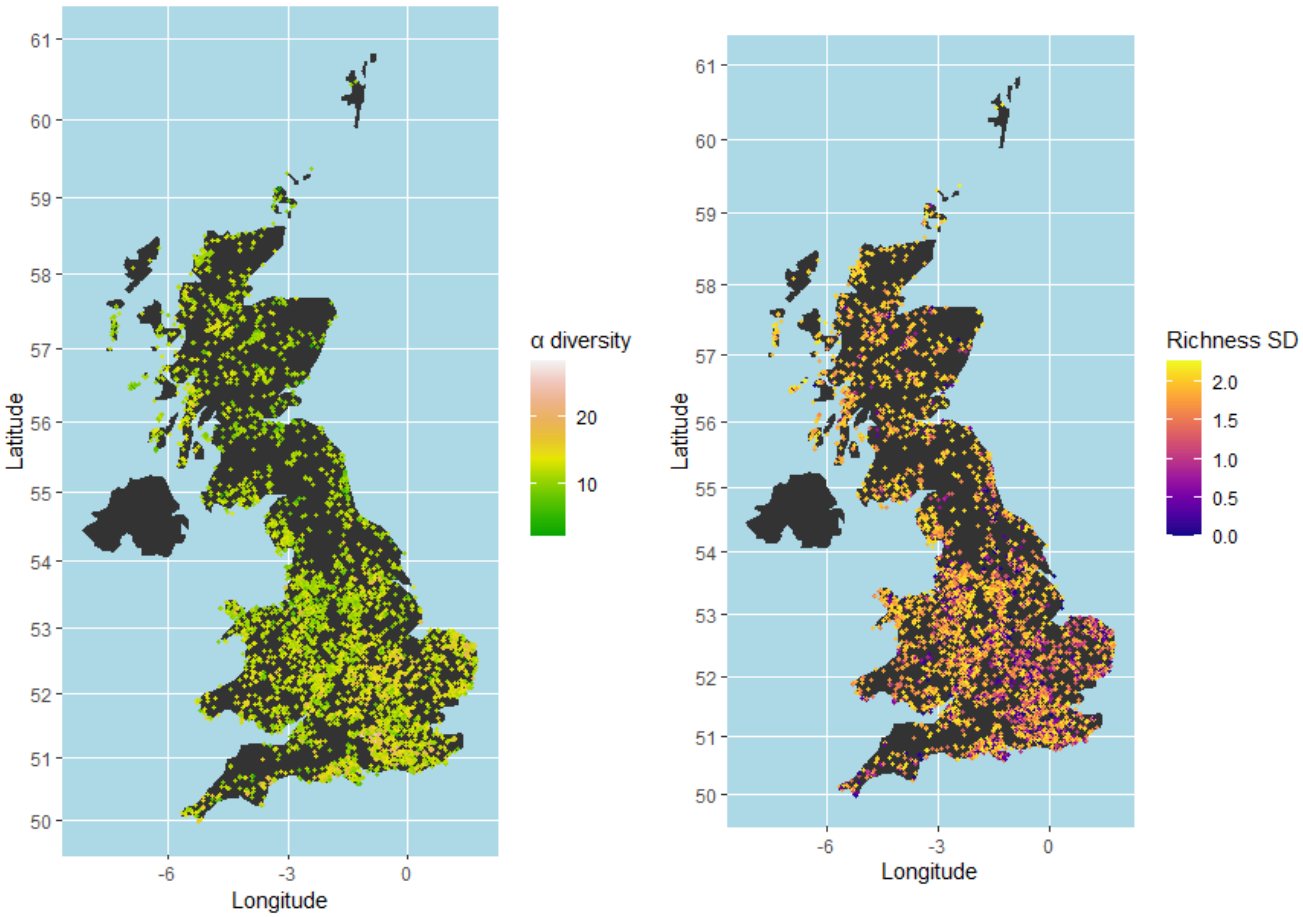


Figure 2.9: Odonata species richness in waterbodies across the UK. Estimated richness posterior mean at each site defined by the presence of a waterbody (left) and estimated standard deviation (right).

2.4.2 Independent normal occupancy model - modelling species specific traits

Since there is no evidence of a strong association between occupancy and detection, species-specific parameters can be drawn from two independent normal distributions. Moreover, a model which assumes that occupancy and detection are linked through the number of different habitats each species occupies can be specified as follows:

$$\begin{aligned} z_{ij} &\sim \text{Bernoulli}(\psi_i) && \text{State process} \\ \sum_{K_j} y_{ij\cdot} | z_{ij} &\sim \text{Binomial}(K_j, p_i z_{ij}) && \text{Aggregated observation process} \\ \left.\begin{aligned} \text{logit}(\psi_i) &\sim \text{Normal}(\mu_\psi, \phi_{0_\psi}) \\ \text{logit}(p_i) &\sim \text{Normal}(\mu_{p_i}, \phi_{0_p}) \end{aligned}\right\} & && \text{Species heterogeneity model.} \end{aligned} \tag{2.40}$$

Species specific-detection probabilities are defined as a function of species traits covariates (Eqn. 2.41) (see chapter 4 for a model that allow to have flexible terms). Thus, the species heterogeneity model for detection results in a regression with i random intercepts and slopes for the mean hyperparameters, i.e.

$$\mu_{p_i} = \alpha_0 + \alpha_1 \times \log(\text{flight period})_i + \alpha_2 \times \log(\text{body size})_i + \alpha_3 \times \text{num. of habitats}_i \tag{2.41}$$

Normal(0, 0.01) priors where used for $\boldsymbol{\alpha}$ and conjugate inverse-gamma priors for ϕ_{0_ψ} and ϕ_{0_p} precision parameters. Table 2.4 shows the estimates for site-level occupancy and species-specific detection hyperparameters.

Table 2.4: Estimates and 95% CRI for the random effect occupancy model.

	mean	2.5%	97.5%
μ_ψ	-1.577	-2.316	-0.820
α_0	-0.467	-2.273	1.259
α_1	0.541	0.004	1.123
α_2	-0.554	-0.1039	-0.0775
α_3	0.122	0.007	0.234
ϕ_{0_ψ}	1.684	1.247	2.146
ϕ_{0_p}	2.811	-4.127	9.622

On average, larger species are less detectable than smaller species (Fig. 2.10). This could be related to smaller body-size species being, on average, more widespread than larger body-size species (Stephen Brooks, personal communication, 2018), which is consistent with the initial exploratory analysis presented in chapter 1. In consequence, spatial variation and heterogeneity in detection probabilities for small-sized species will be larger, as indicated by the uncertainty regions of the posterior draws in Figure 2.10. Moreover, large body sizes species distributions could be constrained by local environmental conditions such as temperature (a further discussion of the relationship between Odonata species distributions and temperature and how this is related with the species body size is provided in chapter 5).

Fig. 2.11 illustrates how detection probabilities increase significantly for those species who have longer flying periods. The longer the species flight period is, the more likely it is for the species to be detected. This is also related to scarce-distributed northern species exhibiting shorter flying periods due to environmental constraints such as temperature (see discussion in chapter 5 for further details).

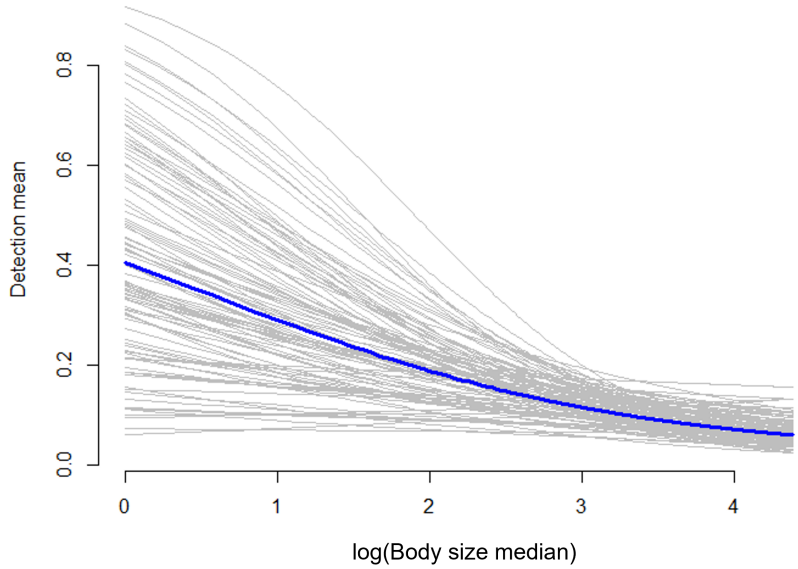


Figure 2.10: Relationship between detection probability and log(size). The gray lines show a random sample (100 draws) from the posterior, and the blue lines indicate the posterior mean.

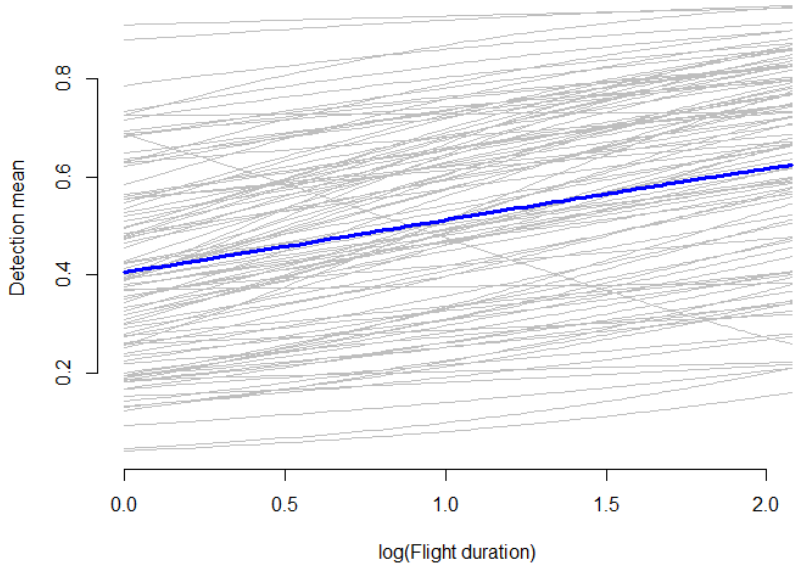


Figure 2.11: Relationship between detection probability and flight duration on log scale. The gray lines shows a random sample (100 draws) from the posterior, and the blue lines indicates the posterior mean.

The proposed model (Eqn. 2.40) diagnostics plots shows evidence of convergence as shown by the traceplots on figures 2.12 and 2.13 which indicates overall good mixing.

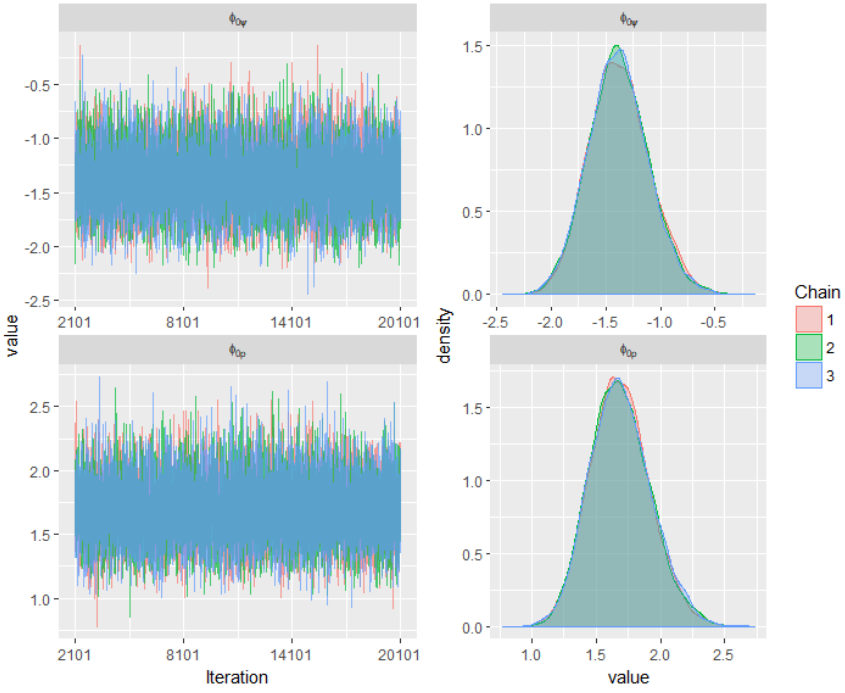


Figure 2.12: Traceplots and posterior densities for occupancy model variance parameters with two independent normal distributions.

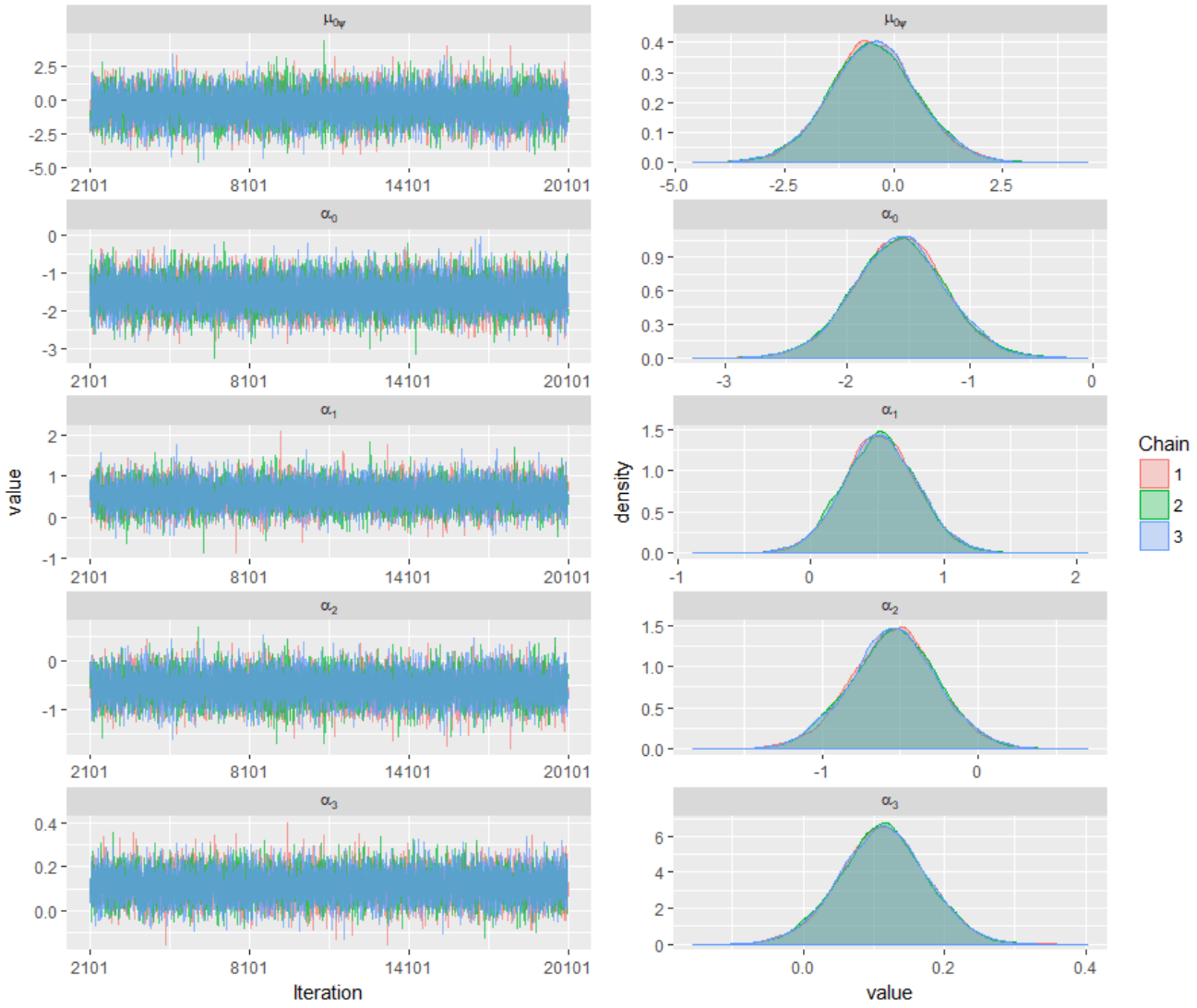


Figure 2.13: Traceplots and posterior densities for occupancy model mean occupancy and detection intercept and slopes parameters with two independent normal distributions.

There is no evidence of positive autocorrelation, and thus, the amount of thinning properly deals with the correlation between MCMC samples (Fig. 2.14).

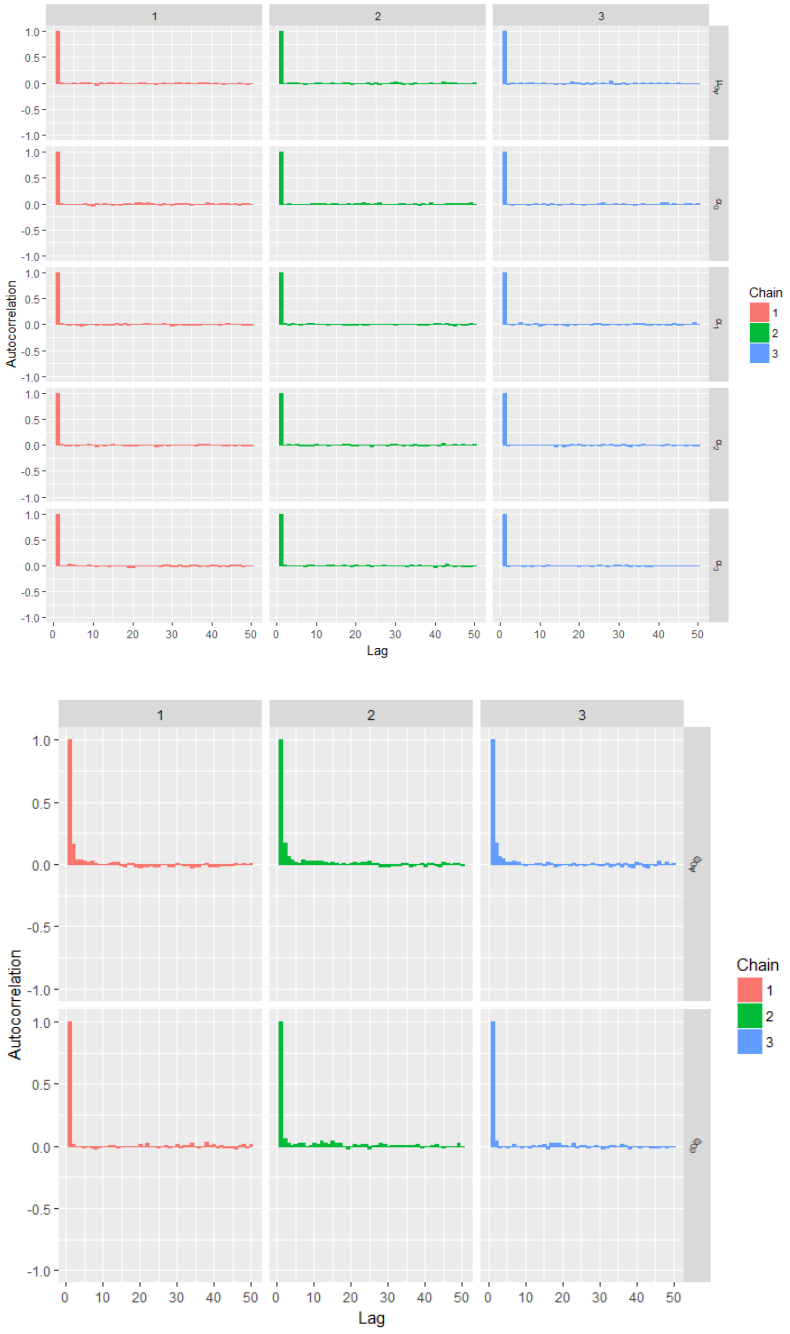


Figure 2.14: Autocorrelation plots for occupancy model mean occupancy and detection intercept, slopes (top) and variance parameters (bottom) with two independent normal distributions.

The Gelman-Rubin convergence diagnostic which compares the estimated between within-chain variances for each parameter indicates that the shrink factor is close to one. This means that chains within and between variances are similar (Fig. 2.15 and Fig. 2.16).

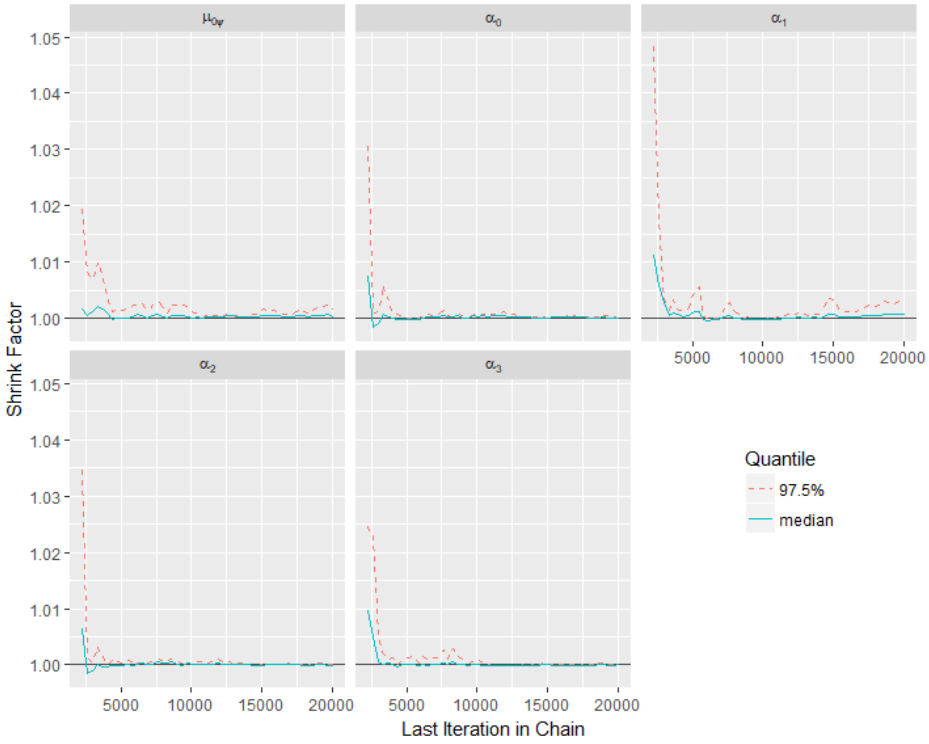


Figure 2.15: Gelman diagnostics for occupancy model mean occupancy and detection intercept and slope parameters with two independent normal distributions.

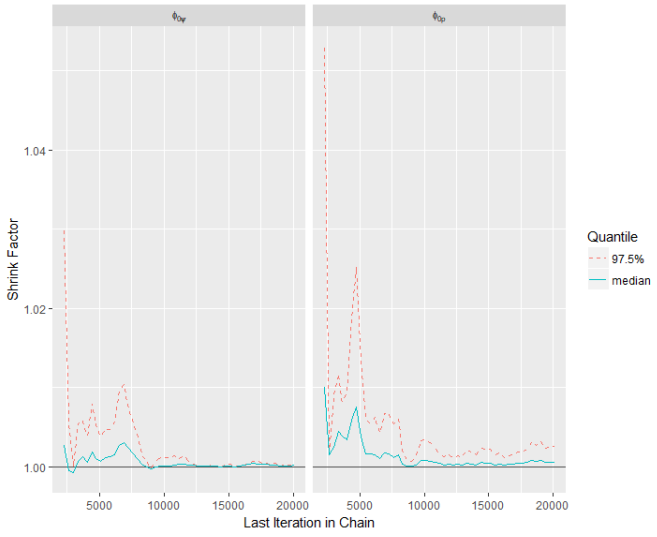


Figure 2.16: Gelman diagnostics for occupancy model variance parameters with two independent normal distributions.

In Summary, the odonata case so far represents a challenging scenario where individual species occupancy ranges widely within the community and detection probabilities are extremely low for some species (e.g. *A. affinis*). This emphasizes the importance of accounting for detection bias when modelling species odonata occurrence. Fitting the multispecies occupancy model and allowing for detection probabilities to be defined by species-specific covariates, decreases the uncertainty associated with each individual species. At this stage, site-level covariates have not been explored yet to explain the observed occupancy patterns. However, the observed patterns in local richness suggests possible environmental drivers that limit the distribution of species, especially in higher latitudes (e.g. temperature). Therefore, it is important to understand how species distribution and richness is affected by interacting driver and pressures at different spatial scales.

2.5 Final insights on modelling species detection bias

How many, where and why species occur, are the main questions Ecology tries to answer by studying how organisms' distributions and abundances are shaped by their interactions with the environment (Krebs, 1972). Whether the goal is to estimate parameters from a hidden state process that cannot be directly observed, or to quantify the importance of interactions or to predict how species' distribution patterns will be affected due to environmental changes, hierarchical models have become an important tool for ecology as they allow these complex scenarios where different sources of uncertainty may be involved to be described (Cressie et al., 2009). However, many challenges arise when building such complex models such as parameter identifiability and computational efficiency. The occupancy model discussed in this chapter was originally developed by (MacKenzie et al., 2002) to identify the unobserved process of a site becoming occupied by a given species from the observational process associated with the detection of that particular species. To estimate the parameters associated with these two processes, one must conduct several surveys at each of the sites in the sample. The simulation study in this work indicates that when detection probability is low (<25%) the amount of sites and visits to each of the sites required to reduce estimators' uncertainty increases dramatically (> 500 sites with ≈ 5 visits each), which will severely increase the costs for conducting the study. Because of this, species occurrence records are often constrained to opportunistic surveys at locations (or biological collections such as museum or herbariums) where individual species have been observed and no information regarding the absences is available. These presence-only records introduce a sampling bias since the inferred occupancy patterns can be clustered based on the localities in which human activities are concentrated instead of the sites that the species actually occupy, i.e. the estimates could be more related to the human activity centres rather than the true species distributions (Fithian and Hastie, 2013). Moreover, monitoring programs usually provides occurrence records for multiple species at the same sites. Thus, the estimators uncertainty will be greater for those elusive and rare species. However, the approach taken here based on the work by Dorazio et al. (2006), allows for species-specific effects to be drawn from the same prior distribution using species-specific covariates, which facilitates information exchange between species to reduce the

uncertainty for these estimates, i.e. inconspicuous species "borrow" information from the other species to provide more accurate estimates.

Bayesian MCMC methods become then a natural approach to fit these models due to their hierarchical structure. Fitting these models, however, can be computationally expensive and the results in this works suggest that `JAGS`, which is a popular Gibbs sampler to draw samples from the posteriors, is not very efficient, but `nimble` arises as an alternative to compile BUGS written code for which the computational time is equivalent to a C++ written sampler.

Regardless of the chosen approach to fit these models, one must acknowledge that the interpretation of the occupancy models' parameters depends on the spatial resolution of the study. This can be illustrated as follows, let the likelihood for a single-season occupancy be defined as:

$$L(\psi_j, p_{jk}) = \left[\prod_{j}^{J} \psi_j \prod_{k}^{K} p_{jk}^{y_{jk}} (1-p_{jk})^{1-y_{jk}}\right] \times \prod_{J+1}^{M} \left[\psi_j \prod_{k}^{K} (1-p_{jk}) + (1-\psi_j)\right]. \tag{2.42}$$

The first part of the likelihood corresponds to sites that had at least one detection. The second part refers to those $M-J$ sites where no species was detected ($y_{jk}=0$) in a particular chosen spatial scale. However, if a different spatial grid was chosen, the interpretation of the parameters made at the original scale would not be valid at the second spatial scale. This scale dependency can mislead the interpretation of the occupancy probabilities when the spatial scale used in the study is not specified (Fithian and Hastie, 2013). The occupancy models considered in this chapter were built under several assumptions that might not hold true under different scenarios such as the one of a closed system where the occupancy probability remains unchanged across different surveys and no local extinctions or colonization occur at the different sites. This lays important groundwork for exploring different extensions of this model that addresses different ecological questions of interest such as modelling spatial-temporal changes in occupancy patterns (Royle and Dorazio, 2008) by incorporating the spatial-temporal dependency between sites (Johnson et al., 2013), accounting for heterogeneous detection probabilities (MacKenzie et al., 2017), metapopulation modelling (Dorazio et al., 2006) and identification of hotpots for vulnerable species (Fattorini, 2009). The later question of interest will be the focus of the following chapter by identifying and quantifying those rare species that are more likely to be threatened by environmental change, while accounting for detection bias. Thus, the next chapter proposes a new approach to identify potentially threaten conspicuous species from those elusive but common species. Then, in chapters 4 and 5, the closure assumption will be relaxed by developing a spatio-temporal model to describe species distributions temporal changes while accommodating imperfect detection. Further discussion about modelling assumptions such as independence of occupancy and detection probabilities across sites will be further explored in chapter 5 (e.g. an approach to investigate spatial autocorrelation in both the state and observational models will be explored) and chapter 6 discusses some potential issues with these assumptions and possible areas of development in that regard.

Chapter 3

A new approach for studying rare species distribution in ecological communities

Understanding how species distributions are influenced by environmental changes has led to the development of a wide range of species distribution models that allow identification of the most important areas for biodiversity conservation (Cressie et al., 2009). These methods are usually based on scoring procedures that describe different attributes of an assemblage of species in a community (Leroy et al., 2012). However, individual species responses may vary widely within a community, making the task of identifying the ecological processes of interest that drives any observed occupancy patterns and changes difficult, especially for rare and elusive species (Bailey et al., 2014).

Rare species are often a point of interest for conservationists as they represent the most vulnerable species to environmental changes (Leroy et al., 2013). A species is considered to be rare when the range of their distribution or their abundances are low with respect to the range or distributional properties of other species within a comparable taxa (Blackburn and Gaston, 1997). Thus, species rarity is considered to be a scale dependent relative emerging property for a set of species (Flather and Sieg, 2007).

Species rarity is often used as a measurement to assess the risk of extinction of species (Hartley and Kunin, 2003). Species with low abundances or limited distribution ranges are prone to experience local extinctions as their genetic pool may be reduced due to population isolation. This can lead to higher rates of inbreeding, expression of deleterious genes and outbreeding depression that reduces the species' ability to adapt to changes in environmental conditions (Ellstrand and Elam, 1993; Karron, 1997). Thus, several studies have developed methods to quantify the rarity of a species at the community level, but none has accounted for detectability bias. For instance, Rabinowitz (1981) proposed a classification system that has been applied in several conservation studies (Broennimann et al., 2005; Isaac et al., 2009; Yu and Dobson, 2000) to categorize species by different types of rarity and commonness based on the density of their local populations, the area of the species range and, the number of different types of habitats each species occupies. However, abundance and habitat specificity are not always available, specially for invertebrates such as dragonflies (Leroy et al., 2013), which makes the task of identifying rare species by this method difficult. Fagan et al. (2005) and Fattorini (2009) also proposed methods to

assess species rarity and determine hotspots by combining information from multiple scales about species occupancy and the International Union for Conservation of Nature (IUCN) red list status of species (which is often lacking for most invertebrates species). Hence, to quantify rare species in a community, Leroy et al. (2012) proposed a multiscale index of relative rarity (IRR) that enables comparison of species rarity regardless of the spatial scale, geographic area or taxonomic group. In this chapter, Leroy's index of relative rarity (IRR) will be developed in a Bayesian setting while accounting for detection bias. Furthermore, a Bayesian occupancy mixture model is proposed as an novel alternative method to quantify species rarity to identify areas relevant for conservation. Finally, an novel 2-stage modelling approach was developed to estimate the effect that connectivity metrics and anthropogenic stressors have on the compositions of rare species while accounting for detectabilty.

3.1 Bayesian Index of relative rarity

To quantify rare species in a community, Leroy et al. (2012) proposed an index of relative rarity (IRR) that allows for the composition of rare species in different assemblages to be compared. The index is computed by assigning a rarity weight to each species based on the observed occupancy and a rarity cut off point which is determined by the user. Rarity weights increase exponentially for those species for which occurrences fall below the chosen cut-off point (Fig.3.1). In this research, we adopt a Bayesian approach to estimate the i-th species' rarity weight as a derived quantity of the multispecies occupancy model. For this, let $Q_{is} = \sum_j z_{ijs}$ the total number of sites where species i is estimated to be present (z_{ij} denotes the latent variable for the true presence of species i on site j) for the sth MCMC sample. Then, rarity weights are assigned to each species based on Q_{is} and a rarity cut off point (Eqn. 3.1).

$$w_{is} = exp\left(-\left(\frac{Q_{is} - Q_{s\min}}{r_s \times Q_{s\max} - Q_{s\min}} \times 0.975 + 1.05\right)^2\right), \tag{3.1}$$

where $Q_{s\min}$ and $Q_{s\max}$ are the lowest and highest occurrences found among species for each s sample and r_s is the cut off point suggested by Leroy et al. (2013) to be the point for which the mean proportion of rare species is 25%, i.e. $\text{quantile}_{25}(Q_{is})/Q_{s\max}$. The constants in the index adjust the species weight to be equal to 5% of the maximum weight at the rarity cut-off point.

The rarity index (IRR) in equation (3.2) is then computed as the weighted sum of each individual species' weight over the observed local species richness $R_{js} = \sum_i z_{ijs}$ for each sample, i.e.

$$IRR_{js} = \frac{\frac{1}{R_{js}} \sum_i \left[w_{is} \times \mathbb{I}_{z_{ijs}=1}\right] - w_{js\min}}{w_{js\max} - w_{js\min}}, \tag{3.2}$$

where $w_{js\min}$ and $w_{js\max}$ are the minimum and maximum of the weights of the species estimated to be present at site j, $\mathbb{I}_{z_{ijs}=1}$ is an indicator variable that takes the value of 1 if species i is present at site j and zero otherwise. An IRR value close to 0 means that all species of the assemblage have the minimum weight, i.e. the site is comprised of widespread species and in consequence, an IRR close to 1 indicates

that the site's assemblage is composed by very rare species. Thus, a novel contribution of this work is to model the IRR as a derived parameter from the Bayesian occupancy model state process that can be used to identify sites where rare species assemblages are dominant while accounting for detectability.

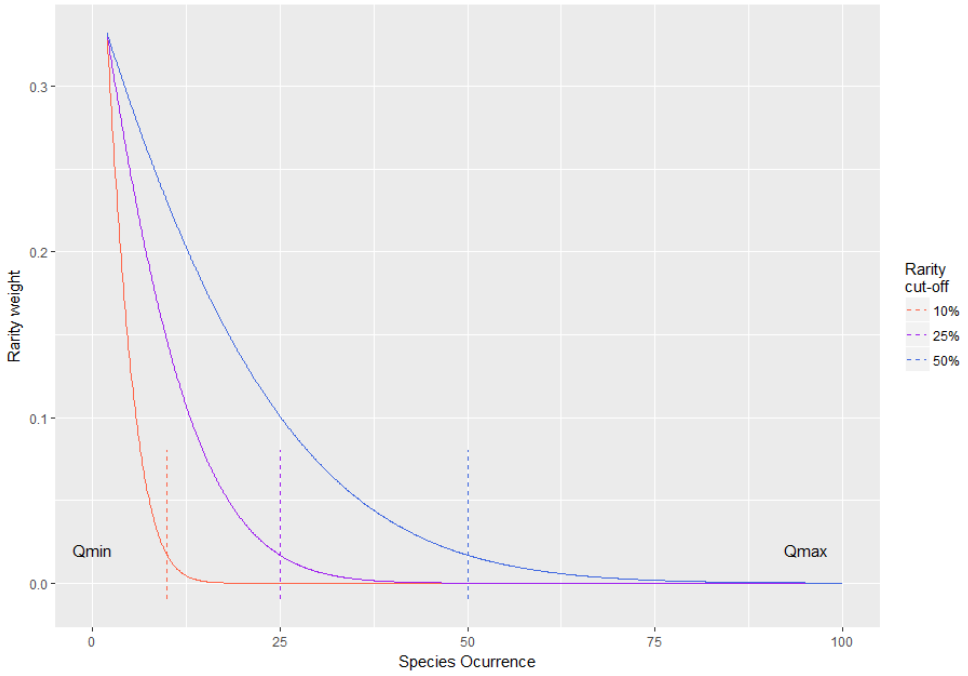


Figure 3.1: Weight assignation curves adjusted to different rarity cut-off points modified from Leroy et al. (2013).

Detection probability plays an important role in computing the IRR, this is illustrated in figure 3.2 by simulating a community of 50 species that occur at 300 sites surveyed 3 times with a constant occupancy probability of 0.8 and varying detection probabilities (0.25, 0.50, 0.75). As detection probability decreases, the difference between the IRR computed on the observed occurrences and that from the true occurrences becomes greater. However, choosing the right cut-off point that better describes the rarity in the system might be difficult when no prior information regarding the species rarity status across a community is available. This is portrayed in Figure 3.3 by estimating the IRR with the aforementioned methodology under different cut-off thresholds (non-informative Normal (0, 0.001) priors were specified for mean occupancy, detection and variance parameters, a total of 10000 iterations, 3 chains and a burnin period of 2000 were used to fit the occupancy model). The estimated IRR are closer to the equality line when a cut-off of 25% is specified and move further away for the 15% and 50% thresholds. This highlights the important role that the cut-off parameter plays when modelling the index; although the 25% threshold suggested by Leroy et al. (2012) seems to work for these simulations, this might not be always the case for real communities where the proportion of rare and widespread species is unbalanced. Thus, in the next section, odonata rare species composition is going to be analysed using the Bayesian IRR to identify possible hot spots where rare species occur and then, a finite mixture model alternative is proposed to classify and quantify species rarity assemblages based on their estimated occupancy and detection probabilities.

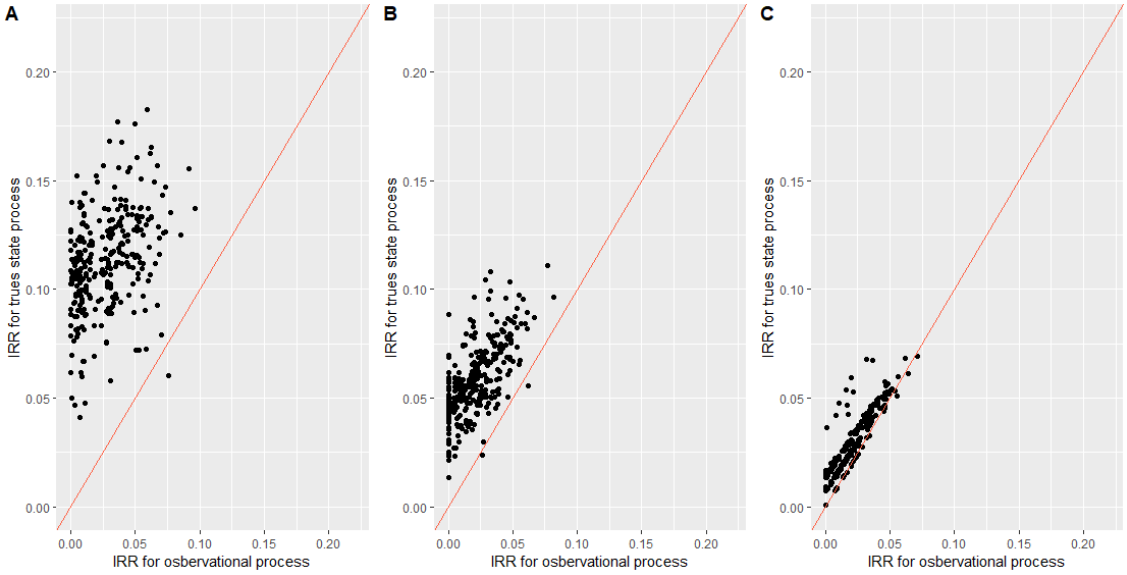


Figure 3.2: IRR computed based on observed occupancy and true occupancy for varying detection probabilities. Plot A shows the relationship for an IRR when detection probability is 25%, plot B when detection probability is 50% and plot C when probability of detection is 75%.

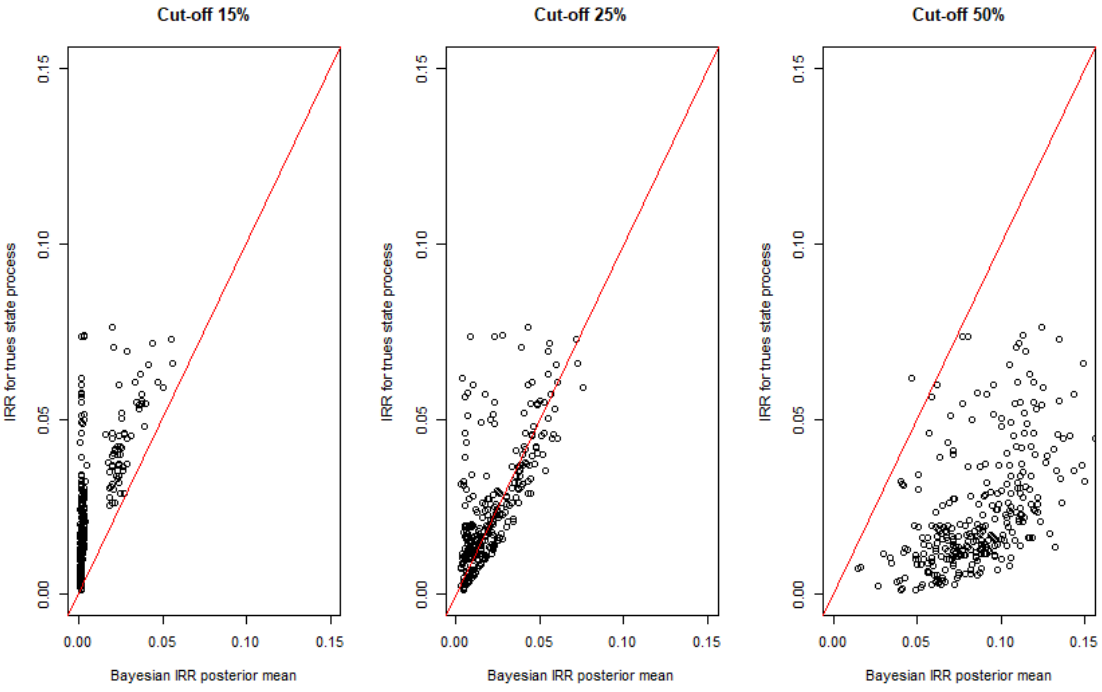


Figure 3.3: Relationship between IRR computed on the true occurences and estimated IRR based on occupancy model for a simulated community of 50 species that occur at 300 sites surveyed 3 times with a constant occupancy probabilty of 80% and a detection probability of 75%.

3.1.1 Index of relative rarity to describe odonata rare species composition

The Odonata case study reviewed in chapter 1, represents a clear example in which a community consists of a broad range of rare, elusive and widespread species. Thus, it is important to account for imperfect detection in order to describe the occupancy patterns for those rare species. The approach developed in this work to incorporate species detection uncertainty into the index of relative rarity proposed by Leroy et al. (2013) enables the composition of rare species across different sites in the UK to be explored. To do so, a species rarity weight w_{ij} (Eqn. 3.1) was computed based on the number of j sites each species i is estimated to occupy ($\sum_j z_{ijs}$) for each of the s MCMC samples derived from the multispecies occupancy model (Eqn. 2.40) defined in section 2.4.2. Table 1 (Appendix B.2) shows the estimated weights for each species based on the MCMC samples used to compute the rarity index IRR (Eqn. 3.2) for each site assemblages (i.e. occurring species at each site). Figure 3.4 shows the estimated IRRs for each grid cell. The estimated IRRs are quite low and homogeneous for all sites meaning that most of the sites are occupied by widespread species. This however, as shown by simulations above, may be the result of specifying a cut-off that is too restrictive in which species should receive a higher rarity weight. Thus, Figure 3.5 shows estimated IRRs with different cut-off points. Even with a cutoff point of 50%, most of the sites appear to be dominated by the occurrence of widespread species.

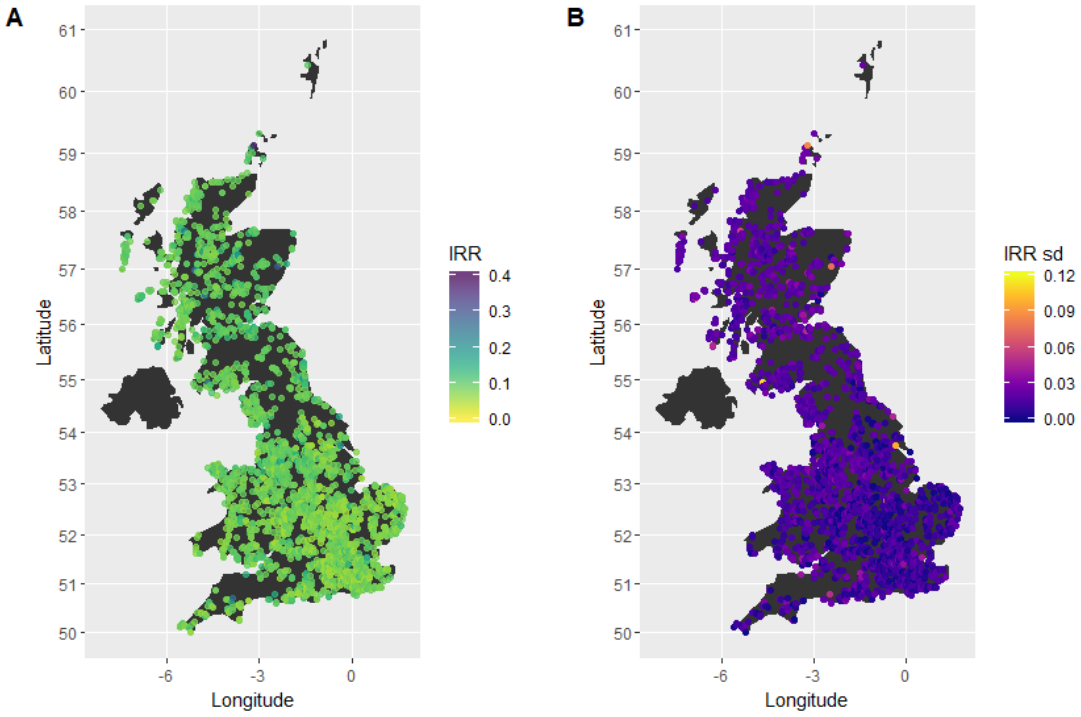


Figure 3.4: Estimated IRR (left) and posterior standard deviation (right) for each grid cell for odonata occupancy in the UK .

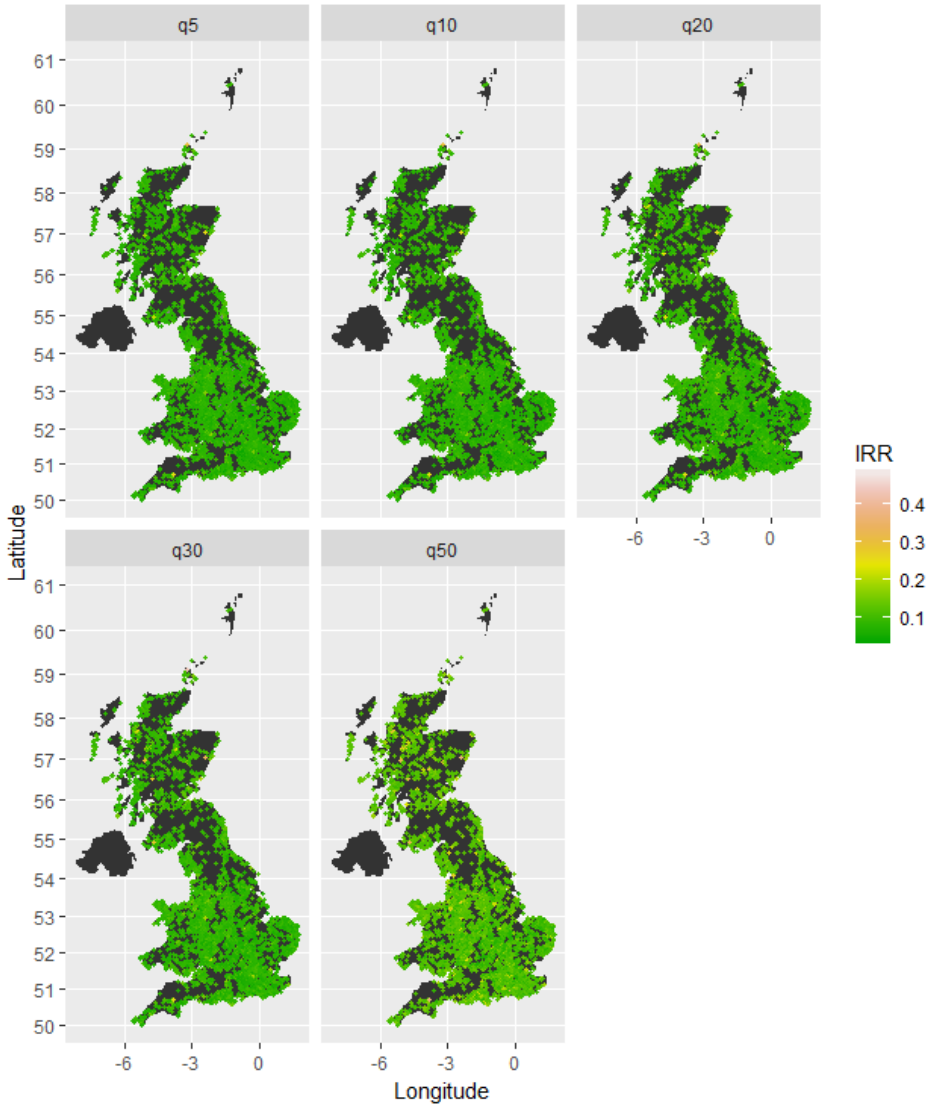


Figure 3.5: Estimated IRR for each grid cell for Odonata occupancy in the UK under different cutoff points ($q_i \quad i \in 1,5,10,20,30,50$ for 5%, 10%, 20%, 30% and 50% cut-off points respectively).

3.2 Bayesian occupancy mixture model

Development and application of mixture models have been explored in a wide range of fields because of their flexibility to model situations in which the population of interest is a mixture of sub-populations for which the components' identity information is not known. Thus, mixture models have gained increasing attention in cluster analysis, particularly in latent class analysis in which the population is a finite mixture of latent classes (Geary, 1989). Finite mixture distributions assume that the population is distributed among H different non-overlapping groups and various methods have been developed to estimate the mixture model parameters that define these groups (McLachlan and Krishnan, 2007). It is Bayesian methods, however, that have become a popular tool for fitting these models (Li et al., 2018). Particularly, finite mixtures have been developed for occupancy models to account for heterogeneous detection probabilities between surveys (MacKenzie et al., 2017). Thus, a natural extension is to perform a latent class analysis based on the occupancy state for each species to identify those rare species in a community. Let a finite mixture model be defined as,

$$f(\mathbf{x}|\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{h=1}^{H} \pi_h f(\mathbf{x}|\theta_i), \tag{3.3}$$

where $\mathbf{x} = x_1,\ldots,x_n$ are assumed to arise from H different classes such that $\sum_{h=1}^{H}\pi_h = 1$ with a probability given by π and $f(\mathbf{x}|\theta_i)$ be a probability density function with unknown parameters θ_i. Bayesian estimation using MCMC methods is based on specifying a latent class model that associates each x_i with a latent variable $\zeta_i \in \{1,\ldots,H\}$. The latent class variable ζ_i is sampled from a categorical distribution such that:

$$\begin{aligned} \zeta_i &\sim \text{Categorical}(H,\pi), \\ \zeta_i|\pi &\sim \text{Multinomial}(\pi_1,\ldots,\pi_H). \end{aligned} \tag{3.4}$$

Thus, the likelihood in Eqn. (3.3) which depends on the ζ_i becomes:

$$f(x_i,\zeta_{hi}|\pi_h,\theta_{hi}) = \prod_{h=1}^{H}\left[\pi_h f(x_i|\theta_{hi})\right]^{\zeta_{hi}}. \tag{3.5}$$

This leads to the joint posterior:

$$p(\pi,\theta|x,\zeta) \propto p(x,\zeta|\pi,\theta)p(\pi)p(\theta), \tag{3.6}$$

from which a Gibbs sampler can be implemented by setting priors for θ and π (see section 3.2.1 for details). The hierarchical structure of the occupancy model allows incorporation of the latent variable $\zeta_i = h$ for those species belonging to a class $h \in (1,2,\ldots,H)$.

By grouping species in classes, the relative frequency of the number of species in each class can be expressed as a proportion of the estimated local species richness (see equation (3.10)).

As discussed in chapter 2, species-level parameters in multispecies occupancy models are drawn from a common distribution with hyper-parameters representing the parameter's average value among all the individual species in a community Dorazio and Royle (2005); Dorazio et al. (2006). Species heterogeneity in occupancy and detection probabilities is usually modelled using logit-normal model (Royle, 2006). In addition to the species heterogeneity described by the logit-normal model, adding the finite mixture component enables occupancy parameters of the ecological process to be defined based on the latent classes to which each species belongs. This allows the composition of a community to be characterized in terms of how common or rare species are. Thus, a multiple species occupancy model can be formulated as follows:

$$
\begin{aligned}
\left.\begin{aligned} \zeta_i &\sim \text{Categorical}(H, \pi_h) \\ z_{ij}|\zeta_i, \psi &\sim \text{Bernoulli}(\psi_{hi}) \end{aligned}\right\} \quad &\text{(a)} \\
\sum_{K_j} y_{ij\cdot}|z_{ij}, \zeta_i, p_i \sim \text{Binomial}(K_j, p_i z_{ij}) \quad &\text{(b)} \\
\left.\begin{aligned} \text{logit}(\psi_{hi}) &\sim \text{Normal}(\mu_{\psi_h}, \sigma^2_{\psi_h}) \\ \text{logit}(p_i) &\sim \text{Normal}(\mu_p, \sigma^2_{p_i}) \end{aligned}\right\} \quad &\text{(c)}
\end{aligned}
\tag{3.7}
$$

Equation 3.7 (a) denotes the ecological/state process where the latent variable ζ is a categorical random variable relating the i-th species' occupancy (z_{ij} for $i = 1,\ldots,S$ species and $j = 1,\ldots,M$ sites) given by the occupancy probability ψ_{hi} to a specific class h such that $\Pr(\zeta = h) = \pi_h$ for $h = 1,\ldots,H$ and $\pi = \pi_1,\ldots,\pi_H$. Then, the observational process driven by the detection probability p_i (equation 3.7 (b)) can be formulated as the total number of times species i was detected at site j across K_j visits (where the number of visits are defined by the different sampling occasions in which a species was recorded). Finally, the species heterogeneity model (equation 3.7 (c)) is characterized by the individual species logit-scaled occupancy and detection probabilities drawn from the same normal prior distribution with hyperparameters describing the overall community response. It is important to notice that logit-scale Normal distributions can overlap to different degrees depending on the mean and variance of each class which affects the model's efficiency to correctly identify the members of each latent class (Sollmann et al., 2021). Issues with the model's performance can also rise if the density mass is heavily skewed due to an imbalanced number of observations allocated to certain classes. Thus, section 2.3 introduces a simulation study to investigate the effect of different degrees of overlapping and class imbalance on the model's performance.

Mixture models parameters also suffer from lack of identifiability due to the invariance of the posterior distribution to permutations of the group labels (Redner and Walker, 1984; Richardson and Green, 1997). To make model parameters identifiable and avoid label-switching issues, components means can be ordered, i.e. $\mu_{\psi,1} < \mu_{\psi,2} < \ldots < \mu_{\psi,H}$. Then, by tracking ζ on each MCMC sample s draw, the posterior probability of each species being classified into the h-th category is given by:

$$\begin{aligned} \Pr(\text{species } i \text{ belongs to } h) &= \Pr(\zeta_i = h|\pi, z, \psi) \\ &\approx \frac{1}{S}\sum_{s}^{S} \mathrm{I}(\zeta_i^{(s)} = h), \end{aligned} \tag{3.8}$$

where $\mathrm{I}(\zeta_i^{(s)} = h)$ denotes an indicator variable that takes the value of one if species i belongs to cluster h and zero otherwise. Species can be clustered by assigning them to a class based on these posterior probabilities as follows:

$$\text{argmax}_h \Pr(\zeta_i = h|\pi, z, \psi) \tag{3.9}$$

Additionally, the class relative frequency at each site (η_{hj}), expressed as a proportion of the local species richness (i.e. $\sum_i z_{ij}$), can be computed as a derived quantity by tracking the h-th class membership and the occupancy status for every species on each MCMC draw:

$$\eta_{hj}^{(s)} = \sum_i z_{ij}^{(s)^{-1}} \sum_i z_{ij}^{(s)} \mathrm{I}(\zeta_i^{(s)} = h), \text{ such that } \eta_{hj} \in [0,1]. \tag{3.10}$$

By adopting a Bayesian inference approach, the proportion of species in each class can be computed as a derived quantity of the occupancy model at each of the sampled sites in the study. In the Odonata case study, this enables inference about the occupancy pattern of rare species to be limited to only the sites in the sample. This is of particular interest since these sites represent lakes and ponds, which are key components of the hydrological network in the UK.

3.2.1 Introducing species-specific effects and site-level covariates

The Odonata case study contains information about species-specific traits that could be associated with the species detection. Thus, these species-specific effects can be specified as a linear model for the mixture occupancy model detection hyperparameters as follows:

$$\mu_p = \gamma_0 + \sum_{m=1}^{T} \gamma_m (m\text{-th species trait covariate}). \tag{3.11}$$

Where the mean detection probability μ_p of each species is a linear function of $T = 3$ regression coefficients γ_t associated with each species traits and an intercept γ_0 representing the baseline detection probability.

The mixture occupancy model described in Eqn.(3.7), has great flexibility and can be adapted to different scenarios depending on the question of interest. For example, time and space varying occupancy and detection probabilities can be incorporated through the logit function in Eqn. 3.7 (c) by either including a site and year random effects (see Outhwaite et al. (2018)) or by having distinct site-level predictors that affect species occupancy and detection Wintle and Bardos (2006) (e.g. $\text{logit}(\psi_{hij}) = \beta_{0hi} + \sum_{m=1}^{P} \beta_{him} x_{jm}$ where logit-scaled occupancy probabilities are defined as a function of P site-level covariates).

A similar approach by Dunstan et al. (2011), implemented finite mixture models in a frequentist setting to capture heterogeneity in species responses to environmental gradients among different latent classes. However, by adopting a Bayesian inference approach the latent variables can be retained while accommodating imperfect detection. Moreover, the hierarchical structure of the occupancy model enables information among the species in the community to be exchanged within each class. Therefore, assuming all species within a particular class (e.g. rare species) are related to one another by being part of the same biological community facilitates the estimation of parameters of species with sparse occurrence records (Dorazio et al., 2011).

To estimate the occupancy mixture model parameters, Dirichlet conjugate prior are specified for π (i.e. $p(\pi) \propto \prod_i \pi_h^{\alpha_h - 1}$) such that the posterior is sampled from $\pi|\zeta$ Dirichlet$(\sum_i \zeta_{1,i} + \alpha_1, \ldots, \sum_i \zeta_{H,i} + \alpha_H)$, where $\sum_i \zeta_{h,i}$ is the number of species assigned to each class). Then, vague normal priors, logistic(0,1) or weakly informative zero centred t-distributed priors with scale parameter of 1.566 and degrees of freedom 7.763 can be specified for the mean hyperparameters (see chapter 4 for a detailed discussion on prior specifications) and inverse-gamma (conjugate prior), Uniform(0,5) or Half-Cauchy priors for the variance hyperparameters (Outhwaite et al., 2018).

3.2.2 Fitting a two-class finite mixture

Note that for this work, only a binary classification problem will be addressed since the aim is to distinguish rare from common species (under the assumption that the number of classes is fixed and known before conducting the analysis, and that the choice of the number of classes is based on the ecological context of the problem only but the method can be easily generalized to multi-class problems). Thus, the likelihood in (3.5) can be simplified to

$$f(x_i) = \pi f(x_i|\theta_{1i}) + (1-\pi) f(x_i|\theta_{2i}), \tag{3.12}$$

allowing Beta (a_1, a_2) priors for π to be specified.

For this work, the sensitivity of the priors was tested by comparing our model results under the different aforementioned priors parametrizations. Results were consistent when either logistic(0,1) and zero-centered t($\sigma = 1.566, \nu = 7.763$) distributed priors were specified for the mean hyperparameters. Specifying such priors instead of vague normal priors avoids the need for calibrating the normal prior precision parameter which can often be a problem when logit scale transformation is used, as vague normal priors (e.g. Normal(0,500)) lead to a high probability density around zero and one on a probability scale. Moreover, Uniform(0,5) and Half-Cauchy priors for the variance hyperparameters showed an overall better mixing and lower autocorrelation compared to inverse-gamma priors. Finally, Dirlichet (10,10) priors were specified for the mixing parameters. Convergence graphical diagnostics for this analysis are presented in the appendix B.2.2.

For each analysis a total of 50,000 iterations were run with a burnin period of 10,000 and a thinning of 10 (for memory optimization purposes only) on three independent Markov chains. The algorithm's convergence was assessed through conventional graphical diagnostics (i.e. posterior samples traceplot) showing overall good mixing with low posterior autocorrelation, and Gelman-Rubin between-within chains variance ratio < 1.1.

All of the modelling work and data manipulation was implemented in R while using the `ggplot` and `ggmcmc` packages for visualization of model results and graphical examination of the algorithm convergence (Fernández-i Marín, 2016; Hadley, 2009). The proposed models were run in `nimble` (de Valpine et al., 2017).

3.2.3 Simulation study: evaluating method performance for quantifying species rarity

Based on the Odonata case study, a simulation study was designed to test the occupancy mixture model performance in which four distinct species classes were defined to simulate a community integrated by a combination of common, rare, elusive, and non-elusive species. The model efficiency to differentiate between common and rare species was tested under varying occupancy, detection, and mixing probabilities by assessing if the proposed model could identify correctly those elusive-common species that could be mislabeled as rare due to their low detection probabilities.

To generate the aforementioned community, a total of $S = 50$ species were simulated in $M = 300$ sites visited on 4 different occasions. Species-specific occupancy and detection probabilities were drawn from a multivariate normal distribution as follows:

$$\Omega \sim \sum_{h=1}^{4} \pi_h \text{Normal}\left(\mu_h = \begin{pmatrix} \mu_{\psi_h} \\ \mu_{p_h} \end{pmatrix}, \Sigma_h\right), \tag{3.13}$$

where the species specific occupancy and detection probabilities denoted by $\text{logit}^{-1}(\Omega) = (\psi_{hi}, p_{hi})$ and π_h mixing probabilities that determine the proportion of species belonging to each class. These species-specific parameters were drawn from the community logit-scaled baseline occupancy and detection probabilities μ_{ψ_h} and μ_{p_h} respectively. Note that the simulated occupancy and detection probabilities ranged between 0.05 and 0.95 across species. This captures a reasonably wide span of different species responses that matches what we observed in the Odonata case study. Σ_h is the covariance-variance matrix for the h-th class with diagonal elements $(\sigma^2_{\psi_h}, \sigma^2_{p_h})$ corresponding to variances for the logit-scaled community mean occupancy and detection probabilities respectively. The off-diagonal elements of the covariance-variance matrix Σ_h were set to zero to remove the abundance-induced detection effect (i.e. when detection probabilities are influenced by the widespread species high abundances), which is assumed to be captured by specifying the number of different habitats each species occupy in our Odonata occupancy mixture model (see equation 3.11 in section 3.2.4 below). Figure 3.6 illustrates the general framework used to simulate the following scenarios.

Simulation scenario 1: Non-overlapping with constant variance and proportional allocation

First, four well-separated classes were simulated by specifying a reasonable distance between the centroids of each cluster defined by the following community occupancy and detection probabilities:

$$\begin{pmatrix} \mu_{\psi_h} \\ \mu_{p_h} \end{pmatrix} = \begin{pmatrix} 0.15 & 0.15 & 0.75 & 0.75 \\ 0.15 & 0.75 & 0.15 & 0.75 \end{pmatrix}. \tag{3.14}$$

On a probability scale, each of the 4 different groups represent:

- "rare and elusive species" ($\mu_{\psi_1} = 0.15, \mu_{p_1} = 0.15$)
- "rare but non-elusive" ($\mu_{\psi_2} = 0.15, \mu_{p_2} = 0.75$)
- "common and elusive" ($\mu_{\psi_3} = 0.75, \mu_{p_3} = 0.15$)
- "common and non-elusive" ($\mu_{\psi_4} = 0.75, \mu_{p_4} = 0.75$)

Mixing probabilities were set to be $\pi_h = 1/4 \; \forall \; h$ in order to retain the same proportion of species of each class. Homogeneous variances were defined for all classes by setting $\Sigma_h = \Sigma$ with diagonal entries $\sigma^2_\psi = \sigma^2_p = 0.5$.

Simulation scenario 2: Moderate overlapping with variance heterogeneity under constant and varying mixing probabilities

For a second simulation, a moderate degree of overlapping between classes was induced by allowing (i) different variances for each class and (ii) by specifying cluster centroids to be closer to one another. The mean occupancy and detection probabilities that defined the classes centroids were:

$$\begin{pmatrix} \mu_{\psi_h} \\ \mu_{p_h} \end{pmatrix} = \begin{pmatrix} 0.20 & 0.20 & 0.70 & 0.70 \\ 0.20 & 0.70 & 0.20 & 0.70 \end{pmatrix}. \tag{3.15}$$

The model accuracy was assessed under constant and varying mixing probabilities by setting $\pi_h = 0.25 \; \forall \; h$ and $\pi_h = (\pi_1, \pi_2, \pi_3, \pi_4)$ such that $\pi_1 + \pi_2$ is the probability of a species being rare, $\pi_3 + \pi_4$ the probability of a species being common, $\pi_1 + \pi_3$ the probability of a species being elusive and $\pi_2 + \pi_4$ the probability of a species being non-elusive. Model results were compared for this scenario when (1) the odds of a species being common were 4 times the odds of being rare and (2) when the odds of a species being elusive were 4 times the odds of being non-elusive. For example, case (1) was defined as $(\pi_3 + \pi_4)/(\pi_1 + \pi_2) = 4$ such that $\pi_3 = \pi_4$ and $\pi_1 = \pi_2$.

Finally, to induce variance heterogeneity the following variance-covariance matrix was defined: $\Sigma_h = \Sigma \mathbb{I}_{2\times 2} \times U_h$ where $U_h = (u_{1h}, u_{2h})$ are two uniform (0,1) random variables and Σ a diagonal matrix defined by $\sigma_{\psi_h} = 0.25$ and $\sigma_{p_h} = 0.25$.

Simulation scenario 3: Strong overlapping with variance and mixing heterogeneity

For the third scenario, a greater degree of overlapping was induced by setting (i) similar detection and occupancy probabilities for each class, (ii) variance heterogeneity and (iii) a different proportion of species allocated to each class. The mean occupancy and detection probabilities that defined the classes centroids were:

$$\begin{pmatrix} \mu_{\psi_h} \\ \mu_{p_h} \end{pmatrix} = \begin{pmatrix} 0.4 & 0.4 & 0.6 & 0.6 \\ 0.3 & 0.6 & 0.3 & 0.6 \end{pmatrix}. \tag{3.16}$$

The mixing coefficients specified to have unequal number of species of each class and the covariance-variance matrix for each class were the same as the one described for scenario 2.

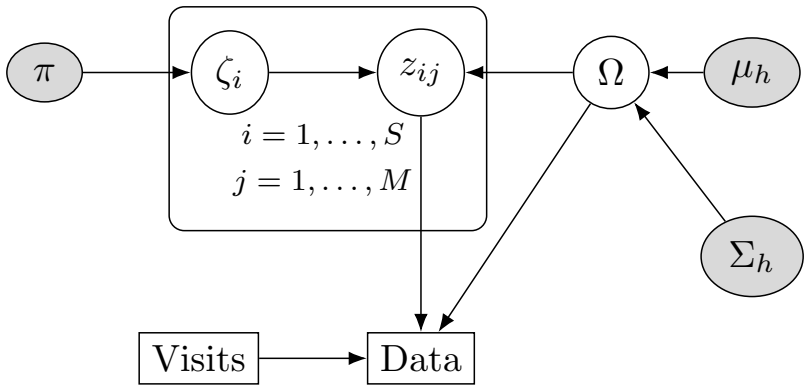


Figure 3.6: Direct acyclic graph illustrating the simulation scheme under varying stochastic parameters indicated by the shaded nodes. The square box represent the model's latent state process with $S = 50$ species and $M = 300$ sites. Square nodes represents simulated data while shaded nodes are the parameters determining each simulated scenario. π are the mixing probabilities, ζ_i and z_{ij} are the latent variables for the ith species rarity class and occupancy state at the jth site respectively and Ω is a multivariate normal distribution with mean μ_h and covariance Σ_h for the hth class.

Model 3.7 was fitted to the simulated data for each scenario using the settings described in section 3.2. The correct classification rate, describing the overall proportion of species that were classified correctly, was then calculated for each simulation as:

$$CCR = \frac{1}{S}\sum_{i}^{S} \mathbb{I}(\zeta_i, \hat{\zeta}_i), \tag{3.17}$$

where $\mathbb{I}$ is an indicator variable such that $\mathbb{I}(\zeta_i, \hat{\zeta}_i) = 1$ if $\zeta_i = \hat{\zeta}_i$ and zero otherwise, $\zeta_i \in 1,2$ is the i-th species true class (rare or common) and $\hat{\zeta}_i$ is the predicted class according to equation (3.8) for $i = 1, \ldots, S$ species. Moreover, additional standard model performance metrics were calculated based on the confusion matrix shown in Figure 3.7. The diagonal elements of the confusion matrix are given by the number of rare and common species that have been correctly predicted as such (true positives (TP) and negatives (TN) respectively). The off-diagonals elements of the confusion matrix contain the errors between classes, i.e. the false positives (FP) - $\Pr(\hat{\zeta}_i = 2 | \zeta_i = 1)$ and false negatives (FN) - $\Pr(\hat{\zeta}_i = 1 | \zeta_i = 2)$. The sensitivity (i.e. the true positive rate TPR) was calculated as TP/(TP+FN) to describe the proportion of rare species that have been classified correctly among the rare class. Similarly, the specificity (true negative rate TNR) characterized by how many species were classified as common out of the total common species was calculated as TN/(TN+FP). Additionally, model precision was described based on (1) the positive predictive value calculated as PPV = TP/(TP+FP), i.e. the proportion of true rare species among all the species that have been predicted to be rare and, (2) the negative predictive value NPV = TN/(TN+FN) describing how many species predicted to be common are truly common. Cohen's Kappa statistic was also calculated as a measure of the model's performance relative to what would be expected by chance. Finally, the F-score (harmonic mean of the sensitivity and PPV) and the Balanced accuracy (arithmetic mean between specificity and sensitivity, useful when comparing classes with an unbalanced number of observations) were used as metrics to assess the overall model performance.

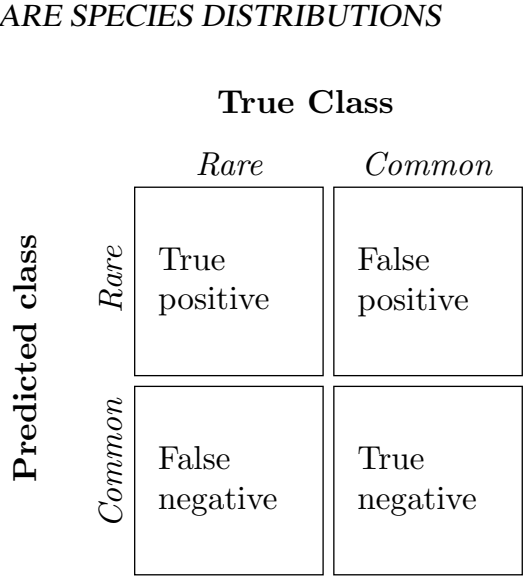


Figure 3.7: Confusion matrix for a two-class classification problem.

Simulation study results in Table 3.2.3 show the standard classification performance metrics for each simulated scenario. Overall, our results suggest good model performance in identifying those rare and widespread species with an $\approx 80\%$ of accuracy across all scenarios. For scenario 1, i.e. well-separated clusters with same variance and equal proportion of species per group, the estimated probability of belonging to either rare or common species was $\hat{\pi}_1 = 0.62$ and $\hat{\pi}_2 = 0.38$, respectively. This model has a 100% accuracy (1 CCR) as it correctly classified all rare and common species with respect to the true categories. Scenario 2, which simulates clusters with a moderate degree of overlap by specifying groups centroids closer to each other with different variance per class, has an CCR of 0.96 when the mixing probabilities are constant for each class. The TPR (sensitivity) of 0.93 indicates that the model correctly predicted 93% of the total number of rare species. Moreover, a PPV of 1 implies that all the predicted rare species were truly rare. The TNR (specificity) on the other hand, indicates that all common species were correctly identified as such. However, an NPV of 0.92 means that very few rare species were predicted as common. The κ statistic, balanced accuracy and F-score suggest a very good performance of our model under this scenario.

The simulation results for scenario 2 also suggest that classification performance is affected by groups with an unbalanced number of species, especially when the proportion of common species is greater than the proportion of rare species. For example, when the proportion of common species was simulated to be 4 times greater than the proportion of rare species, the TNR and PPV were 0.83 and 0.70 respectively, suggesting that 83% of species were classified as common out of the total number of common species. The latter because some common species have been predicted as rare (i.e. only 70% of predicted rare species were actually rare).

Note that none of the species predicted as common were rare (as the NPV of 1 suggests). Also, the TPR indicates that there were no rare species predicted as common (i.e. there were no false negatives).

The F-score shows a lower value compared to CCR and balanced accuracy because it does not consider the true negatives. Thus, the balanced accuracy is a more appropriate metric to assess the scenarios with unequal proportion of species per class as it is based on both the TPR and TNR. For the second scenario the balanced accuracy suggests a very good classification performance when either equal or unequal proportion of species are allocated to each class.

On the third scenario with proportional allocation and stronger overlapping between classes, the accuracy decreased to $\approx 80\%$. The different classification performance metrics were approximately 10-20% lower than the metrics in scenario 2. Likewise, having an unbalanced number of species for each class yields to a much lower accuracy. Particularly, when the number of common species is greater than the number of rare species, both the PPV and TNR suggest an overestimation of the true number of rare species. However, since the proportion of predicted common species that were truly common is reasonably high (0.93) the balanced accuracy is still close to 0.8. Finally, when the number of elusive species was higher than the number of non-elusive species, we can see an important decrease in the TPR, suggesting that several rare species are being classified as common (only 67% of the species are classified as rare out of the total number of rare species). Nevertheless, the PPV indicates that from this set of predicted rare species 86% are truly rare. The TNR and the NPV suggest that proportion of common species predicted as common is reasonably high. In summary, the accuracy, F-score and κ statistic suggest a reasonably good performance of the mixture occupancy model as a classifier when there is an balanced number of observations for each class. When there is an unbalanced number of observations, the balanced accuracy suggest a very good performance of our model for scenario 1 and 2 and a moderate performance under scenario 3. Particularly, our model performance under scenario 3 depends on the different class proportions, i.e. a greater number of common species yields to an overestimation of the true number of rare species, while having more elusive and non-elusive species results in an underestimation of rare species.

Table 3.1: Occupancy mixture model classification performance metrics under different simulated scenarios.

Scenario	Proportion of species per class	Overlapping	CCR	TPR	TNR	PPV	NPV	κ	Balanced Accuracy	F-Score
1	Equal proportion	No	1	1	1	1	1	1	1	1
2	Equal proportion	Moderate	0.96	0.93	1	1	0.92	0.92	0.96	0.96
	More Common than rare		0.88	1	0.83	0.70	1.00	0.74	0.92	0.82
	More elusive than non eluisve		0.94	0.92	0.96	0.96	0.93	0.88	0.94	0.94
3	Equal proportion	Strong	0.80	0.64	1	1	0.69	0.61	0.82	0.78
	More common than rare		0.76	0.86	0.72	0.55	0.93	0.49	0.79	0.67
	More elusive than non elusive		0.78	0.67	0.89	0.84	0.74	0.56	0.78	0.74

Despite the occupancy mixture model 3.7 showing a lower correct classification rate for scenarios with an unbalanced proportion of each class, the next comparison illustrates the advantages of using the mixture occupancy model opposed to the Bayesian IRR reviewed in section 3.1. Both methods were fitted to the simulated scenario 2 with proportional allocation. The mixture occupancy model provides the advantage of assigning a group membership to each individual species whereas IRR only assigns each species a weight based on the highest estimated occurrences among all the species in the sample. Under this scenario, the mixture occupancy model had an accuracy of 96% of correctly classifying rare species. The weights on the other hand, depend on the *i*th species highest occurrence and the cut-off point. Fig. 3.8, shows the histogram for the estimated weights for those species that are known to be rare. As the threshold decreases, more species that are known to be rare receive a lower weight, leading to an underestimation on the true composition of rare species for the site-level IRR.

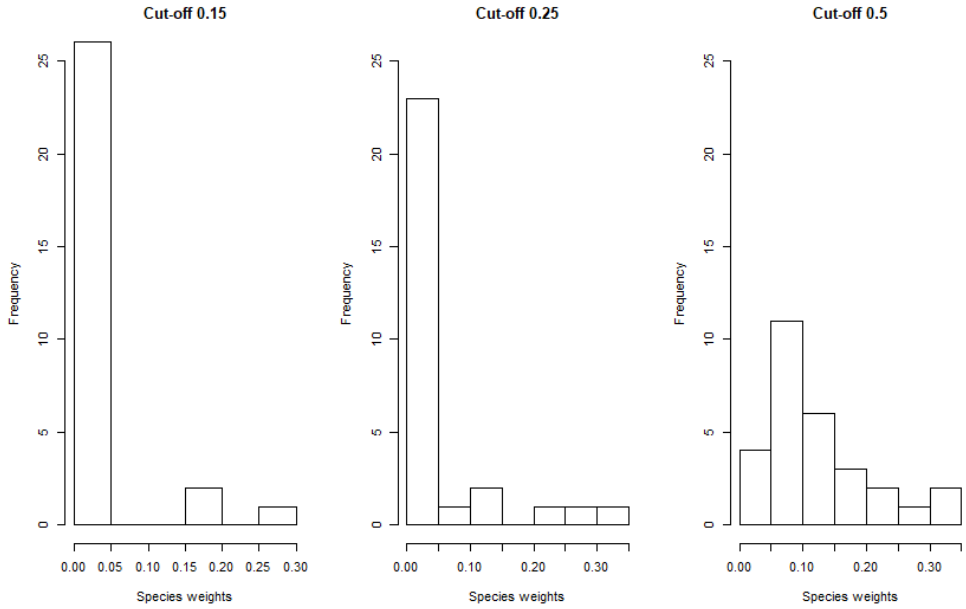


Figure 3.8: Histogram of the estimated species weights which are known to be rare.

Site-level species richness can be computed as a derived quantity of the occupancy model as $\sum_i z_{ij}$ for all species occurring at the jth site. Similarly, the proportion of rare species at each site can be estimated as $\sum_i \mathbb{I}(z_{ij} = h)/z_{ij}$ for species belonging to the h rare class (in this case $h = 1$ if the species is classified as rare and 0 otherwise). Fig. 3.9 indicates an overall good performance of the mixture model for estimating the true proportion of rare species at the different sites.

The estimated IRR, however, indicates an overall homogeneous and scarce distribution of rare species across sites when compared to the true proportion of rare species (Fig. 3.10). As the cut-off points decrease, more species will receive the minimum weight and the IRR becomes more homogeneous across sites. Nevertheless, IRR is not constructed to produce a rarity measurement in absolute values, but to provide a descriptive tool to identify rare species assemblages in relative terms of the highest and lowest occurrences among all the species in the study.

Yet, the threshold level to determine which species receive a weight $\neq 0$ has to be established based on strong ecological support (which often is unknown) and specific IRR values don't have a direct interpretation. Moreover, assigning an exponential increase in each species weight is not fully justified and more research is needed on this matter. Thus, in the next section the Odonata data set will be analysed using the occupancy mixture model approach to classify rare species based on their occupancy and to produce a map for the proportion of rare species that occupy each site.

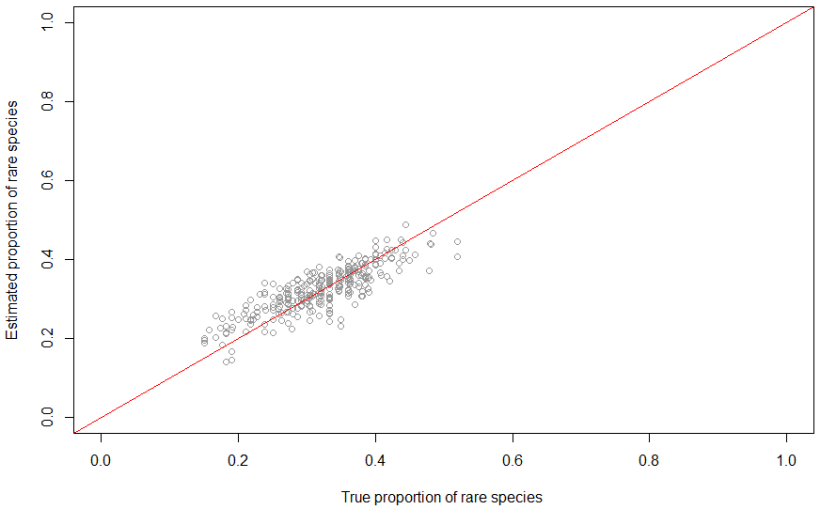


Figure 3.9: True vs mixture model estimated proportion of rare species (Eqn. 3.7) for each site.

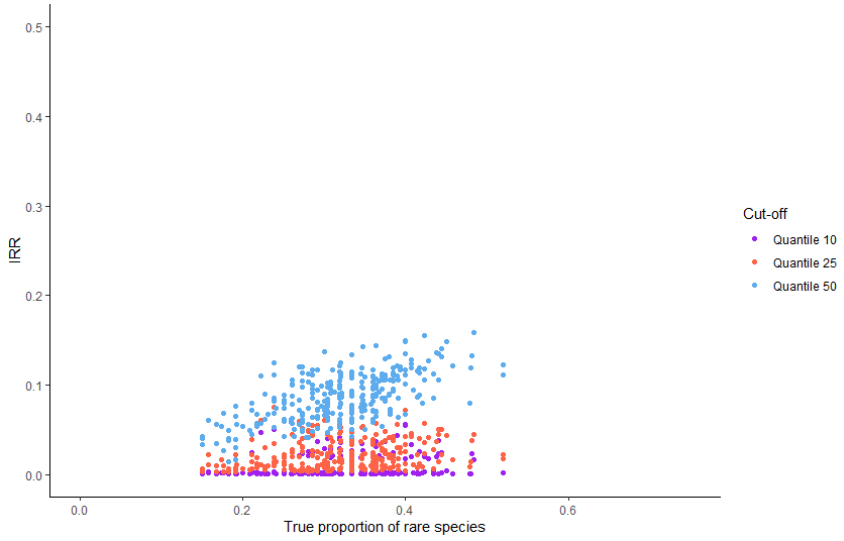


Figure 3.10: IRRs values vs estimated proportion of rare species after fitting model 3.2 with three different cut-off points.

3.2.4 Multispecies Occupancy Mixture model to quantify Odonata rare species in communities across the UK

We fitted the proposed mixture occupancy model (equations 3.3 and 3.11) to quantify the proportion of Odonata rare species across the UK. The proportion of rare species at each site was calculated according to Eqn. 3.10. Figure 3.11 shows the predicted classes for each of the Odonata species, with approximately half of the species being categorized as rare. Note that the credible intervals show that the uncertainty for detection probabilities is larger for rare species than for common species. However, at high levels of occupancy, widespread species estimated occupancy uncertainty becomes larger than for rare species. This pattern could be related to distribution of occupancy being right-skewed (a small number of common species show relatively high occupancy probabilities > 0.75) so that uncertainty around estimates of species with high occupancy probability will be larger than estimates for species with low occupancy. In addition, when the occupancy is low, we have less information from which the probabilities of detection are estimated resulting in greater uncertainty around these estimates.

Figure 3.12 shows the comparison between the predicted classes from our mixed occupancy model and the preliminary distribution class each of the species was given by Powney et al. (2014). For instance, species assigned with a preliminary limited distribution status (Rare, Scarce, Vagrant and Very rare) are classified as rare species by the mixture model. On the other hand, species with a wider distribution range (very widespread and widespread) are classified as common with the exception of *Calopteryx virgo* for which occupancy probability is below 25% and thus, it has been classified as rare. Note that species defined as local are classified both as rare and common depending on their estimated occupancy probabilities.

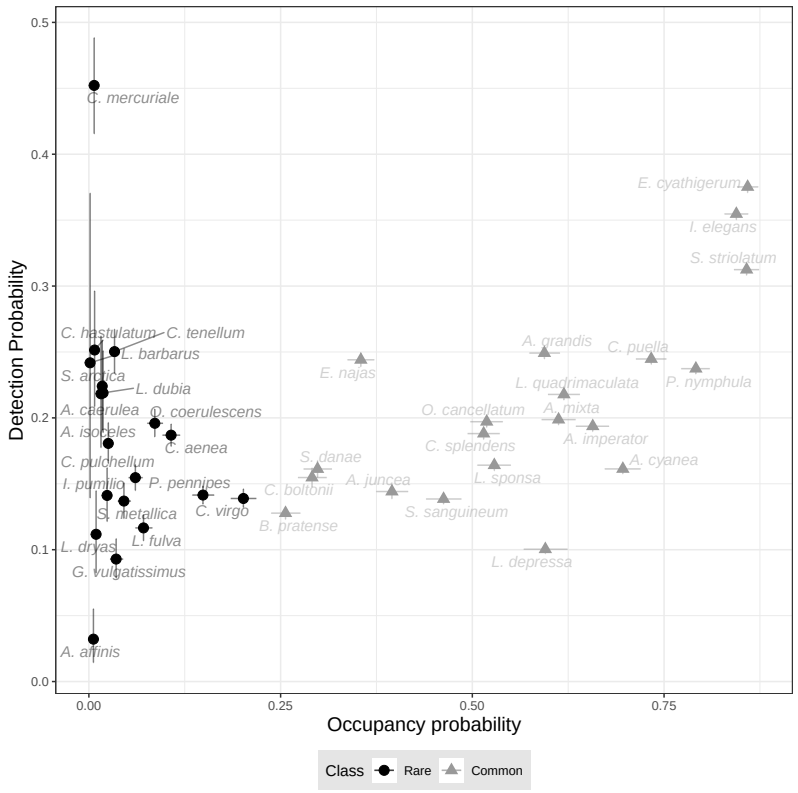


Figure 3.11: Estimated occupancy, detection probabilities and predicted class for the Odonata species. Error bars represent 95% credible intervals for the corresponding individual species occupancy/detection probability from model (3.7).

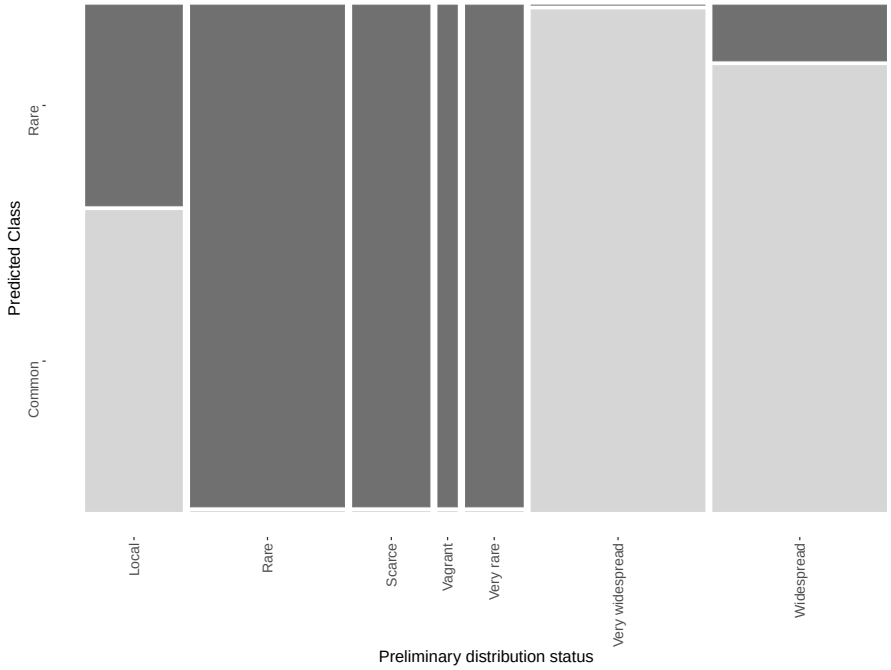


Figure 3.12: Mosaic plot for the number of predicted rare and common classes (Eqn. (3.7) vs the number of species allocated to 5 different assigned preliminary distribution status categories in Powney et al. (2014).

The proportion of rare species across the different sites is shown in Figure 3.13. Overall, the proportion of rare species and its estimated standard deviation are low and homogeneous across space. There are however, some areas up to the north and south-west where composition of rare species is greater than central areas which are dominated almost exclusively by wide spread species. These proportions were calculated with a reasonably small error as the standard deviations (SD) in Figure 3.13 (right) suggest (SD < 0.2 for most sites), though there are some sites where uncertainty is larger possibly due to the very low number of Odonata occurrences records at those sites.

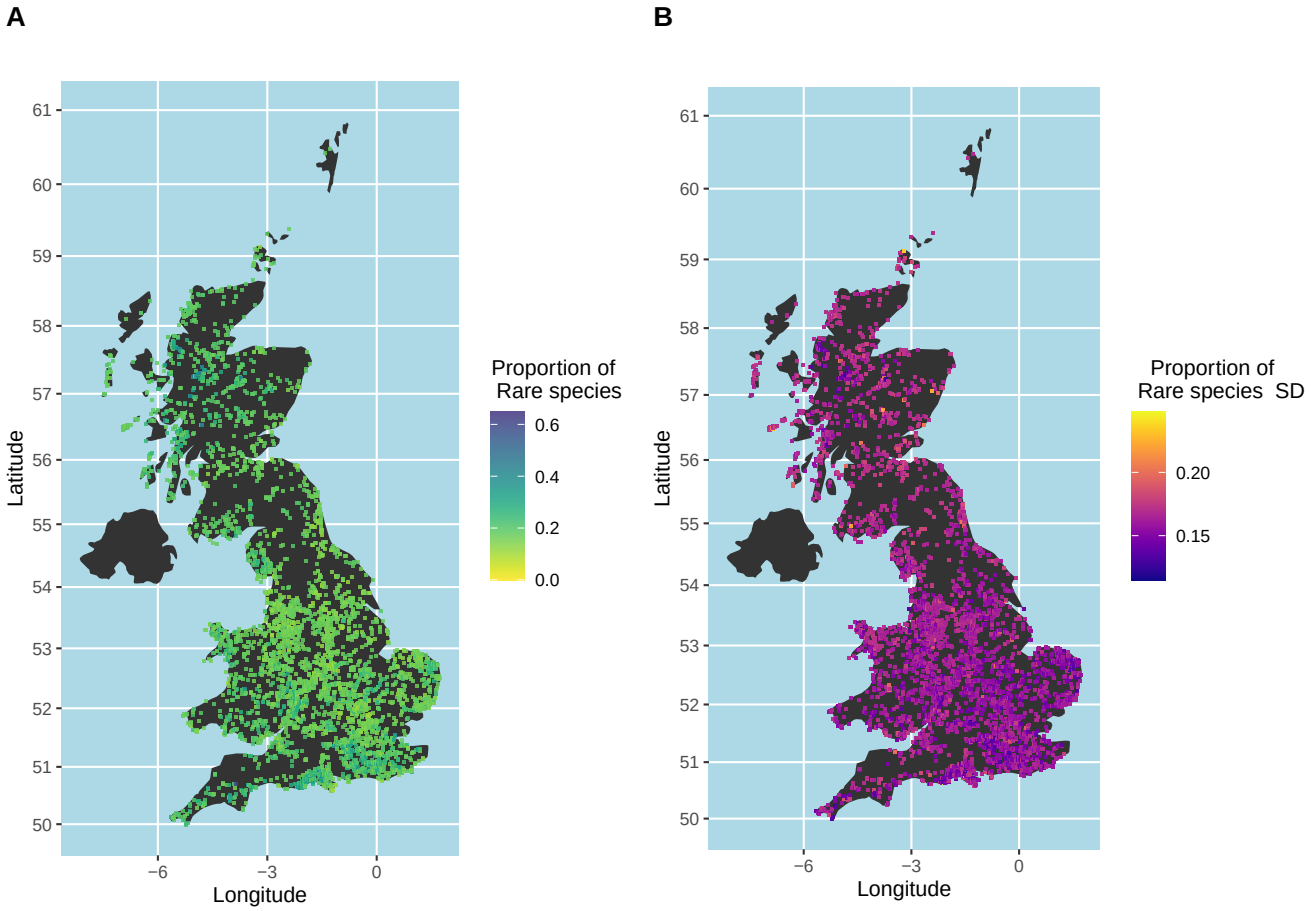


Figure 3.13: Proportion of rare species for each site across the UK with estimates of the mean proportion of rare species at each site (left) and estimated standard deviation (SD) (right) after fitting model (3.7) to the Odonata data set.

Model diagnostic plot shows evidence of convergence as shown by Gelman-Rubin diagnostic variance ratio < 1.1, the traceplots and autocorrelation plots (Appendix Figures 19 and 20) which indicates overall good mixing and no evidence of strong positive autocorrelation.

3.3 Hierarchical rare species distribution modelling across high dimensional nested spatial scales

Accounting for uncertainty in ecological studies is no easy task. It involves challenges like establishing statistical linkages between data sources with different uncertainties, spatial support and variability. Such linkages are used to develop methods to describe and predict how biodiversity responses change in space and time due to environmental changes (Cressie et al., 2009). Several statistical methods have been developed for quantifying and predicting the effect that different environmental drivers will have on species distributions through time and space (Elith and Leathwick, 2009; Guisan and Thuiller, 2005; Schroeder, 2008).

In many cases, the metrics produced by these methods correspond to ecological parameters derivable from partially observed species occurrences that are themselves modeled as a function of different environmental metrics. Thus, to avoid underestimating the uncertainty when modelling these quantities as responses, one must acknowledge the error associated with the estimation of these quantities (McCarthy and Masters, 2005).

Therefore, in this section, a 2-stage statistical modelling framework for analysing how different connectivity metrics interact with land-use change to modify rare species distribution (defined by the proportion of rare species and the IRR values) while accounting for detectability is developed.

At a first stage, site-level rare species composition index and estimated proportion rare species were obtained based on the Bayesian IRR (Eqn. 3.2) and the occupancy mixture model (Eqn. 3.7) results respectively.

In the second step, the effect of site level covariates on species rarity metrics is evaluated, using two sub-steps:

(a) ***Random forests*** **(RF)** to identify the explanatory variables that are "important" to the response (i.e. the most relevant) from a large dataset of potentially relevant variables (sites with missing values were removed from the analysis), with a reduced set selected through minimisation of mean square errors of predictions using `randomForest` library (Liaw and Wiener, 2002). Random forests is a nonparametric tree-based method that can be used to assess variable importance based on the decrease in the out-of-bag (OOB) error due to each variable (Hastie et al., 2005). Accuracy-based importance $\hat{I}_j$ of the jth predictor is computed based on the difference between the MSE on the out-of-bag sample and a $MSE_{(j)}$ defined after a random permutation of the values for each explanatory variable j ($j = 1,\ldots,p$) in the OOB sample. After an initial random forest run, the preliminarily standardized explanatory variables X_{ip} are sorted based on their importance. Random forests are then run on a nested subset of the new ordered covariates X^*_{ip-g} for $g \in [0,10,20,\ldots,G]$, with out-of-bag error calculated for each run. The cutoff point for establishing the "optimal" set was determined by the number of covariates that minimized the MSE (Fig.3.14). This allows for a subset of variables to be selected for further investigation in the next sub-step.

After the random forest selection process, variables measured at different spatial scales showed strong correlations to one another (e.g. % Urban land use at 1.5 km and % of Urban land use at a 2 km buffer). Thus, these redundant variables were removed based on their importance according to the random forest output. A total of 12 covariates were selected after removing redundant predictors (Fig.3.15).

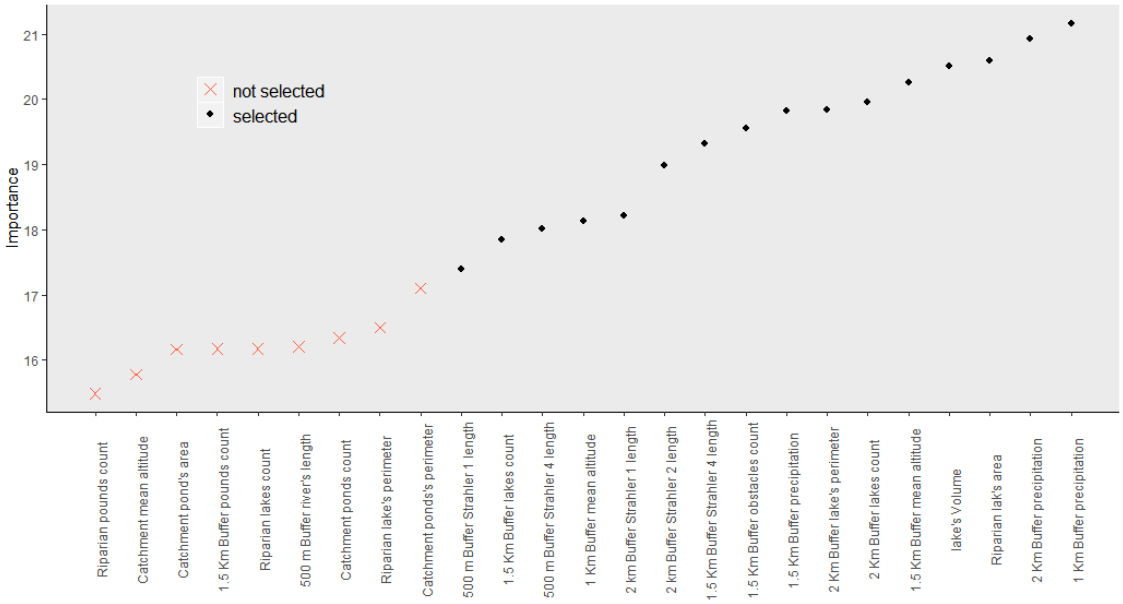


Figure 3.14: Random forest importance plot indicating the point at which variables are to be considered in a flexible regression model due MSE minimization criteria (note that for visualization purposes this plot does not include all the predictors in the data set).

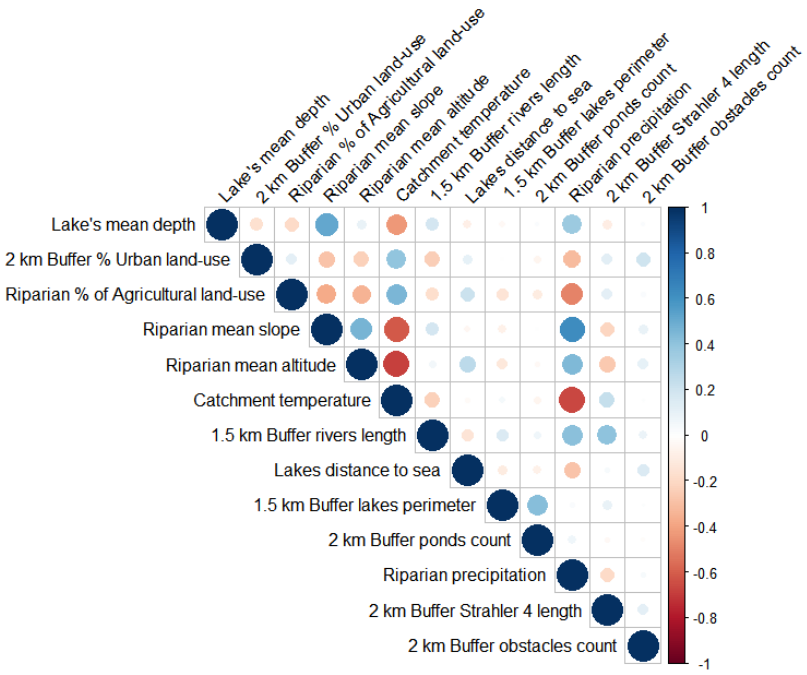


Figure 3.15: Correlation plot for site-level covariates after removing redundant variables.

(b) In the second stage, the reduced set of explanatory variables are considered in a ***generalised additive model (GAM)*** using `mgcv` library (Wood, 2004), allowing for smooth, nonlinear relationships, with interactions modelled using tensor products. The stage 1 uncertainties are incorporated through inverse-variance weighting (Thompson and Sharp, 1999) using the `gamm` function in R. For example, IRR can be modelled as follows:

$$\begin{aligned}\text{logit}(\text{IRR})_j = & f_1(\text{agricultural landuse})_j + \dots \\ & f_{8,5}(\text{lake area, urban landuse}) + u_j + e_j,\end{aligned} \tag{3.18}$$

where $f(x_j)$ holds the connectivity and stressor effects smooth terms. Then, $u_i \sim \text{Normal}(0, \tau^2)$ and $e_i \sim \text{Normal}(0, \sigma^2 v_i)$ such that τ^2 denotes the residual heterogeneity in the true effects, $v_i = \sigma_s^{-2}$ correspond to the sampling variances (IRR posterior standard deviation) and σ^2 is a proportionality constant.

Another additive variance structure that has been considered is $\sigma^2(1 + \exp(\eta) \times v_i)$ where σ^2 is the usual residual variance that scales the overall variance and $\exp(\eta)$ that weights the two variance components while ensuring a positive value for the second parameter (see `https://www.stat.berkeley.edu/~paciorek`).

Together, these two sub-steps allow us to understand the shapes of the effects of the site-level covariates and their interactions on the species rarity metrics, having first identified a small number of the most relevant variables among a large number of potentially relevant covariates, while incorporating the uncertainties from the stage 1 model. This approach was also applied to the site-level proportion of rare species estimated by the mixture occupancy model. The results for both rarity metrics are presented next.

3.3.1 Proportion of rare species modelling

After running RF and model selection, the final model for the proportion of rare species $w_j = \text{logit}(\text{proportion of rare species}_j)$ is:

$$\begin{aligned}w_j = \alpha + & f_1(\text{urban landuse}_j) + f_2(\text{catchment temperature}_j) + f_3(\text{number of obstacles}_j) + \\ & f_4(\text{ponds area}_j) + f_5(\text{Strahler 4}_j) + + f_{1,6}(\text{urban landuse, log. lake volume}_j) + \\ & f_{1,7}(\text{urban landuse, distance to sea}) + f_{1,8}(\text{urban landuse, log. precipitation catchment})\end{aligned} \tag{3.19}$$

Including the first stage estimation uncertainty with either the inverse variance weights (25.4% explained deviance) or additive variance structure (23.7% explained deviance) did not improve model goodness of fit compared to a model with no variance structure (25.3% explained deviance). Moreover, both AIC and R^2 values selected the model with no variance structure (Table 3.2). Similarly, diagnostics plots did not show any major difference between the models that incorporated a variance structure and the model with no variance structure presented in Fig. 3.16 (see appendix B.2 for the model (3.20) diagnostics when the additive variance structure is specified).

This could be due to the homogeneous variability observed in rare species proportion estimates and posterior standard errors across sites (Fig. 3.13). In the inverse-variance method, higher weights are assigned to model estimates associated with low posterior standard errors but because of the aforementioned posterior standard errors homogeneity across space, each site receives a similar weight. Thus adding a variance structure in the model makes no substantial difference to the fit.

The importance that heterogeneity in variances has in the estimation of the residual variance parameter of a second stage model in which the heteroscedastic variance from the first stage is added has recently been discussed in some health meta-analysis studies (Makambi, 2004; Langan et al., 2019). It has been found that the homoscedastic variance parameter in the second stage model can be biased depending on the heterogeneity level of the first stage model uncertainties. Thus, authors have suggested using restricted maximum likelihood (REML) to correct for the bias in MLE variance parameters (Langan et al., 2019). In this work, however, no meaningful differences were found after fitting the aforementioned models using REML, suggesting that the homogeneity observed in the variability of the first-stage estimates results provides little additional information to have a meaningful effect on the model fit. This perhaps is related to the spatial scale at which species rarity is measured (see discussion in the last section of this chapter and revisited in chapter 6). Model 3.20 diagnostic plots indicates that residuals are not optimal but reasonably well behaved, there is not strong evidence against normality and heteroscedasticity (with the exception of one possible extreme value) (Fig. 3.16).

Table 3.2: Goodness of fit for proportion of rare species gam (Eqn. 3.19) under different variance structures.

Model	Deviance explained (%)	AIC	BIC	R^2
No variance structure	25.3	-209.67	100.32	0.24
Inverse variance weights	25.4	-117.64	42.67	0.22
Additive form variance	23.7	-110.22	55.81	0.22

Four different univariate predictors are found to have a significant effect on the proportion of rare species. However, the predicted values along with the % of Deviance explained suggest a weak relationship between the proportion of rare species and the covariates. This is portrayed on Figure 3.17 where the relationship between the proportion of rare species that are estimated to occur on a 1 km scale and the predictors urban land-use change and catchment temperature does not show a strong association with each other. Bivariate effects show a more complex pattern. For instance, Fig. 3.18 suggests that the proportion of rare species decreases as both catchment precipitation and urban land-use levels decrease. On the other hand, the opposite effect is seen when a site's distance to the sea and urban land-use levels increase as the proportion of rare species also increases.

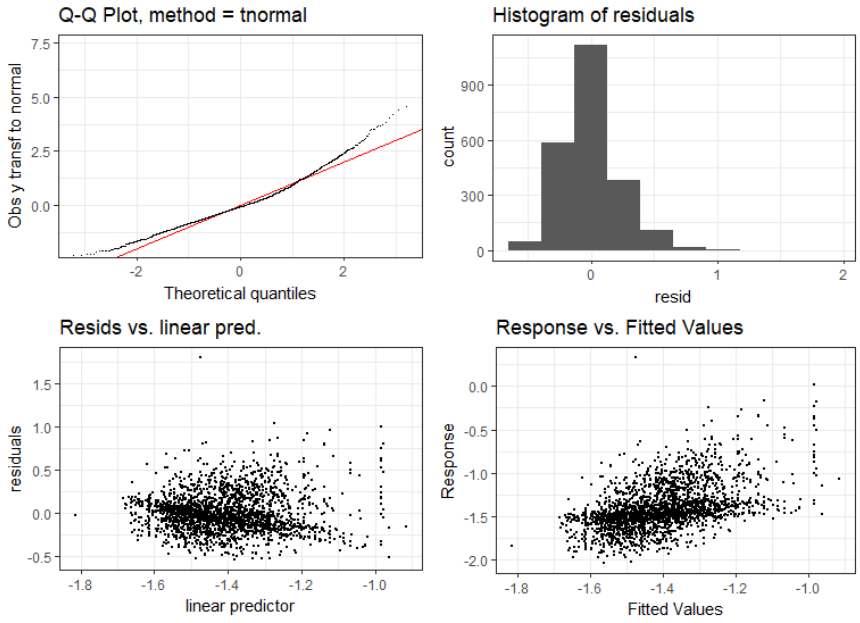


Figure 3.16: Diagnostics plot for proportion of rare species gam (Eqn. 3.19) without variance structure.

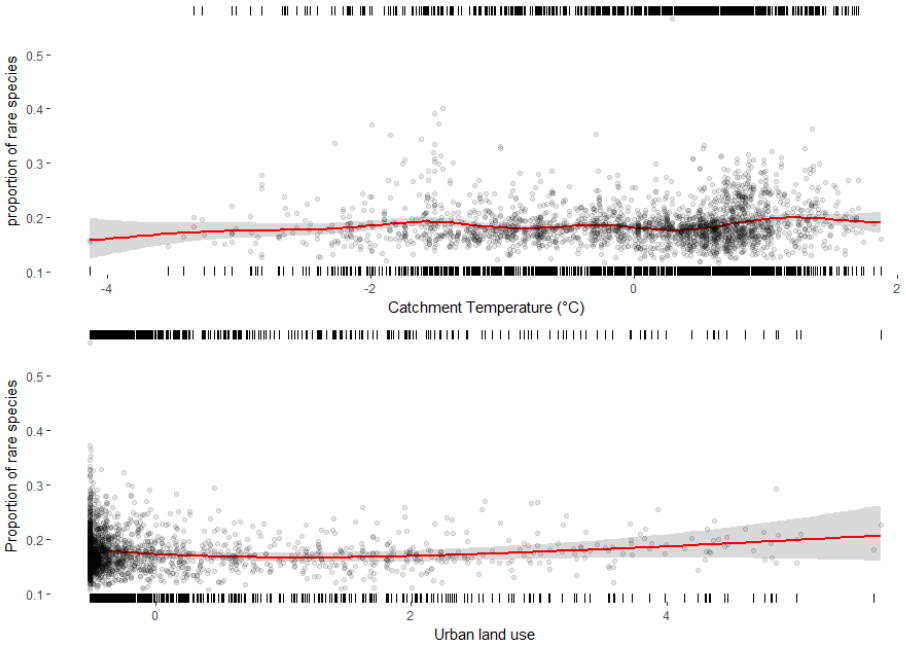


Figure 3.17: Univariate smooth effects for proportion of rare species gam (Eqn. 3.19) without variance structure.

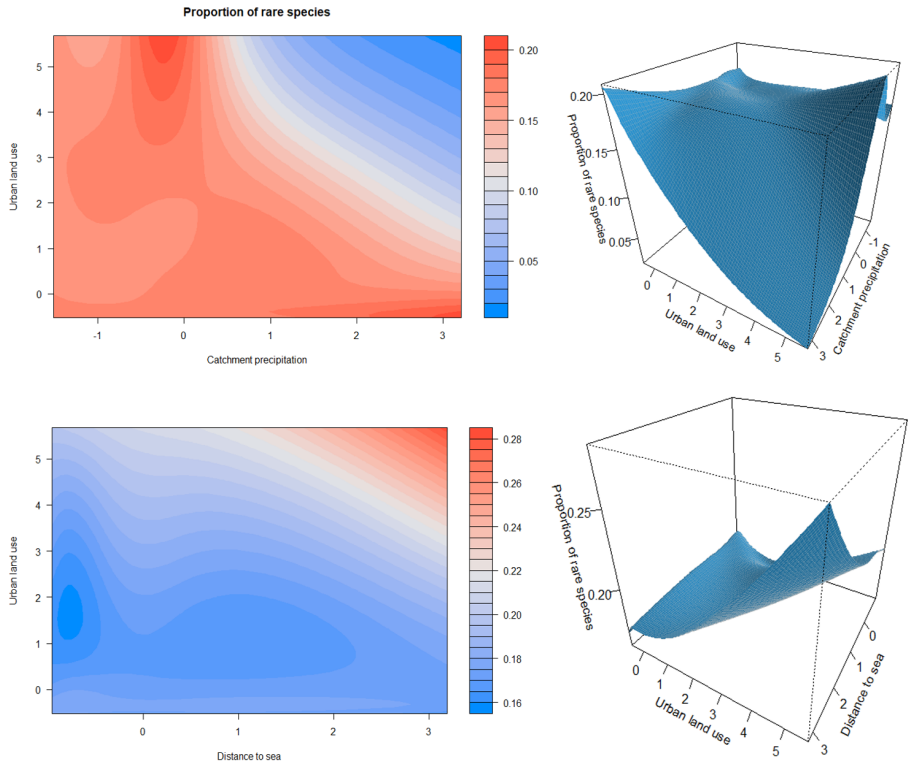


Figure 3.18: Bivariate effects smooths for the interaction between catchment precipitation, distance to sea and urban land-use for the proportion of rare species gam (Eqn. 3.19) without variance structure.

3.3.2 Index of relative rarity modelling

After variable selection, the final model for the index of relative rarity was:

$$\begin{aligned}\text{logit}(IRR_j) = \alpha &+ f_1(\text{agricultural landuse}_j) + f_2(\text{distance to sea}_j) + f_3(\text{ponds area}_j) + \\ &f_4(\text{number of lakes}_j) + f_{2,5}(\text{distance to sea, urban landuse}) + \\ &f_{3,5}(\text{pond area, urban landuse}) + f_{6,5}(\text{depth, urban landuse}) + \\ &f_{7,5}(\text{slope, urban landuse}) + f_{8,5}(\text{lake area, urban landuse}) + \\ &f_{9,5}(\text{number of obstacles, urban landuse}) + f_{6,1}(\text{depth, agricultural landuse}) + \\ &f_{7,1}(\text{slope, agricultural landuse}) + f_{8,1}(\text{lake area, agricultural landuse}). \end{aligned} \tag{3.20}$$

Including the first stage estimation uncertainty improved model goodness of fit compared to a model with no variance structure (a 19.36% in explained deviance was gained when the additive variance structure was specified). Moreover, additive variance structure suggests a better fit for the final model (Table 3.3).

Table 3.3: Goodness of fit for IRR gam (Eqn. 3.20) using different variance structures.

Model	Deviance explained (%)	AIC	BIC	R^2
No variance structure	25.3	-192.25	287.11	0.24
Inverse variance weights	26.5	-193.46	287.89	0.24
Additive form variance	44.66	-219.79	66.48	0.21

Diagnostic plots indicate that residuals are normally distributed and there are no issues regarding heteroscedasticity (Fig. 3.19). Note that more evident departures from normality were observed in the diagnostics of the model with no variance structure (available in appendix B.2), suggesting that for the estimated IRRs, a model with an additive variance provides an overall better fit than a model without it).

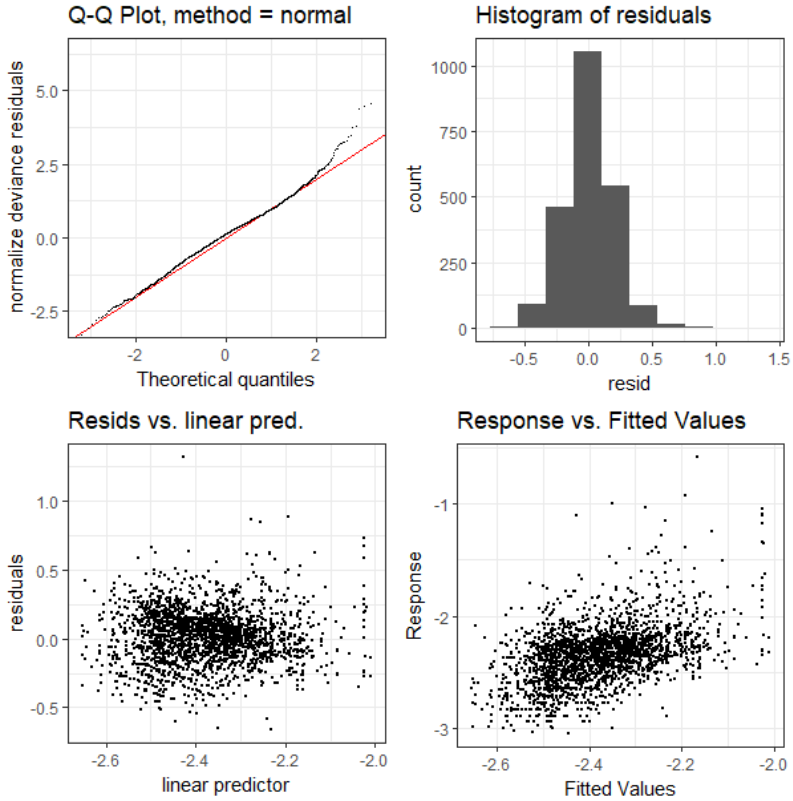


Figure 3.19: Diagnostics plot for IRR gam (Eqn. 3.20) with additive variance structure.

Similar to the results shown by the proportion of rare species gam, the predicted values along with the % of Deviance explained suggest a weak relationship between the IRRs and the covariates. For instance, river length on a 1.5 km buffer has a weak negative relationship with IRR. However, assemblages of rare species increases with catchment temperature (Fig. 3.20). Non linear interactions between stressors and connectivity metrics show more complex patterns (the effect size is rather small to determine a strong association). For instance, the assemblage of rare species is expected to be greater at areas with

increased urban land use coverage where river length is large. For agricultural land use coverage, GAM results suggest that species rareness increases at sites with higher Strahler numbers (measurement of the branching complexity of a stream or river within a river network) where agricultural land-use coverage is low. However, as agricultural land-use increases rare species composition becomes smaller.

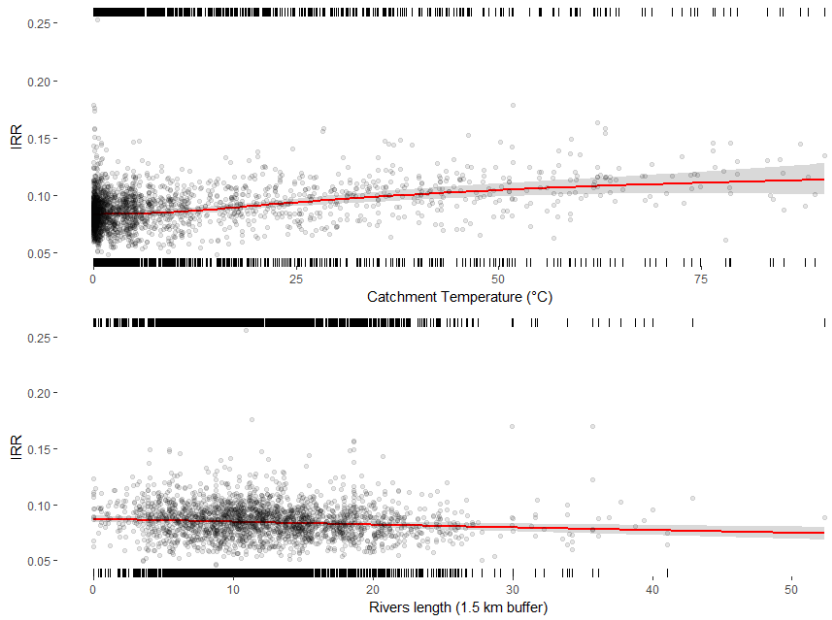


Figure 3.20: Univariate smooth effects for proportion of rare species gam (Eqn. 3.20) with additive variance structure.

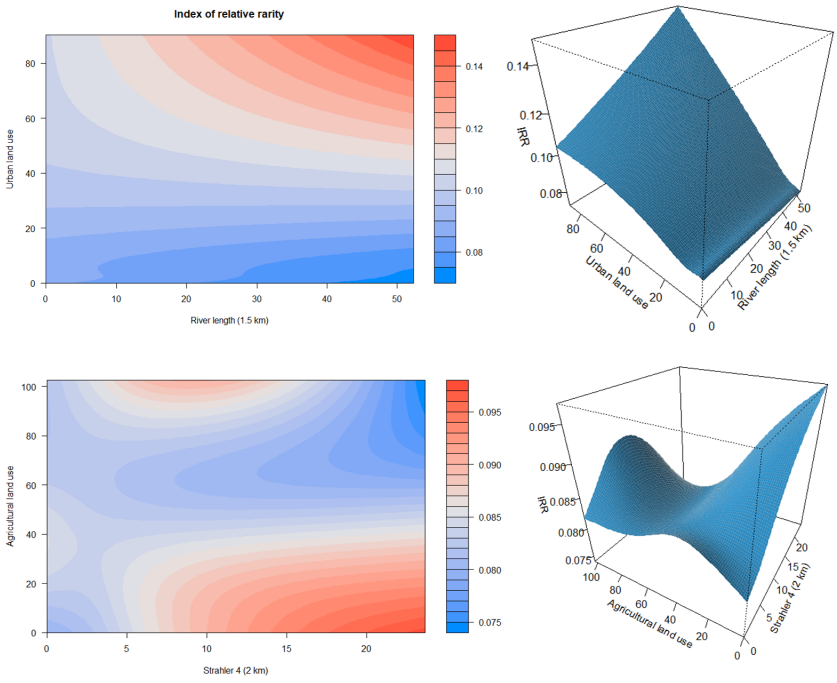


Figure 3.21: Bivariate effects for model 3.19 with additive variance structure.

3.4 Final remarks

In this work, a new approach to identify and quantify the composition of rare species in a community is presented. The proposed model differs from previous methods by considering the detectability bias induced by species false-absences. Moreover, this modelling framework enables using standard occupancy model designs for which information about the species distribution ranges, populations' densities, or habitat specificity are not required to estimate the species distribution. In fact, any species-specific data of this nature could be incorporated in either or both latent state and observational processes of the model as described in section 3.2.4.

The models in this chapter were constructed under the single-season occupancy model closure assumption in which the species distribution is assumed to remain constant across the sampling occasions. The single-season occupancy model was used as the starting point since the primary objective was to model Odonata non-invasive species that show little evidence of major temporal occupancy changes. However, any occupancy model extension could potentially be developed within the proposed mixture modelling approach. For example, our model could be integrated with an explicit spatio-temporal occupancy model (see Rushing et al. (2019)) to estimate the occupancy temporal changes that a species of a specific class experiences. This could be potentially useful for monitoring invasive species occupancy dynamics at different stages after its introduction and to identify the tipping point when a focal species experiences a major increment or reduction in its distribution given by the change in its latent class.

The simulation analysis in this chapter showed reasonably good classification performance of the mixture occupancy model, specifically when the proportion of individual species was similar between classes. However, when proportion of common species was set to be greater than the proportion of rare species, a larger proportion of common species were classified as rare. Under this unbalanced scenario, the posterior mass density from which the species occupancy community mean parameters are drawn will be greater for common than for rare species. Thus, uncertainty around rare species occupancy community mean will be larger and pulled towards the commons species density mass, resulting in some common species to be classified as rare, yet producing precise predictions for the common species (i.e. the sensitivity and the proportion of predicted common species that are truly common will be high). Note that for this simulation analysis, the large difference in the proportion of species belonging to each class could very well portray a real-life situation where biological communities are composed primarily of widespread species. Thus, to avoid the overestimation or underestimation of rare species caused by unbalanced classes, informative Dirichlet priors (or Beta priors for a two-class classification problem) could be specified to reflect our prior judgment about the proportion of species of each class determining the community structure.

It is important to notice that using an informative prior causes species rarity classification to change and hence, introduces a risk of underestimating the number of truly rare species if the prior weight is greater for common species, or overestimating the rare species if the prior assigns more weight to this class. However, as mentioned by Leroy et al. (2012), it is preferable to conduct an analysis to characterize species rarity rather than ignore it and thus potentially overlook hot spots relevant for conservation.

Therefore, the choice of priors depends on (1) the previous knowledge about the target species distribution and (2) the conservation targets. For instance, if the conservation study focuses on rare species, more weight could be assigned to the priors probabilities of a species being classified as rare to ensure that most of the species will be correctly classified as rare.

The proposed mixture occupancy model approach enables the class-membership for multiple species to be estimated based on a fixed number of classes. If the number of classes is unknown, a Bayesian non-parametric mixture model involving Dirichlet processes could be specified (Hjort et al., 2010). For instance, the Chinese Restaurant Process or a truncated stick-breaking Dirichlet process have now been implemented and can be fitted in `nimble` (de Valpine et al., 2017). Modelling the distribution of rare species using the hierarchical structure of occupancy models enables information exchange between species to facilitate the estimation of rare species responses that otherwise would be difficult. Nevertheless, while it is beyond the scope of this work, extremely rare species can still be estimated with our proposed method, but the number of samples needed for the parameters to converge represents a computationally demanding process. Future work will investigate different computationally efficient methods for the specific aim of identifying such species.

In the case study analysed in this work, Odonata rare species composition is generally scarce and homogeneous across space. There are, however, some areas up in north of Scotland where rare species composition is above 50%. This could be explained by the occurrence of species at very local scales such as *Coenagrion hastulatum* and *Somatochlora arctica* which are confined to a few particular lakes in Scotland and south-west Ireland (source: https://british-dragonflies.org.uk). Consequently, the distribution patterns for some of these rare species might not be evident on a national scale and could be limited to regional scale specific habitats subject to different environmental conditions and extinction processes (Hartley and Kunin, 2003). For example, for some of the rare species identified in our study, there are some which have been reported to have completely different responses to environmental stressors such as *Coenagrion mercuriale Platycnemis pennipes* and *Calopteryx virgo*. While the first two are found in small brooks and springs closely by agricultural fields, canals, gravel-pits, and fish-ponds, *C. virgo* is rarely seen in urban areas or in regions with strong agricultural land-use (source: www.iucnredlist.org). Hence, because of the occurrence-based approach implemented here, where the proportion of rare species on a national scale is derived from the relative mean occupancy probabilities, the estimation of rarity depends on the study spatial scale (e.g. species with very low occurrences in northern latitudes will be classified as rare on a national scale but could be classified as common on a regional scale). Previous studies by Hartley and Kunin (2003) and Leroy et al. (2013) have shown how species rarity can be a scale-dependent process and have proposed different multiscale metrics to quantify the species rarity. However, producing such metrics under imperfect detection is a less studied subject for small invertebrates ,Leroy et al. (2012, 2013), like dragonflies and damselflies, and a very interesting area for future research. For example, the proposed modelling framework could be integrated with multi-scale occupancy modelling designs and sampling schemes like the one proposed by Nichols et al. (2008) to estimate rare species occupancy at varying spatial scales under imperfect detection (see further discussion in chapter 6).

Other occurrence-based methods that assess species rarity (such as Leroy et al. (2012) rarity index) rely greatly on the sampling methods, as uneven sampling effort might induce an artificial rarity for those species recorded in poorly sampled sites. Accounting for sampling effort and detection probabilities in the observation process is of major importance, specifically when working with large citizen science data sets where observations are collected without a standardized field sampling protocol for a frequently arbitrary selection of sites. This variation in observation effort causes detection probabilities to vary over time and space (Kéry et al., 2010). The two-component structure in the proposed occupancy model allows for terms that correct for uneven sampling effort to be incorporated. In here, only the number of visits at each site was used as a proxy for the sampling efficiency. However, the day when each site was visited and the daily species lists in combination with occupancy models have also been proposed as a way to correct for unequal observation effort in these type of studies and represent a very interesting area for further exploration (a dynamic occupancy model that incorporates this type of structure will be explored in chapter 5) (van Strien et al., 2010, 2013).

The two-stage approach developed in this chapter presents a computationally efficient method for modelling the effects of high dimensional data on nested spatial scales with imperfect detection. It is a useful tool that allows for an otherwise complex Bayesian hierarchical model to be simplified. It can be applied to other responses such as biological indicators derivable from partially observed occurrences, while retaining the interpretability of GAMs. This allows for relevant areas for biodiversity conservation to be identified when the interest lies in protecting threatened rare species. McCarthy and Masters (2005) discuss the importance of propagating the uncertainty of any ecological scoring method into a second analysis to increase the confidence of the conclusions that are drawn from the study.

The importance that heterogeneity in variances has in the estimation of the residual variance parameter of a second stage model in which the heteroscedastic variance from the first stage is added has recently been discussed in some health meta-analysis studies Makambi (2004); Langan et al. (2019). It has been found that the homoscedastic variance parameter in the second stage model can be can be biased depending on the heterogeneity level of first stage model uncertainties. Authors have suggested using restricted maximum likelihood (REML) to correct for the bias in MLE variance parameters (Langan et al., 2019). However, no meaningful differences were found when fitting the models discussed in sections 3.3.1 and 3.3.2, suggesting that the homogeneity observed in the variability across sites has a little effect towards the overall variance in the model.

In the case study analysed through this chapter, species with high occupancy estimates are well represented across the different Odonata communities in the UK. Uncertainty around the estimates is greater for those rare and elusive species. Incorporating these rarity metrics into a flexible model allows for different site-level covariates to be evaluated. Even though results suggest that rare species may be more frequent in sites where urban land-use is high, strong relationship between these variables could not be completely established through this study.

This can be the result of the wide range of different individual species responses to drivers and pressures. For instance, the two-steps modelling approach results suggest that rare species composition (based

on IRR) and proportion of rare species decreases at those sites with shorter river length and low precipitation respectively. This type of habitat has been reported to be used by some of the rare Odonata species in the UK such as *Lestes barbarus* which occupies ephemeral sites that dry out early in summer, *Libellula fulva*, *Somatochlora metallica* and *Coenagrion pulchellum* which are found in slow-flowing waters such as ponds, rocky lakeshores, canals and sluggish rivers. Other species like *Leucorrhinia dubia* and *Lestes dryas*, have been reported to be influenced by water bodies' alkalinity level, which is available for just a few number sites and not necessarily determines the distribution of other members of the communities. Interestingly, models for IRR and proportion of rare species had urban land-use as an important predictor for the response (although for IRR, urban land-use appear more frequently as an interaction term with different connectivity metrics), suggesting that the interaction of this stressor with some relevant connectivity metrics might have an important role in determining the distribution or composition of rare species. Moreover, the number of connectivity metrics that have a significant effect on the IRR was greater than the number of metrics affecting the proportion of rare species. This could be related to the variability in the response. Recall that inverse variance-weighting had little effect on the goodness-of-fit of model for the proportion of rare species. Thus, the homogeneity in the estimated proportions and low homogeneous variances associated with these estimates can contribute to having a more parsimonious model compared to the model for the IRR (which proved to have a better fit when an additive variance structure was used).

Finally, the models so far assume that species occurrences are independent to one another. However, species co-occurrence might be plausible for a pair of species in this study. First, *Aeshna affinis* occurrences have been reported to be less common in sites where *Aeshna mixta* occur (Utzeri and Raffi, 1983). However, interaction between these two species has not been reported yet and more studies are required to establish co-occurrence between these two species. A more clear case of species co-occurrence is the one of *Lestes dryas* whose distribution is shaped by other Dragonfly predation (particularly by other permanent-water Lestes) (Stoks and McPeek, 2003). However, this type of predation has been reported to occur typically at a larval state and thus, adults' distribution patterns might be independent given habitat-exclusion on their larval state.

The approach presented in this work provides a new method to understand the species rarity assemblage in a community when species are undetected imperfectly and its relationship with the environment. It can be potentially applied to different situations such as multiscale studies and spatio-temporal modeling. The proposed models have been applied on a National scale to estimate Odonata rare species composition; a group (along with other small invertebrates) that has a significant under-representation in studies accounting for species imperfect detection (Kellner and Swihart, 2014; Devarajan et al., 2020). Thus, this work provides a substantial improvement to how rare species are identified and the understanding of how species rarity assemblages and their relationship with their environment can be quantified.

Chapter 4

Developing a flexible occupancy model to evaluate non-linear population dynamics

The occupancy model was originally developed under the assumption of a closed system where the occupancy probability remains unchanged across different surveys and no local extinctions or colonizations occur at different patches (MacKenzie et al., 2002). Thus, colonization and extinction are population dynamic mechanisms that drive the sites' occupancy rate of change over time. Different extensions of the state model that incorporate temporal changes in the occupancy have been explored to allow for open systems where local extinctions and colonization occur (Royle and Dorazio, 2008). MacKenzie et al. (2003) applied a sampling design with multiple sampling seasons which enables colonization and extinction probabilities to be estimated through different patches in long-term studies. This sampling scheme, firstly introduced by Pollock (1982), defines a primary sampling period t in which occupancy state changes for each site j, and a secondary sampling $k(t)$ (nested within each primary period) where the occupancy status is assumed to remain constant. Then, the detection history is given by $\mathbf{Y}_{j,t}$ with $\mathbf{Y}_{j\cdot}$, being the overall detection history for each site.

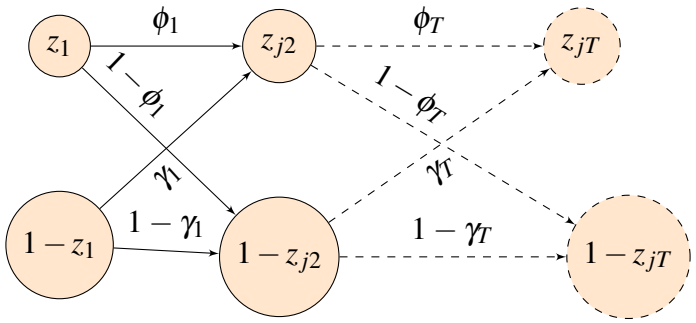


Figure 4.1: Multiple seasons occupancy model diagram showing how the j site's occupancy state (z_{jt}) changes over T time periods given the colonization probability γ_t and survival probability ϕ_t.

The multiple seasons occupancy model (MSOM) portrayed in Figure 4.1, enables estimation of occupancy probabilities in long-term studies where occupancy state changes over time, given the following assumptions:

- Let $t \in 1,\ldots,T$ denote the primary sampling period in which the binary occupancy state (z_{jt}) changes for each site, and a secondary sampling (number of visits within t) where the occupancy status is assumed to remain constant.

- γ_t is the colonization probability of an unoccupied site at primary period t to become occupied at $t+1$.

- ϕ_t is the survival probability for an occupied site at period t to still be occupied at $t+1$ (i.e. not going locally extinct). Thus, local extinction can be written as $\varepsilon_t = 1-\phi_t$.

- ψ_1 is the initial occupancy probability. Thus, the occupancy probability at time t for any site can be defined recursively as: $\psi_t = z_{t-1}(1-\varepsilon_{t-1}) + (1-z_{t-1})\gamma_{t-1}$.

Royle and Kéry (2007) provided a Bayesian hierarchical representation of this model which enable latent variables to be retained. As discussed by these authors, a Bayesian analysis of the hierarchical parametrization of the MSOM allows inference to be made about the sampled sites (opposed to likelihood-based approaches that focuses on a larger population of sites). It also enables random effects to be incorporated to accommodate different variance terms for spatial or temporal structures. In this work, this flexibility will be used to develop a model that incorporates non-linear relationships in colonization and survival probabilities.

The MSOMs have set a solid line of research for contrasting different ecological hypotheses and thus, have been applied to different scenarios such as predicting metapopulation dynamics after species reintroduction (Chandler et al., 2015), evaluating the effect that the patch size and neighborhood occupancy have on colonization and survival (Eaton et al., 2014) and integrating population demographics to model species stage-specific colonization and extinction rates (Sutherland et al., 2014). The majority of the work proposed for modelling species' occupancy temporal changes has been focused on single species studies and with the exception of the aforementioned work by Sutherland et al. (2014), which uses an exponential dispersal kernel for defining the colonization probability, most of the work assumes a linear relationship between occupancy and environmental covariates.

However, in many situations, the relationships between an ecological response (e.g. occupancy) and covariates departs from linearity and involves nonlinear effects that cannot be expressed in specific formulae (Pedersen et al., 2019). In such cases, a more flexible modelling frameworks has to be developed to handle such nonlinear relationships without making the strict assumption that the data generating process is known up to a finite number of parameters. Thus, the aim of this chapter is to develop a non-parametric MSOM that incorporates non-linear colonization and survival terms for multiple species. Very often, the model's performance in predicting the response depends on the quality of the data and the complexity of the relationship between the response and the explanatory variables. Thus, another aspect

that will be considered is the effect that (1) degree of smoothing and (2) multiple sampling and have on the model estimates through a simulation study. For instance, very often the data available for fitting these models lacks repeated measurements for which different approaches have been tried such as using spatial replicates as a substitute for temporal visits (Guillera-Arroita et al., 2011; Sadoti et al., 2013). However, assuming spatial replicates have the same variability and species occupancy might not be feasible due to distinct environmental features across sites (Peach et al., 2017). Thus, a different alternative to estimate model parameters in the absence of replicate surveys will be discussed in the following section.

4.1 Modelling occupancy with presence only data

Occupancy models have become a major tool for studying species distributions when detection probabilities are less than one. Most of these methods rely on planned systematic surveys where the observer visits a patch or a site for a specific amount of time and takes a record of how many species are present at that location. Unsuccessful surveys where species are not detected are denoted as absences. Then, each location is revisited in k occasions and a detection history for each site is produced. This presence-absence occupancy model assumes that the occupancy probability remains constant throughout different visits and both occupancy and detection only vary in terms of site level covariates. Unfortunately, the sampling effort involving these types of survey schemes results in highly costly studies, especially for rare or elusive species (Lele et al., 2012). Thus, species occurrence records are often constrained to opportunistic surveys at locations (or biological collections such as museums or herbariums) where individual species have been observed and no information regarding the absences is available (Koshkina et al., 2017). These presence-only records introduce a sampling bias since the inferred occupancy patterns can be clustered based on the localities in which human activities are concentrated instead of the sites that the species actually occupy, i.e. the estimates could be more related to human activities rather than the true species distributions (Fithian and Hastie, 2013).

Accounting for this bias when the only available source of data comes from single visits has been addressed by Peach et al. (2017) and Lele et al. (2012) by following MacKenzie and Royle (2005)'s removal sampling design which incorporates the survey effort as a predictor to model the probability of detecting a species at least once during repeat surveys. MacKenzie et al. (2003) and Lele et al. (2012) showed that estimating a site occupancy probability in the presence of detection error in a single survey sampling scheme is possible when distinct covariates that affect occupancy and detection probabilities are specified (usually survey effort is used as a linear predictor for the observational model). This approach was extended by Peach et al. (2017) to model species temporal occupancy dynamics for single - visit surveys schemes.

The initial occupancy probability at the first primary sampling period $\psi_{1,j}$ can be modeled as follows:

$$\text{logit}(\psi_{1,j}) = \beta_{0_o} + \beta_{1_o} x_{1j} + \ldots + \beta_{U_o} x_{Uj}, \tag{4.1}$$

where $[\beta_{0_o}, \ldots, \beta_{U_o}]$ are a set of unknown coefficients for each site-level covariate $x_{1j}, \ldots, x_{Uj}$. Supposing two primary sampling periods, the occupancy at the second primary sampling period is described by following colonization (γ) and extinction (ε) probabilities for each site j (Eqn 4.2).

$$\begin{aligned} \text{logit}(\gamma_j) &= \beta_{0_c} + \beta_{1_c} x_{1j} + \ldots + \beta_{U_c} x_{Uj} \\ \text{logit}(\varepsilon_j) &= \beta_{0_e} + \beta_{1_e} x_{1j} + \ldots + \beta_{U_e} x_{Uj} \end{aligned} \tag{4.2}$$

To make parameters in the model identifiable, occupancy and detection probabilities must depend on covariates and those covariates affecting the occupancy must differ by at least one variable from the set of covariates that affect detection.

Peach et al. (2017) suggests incorporating sampling effort as a power term to account for the nonlinear relationship between sampling effort (time spent in the area) and the detection.

Let p^* be the probability of a species being detected at least once during k consecutive surveys of an occupied site and p the probability of detection, which can be written as a function of a site level covariate x_{2j}

$$\text{logit}(p_j) = \alpha_0 + \alpha_1 x_{2j}. \tag{4.3}$$

Assuming a constant detection probability across sites simply implies that $\alpha_1 = 0$. Denote K a random variable for the number of visits needed to detect a species for the first time. Then, the probability that the first detection requires k independent visits, (each with a probability of success p) is given by $\Pr(K = k) = (1-p)^{k-1} p$ for $k = 1, 2, \ldots$. Thus, the probability of a species being first detected after k independent visits is given by the geometric cdf:

$$\begin{aligned} p^* &= \Pr(K \leq k) \\ p^* &= 1 - \Pr(K > k) \\ &= 1 - \sum_{w=k+1}^{\infty} p(1-p)^{w-1} \\ &= 1 - (1-p)^K. \end{aligned} \tag{4.4}$$

The number of K visits is then substituted by units of effort per site (e.g. number of hours), i.e.

$$p^* = 1 - (1-p)^{\text{effort}_j}. \tag{4.5}$$

Non linearity in occupancy models is usually accounted for by including power terms on the different site/survey-level covariates (Wintle and Bardos, 2006; Cressie et al., 2009; Kéry and Royle, 2009). Up to very recently, little research has been made available to extend occupancy models to estimate non-linear effects of covariates (Rushing et al., 2019). For example, Collier et al. (2012) were the first to introduce radial basis penalized splines to allocate a spatial smooth effect for each surveyed site in a single season occupancy model for the American Warbler endangered species. More recently, Rushing et al. (2019) used a spatial smoothing function in conjunction with time-varying covariates to model temporal and spatial variation in the occupancy of 10 eastern North American birds. Bled et al. (2013) also used smooth terms to describe the relationship between the occupancy of *Bostrychia hagedash* species and the habitat structure given by vegetation coverage data.

Developing a flexible dynamic occupancy model will enable to describe complex relationships between environmental covariates and population dynamics of colonization and survival. Thus, the next section of this chapter introduces flexible model for colonization and survival parameters that will be assessed with a simulation study to investigate potential issues when data from a single and multiple visits are used and will be compared with the parametrization proposed by Peach et al. (2017) for a single species. Simultaneously, the simulation analysis will explore the potential of the proposed modelling approach under simple and complex smooths. The central hypothesis is that by specifying a non-parametric formulation of the model, the nonlinear components in the model will be handled automatically regardless of the relationship's complexity (but might be affected on the quality of the data). This will be particularly useful since commonly used transformations (or power terms specification) often require a good expertise and time Wood (2004). Finally, a multispecies model that also incorporates flexible terms for the detection parameters will be developed to capture complex non-linear effects on both state and observational stages of the model.

4.2 Modelling non linear relationships for state variables

In Royle and Kéry (2007)'s MSOM hierarchical Bayesian formulation, colonization and survival are latent variables assumed to follow a linear relationship with the site-level covariates through the logit function. However, while this affects the relationship between the responses and the explanatory variables $\mathbf{X}$ across all the range of $\mathbf{X}$, the relationship between the variables is often specific to a certain local region of $\mathbf{X}$ and exhibits a more complex non-linear relationship (Ritz and Streibig, 2008). Crainiceanu et al. (2005) shows how penalized splines can be fitted within a mixed model framework in WinBugs to perform different non-parametric Bayesian analysis.

Define a regression model as:

$$y_i = f(x_i) + \varepsilon_i, \tag{4.6}$$

where $f(\cdot)$ is a smooth function and $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. Authors suggest using low-rank thin-plate splines because posterior correlation is lower between parameters which improves the mixing:

$$f(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k |x - \kappa_k|^3, \tag{4.7}$$

where $\boldsymbol{\theta} = \beta_0, \beta_1, u_1, \ldots, u_K$ is the vector of regression coefficients for $k_1, \ldots, k_K$ fixed knots. The basis functions can then be defined accordingly as:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ and } \mathbf{Z} = \begin{bmatrix} |x_1 - \kappa_1|^3 & \ldots & |x_1 - \kappa_K|^3 \\ \vdots & \ddots & \vdots \\ |x_n - \kappa_1|^3 & \ldots & |x_n - \kappa_K|^3. \end{bmatrix}. \tag{4.8}$$

Thus, $\hat{\theta}$ minimizes the expression :

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}||^2, \tag{4.9}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ and $\mathbf{u} = (u_1, \ldots, u_K)^T$. The choice of knots and their position influence the shape of the spline function. Thus, to avoid overfitting (or alternatively over smoothing) penalized splines are used:

$$\hat{\theta} = \operatorname{argmin} \left[\sum_{i}^{n} (\mathbf{y} - f(x, \boldsymbol{\theta})) + \lambda \mathscr{J}_\theta \right]^2, \tag{4.10}$$

where $\mathscr{J}_\theta$ is the roughness penalty measuring the smoothness of $f(\cdot)$ and λ the tuning parameter controlling the trade-off between smoothness and model fit.
This can be written in terms of a semi-definite penalty matrix $\mathbf{D}$, i.e. $\mathscr{J}_\theta = \boldsymbol{\theta}^T \mathbf{D} \boldsymbol{\theta}$. Such that,

$$D = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & (\Omega_K^{1/2})^T \Omega_K^{1/2}, \end{bmatrix}$$

where the l,k-th entry of matrix Ω_K is $|\kappa_l - \kappa_k|^3$ and divided by σ_ε^2 the penalized spline becomes:

$$\frac{1}{\sigma_\varepsilon^2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}||^2 + \frac{\lambda^2}{\sigma_\varepsilon^2} \mathbf{u}^T (\Omega_K^{1/2})^T \Omega_K^{1/2} \mathbf{u}. \tag{4.11}$$

Crainiceanu et al. (2005) specifies the smoothing parameter to be λ^{-1} instead. Then, let $\sigma_u^2 = \lambda \sigma_\varepsilon^2$, $\boldsymbol{\beta}$ a vector of fixed parameters and $\mathbf{u}$ a set of random coefficients such that $\mathbb{E}(\mathbf{u}) = 0$ and $\mathrm{cov}(\mathbf{u}) = \sigma_u^2 \Omega_K^{-1/2} (\Omega_K^{-1/2})^T$.

The penalized regression spline as a mixed model of the form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \qquad \text{Cov} = \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}. \tag{4.12}$$

Thus, $\mathbf{Y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}^T G\mathbf{Z} + R)$. With $R = \sigma_\varepsilon^2 \mathbb{I}_n$. Alternatively, by setting $\mathbf{b} = \Omega_K^{1/2}\mathbf{u}$ and $\mathbf{Z}^* = \mathbf{Z}\Omega_K^{-1/2}$ the LMM is equivalent to:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{b} + \boldsymbol{\varepsilon} \qquad \text{Cov} = \begin{bmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 \mathbb{I}_K & 0 \\ 0 & \sigma_\varepsilon^2 \mathbb{I}_n \end{bmatrix} \tag{4.13}$$

Classical approaches to fit this model rely on using the Best linear Unbiased Predictor (BLUP), i.e. for arbitrary $1 \times n$ vectors $\mathbf{s}$ and $\mathbf{t}$, to find $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ to minimize the prediction error

$$\mathbb{E}\left[(\mathbf{s}^T\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{t}^t\mathbf{Z}^*\tilde{\mathbf{b}}) - \mathbf{s}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{t}^t\mathbf{Z}^*\mathbf{b})\right]^2, \tag{4.14}$$

subject to

$$\mathbb{E}\left[(\mathbf{s}^T\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{t}^t\mathbf{Z}^*\tilde{\mathbf{b}})\right] = \mathbb{E}\left[\mathbf{s}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{t}^t\mathbf{Z}^*\mathbf{b})\right] \tag{4.15}$$

then,

$$\begin{aligned} \text{BLUP}(\boldsymbol{\beta}) &\equiv \tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} \\ \text{BLUP}(\mathbf{b}) &\equiv \tilde{\mathbf{b}} = \sigma_b^2\mathbb{I}_K\mathbf{Z}^{*T}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \end{aligned}$$

where $\mathbf{V} = \mathbf{Z}^{*T}\sigma_b^2\mathbb{I}_K\mathbf{Z}^* + \sigma_\varepsilon^2\mathbb{I}_n$.

Most common software implements this approach by substituting in the ML or REML estimates of $\hat{\mathbf{V}}$ and $\hat{G} = \hat{\sigma_b^2}$, producing EBLUPs (estimated BLUPs) that have two sources of uncertainty due to estimation of $(\boldsymbol{\beta}, \mathbf{u})$ and $(\mathbf{G}, \mathbf{V})$. Accounting for this two step estimation error is not straightforward. However, Bayesian methods produce credible intervals that do not use the plug in method as they account for the variability of each parameter.

Following the work of Bled et al. (2013), state variables in the dynamic occupancy model can be formulated in a generalized penalized spline model to accommodate non-linear relationships between colonization and survival. These parameters are assumed to be constant across all years and only vary between sites (given the site-level effects are present). However, if site level covariate changes over time are known to influence species survival or colonization probabilities, then yearly specific parameters can be specified. If that is not the case, then the generalized penalized spline dynamic occupancy model can be formulated as follows:

$$\begin{aligned} \psi_{jt} &= z_{jt-1}\phi_j + (1 - z_{jt-1})\gamma_j \\ \text{logit}(\phi_j) &= \mathbf{X}\boldsymbol{\phi} + \mathbf{Z}^*\mathbf{u} \\ \text{logit}(\gamma_j) &= \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}^*\mathbf{v}, \end{aligned} \tag{4.16}$$

where $\boldsymbol{\phi} = (\phi_0, \phi_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are both the logit-scale intercepts and slopes for the effect that covariate x has on colonization and survival respectively. $\mathbf{X}$ is the design matrix for the fixed effects with the jth row $\mathbf{X}_j = (1, x_{1_j})$. $\mathbf{Z}^* = \mathbf{Z}\Omega_k^{-1/2}$ is the design matrix for the random effects with coefficients given by $\mathbf{u} = (u_1, \ldots, u_K)$ and $\mathbf{v} = (v_1, \ldots, v_K)$. The initial occupancy state z_{j1} of site j in the first time period is defined as:

$$\begin{aligned} z_{j1} &\sim \text{Bernoulli}(\psi_{j1}) \\ \text{logit}(\psi_{j1}) &= \beta_0 + \beta_1 x_{1j}. \end{aligned} \tag{4.17}$$

In this work, model (4.16) will be extended to model a site's occupancy state with multiple species. Thus, if the true occupancy state is observed imperfectly, the hierarchical structure becomes:

$$\begin{aligned} z_{ij1} &\sim \text{Bernoulli}(\psi_{ij1}) \\ \text{logit}(\psi_{j1}) &= \beta_{0i} + \beta_{1i} x_{1j} \\ z_{ijt} &\sim \text{Bernoulli}(\psi_{ijt}) \quad \text{for } t = 2, \ldots, T \\ \psi_{ijt} &= z_{ijt-1}\phi_{ij} + (1 - z_{ijt-1})\gamma_{ij} \\ \text{logit}(\phi_{ij}) &= \phi_{0i} + \phi_{1i} x_{1j} + \sum_{k=1}^{K} u_{ik} z^*_{jk} \\ \text{logit}(\gamma_{ij}) &= \gamma_{0i} + \gamma_{1i} x_{1j} + \sum_{k=1}^{K} v_{ik} z^*_{jk} \\ y_{i,j,n,t} | z_{ijt} &\sim \text{Bernoulli}(p_{i,j,n,t}) \quad \text{for } n = 1, \ldots, N \\ \text{logit}(p_{i,j,n,t}) &= \alpha_{0i} + \alpha_{1i} x_{2jnt}, \end{aligned} \tag{4.18}$$

where z^*_{jk} is the (j,k)-th entry of the design matrix $Z^* = \mathbf{Z}\Omega_K^{-1/2}$ (Eqn. 4.13) for the random effects with coefficients given by $\mathbf{u}_j = (u_{i1}, \ldots, u_{iK})$ and $\mathbf{v}_i = (v_{i1}, \ldots, v_{iK})$ for fixed K number of knots. The latent variables β_{0i} and β_{1i} are the species specific initial occupancy intercept and slope, γ_{0i} and ϕ_{0i} are the species-specific baseline colonization and survival logit-scale probabilities, and γ_{1i} and ϕ_{1i} are the species-specific slopes for the effect that covariate x has on colonization and survival.

The observation process for the individual detection at site j of species i during the nth visit ($n \in 1, \ldots, N$) on time t is given by the detection probability defined by the species-specific intercepts $\alpha_{0,i}$ and slopes $\alpha_{1,i}$. The species-specific parameters are drawn from a common prior distribution. Choice of priors and model fitting details are discussed in the simulation analysis work below. Note that under this framework, non-linear relationships could also be included for any covariate on the observational model by defining a set of $m_{it1}, \ldots, m_{itK}$ random coefficients, where the design matrix $\mathbf{G}\Omega_k^{-1/2}$ becomes now an $l \times K$ matrix for K fixed knots for each t time period. The l-th entry is written in terms of the number of visits per site (j,n) such that $l = N(j-1) + n$.

Thus, the detection probability can be formulated as:

$$\text{logit}(p_{i,l,t}) = \alpha_{0i} + \alpha_{1i}x_{2lt} + \sum_{k}^{K} m_{itk}g_{lkt}. \tag{4.19}$$

Multiple species occupancy models are usually parametrized in an autologistic form, i.e. (Banner et al., 2019; Royle and Kéry, 2007; Royle and Dorazio, 2008) :

$$\begin{aligned}
\psi_{ijt} &= z_{ijt-1}\phi_i + (1 - z_{ijt-1})\gamma_i \\
&= \gamma_i + z_{ijt-1}\phi_i - z_{ijt-1}\gamma_i \\
&= \gamma_i + z_{it-1}(\phi_i - \gamma_i) \\
\text{logit}(\psi_{ijt}) &= a_i + b_i(z_{ijt-1}).
\end{aligned} \tag{4.20}$$

Including covariate effects equation (4.20) becomes:

$$\text{logit}(\psi_{ijt}) = a_i + b_i(z_{ijt-1}) + \beta_{1i}x_j + \beta_{2i}x_j z_{ijt-1}, \tag{4.21}$$

where β_1 is the effect of x on the logit of colonization probability (since $b_i = \beta_{2i} = 0$ when $z_{ijt} = 0$) and $\beta_{1i} + \beta_{2i}$ is the effect of x on the logit of survival probability. a_i and $(b_i + a_i)$ are the species baseline colonization and survival probabilities on the logit scale.

It can be shown that the state part of the Markov model (Eqn. 4.18) developed in this chapter is equivalent to the aforementioned autologistic model. Ignoring smooth terms and based on equation (4.20) this is:

$$\psi_{ijt} = \text{logit}^{-1}(\gamma_{i0}) + z_{it-1}(\text{logit}^{-1}(\phi_{i0}) - \text{logit}^{-1}(\gamma_{i0})) \tag{4.22}$$

Including site-level covariates and conditioning the logit-scaled parameters on the latent occupancy state to avoid β_1 and $beta_2$ to be confounded yields to:

$$\text{logit}(\psi_{ijt}) = a_i + b_i(z_{ijt-1}) + \beta_1 x_{1j}(1 - z_{ijt-1}) + \beta_2 x_{1j} z_{ijt-1}. \tag{4.23}$$

Note that $b_i = \phi_{0i} - \gamma_{0i}$. Thus,

- $a_i = \gamma_{0i}$ logit baseline colonization
- $a_i + b_i = \phi_{0i}$ logit baseline survival
- $\text{logit}(\psi_{ijt})|(z_{ijt-1} = 0) = a_i + \beta_1 x_{1j} \equiv \gamma_{0i} + \gamma_{1i}x_{1j}$
- $\text{logit}(\psi_{ijt})|(z_{ijt-1} = 1) = a_i + b_i + \beta_2 x_{1j} \equiv \phi_{0i} + \phi_{1i}x_{1j}$

Including a smooth term to model the survival probability for example, leads to:

$$\text{logit}(\psi_{ijt}) = a_i + b_i(z_{ijt-1}) + \beta_{1i}x_j(1 - z_{ijt-1}) + \beta_{2i}x_j z_{ijt-1} + (z_{ijt-1})\sum_{k=1}^{K} v_{ik}B_k(X). \tag{4.24}$$

In which case the $B_k(X)$ represents the basis system for the flexible survival terms. According to Royle and Kéry (2007), autologistic formulation of the MSOM leads to a more efficient Bayesian implementation since Markov-chains exhibit lower autocorrelations. However, the direct interpretation of the covariate effects on survival and colonization is not possible due to effects being confounded within autologistic parameters a_i and b_i. Under a generalized penalized splines model the parameters' interpretability becomes a very important feature that the Markov formulation retains and thus, this will be the approach to be taken in the following section.

4.3 Simulation study

In this section a simulation analysis of the generalized penalized splines MSOM introduced in the previous section of this chapter will be presented in order to provide and discuss the methodological aspects that need to be considered to apply such models in real-life situations such as the odonata case of study. The main questions that will be addressed with these simulations are:

- to what extent does the proposed model captures the different degrees of complexity in the relationship between the response and the predictor?
- how does the number of visits affects the model's efficiency to estimate complex nonlinear relationships?
- how close do the estimated community diversity metrics are to the true community parameters?
- what potential issues or considerations may arise when a dynamic occupancy model is formulated within a flexible framework?

First, to illustrate the flexibility of model 4.16 to estimate non-linear relationships in colonization and survival, different survival and colonization probabilities curves will be simulated for a single species. To do so, zero-centered site-level and survey-specific covariates (x_1, x_2) are simulated for a fixed number of 500 sites and 4 primary sampling periods (from now on referred to as time periods).

The initial occupancy probability at time $t = 1$ is defined by the following logit-scale intercept and slope (Fig. 4.2):

$$\text{logit}(\psi_{j1}) = -1 + 1.5 \times x_{1j} \quad \text{for } j = 1, \ldots, 500. \tag{4.25}$$

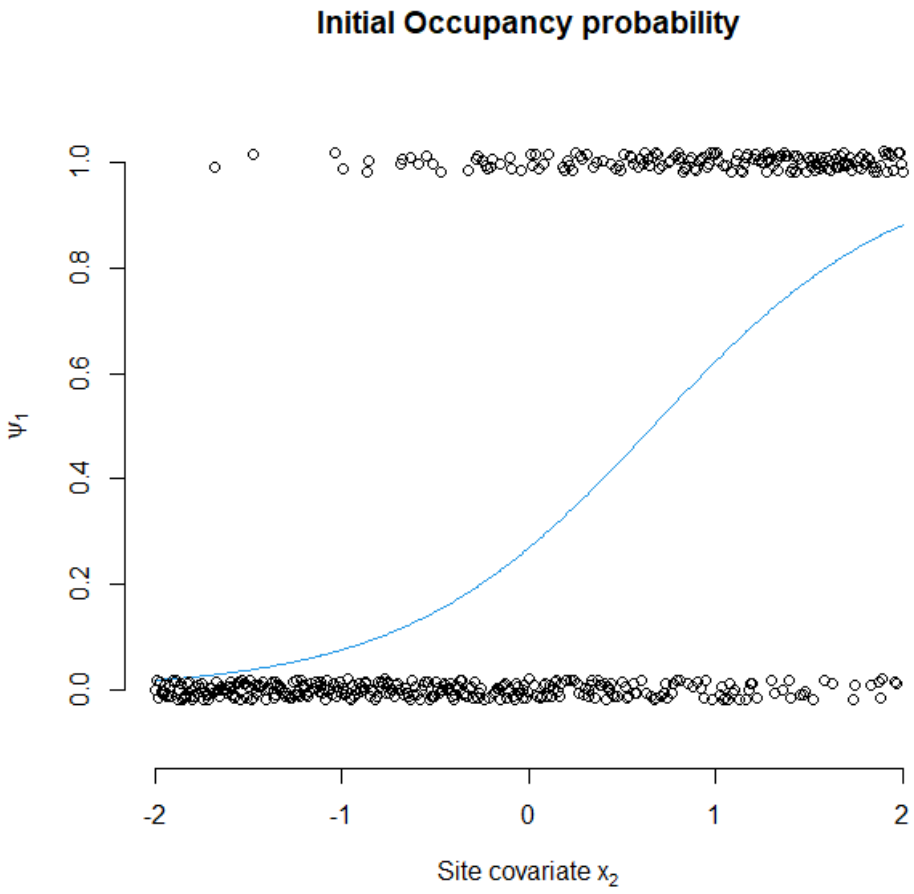


Figure 4.2: Simulated initial occupancy probability for a single species as a function of a site level covariate

.

Subsequent occupancy states at time $t+1$ are defined based on colonization and survival probabilities. For a first scenario, logit-scale colonization probability will be modelled using a cubic term for the site level covariate effect ($\text{logit}(\gamma_j) = 1.39 - 1.5x_{2j} - 1.5x_{2j}^2 + 1.5x_{2j}^3$) and survival probability will be modelled using a quadratic term ($\text{logit}(\phi_j) = 1.39 - 1.5x_{2j} - 2.5x_{2j}^2$).

The second scenario simulates non-linear responses by creating a set of b-spline basis functions using the `splines` package in R to allow for different degrees of smoothness between colonization and survival. Colonization curves were simulated by using a set of b-splines of degree $d = 2$ and $k = 5$ equidistant inner knots defined by the sample quantiles. Survival curves on the other hand, were simulated by setting a b-spline basis of degree $d = 3$ and $k = 8$ equidistant knots. Boundary knots were defined by the maximum and minimum value of x_1. These values were chosen after a thorough investigation of different curve shapes and patterns to ensure reasonable coverage of the colonization and survival curves range and complexity within a reasonable amount of time.

Colonization and survival parameters (γ_1, ϕ_1) and smooth term coefficients (u_k, v_k) were drawn from normal distributions centered at zero with a standard deviation of one. Baseline colonization and survival probabilities were set to $\gamma_0 = \phi_0 = 0.8$.

The state process is then simulated recursively for the occupancy states z_{jt} at time period $t > 1$ at each site j according to:

$$\begin{aligned} z_{j,t} &\sim \text{Bernoulli}(\psi_{j,t}) \\ \psi_{j,t} &= z_{j,t-1} \times \phi_j + (1 - z_{j,t-1}) \times \gamma_j. \end{aligned} \tag{4.26}$$

Thus, if site j is unoccupied $(1 - z_{ijt})$ on t then it can be colonized on $t + 1$ with probability γ_j. Likewise, if site j is occupied (z_{jt}) in t, it remains occupied (survives) with a probability of ϕ on $t + 1$. Then, the observation process can be simulated by defining by the following relationship:

$$\begin{aligned} y_{j,n,t} &\sim \text{Bernoulli}(p_{j,n,t}) \\ \text{logit}(p_{j,n,t}) &= \alpha_0 + \alpha_1 x_{2jnt}, \end{aligned} \tag{4.27}$$

where $y_{j,n,t}$ are the individual detections for the j-th site on time t for the n visit or sampling occasion. $\alpha_0 = -2$ is the logit-scale baseline detection and $\alpha_1 = 3$ is the coefficient associated with the effect that survey-level covariate x_{2jnt} has on the detection. In both scenarios, non-linear functions for colonization and survival will be simulated and compared when data from a single visit ($N = 1$) and three multiple visits ($N = 3$) are used. For simplicity it is assumed that every site had the same number of visits, however this could be relaxed by indexing the total number of visits by each site, i.e. N_j.

The number of visits was selected based on previous simulations results (see simulation analysis in 2) that suggest three visits be sufficient for estimating occupancy probabilities with a reasonably low error. Additionally, sampling locations in long-term ecological studies are typically visited between 1 to 3 times a year due to sampling costs (Kellner and Swihart, 2014). This is also the case in the Odonata case study where the number of visits for a large collection of sites is determined by the presence of a few or even a single species (this will be further discussed in chapter 5).

Model (4.16) was fitted in `nimble` because of its computational efficiency (which has been discussed previously on chapter 2). Three independent chains with a total of 50,000 iterations, a burnin period of 2000 and a thinning of 10 were specified to sample from the posterior distribution.

The choice of priors for the logit-scale parameters (for both the state/ecological and observational processes in (4.16)) should assign low probabilities outside the range of (-5,5) as an increase of 5 on the logit-scale is equivalent to a shift of 0.5 on the probability scale. Several authors have discussed different priors specification for logistic regression and occupancy models. For instance, Gelman et al. (2008) suggested using Student-t distributed priors with scale $\sigma^2 = 2.5$ and 1 degree of freedom to obtain relatively flat values for most of the (0,1) probability range. Dorazio et al. (2011) proposed using Student t distributed priors with scale parameter of 1.566 and degrees of freedom 7.763 instead. Other priors such as logistic (0,1) (Dorazio, 2016) and Normal ($0,2.25^2$) (Broms et al., 2016) have also been found to be suitable for occupancy models. A further discussion of prior choice is provided at the end of this chapter.

After a sensitivity analysis to the aforementioned priors, weakly informative zero centered Student t distributed priors with scale parameter of 1.566 and degrees of freedom 7.763 were chosen for $\beta_0, \beta_1, \gamma_0, \gamma_1, \phi_0, \phi_1, \alpha_0$ and α_1 hyperparameters. As for the smooth coefficients, the specified priors were $u_k \sim \text{Normal}(0, \tau_u)$ for colonization and $v_k \sim \text{Normal}(0, \tau_v)$ for the survival probability. Half-Cauchy priors were specified for the precision parameters. This choice of priors was found to be the most stable among different priors such as inverse-gamma for variance terms or normal distributed priors for the means (important considerations will be discussed at the end of this chapter).

The number of knots used to establish the basis dimension was $K = 35$ for both smooth functions. This is considered to be large enough to avoid over-smoothing as the exact choice of K is not relevant because the smoothness degree is determined by the penalty parameter λ (Eqn. 4.10), which is estimated through the precision parameters of the smooth terms (Eqn. 4.13). A large value of λ produces more constrained priors and thus, smoother estimates of the random coefficients (Rushing et al., 2019). Model diagnostics showed a reasonable convergence of the MCMC samplers for both scenarios (example for some the parameters traceplots are shown in appendixC.3 - figures23 and 24).

The results of fitting the generalized penalized splines MSOM (Eqn. 4.16) when species detection is simulated under a 3 repeated visits scheme are shown on Figures 4.3 and 4.4. The model estimates lie very closely to the true values when quadratic or cubic relationships are simulated for the state variables (Fig. 4.3). However, when b-splines are used to simulate non linear responses, the smaller peaks and troughs are not captured by the model and a rather smooth curve is estimated instead. When data from a single visit were used, the estimated colonization and survival probabilities are still close to the true values. However, the uncertainty around these estimates becomes larger (especially for the observational process) (Figs. 4.5 and 4.6). The results shown in Figure 4.6 indicate that as the smooth degree of the true curves decreases the estimated curves become smoother compared to the quadratic and cubic simulations (Fig. 4.5), suggesting that the number of visits greatly affects the occupancy estimates depending on the smoothness degree. Simulation results indicate that the uncertainty around the estimates becomes larger when data from a single visit are used and when detection probabilities are close to zero. Thus, the following simulation explores the parametrization of the observational process from Peach et al. (2017) that accounts for sampling effort as a power term to model non linear effects on detection when data from a single visit are used.

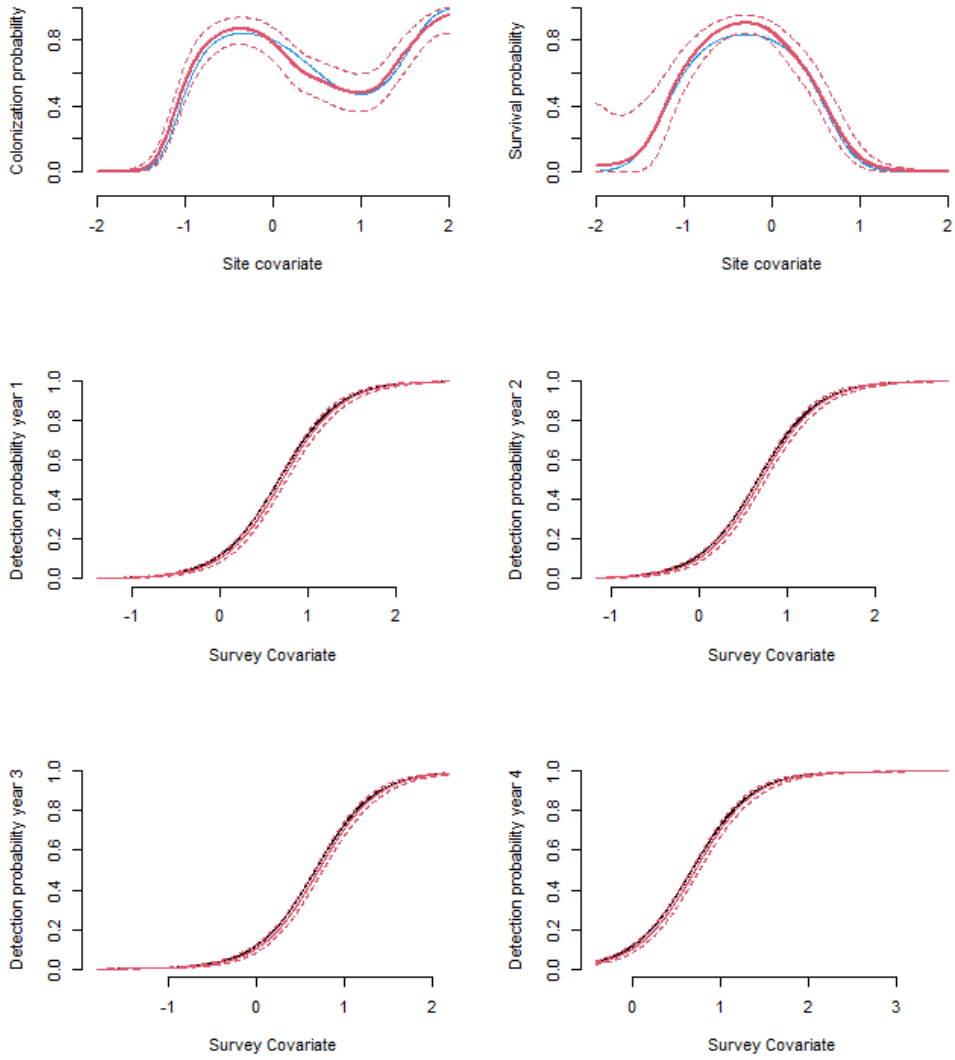


Figure 4.3: True cubic vs estimated colonization and quadratic survival curves for $N = 3$ repeated visits from fitting model 4.18 . Blue line represent the true relationship between parameters and site/survey-level covariates. The red solid lines represent the estimated probabilities with 95% Credible intervals indicated by the red dashed lines.

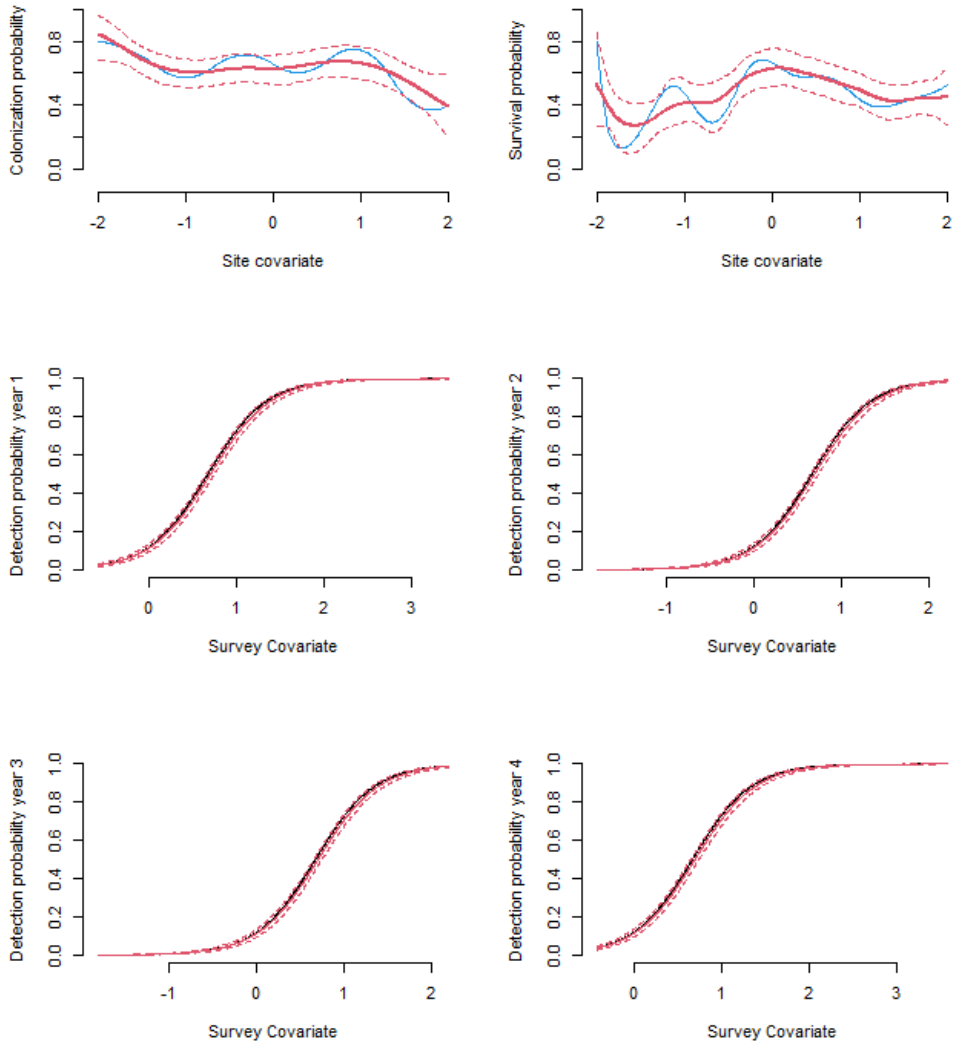


Figure 4.4: True vs estimated splines-based flexible colonization (knots = 5, degree = 2) and survival (knots = 8, degree = 3) curves for $N = 3$ repeated visits from fitting model 4.18. Blue line represent the true relationship between parameters and site/survey-level covariates. The red solid lines represent the estimated probabilities with 95% Credible intervals indicated by the red dashed lines.

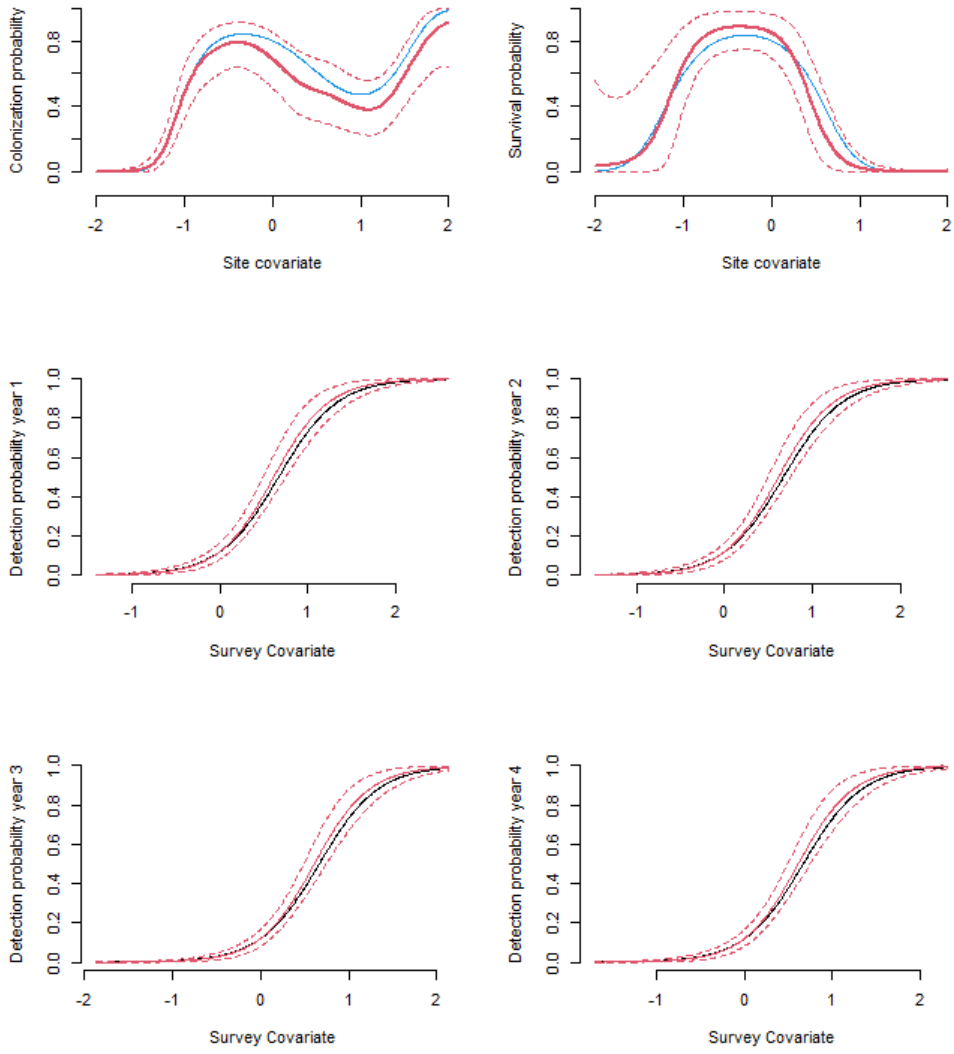


Figure 4.5: True vs estimated cubic colonization and quadratic survival curves for a single visit from fitting model 4.18. Blue line represent the true relationship between parameters and site/survey-level covariates. The red solid lines represent the estimated probabilities with 95% Credible intervals indicated by the red dashed lines.

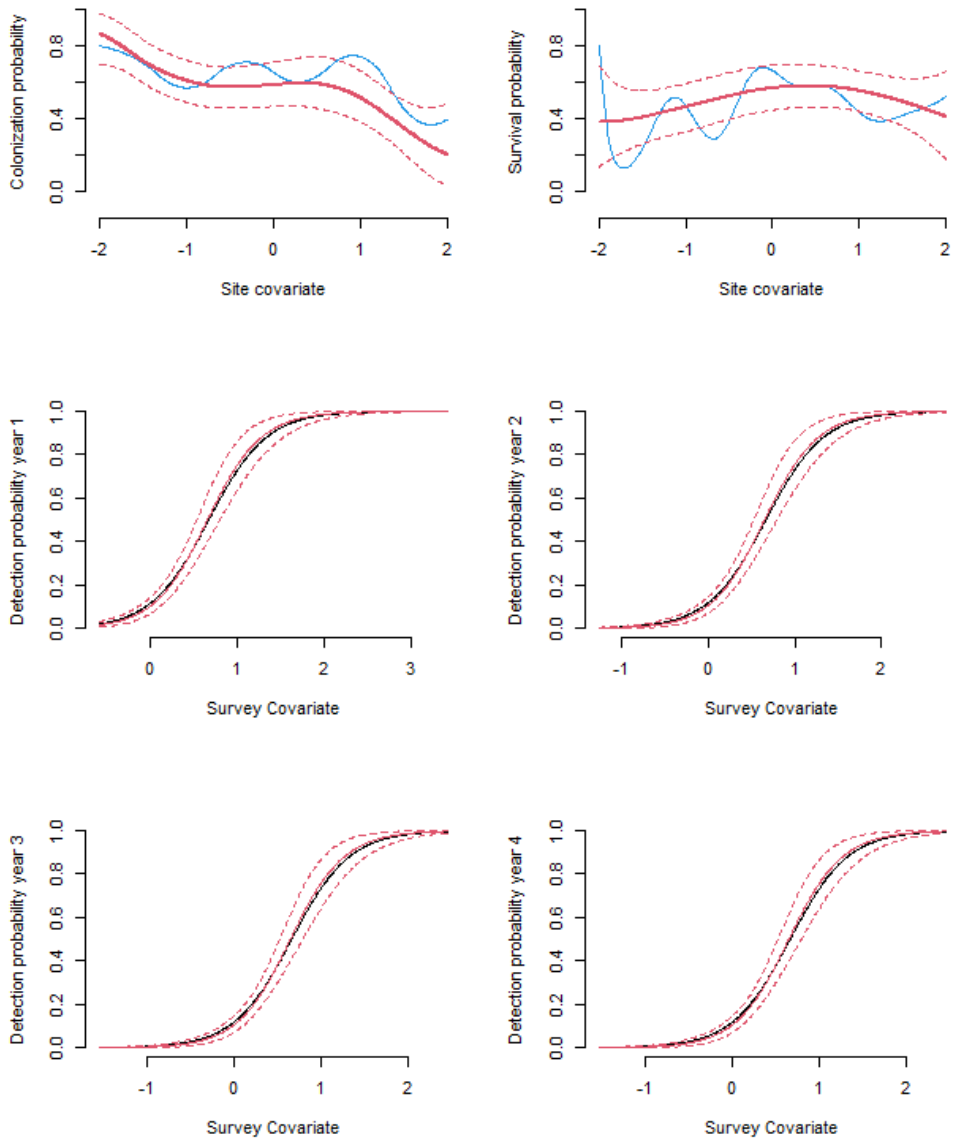


Figure 4.6: True vs estimated splines-based flexible colonization (knots = 5, degree = 2) and survival (knots = 8, degree = 3) curves for a single visit from fitting model 4.18. Blue line represent the true relationship between parameters and site/survey-level covariates. The red solid lines represent the estimated probabilities with 95% Credible intervals indicated by the red dashed lines.

To explore how the smoothness degree affects the model fit under a single visit sampling scheme, different survival probability curves will be simulated while keeping colonization parameters linear ($\text{logit}(\gamma_j) = -2 + 3 \times x_{1j}$). The scenarios that will be compared encompass the different non linear functions stated for the previous simulations. Then, the detection probability under a removal sampling scheme is defined for a single survey by simulating a survey-specific effort covariate. To achieve this, the observational process (Eqn. 4.27) is defined by the probability of spending n effort units before a detection is made, i.e.

$$\begin{aligned} y_j &\sim \text{Bernoulli}(p^*) \\ p^* &= F(\text{effort}|p) \\ &= 1-(1-p)^{\text{effort}_j} \qquad \text{for effort} = 1,\ldots,N, \end{aligned} \tag{4.28}$$

where $p = 0.025$ is the baseline detection probability (as an unit of effort increases, the probability of detecting a species increases by 0.025). This parametrization enforces the detection to be determined only by the baseline detection probability per unit effort. This could be useful in scenarios where survey-specific covariates are not available and the only source of information regarding the species detectability is the sampling effort. However, results indicate that uncertainty around estimated non-linear relationships with state variables is much greater when detection probabilities are close to zero (Fig.4.7). Moreover, this parametrization shows the same limitation identified on the previous simulation by not capturing the “wiggly" curve patterns in certain regions as the smoothness of the true curve decreases (Figs. 4.8 and 4.9). Another aspect to be considered is that the suggested parametrization would not be appropriate for describing scenarios in which survey-level covariates or sampling effort measurements show a more complex non linear relationship with the species detection. Thus, including the proposed smooth components into the observational process structure (as described in equation 4.19) enables specification of a more flexible model that can describe complex relationships between the detection process and the information available.

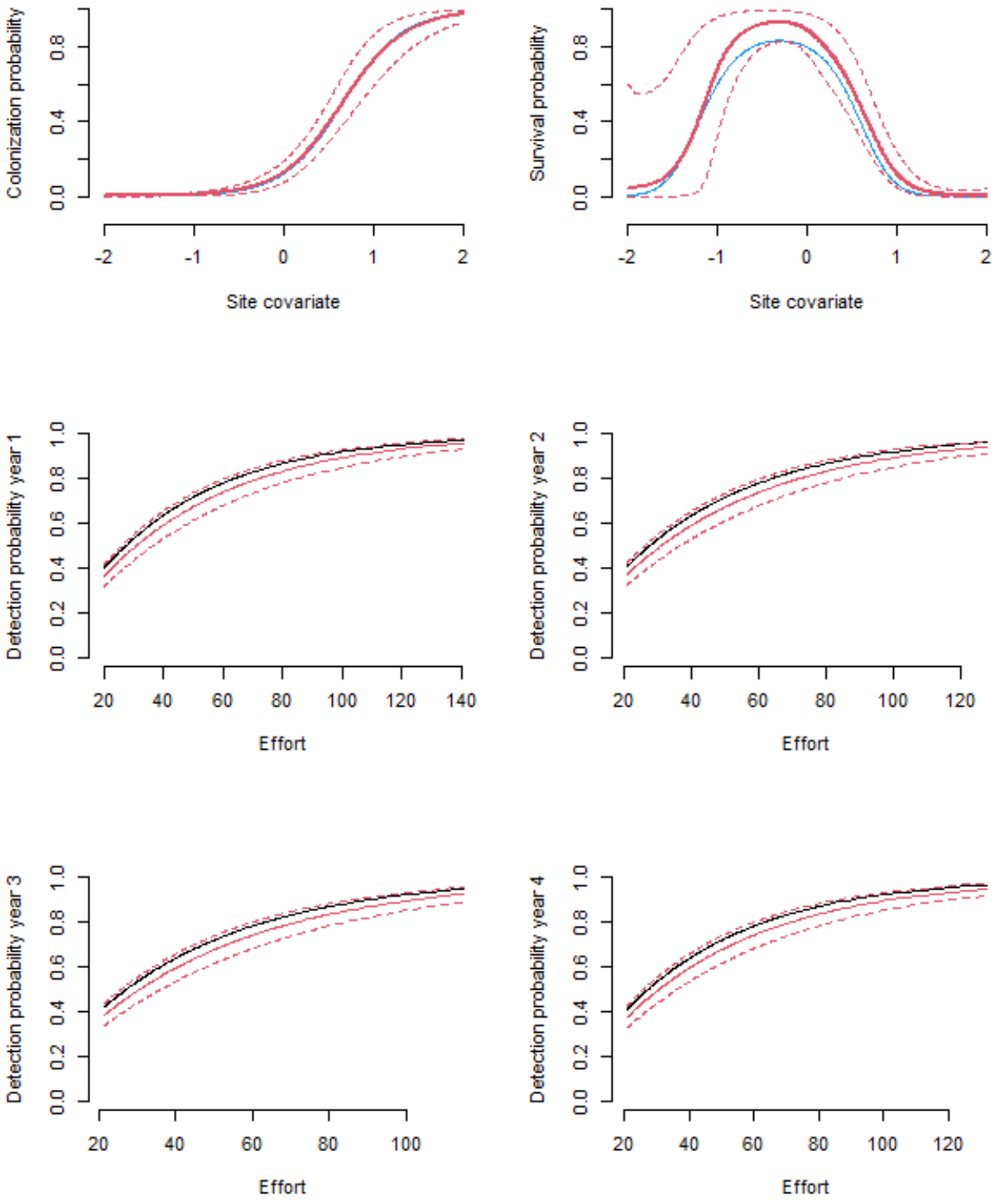


Figure 4.7: True vs estimated survival quadratic term, colonization and detection probabilities for a **single visit** from fitting model 4.18. Blue line represent the true relationship between parameters and site/survey-level covariates. Red line represents the estimated probabilities with 95% Credible intervals indicated by the red dashed lines.

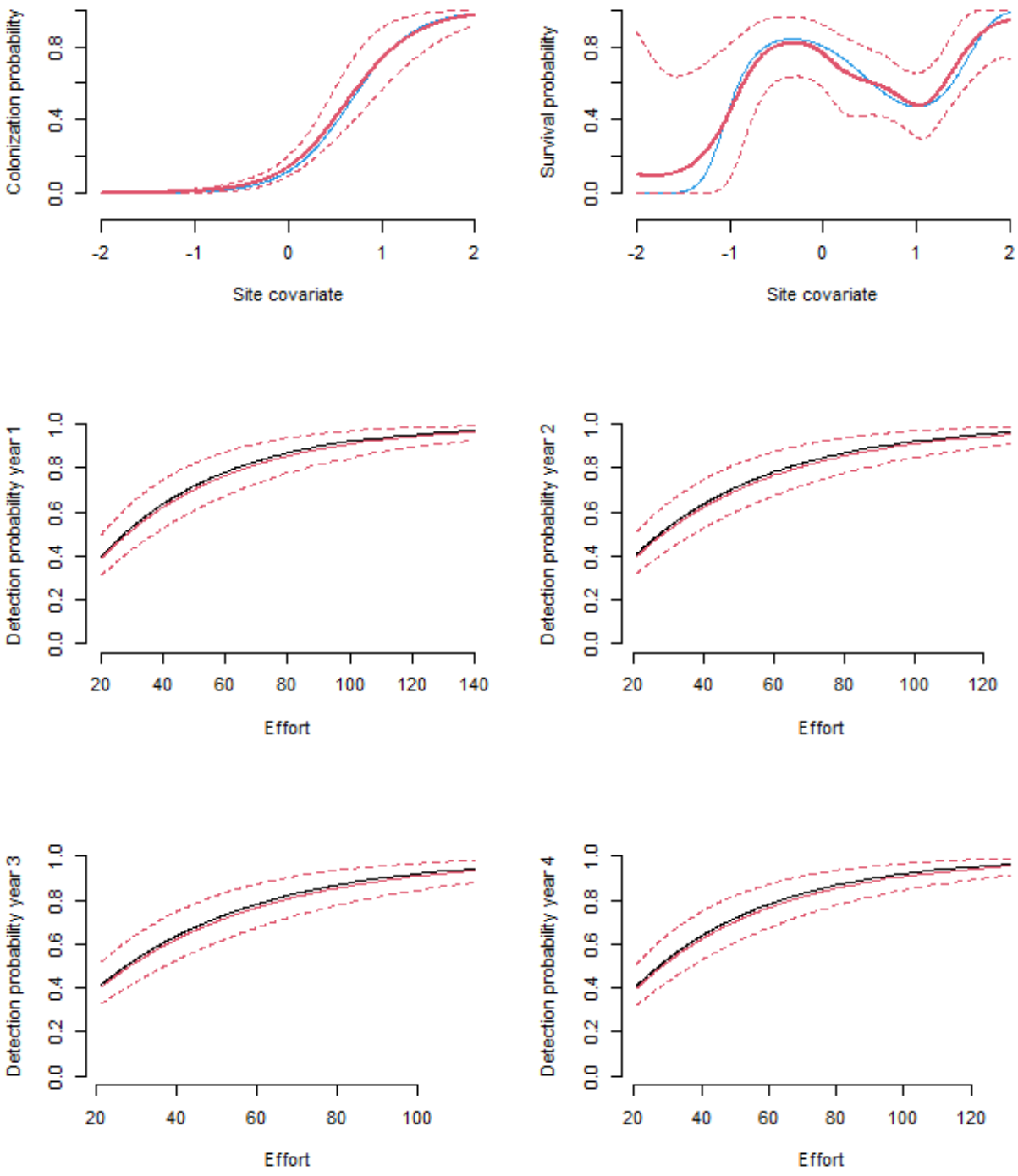


Figure 4.8: True vs estimated survival cubic term, colonization and detection probabilities for a **single visit** from fitting model 4.18. Blue line represent the true relationship between parameters and site/survey-level covariates. Red line represents the estimated probabilities with 95% Credible intervals indicated by the red dashed lines.

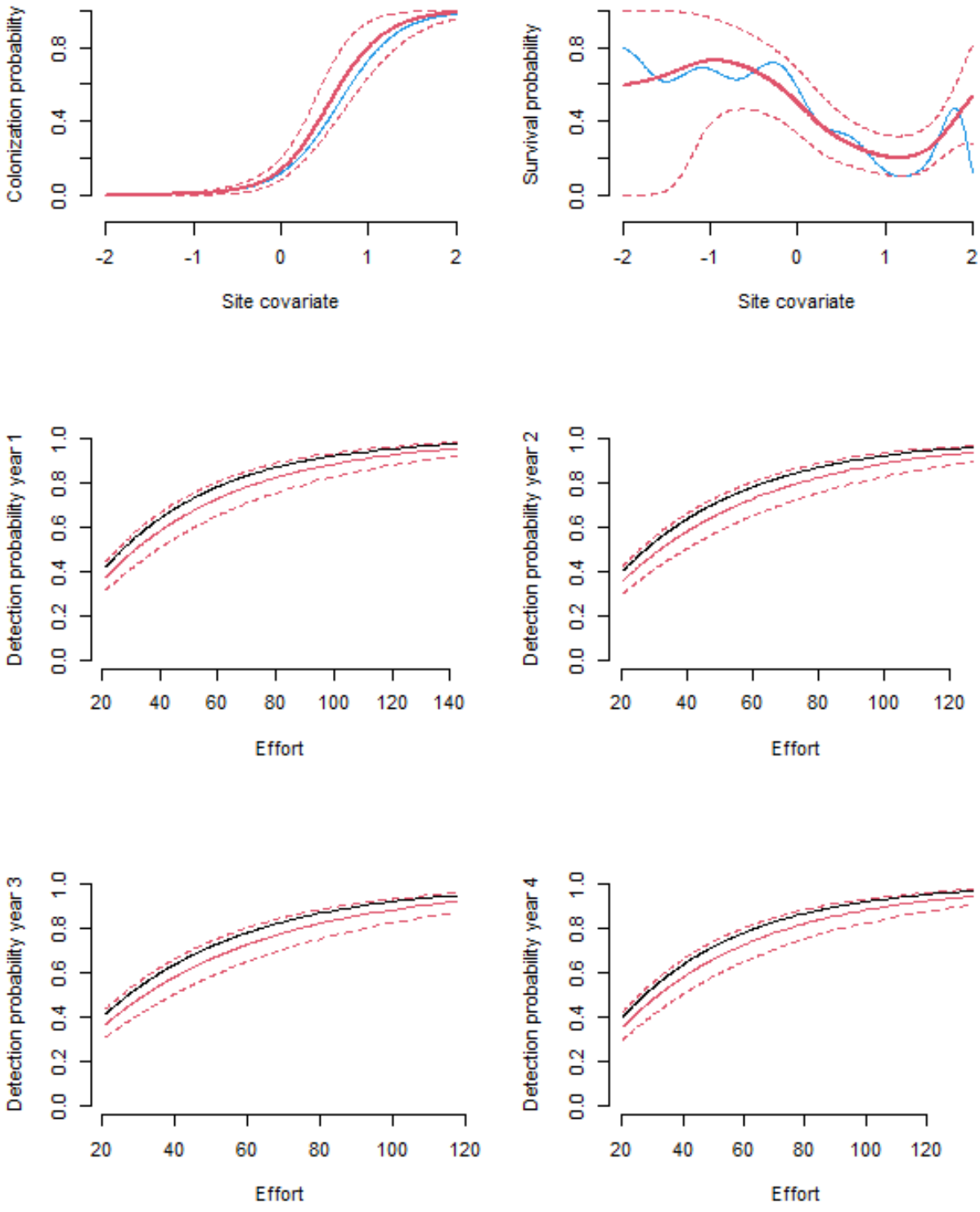


Figure 4.9: True vs estimated survival flexible term (knots = 8, degree = 3), colonization and detection probabilities for a **single visit** from fitting model 4.18. Blue line represent the true relationship between parameters and site/survey-level covariates. Red line represents the estimated probabilities with 95% Credible intervals indicated by the red dashed lines.

Community models enable estimation of occupancy patterns over a collection of sites for a whole group of species. Those patterns are the result of community level processes across sites and individual species-specific responses that affect both occupancy and detection probability. Therefore, the final simulation explores how the variety of individual species responses affect the model estimates when different non-linear relationships for colonization, survival, and detection are simulated for the different species within a community under a single visit or multiple visit sampling schemes. To assess the communities response under the different scenarios, the mean species richness across sites and the proportion of occupied sites across time will be compared. Besides these derived parameters, two other occupancy derived community parameters will be evaluated. First, the average growth rate per year defined as:

$$\bar{\lambda}_{it} = \frac{1}{J}\sum_{j}^{J} \psi_{ijt} / \frac{1}{J}\sum_{j}^{J} \psi_{ijt-1} \text{ for } t = 2, \ldots, T,$$

describes the average inter-annual changes in occupancy over the different years for each species (MacKenzie et al., 2003). Then, the total growth rate defined by Banner et al. (2019) as

$$\lambda_{tot} = \frac{1}{J}\sum_{j}^{J} \psi_{ij1} / \frac{1}{J}\sum_{j}^{J} \psi_{ijT},$$

to evaluate the net change in occupancy across time. To simulate multiple species occupancy states through time, species-specific initial occupancy probabilities were simulated according to model 4.18. Initial occupancy logit-scale intercepts and slopes were drawn from $\beta_{0i} \sim \text{Uniform}(-3,1)$ and $\beta_{1i} \sim \text{Uniform}(1,2)$. These values were chosen based on different simulations to avoid occupancy probabilities approaching zero, resulting in a sparser observed detection and thus, enforcing the sample size to be even larger in order to identify the changes in occupancy. The way in which different sample sizes affect occupancy model estimates is beyond the scope of this simulation study. However, general aspects on the effect of sample size on species occupancy have been presented previously in chapter 2 and can be found in a more recent simulation study by Banner et al. (2019) within the context of MSOM.

The initial occupancy probability at first time period $t = 1$ for $j = 1, \ldots, 500$ sites across $i \in 1, \ldots, 10$ species is given by $\text{logit}(\psi_{ij1}) = \beta_{0,i} + \beta_{1,i} x_{1j}$. Occupancy states at the following time periods $t \in t = 2, \ldots, 4$ are simulated by drawing different non-linear species-specific survival and colonization curves. Each species was randomly assigned to one out of four different functions: (1) simple linear model, (2) quadratic model, (3) cubic model and (4) b-spline smooth model of degree $d \in 2,3$ and $k \in 3, \ldots, 8$ inner knots. Baseline-logit scale intercepts for colonization and survival were drawn from $\phi_{0i}, \gamma_{0i} \sim \text{Uniform}(0.15, 0.8)$ to avoid probabilities being close to one or zero. Then, slope parameters were drawn from Uniform (-3,3) distributions for linear, quadratic and cubic forms and Normal (0,1) for b-splines smooth coefficients. This choice of parameters enables a heterogeneous community to be simulated to resemble a real biological community in which species responses vary widely to local environmental conditions. The different simulated curves are presented in Figure 4.10 and portray the heterogeneity in species-specific responses within the simulated community.

This same approach was then used to simulate species-specific detection probabilities as non linear functions of survey level covariate x_{2jnt} for $n = 1,3$ visits within time period t. Finally, observed detections were simulated according to equation 4.19 (where the basis system dimension is defined based on the degree of the simulated polynomial curve). Model 4.18 was fitted to the simulated data set in `nimble` using the same previous simulation settings. Species-specific parameters $\theta_q = \beta, \phi, \gamma, \alpha$ were drawn from common prior distributions Normal(μ_q, τ_q). Smooth coefficients u_{ik}, v_{ik}, m_{itk} where drawn from Normal$(0, \tau_u)$, Normal$(0, \tau_v)$, Normal$(0, \tau_m)$ respectively, with precision parameters drawn from Gamma(0.01,0.01).

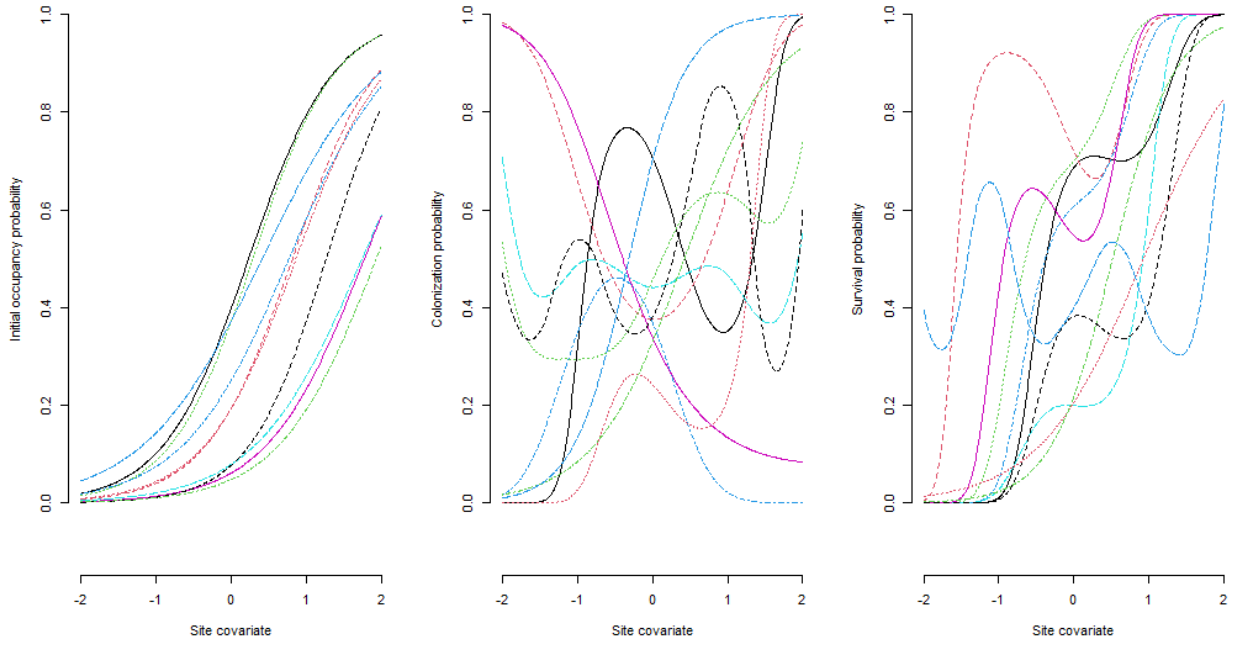


Figure 4.10: Simulated initial occupancy (left), survival (middle) and colonization (right) probabilities for 10 different species. Survival and colonization probabilities are defined as as a non linear function of site level covariate x_{1j} for $j = 1, \ldots, 500$ sites.

Estimated individual species curves lie closely to the true species curves for both colonization (Fig. 4.11) and survival (Fig. 4.12) probabilities when data from multiple visits are used. Uncertainty around the estimated curves is reasonably small with some exceptions in which probabilities lie closely to zero over some regions of x_1 (e.g survival probabilities for species 8 and colonization probability for species 1 and 3). Likewise, the observational process estimated curves capture the overall shape of the true curves with the exception of species 10 for which the smoothness degree of the simulated curve seems too low for the model to capture (Fig. 4.13).

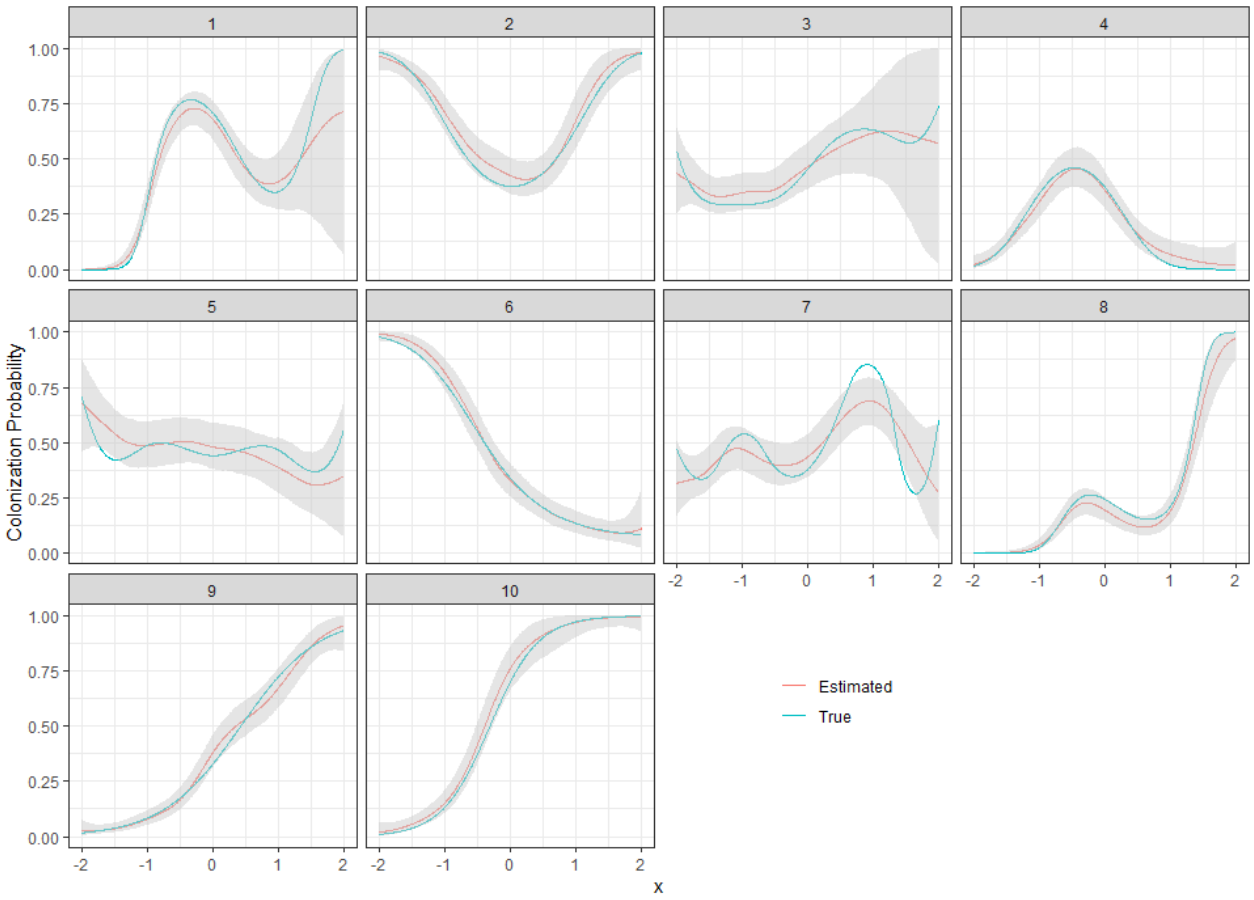


Figure 4.11: Generalized penalized splines multiple species MSOM (model 4.18) estimated species-specific colonization probabilities under a repeated visits sampling scheme. Red line representing the estimated curve with 95% Credible intervals denoted by the shaded region. Solid blue line indicates the true relationship between colonization and site-level covariate for each of the 10 simulated species.

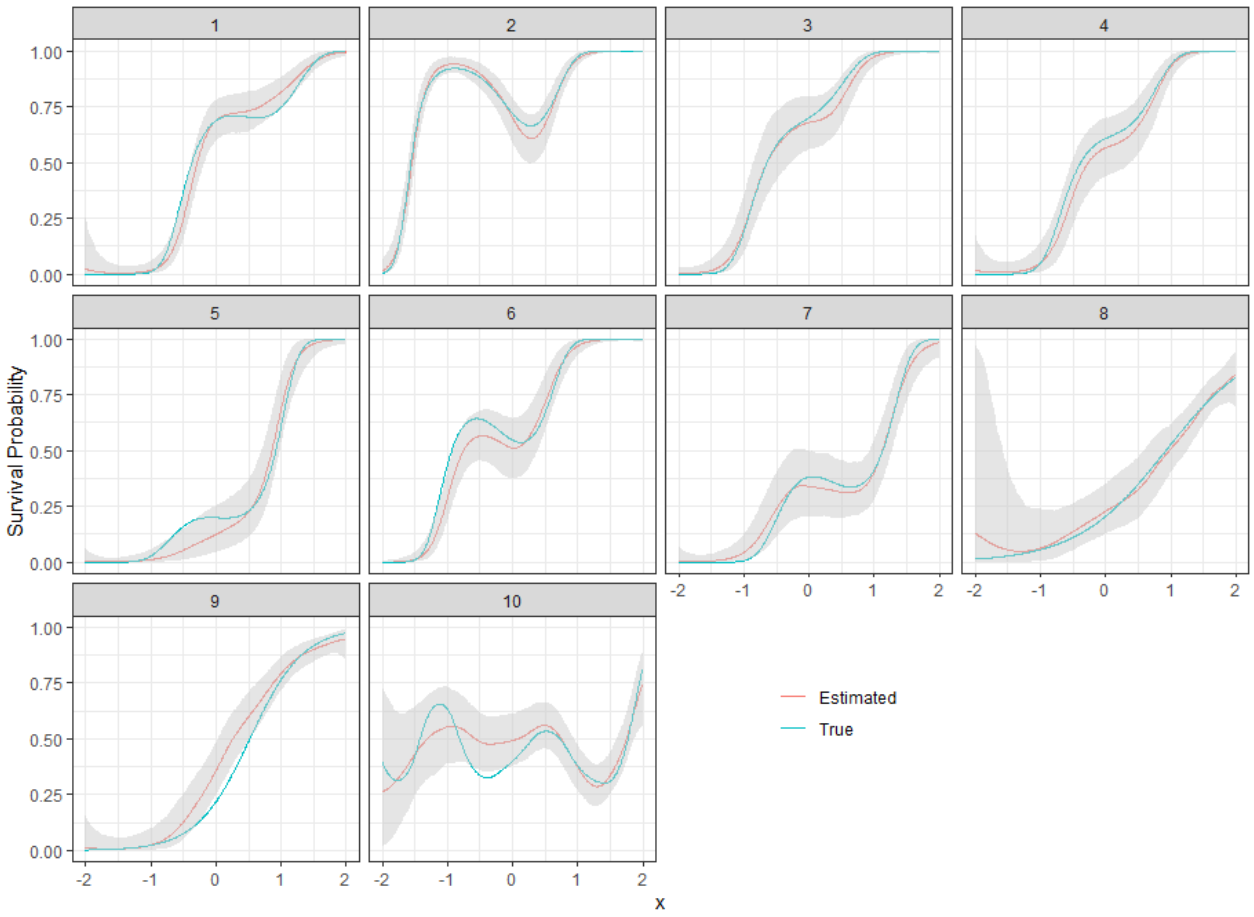


Figure 4.12: Generalized penalized splines multiple species MSOM (model 4.18) estimated species-specific survival probabilities under a repeated visits sampling scheme. Red line representing the estimated curve with 95% Credible intervals denoted by the shaded region. Solid blue line indicating the true survival curve for each of the 10 simulated species.

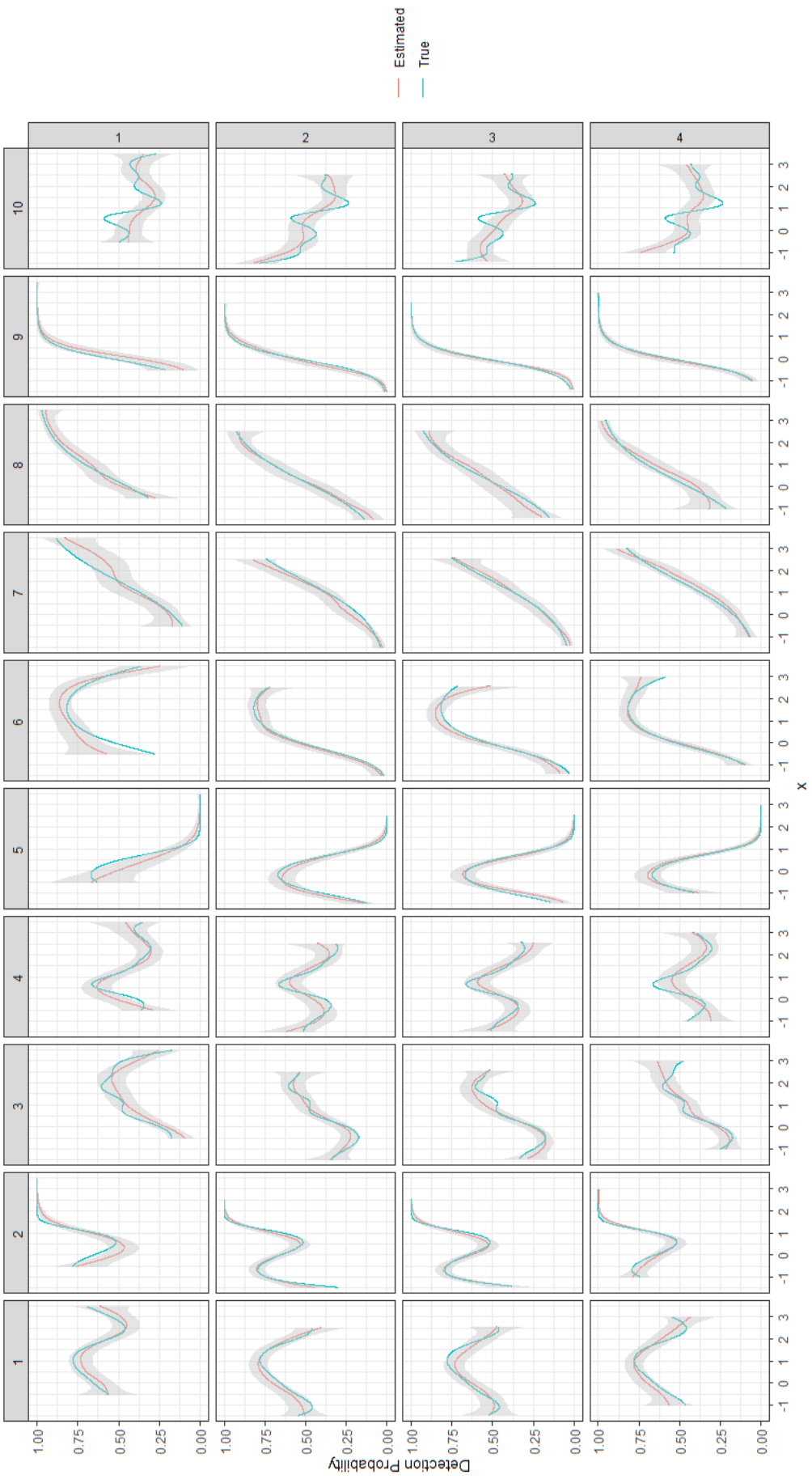


Figure 4.13: Generalized penalized splines multiple species MSOM (model 4.18) estimated species-specific detection probabilities under a repeated visits sampling scheme. Red line representing the estimated curve with 95% Credible intervals denoted by the shaded region. Solid blue line indicating the true detection curve for each of the 10 simulated species (columns) during 4 primary sampling periods (rows).

The results for fitting this model when data from a single visit are used are consistent with the previous simulation results. As the complexity of the curve increases, the uncertainty around the estimates becomes larger and the estimated curve becomes smoother because there is substantially less data to provide information that allows us to distinguish between the different sources of variation. This can be seen, for example, by comparing the species 2 estimated curves when data from a single visit are used. The estimated detection curve for this species (Fig. 4.16) is rather smooth compared to its original shape. Thus, the uncertainty associated with the estimated colonization and survival curves for this species (Figs. 4.14 and 4.15) will be larger than the estimated curves when data from multiple visits are available (Figs. 4.11 and 4.12). Nevertheless, very often in ecology the main quantities of interest are metrics that summarize the general patterns of different species' occupancy states within community, rather than individual species responses. Examples of such occupancy derived parameters (e.g. species richness and average growth per year) are calculated by model 4.18 under the single and multiple visits sampling schemes.

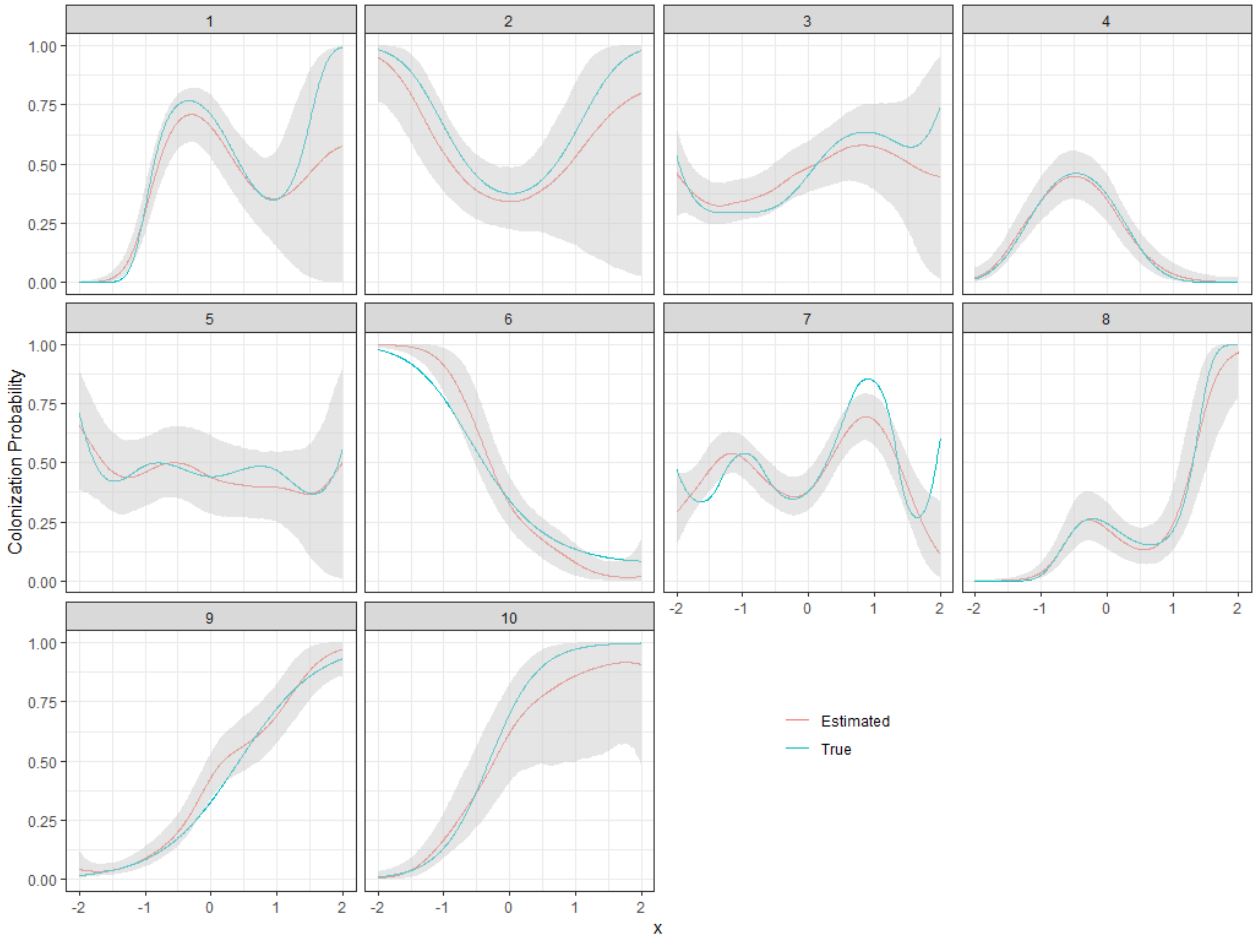


Figure 4.14: Generalized penalized splines multiple species MSOM (model 4.18) estimated species-specific colonization probabilities under a single visit sampling scheme. Red line representing the estimated curve with 95% Credible intervals denoted by the shaded region. Solid blue line indicates the true relationship between colonization and site-level covariate for each of the 10 simulated species.

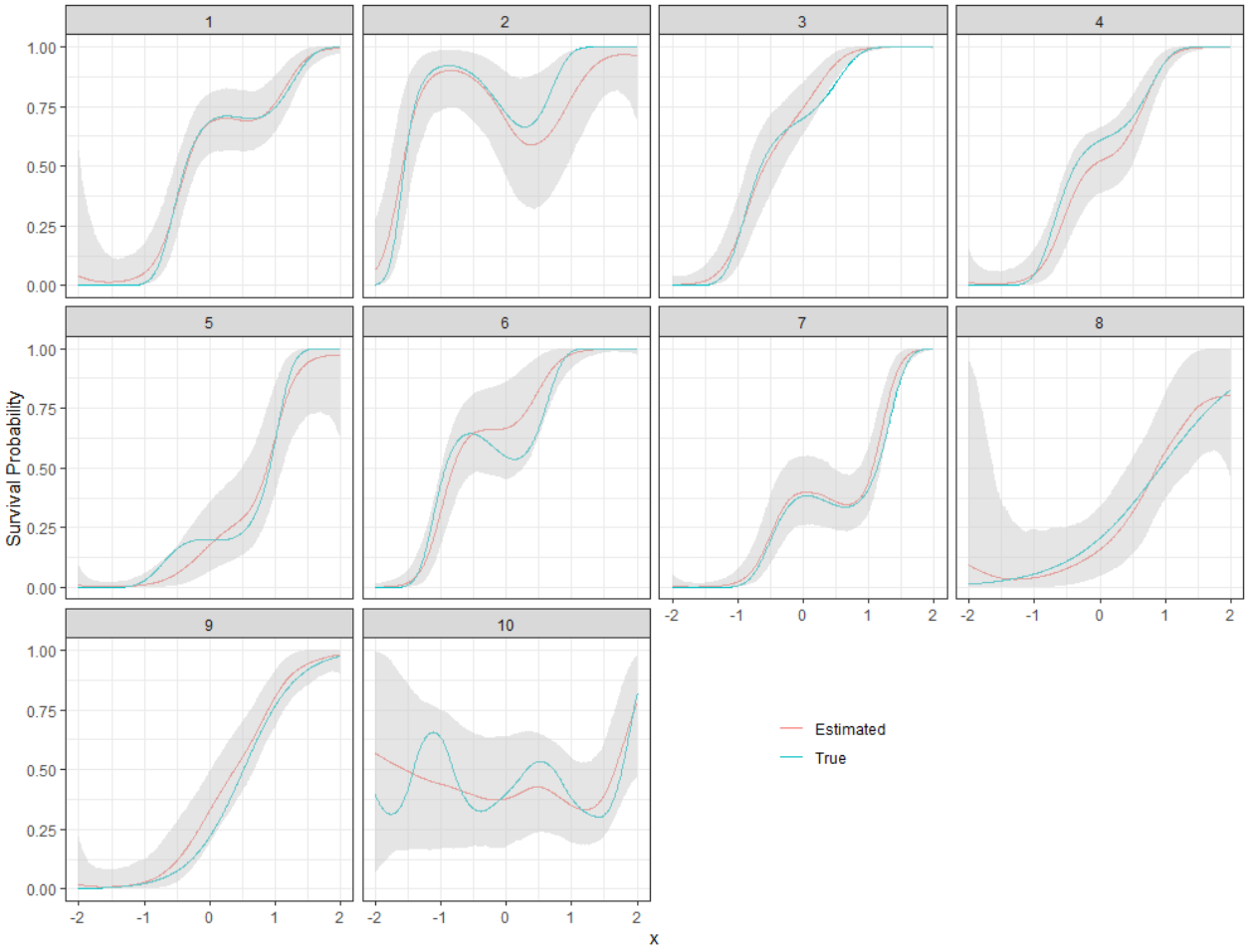


Figure 4.15: Generalized penalized splines multiple species MSOM (model 4.18) estimated species-specific survival probabilities under a single visit sampling scheme. Red line representing the estimated curve with 95% Credible intervals denoted by the shaded region. Solid blue line indicating the true survival curve for each of the 10 simulated species.

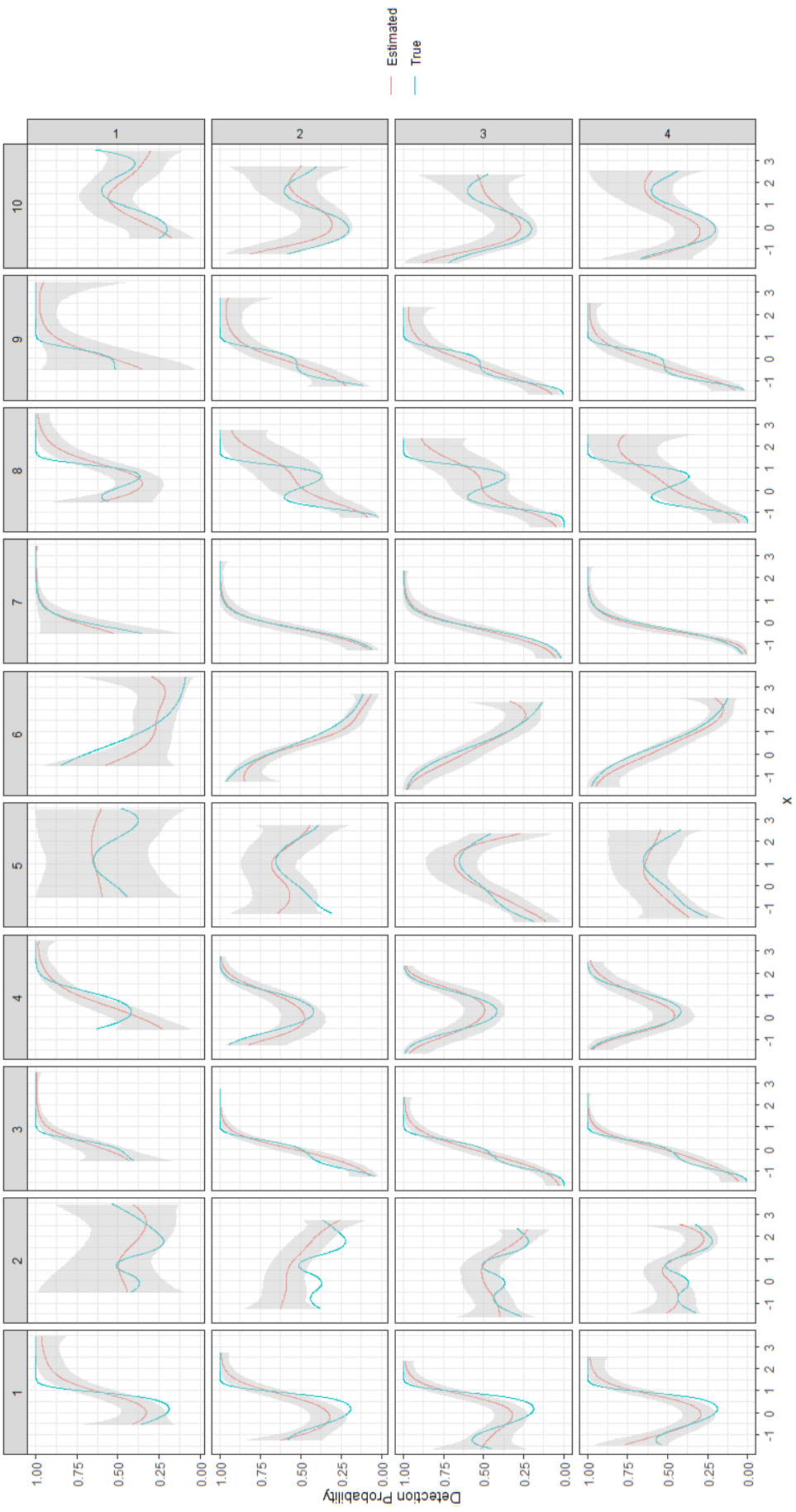
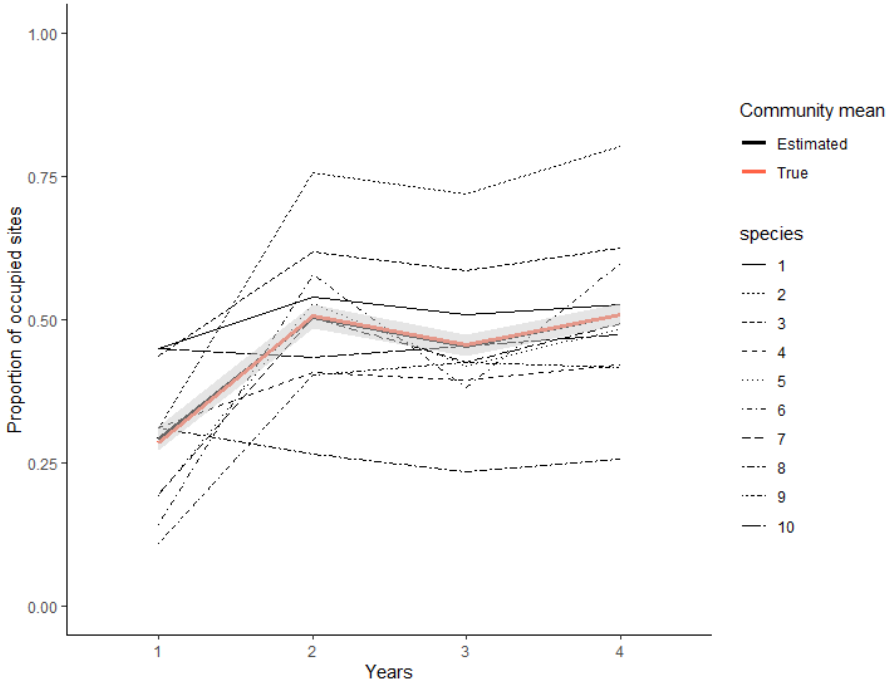
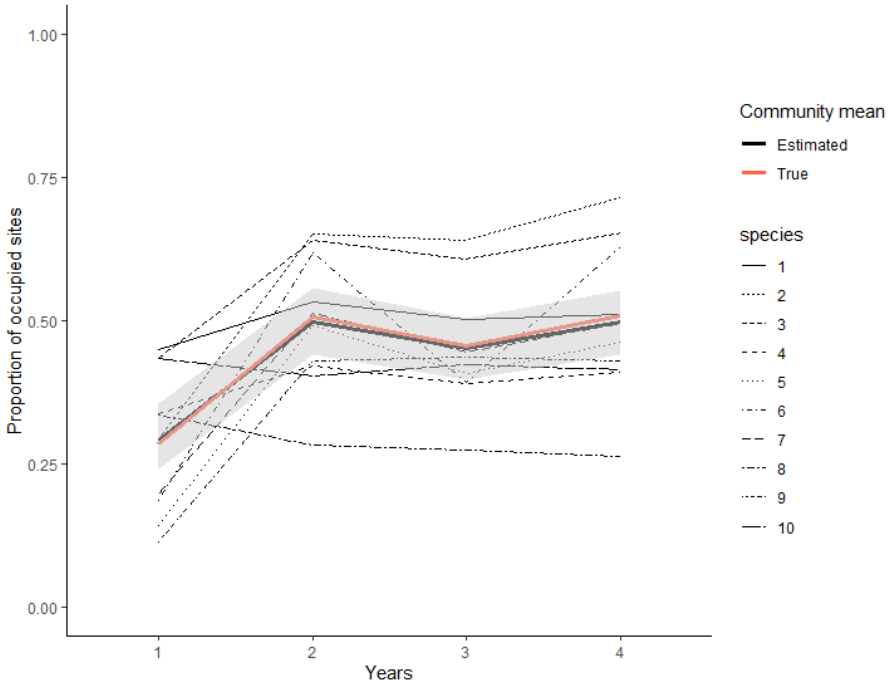


Figure 4.16: Generalized penalized splines multiple species MSOM (model 4.18) estimated species-specific detection probabilities under a single visit sampling scheme. Red line representing the estimated curve with 95% Credible intervals denoted by the shaded region. Solid blue line indicating the true survival curve for each of the 10 simulated species (columns) during 4 primary sampling periods (rows).

Simulation results indicate that estimated and true proportion of occupied sites across all species are very close to each other in both multiple visits (Fig. 4.17 top) or single visits (Fig. 4.17 bottom) scenarios. In which the latter, shows slightly wider credible intervals around the community average.



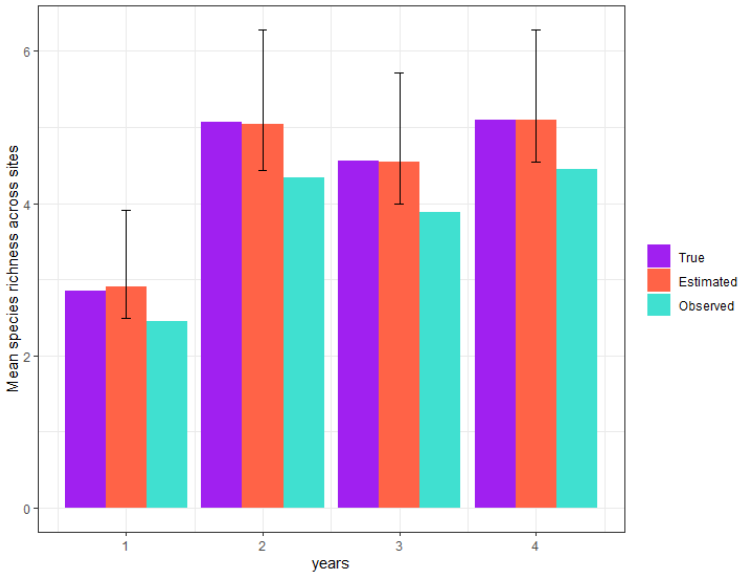
(a) Estimated proportion of occupied sites when data from three different visits is used.



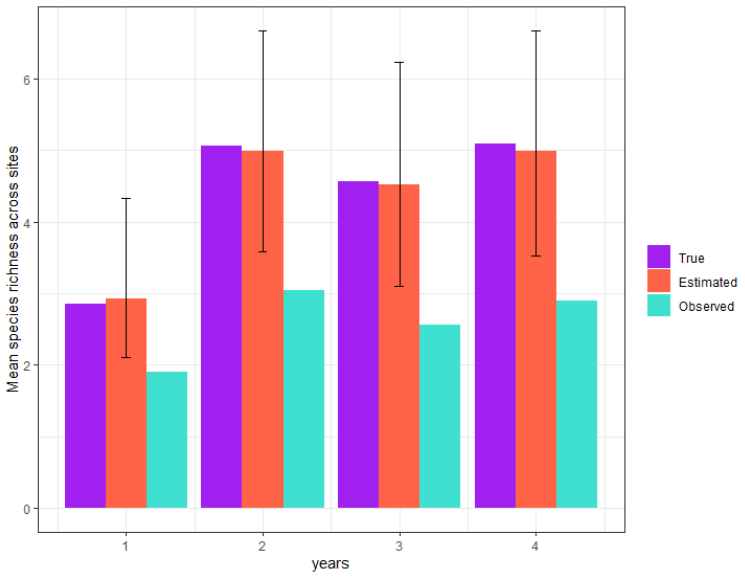
(b) Estimated proportion of occupied sites when data from a single visit is used.

Figure 4.17: Proportion of occupied sites for each species (indicated by different line types) over time. Red solid line indicates the true average across all species for each year. Black solid line is the estimated mean proportion of occupied sites over time with 95% credible intervals indicated by the grey areas.

Average estimated species richness across all sites are close to the true mean richness, with credible intervals being wider for the single visit sampling scheme (Fig. 4.18). However, Figure 4.18 (b) highlights the importance of accounting for imperfect detection, even when the only source of information comes from a single visit survey scheme, since the observed species richness is lower in both scenarios, especially when single visits data are used. This is an important aspect to be considered because the smoothness of detection curve can portray different heterogeneous detection patterns between sites that limit the information available to confirm a particular species presence at a given site.



(a) Mean species richness comparison with the true observed and estimated means under three visits sampling scheme.



(b) Mean species richness comparison with the true observed and estimated means under a single visit sampling scheme.

Figure 4.18: Observed, true and estimated species richness mean across all site for each year with error bars representing 95% credible intervals.

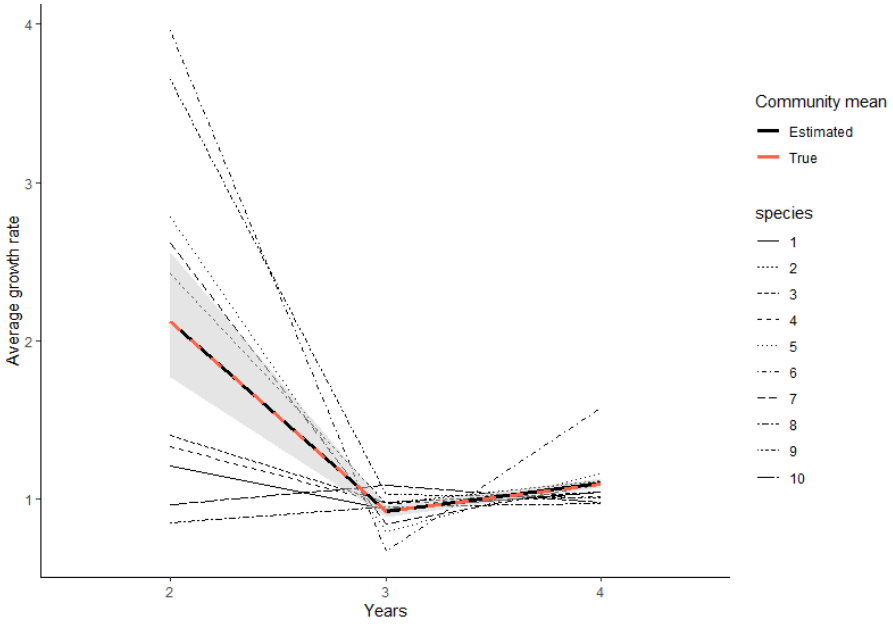
The mean growth rate $\bar{\lambda}_{it}$ indicates the average change in occupancy after the first time period. It summarizes the inter-annual changes in the average occupancy over all sites through the subsequent time periods. Figure 4.19 shows the estimated average growth per year for each of the species, and the community mean (average over all species, i.e $\frac{1}{S}\sum_i^S \bar{\lambda}_{it}$) which lies closely to the true value for both single-visits and multiple-visits scenarios (with the former showing wider credible interval, Fig. 4.19 (b)). In addition to the mean growth rate, the total growth rate λ_{tot}, defined by Banner et al. (2019) as the ratio between the mean occupancy in the last year and first year, is also estimated as a model derived parameter. This parameter estimates the net change in the total number of sites where species occur over the time period of the study. The total growth rate is commonly used to describe the long-term trend in occupancy from the start of a monitoring program (Banner et al., 2019). Table 4.1 shows the estimated total growth rate for each species under repeated sampling. Compared to the estimates from a single visit sampling scheme (Table 4.2), estimates from both scenarios are close to the true values with similar credible intervals for several species (i.e. species 1, 3, 4, 7, 8 and 10).

Table 4.1: True vs. estimated total growth rate from fitting model 4.18 for 10 simulated species when observed detection are simulated for three repeated visits.

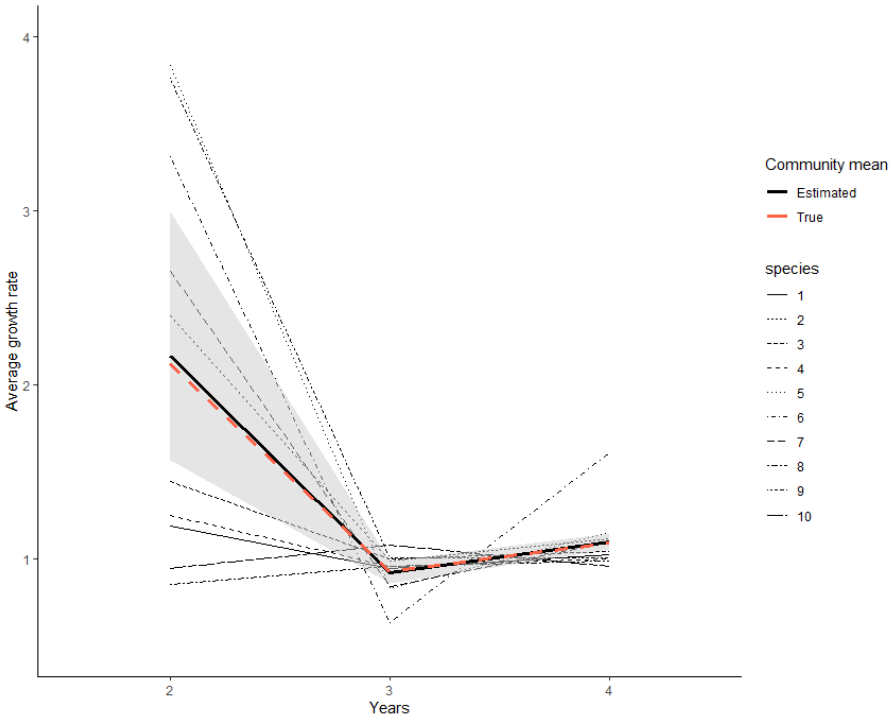
Species	λ_{tot}	$\hat{\lambda_{tot}}$	95% Credible interval
1	1.18	1.18	[1.08 , 1.29]
2	2.57	2.63	[2.38 , 2.91]
3	1.45	1.44	[1.28 , 1.60]
4	1.24	1.31	[1.14 , 1.51]
5	3.06	2.59	[1.83 , 3.62]
6	4.23	4.17	[3.46 , 5.02]
7	2.32	2.42	[1.98 , 2.92]
8	0.85	0.79	[0.69 , 0.89]
9	3.35	3.77	[3.02 , 4.75]
10	1.06	1.03	[0.90 , 1.16]

Table 4.2: True vs. estimated total growth rate from fitting model 4.18 for 10 simulated species when observed detection are simulated for a single visit.

Species	λ_{tot}	$\hat{\lambda_{tot}}$	95% Credible interval
1	1.18	1.15	[0.99 , 1.33]
2	2.57	2.66	[1.54 , 4.94]
3	1.45	1.51	[1.37 , 1.66]
4	1.24	1.20	[1.05 , 1.37]
5	3.06	3.68	[1.64 , 6.09]
6	4.23	3.40	[2.63 , 4.46]
7	2.32	2.44	[2.08 , 2.88]
8	0.85	0.81	[0.62 , 1.06]
9	3.35	3.81	[2.79 , 5.14]
10	1.06	0.97	[0.69 , 1.37]



(a) Estimated vs true average growth rate mean under three visits sampling scheme.



(b) Estimated vs true average growth rate mean when data from a single visit is used.

Figure 4.19: Mean growth rate for each species (indicated by different line types) over time. Red solid line indicates the true average growth rate mean across all species for each year. Black solid line is the estimated average growth rate mean over time with 95% credible intervals indicated by the grey areas.

4.4 Discussion and considerations

The simulation analysis of the flexible dynamic occupancy model developed in this chapter provides some important considerations in terms of the complexity of the relationship between state variables and exploratory variables. The simulation results suggest that the proposed model is able to estimate complex non linear relationships to some extent, depending on the smoothness degree of the curve. Very wiggly curves will result in more constrained priors that produce smoother coefficients. Whilst posterior means lie closely within the true shape of the curve, the uncertainty around the estimates depend on the number of visits from which the observational process is simulated.

Results from fitting these models should be interpreted with caution when presence-only data are used, as the information available might not be enough to produce low-uncertainty estimates for some species exhibiting complex non-linear relationships and low detection probabilities. For example, Banner et al. (2019) recently presented a simulation-based power analysis of the MSOM in which different spatial extensions, sample sizes, sampling intensities and distinct initial occupancy and colonization probabilities were assessed. In this work, Banner et al. (2019) discuss how the sample size required to capture model trends can become extensively large when occupancy probabilities approach zero. They also reported a higher power for larger sample sizes (number of sites) and sampling intensities (percentage of grid cells surveyed). Even though the simulation work in this chapter highlights the influence of the sampling scheme when dealing with flexible components in the occupancy model, another aspect that should also be considered is the initial occupancy, colonization and survival probabilities.

Banner et al. (2019) approached this by specifying low and high initial occupancy and colonization probabilities. They reported that the power was found to be greater when higher baseline initial occupancy and lower initial colonization probabilities were specified ($\text{logit}(\psi) = 1.4$ vs. $\text{logit}(\psi) = 0$; $\gamma = 0.01$ vs $\gamma = 0.2$), emphasizing that very low occupancy probabilities produce sparser observed detections, imposing very large sample sizes to estimate changes in occupancy. Overall, the simulation results in this chapter suggest that model fit improves when data from repeated visits are available. However, the parametrization of the observational process in Peach et al. (2017) is presented as an alternative that uses sampling effort instead of repeated visits when presence-only data is the only source of information available. Although the formulation in Peach et al. (2017) was originally proposed as a method to account for non-linear effects on the detection probability, the uncertainty around the estimates of non-linear terms is still greater than those from repeated visits. Moreover, individual species occupancy and detection patterns within a community can vary widely. Thus, to account for more flexible scenarios in which detection probabilities show more complex non-linear patterns for a group of species, smooth terms were introduced into the observational process.

The simulation analysis of the multiple species flexible dynamic occupancy model proved an overall good fit to individual-species responses, even when presence-only data were used. These results highlight an important feature of the flexible model in capturing the usual logistic parametric shape assumed by standard occupancy models while also providing evidence of more complex relationships across species. Hence, the proposed model should be more robust than standard methods that assume a parametric form

in the relationship. Note posterior predictive checks (see definition in chapter 5) could be used to asses lack of linearity by plotting binned residuals against potential exploratory variables. This can be useful to identify and select important nonlinear effects and avoid fitting an overparametrized model that would otherwise be computationally expensive to fit (see a discussion on computationally efficiency in chapter 6). When summarizing community metrics through different derived parameters, community trends over time were accurately estimated under both presence-only and multiple visits scenarios.

Unlike the models developed in previous chapters, the flexible occupancy model include covariates on the occupancy process. The prior specifications for these covariates is still a subject of research and some considerations have been pointed out in the literature. Conventional vague normal priors (Normal$(0,\sigma)$)) are usually chosen for the intercept and slope parameters (Gelman and Hill, 2006). However, as reported by Northrup and Gerber (2018), normally distributed priors are a sensible choice when non-linear transformations are used (such as the logit function in occupancy models).

In the particular case of the logit function, normal priors are not invariant to this transformation because when large negative or positive value are transformed into the probability scale, the transformed probabilities approach either zero or one. Thus, very vague priors (e.g. Normal$(0,500)$) lead to high probability densities around zero and one (Northrup and Gerber, 2018). Choosing too vague non-informative priors on the logit scale could then have an important effect on the posterior distribution when it is taken to the probability scale. Gelman et al. (2008) proposed using Student-t prior distributions with scale $\sigma^2 = 2.5$ and 1 degree of freedom on the coefficients of standardized covariates. This specification allows to constrain the prior probability mass points to lie within the interval $[-5,5]$ (since a difference of 5 in the logit scale is equivalent to a shift of ≈ 0.5 on the probability scale). Dorazio et al. (2011) discussed how these priors assign high probabilities in the boundaries where occupancy or detection probabilities are close to zero or one. Thus, they suggest using Student-t distributions with $\sigma = 1.566$ and degrees of freedom $\nu = 7.763$ as priors for each logit-scale parameter. This distribution assigns low probabilities to any value outside the $[-5,5]$ interval and approximates to a Uniform(0,1) (Dorazio et al., 2011). Another alternative prior that has been used for Bayesian logistic regression because of the equivalence to an Uniform(0,1) is the logistic distribution centered at 0 with scale parameter 1 (Dorazio, 2016). Northrup and Gerber (2018) reported that both logistic and t distributed priors produced occupancy estimates that were invariant to sample size while conventional normal vague priors had an important effect on the posterior distribution when a larger prior standard deviation (> 100) was specified for small sample sizes.

Occupancy posterior distributions can be strongly influenced by normally distributed priors with large standard deviation or small precision. Hence, Hobbs and Hooten (2015) and Northrup and Gerber (2018) recommend using Normal$(0,2)$ (i.e. centred at zero with variance $\sigma^2 = 2$) to reduce the priors' influence on the logit-scale parameters.

The effect of a prior could be more pronounced when little data are available which could be the case when the occurrence of rare and elusive species is being analyzed (Northrup and Gerber, 2018). Thus, species-specific parameters can be treated as unknown parameters drawn from a common hyperprior

distribution that allows for information exchange (see chapter 1). This approach has been used to allow for the overall mean and variation in covariate effects to adapt to the data (Outhwaite et al., 2018). For instance, Yamaura et al. (2011) used conventional vague Normal (0,1000) and Gamma (0.01,0.01) hyperpriors for the mean and precision parameters from which species specific effect were drawn.

For the simulation study results presented in this chapter Student-t distributed hyperpriors were used (Figs. 4.20, 4.21, Gelman $\hat{R} < 1.1$). These results were consistent when both logistic and Student-t distributed hyperpriors were specified. Non-informative Normal (0,1000) hyperpriors were also tested, however mixing under t-Student and logistic weakly informative priors was slightly better. As for the precision parameters, Outhwaite et al. (2018) and Gelman et al. (2006) recommendations of using Half-Cauchy priors (equivalent to a Student-t distribution with 1 degree of freedom) for the variance parameters were followed. Inverse gamma priors on the variance were also tested however posterior samples showed some degree of correlation. This has also been reported by Outhwaite et al. (2018) when inverse gamma priors were used as hyperprior for the precision parameters. Furthermore, authors compared half Cauchy priors with uniform (0,5) and suggested using Half-Cauchy priors to avoid boundary effects that uniform priors have.

Although the sensitivity analysis showed consistent results for the simulation work on this chapter when different priors were used, the noise in real data and the specification of multiple covariates with varying sample sizes can contribute to posterior distributions to be influenced by the choice of priors (Northrup and Gerber, 2018). Thus, a thoughtful examination of prior choices will be covered in the next chapter where the methods developed here will be applied to the Odonata case study.

Beside the methodological assumptions under which these methods have been developed, it is also important to be aware that the population dynamics that can be described by these models have to be interpreted within a spatio-temporal ecological context. Metrics that describe how occupancy states change over time can be attributed not only to colonization or local extinction dynamics but also could be the result of organisms movement to suitable habitats outside the study region (Noon et al., 2012).

Real ecological data can be complex, especially when dealing with a group of multiple species with potentially very different responses towards environmental drivers. Therefore, the methods developed in this chapter and the simulation study provide a useful framework to explore a complex real-life situation to test different ecological hypotheses on how environmental drivers affect population dynamics across different biological communities. It can also help ecologists and conservationists to set the basis for monitoring programs and refine sampling designs that aim to quantify how the dynamic structure of a biological community evolves over time given the environmental pressures they are exposed to. Thus, in the next chapter, the flexible dynamic occupancy model will be further developed and applied to the Odonata case study to describe how the occurrences of dragonflies and damselflies communities change through time and to identify the statistical challenges that might open new lines of research.

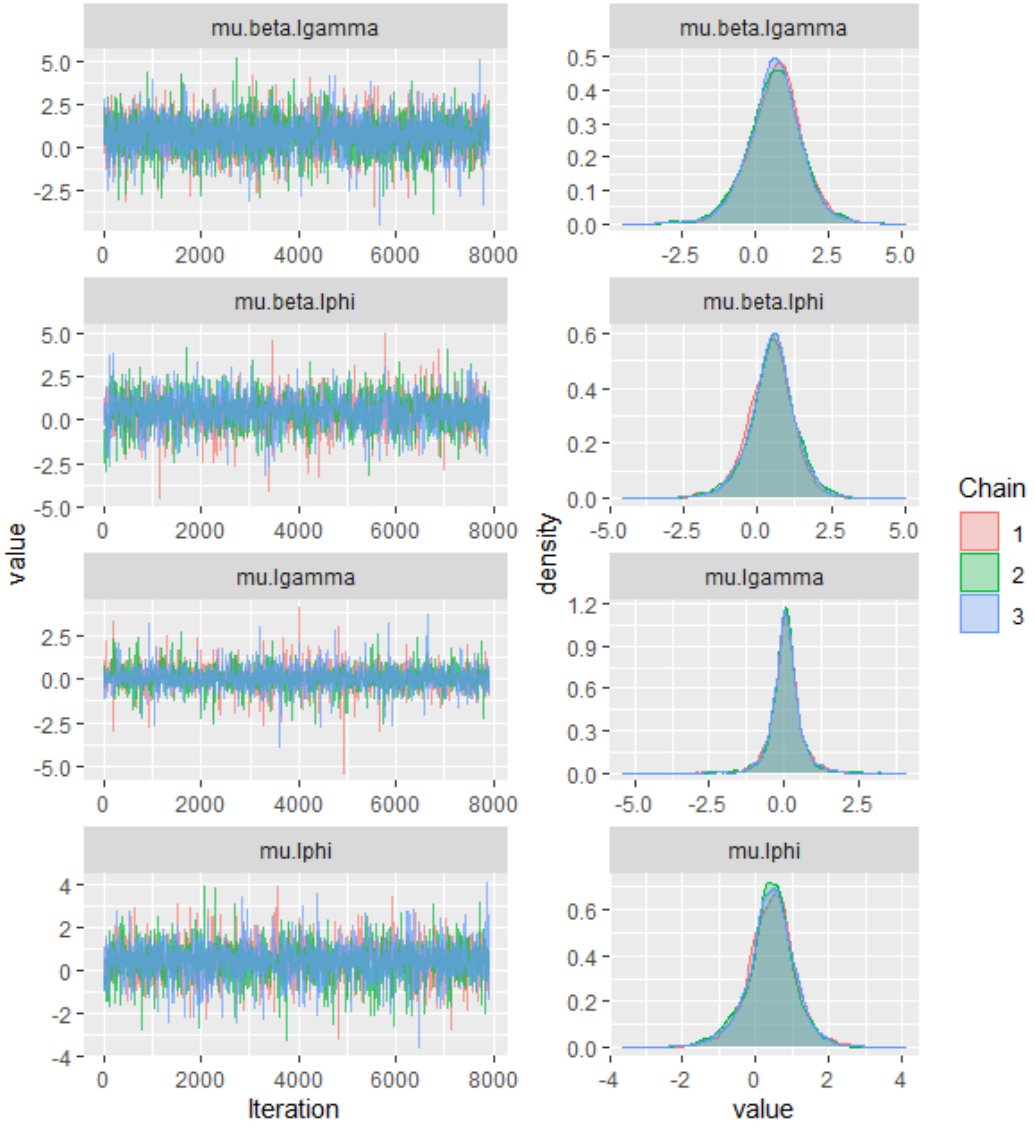


Figure 4.20: Trace plots and posterior densities for generalized penalized splines MSOM hyperparameters for multiple species under a repeated sampling scheme.

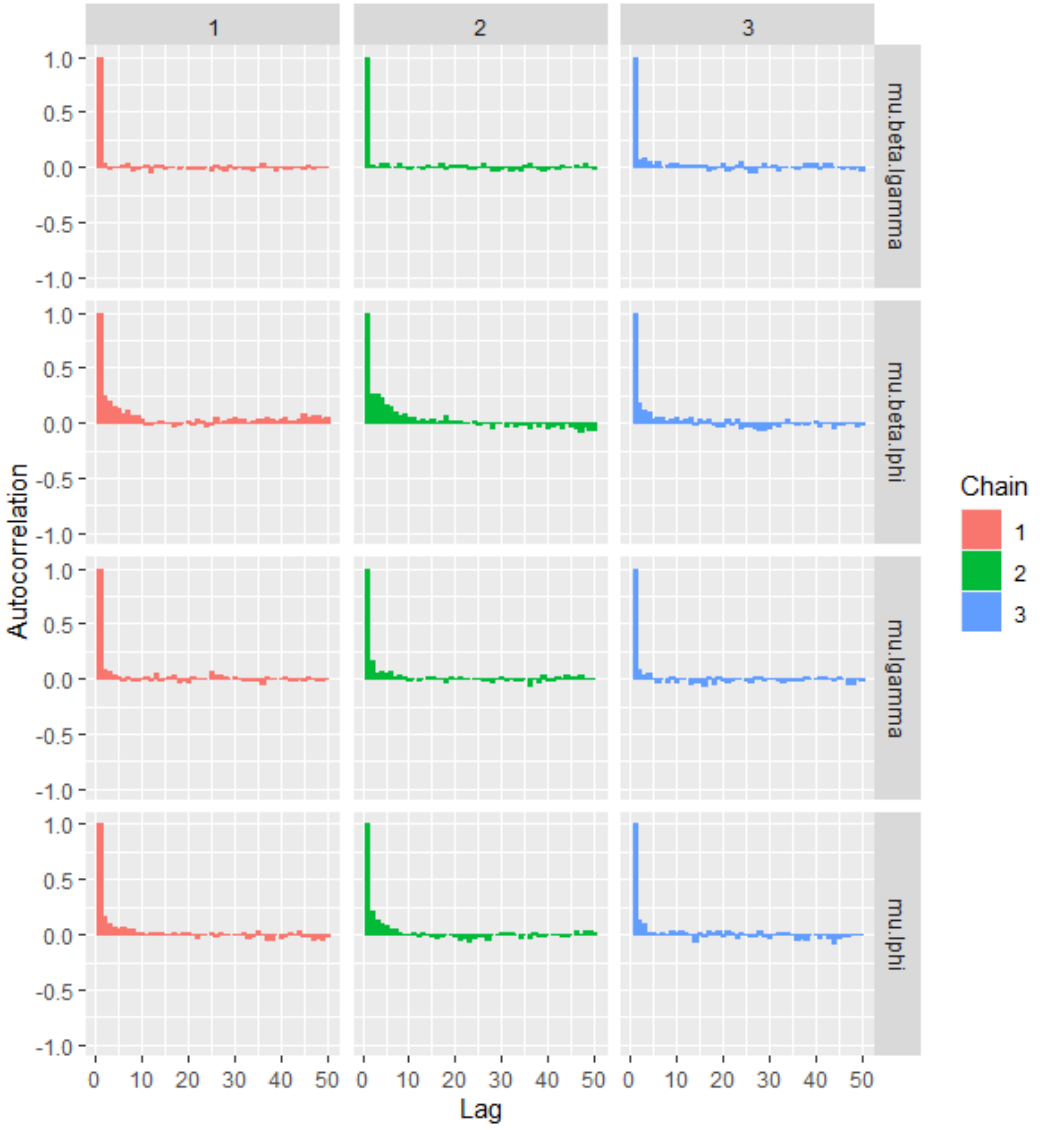


Figure 4.21: Autocorrelation plots for generalized penalized splines MSOM hyperparameters for multiple species under a repeated sampling scheme.

Chapter 5

Application of flexible dynamic Odonata occupancy model

The Odonata case study explored in chapters (1-3) has been analyzed under the closure assumption in which the occupancy probabilities are constant across time. However, the observed invasive species' occurrence patterns suggest an increasing trend over the years across sites (Fig. 5.1). As for the non-invasive species, the models developed so far do not account for local extinctions or colonization dynamics. Thus, the penalized splines dynamic occupancy model developed in chapter 4 can be used to analyze the Odonata population dynamics and community trends over time while accounting for non-linear effects of site-level covariates on detection, colonization and survival probabilities. This case study has some important challenges that need to be considered in order to develop a flexible occupancy model under a temporal framework.

Figure 5.1: *C. viridis* and *E. viridulum* observed occurrences (red dots) during 2000, 2007 and 2012.

First, the primary sampling periods in which the occupancy state is assumed to be constant have to be determined. Odonata species occurrences were recorded mainly during late Spring and early Autumn (Fig. 5.2). This could be related to life-expectancy of adults which has been reported to be just a couple of weeks for some species and potentially longer (6-8 weeks) for others (Smallshire and Swash, 2018). Therefore, it is reasonable to assume that the occupancy state does not change between the sampling occasions within a year, and that the primary sampling periods in which occupancy state changes can be defined by the different years as the occurrences records arise from a new generation of individuals each year.

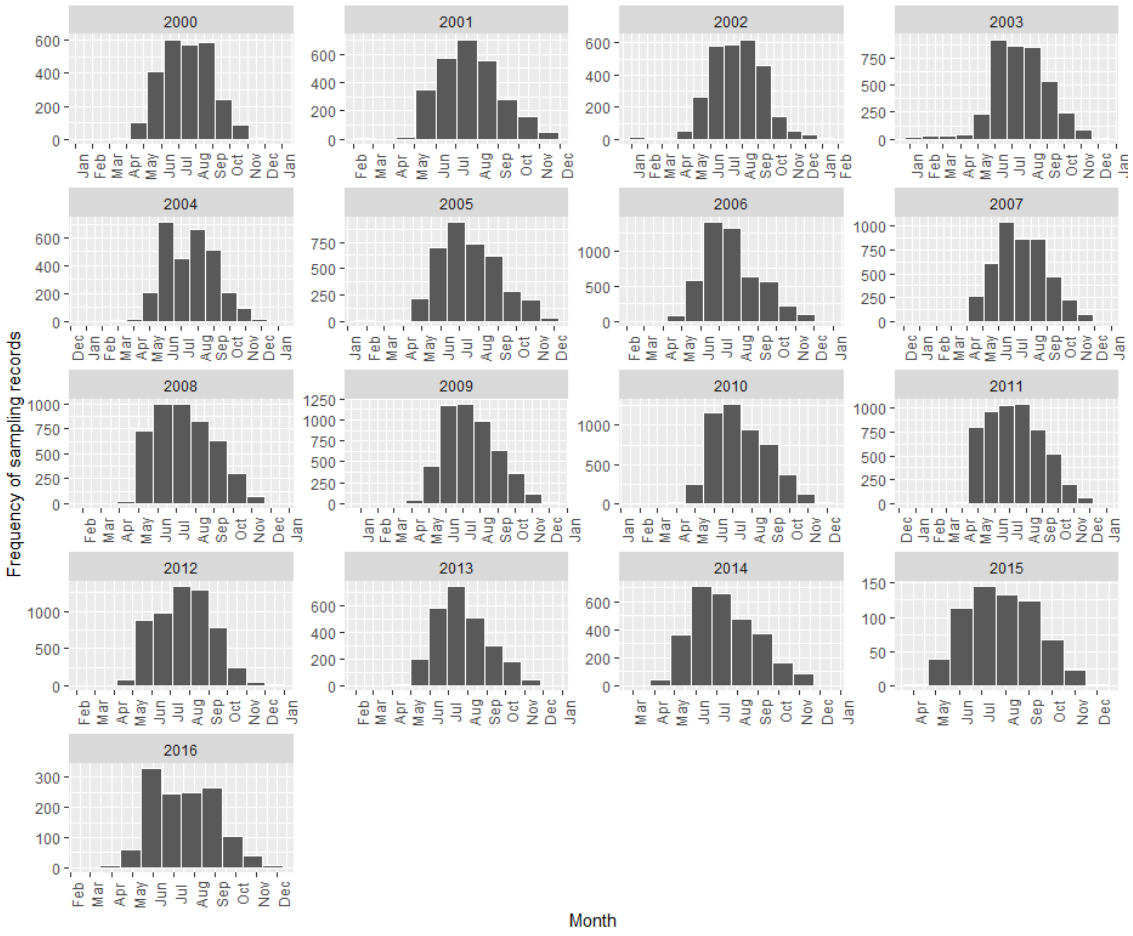


Figure 5.2: Histogram for the dates in which Odonata occurrences where recorded from 2000 to 2016.

Another key point to be considered is the fact that Occupancy models are fitted based on presence/absence data collected from individual species records during repeated visits. The Odonata case study consists mainly of presence-only records where there is no information regarding species absences. Thus, individual species absences were inferred based on the information of sightings of other dragonfly and damselflies species (as described in chapter 1 by following van Strien et al. (2010), van Strien et al. (2013) and Termaat et al. (2019)), i.e. Species *i* sighting confirmed its presence at a given site, and was deemed as undetected at those sites where any species other than *i* were recorded within each year.
Under a temporal framework, the information about Odonata presence/inferred-absences is incomplete

because some species sightings have only occurred for some sites in very few years, i.e. there are sites in certain years where none of the species were recorded. Thus, there is no information about any other species occurrences across the other years.

The time period for which Odonata records are available is from 2000 - 2016. However, the first and last couple of years (i.e. 2000-2002 and 2012-2016) have a considerable large number of sites where neither dragonflies nor damselflies were recorded ($> 90\%$), and thus, were excluded from any posterior analysis. Nevertheless, the Odonata occurrences records from 2000-2012 still have a large number of locations where no Odonata species have been recorded. Thus, only sites which had less than 50% of non-detection records of any species were kept in order to reduce the influence that priors have on the posterior distribution of occupancy models estimates.

The reduced set of occurrences consist of Odonata records across approximately 700 sites over a 10 years time period. The selected sites still have a reasonable coverage of the UK as shown on Figure 5.3. To account for sampling effort, the number of visits each site had was calculated for each year based on the different dates when a species was recorded in a given site according to the methods described in chapter 1 section 1.4 (different species sightings on the same site during the exact same date were taken as a single visit).

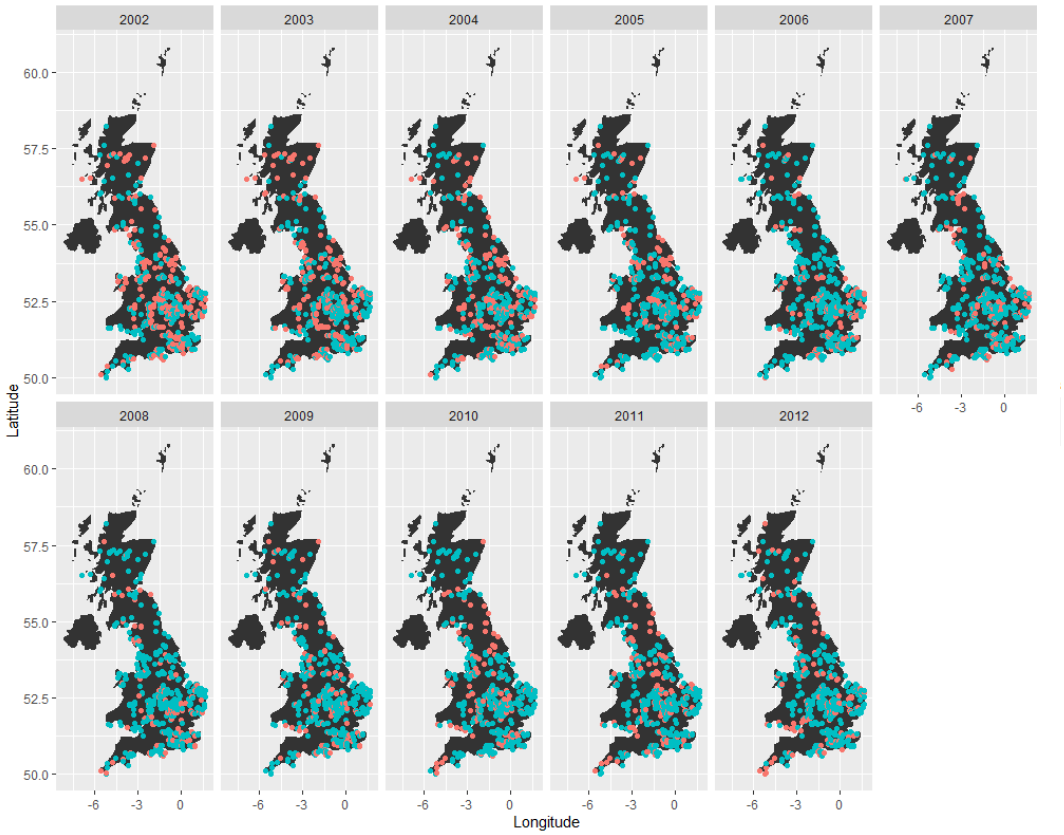


Figure 5.3: Odonata occupancy maps from 2002 to 2012. Blue points indicates sites where at least one dragonfly or damselfly species has been recorded during a specific year. Red points indicate sites where neither dragonflies or damselflies were recorded

5.1 Methods: Odonata flexible dynamic occupancy model

A dynamic flexible occupancy model based on the work developed in chapter 4 is proposed for analyzing Odonata population dynamics. First, the state process describing Odonata occupancy changes over time can be formulated as follows.

Let z_{ij1} denote the occupancy state of site j by species i during year 1 such that $z_{ij1} \sim \text{Bernoulli}(\psi_{ij1})$, where ψ_{ij1} is the initial occupancy probability defined on the logit scale as:

$$\text{logit}(\psi_{j1}) = \beta_{0i} + \beta_{1i}x_{1j} + \ldots + \beta_{mi}x_{mj}. \tag{5.1}$$

In here, $\boldsymbol{\beta}_i$ are the species-specific initial occupancy intercept and slopes for m different site-level metrics $x_{1j}, \ldots, x_{mj}$ for $j = 1, \ldots, J$ sites. The occupancy state across years (z_{ijt} for $t > 1$) can be defined recursively as follows:

$$\begin{aligned} z_{ijt} &\sim \text{Bernoulli}(\psi_{ijt}) \quad \text{for } t = 2, \ldots, T \\ \psi_{ijt} &= z_{ijt-1}\phi_{ijt-1} + (1 - z_{ijt-1})\gamma_{ijt-1} \\ \text{logit}(\phi_{ijt}) &= \phi_{0it} + f_i(x_{1j}) + \ldots + f_i(x_{mj}) \\ \text{logit}(\gamma_{ij}) &= \gamma_{0it} + g_i(x_{1j}) + \ldots + g_i(x_{mj}), \end{aligned} \tag{5.2}$$

where γ_{0it} and ϕ_{0it} are the species-specific baseline colonization and survival logit-scale probabilities during year t, $g_i(\mathbf{x}_j)$ and $f_i(\mathbf{x}_j)$ are the species-specific smooth functions describing the non-linear effect that site-level covariates have on colonization and survival.

Weakly informative priors were chosen for the state process parameters based on the discussion in chapter 4. Species-specific effects were drawn from zero-centered t-distributed hyperpriors with a scale parameter of 1.566 and degrees of freedom 7.763 for the means, and half - uniform (0, 5) for the variances. Species-specific intercepts hyperpriors on the other hand, were chosen following Outhwaite et al. (2018) by specifying a random walk model (Eqn. 5.3) that enables systematic changes from year to year. This allows for colonization and survival probabilities to be similar from one year to the next.

$$\gamma_{0it} \sim \begin{cases} \text{Normal}(\mu_c, 0.01) \text{ for } t = 1 \\ \text{Normal}(\gamma_{0it-1}, \sigma_c^2) \text{ for } t > 1 \end{cases} \qquad \phi_{0it} \sim \begin{cases} \text{Normal}(\mu_s, 0.01) \text{ for } t = 1 \\ \text{Normal}(\phi_{0it-1}, \sigma_s^2) \text{ for } t > 1, \end{cases} \tag{5.3}$$

where $[\mu_c, \mu_s] \sim \text{t}(\sigma = 1.566, \nu = 7.763)$ and $[\sigma_c, \sigma_s] \sim \text{Uniform}(0, 5)$. Smooth coefficients were drawn from independent Normal priors as suggested by Crainiceanu et al. (2005) with variance parameters drawn from half-uniform hyperpriors. A total of $K = 5$ fixed knots were specified to construct the thin-plate basis system based on the work proposed by Crainiceanu et al. (2005) and after thorough examination of the convergence diagnostics where high dimensionality induced high levels of autocorrelation and bad mixing .

Observational model 1: Aggregated number of species records

Species detection is proposed to be modeled by aggregating the total number of species presence records given the total number of visits to each site per year. Detection probabilities were defined as a smooth function of the human connectivity index produced by Chapman et al. (2020) to account for the effect that human activities have on the observed species occurrences (Eqn. 5.4). The chosen index accounts for all types of visits, such as recreation, fishing and, water sports which have proven to be an influential factor for the non-native species dispersion of different biological groups (Chapman et al., 2020). Thus, the observational process is formulated as follows:

$$\begin{aligned} \sum_{N_{jt}} y_{ijt} | z_{ijt}, N_{jt} &\sim \text{Binomial}(p_{ijt}) \\ \text{logit}(p_{ijt}) &= \alpha_{0it} + h_i(\text{human activities}_j), \end{aligned} \tag{5.4}$$

where N_{jt} are the number of visits on site j during year t, $h_i(\text{human activities}_j) = \alpha_{1i} + \sum_k^K e_{ik} Z_{jk}$ is the smooth function for each species (e_{ik} and Z_{jk} are the smooth coefficients and basis system respectively). The annual species-specific intercepts α_{0it} are introduced here as a random effect drawn from the same community-level hyperprior defined as a linear function of the species-specific traits with hyperparameters $\boldsymbol{\delta}$ drawn from $\text{t}(\sigma = 1.566, \nu = 7.763)$ (Eqn. 5.3). The species-specific baseline detection probabilities for $t > 1$ are drawn by using the aforementioned random walk model (Eqn. 5.3), i.e.,

$$\begin{aligned} \alpha_{0it} &\sim \begin{cases} \text{Normal}(\mu_{p_i}, 0.01) & \text{for } t = 1 \\ \text{Normal}(\alpha_{0it-1}, \sigma^2_{det}) & t > 1 \end{cases} \\ \mu_{p_i} &= \delta_0 + \delta_1 \times \log(\text{flight period})_i + \delta_2 \times \log(\text{body size})_i + \delta_3 \times \log(\text{num. of habitats})_i. \end{aligned} \tag{5.5}$$

In here, σ^2_{det} is drawn from half - uniform (0,5) prior and is the inter-annual variance controlling the degree of similarity between years which enable for sequential changes in detection probabilities.

Unfortunately, site-level covariates that explain variation in detection probabilities due to uneven sampling effort are not always available. Thus, the date each site was visited and the species list length have been proposed as predictors to account for uneven sampling effort (van Strien et al., 2010). This information can be easily obtained from citizen science data sets, thus it is of interest to compare the output from the model using the aggregated number of species records and a model that uses only the information available from sampling records.

Observational model 2: Presence-absence binary records

The observational model can be formulated as a binary outcome of whether a species is detected or not in a specific visit with a detection probability defined as a function of covariates. Termaat et al. (2019) proposed a dynamic occupancy model for multiple Dragonfly species in Europe that defines detection probabilities as function of the Julian date when a species is observed. Including the Julian date as a covariate accounts for changes in population size during the species' flight period (van Strien et al., 2013).

This formulation, originally proposed by van Strien et al. (2010), also incorporates the species list at each site for each year as a proxy for the sampling effort. Thus, the observational model to estimate the yearly detection can be written as:

$$\text{logit}(p_{ijlt}) = \alpha_{0it} + h_i(\text{date}_{jlt}) + \alpha_1 \times \log(\text{Species List})_{jlt} + \alpha_2 \times \log(\text{Num. Visits})_{jlt}. \tag{5.6}$$

Termaat et al. (2019) specified a common intercept for all species (i.e. α_{0t}) and introduced a quadratic term to capture the non-monotic trend in species detection. Thus, the definition for α_{0it} in equation 5.6 enables estimation of systematic changes in species detection through time while incorporating the species-specific trait information. The flexible occupancy model developed in the previous chapter can incorporate a smooth function $h_i(\text{date}_{jlt})$ to model the non-linear effects of the date of visit l at site j in year t.

The detection model proposed by Termaat et al. (2019) assesses the sampling effort by making a distinction between species lists with a single species record during a visit and species lists with multiple sightings made from the same observer. However, for the Odonata case study, this distinction cannot be made since there is no information about whether multiple species are recorded by the same observer or not (thus it is assumed that occurrences from different species at the same site on the same date were recorded on a single visit since the identity of the observer(s) is unknown). Thus, to account for the sampling effort, the effects that the species list length (num. of species) (α_1) and number of visits per site (α_2) have on the detection probability are introduced.

Occupancy model residuals and spatial autocorrelation

To assess spatial autocorrelation, the approach of Wright et al. (2019) was taken by defining separate residuals for the state and observational model respectively and then calculating Moran's I for each species. The state process residual for site j and species i during year t is defined as:

$$o_{ijt}^{[s]} = z_{ijt}^{[s]} - \psi_{ijt}^{[s]}, \tag{5.7}$$

where $\mathbf{z}^{[s]}$ is a draw from the posterior distribution of the occupancy state. Then, for each draw s, Wright et al. (2019) define the detection residual as:

$$\left[d_{ijlt}^{[s]} | z_{ijt}^{[s]} = 1\right] = y_{ijlt}^{[s]} - p_{ijlt}^{[s]}. \tag{5.8}$$

These residuals are conditioned on the posterior occupancy state for every draw s for each site j during visit l and can be derived directly from the observational model 5.6.

Residuals for the observational model 5.4 can be calculated as

$$\left[d_{ijt}^{[s]}|N_{jt}, z_{ijt}^{[s]} = 1\right] = y_{ijt}^{[s]} - \mathbb{E}(y_{ijt}^{[s]}), \tag{5.9}$$

where $y_{ijt}^{[s]}$ are the total number of detections for species i at each draw. By conditioning on the occupancy state, sites with no detection (i.e. whenever $y_{ijt}^{[s]} = 0$) contribute to the residuals if the site is occupied.

The efficiency of using these residuals to identify the underlying spatial structure which has not been accounted for by a fitted occupancy model on both state and observational process was verified using the simulation approach of Wright et al. (2019). This approach uses Moran's I to produce correlograms across different distance classes to distinguish which model component contains remaining underlying spatial structure. Moran's I statistic is calculated for every posterior draw as follows:

$$I^{[s]} = \frac{J}{\sum_j \sum_{j'} w_{jj'}} \frac{\sum_j \sum_{j'} w_{jj'} (R_j^{[s]} - \bar{R}^{[s]})(R_{j'}^{[s]} - \bar{R}^{[s]})}{\sum_j (R_j^{[s]} - \bar{R}^{[s]})^2} \tag{5.10}$$

where J is the total number of sites, $w_{jj'}$ is an indicator of whether sites j and j' are neighbours (residuals are considered neighbours if the distance between them is within a given distance class), R is the site j-th residual ($R = o_{ijt}$ for the state model and $R = d_{ijt}$ for the detection model) and $\bar{R}$ is the residual mean at each posterior sample. Then, the correlogram is created by defining a set of increasing distance classes for which Moran's I statistic is calculated based on the set of neighbours defined by each class.

5.2 Results: Odonata population dynamics and detectability changes over time

Model 5.2 was fitted in `nimble` compiler using temperature as the only smooth term in the state process. While the model developed in chapter 4 can accommodate several smooth terms for a distinct set of covariates, the small sample size in the Odonata case study (number of sites and visits) constrained the number of terms that could be included in the model (a modelling selection and dimensionality reduction approach is discussed in chapter 6). Thus, site-level temperature (° C) measured at a 2.5 Km Buffer was chosen as the only predictor for Odonata species distribution after discussing with experts from the British Dragonfly Society. Temperature has proven to be a key environmental condition that drives Odonata occupancy patterns (Collins and McIntyre, 2015), and thus, it is of interest to investigate whether Odonata population dynamics are related with temperature in a non-monotic fashion. A total of 200,000 iterations with a burnin period of 50,000 and a thinning of 25 were specified (similar settings can be found in Broms et al. (2016)).

The estimated species-specific colonization probabilities showed heterogeneous non-linear responses that vary widely within the community. For instance, Figure 5.4 illustrates some species for which colonization probability is higher at a mid-range temperature (e.g. *C. viridis*, *E. viridilum* and *L. sponsa*). Note that *C. viridis* and *E. viridulum*, the two invasive species identified in chapter 1, have a quadratic relationship with the temperature with the latter, *C. viridis*, showing a clear gradual increment of the colonization probability over the years. There are also species with low colonization probabilities at low temperatures that gradually increases with temperatures above 10° C (e.g. *S. striolatum*). Moreover, species like *A. juncea* with very sparse occurrences at sites with low temperatures yields colonization probabilities estimates with a high degree of uncertainty. A larger sample of the species-specific estimated relationships between temperature and colonization across all years are presented in the appendix D.4.1.

Survival probabilities have a less complex relationship with temperature compared to colonization probabilities for most of the Odonata species (Figure 5.5). For instance, *C. viridis* and *E. viridulum* survival probabilities are close to 1 and almost constant across the temperatures range, suggesting a strong permanence for these species once a site is occupied. Other species like *S. striolatum* show a typical logistic increasing function (similar patterns can be observed for *C. puella* and *I. elegans* survival probabilities in the appendix D.4.1). Furthermore, negative non-linear relationships between survival probabilities and temperature have been estimated for some species such as *A. juncea* (see also *E. cyathigerum* and *L. sponsa* in the appendix D.4.1). While convergence was achieved for most of the species, the high number of unknown parameters made MCMC convergence computationally expensive, and poor mixing was found for some rare species at sites with no visits in certain years (e.g. the third species on Figure 27 in appendix D.4.1). Similar convergence issues have been reported by Broms et al. (2016) when too many covariates are included or when the number of rarely detected species in the model is high, resulting in extremely long MCMC runs for which the algorithm convergence usually fails (an unfeasible large number of iterations would be required to achieve convergence).

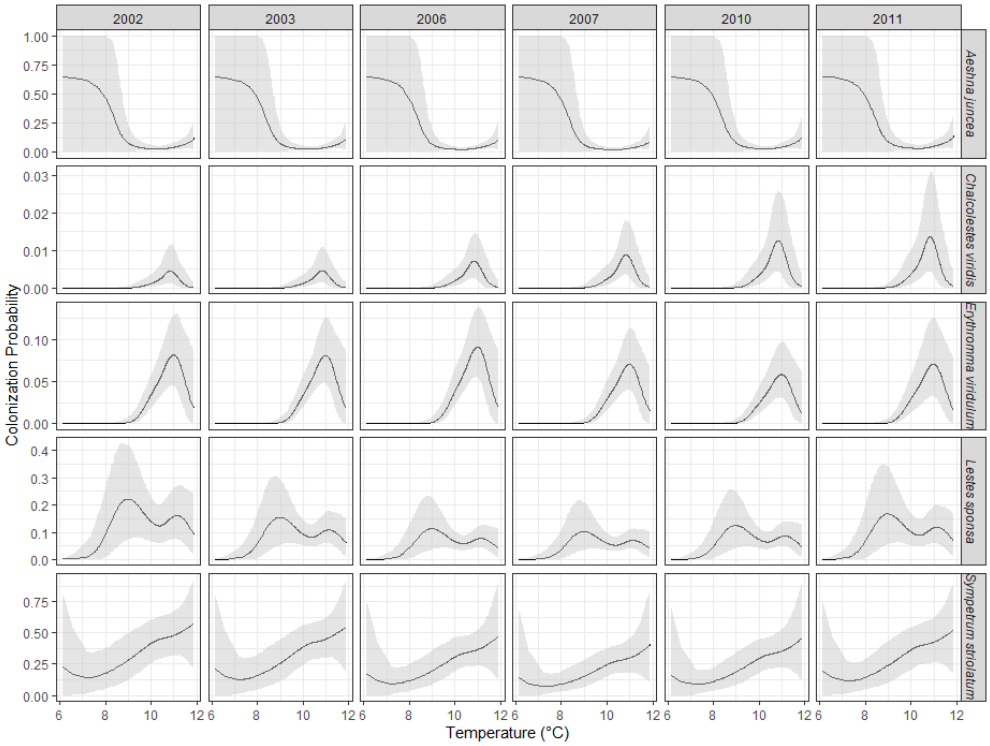


Figure 5.4: Relationship between colonization probabilities and temperature (° C) for Odonata species. Shaded gray area represents 95% credible intervals.

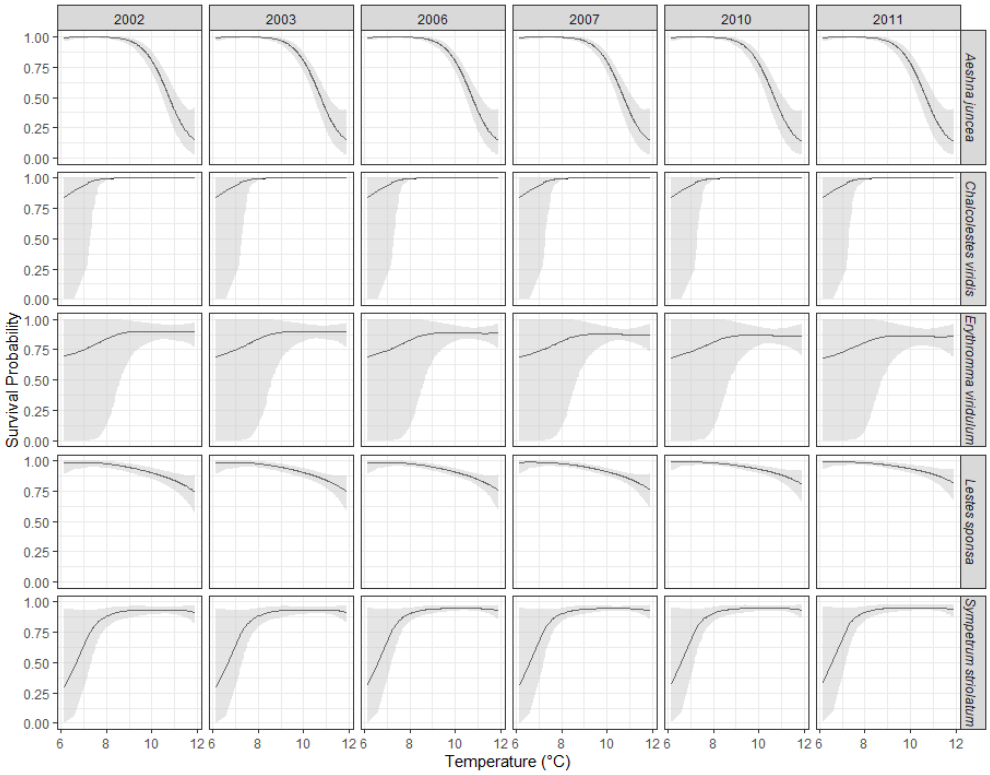


Figure 5.5: Relationship between survival probabilities and temperature (° C) for Odonata species. Shaded gray area represents 95% credible intervals.

An important number of species show a low detection probabilities that relates in a non-linear way with human activities (Fig. 5.6). Note that most of the species detection probabilities are estimated to be less than 0.5, with some very rare species having detection probabilities less than 0.2 (e.g. *L. sponsa* and *A. imperator*; see appendix D.4.1 for estimated detection probabilities for a larger number of species).
The reduced number of sites along with the sparse species' occurrences with low detection probabilities limit the information that can be obtained for some species. However, the primary goal of a multispecies occupancy model is not to track individual species responses, but to make inference about different species assemblages to describe the overall community occupancy patterns across time and space.

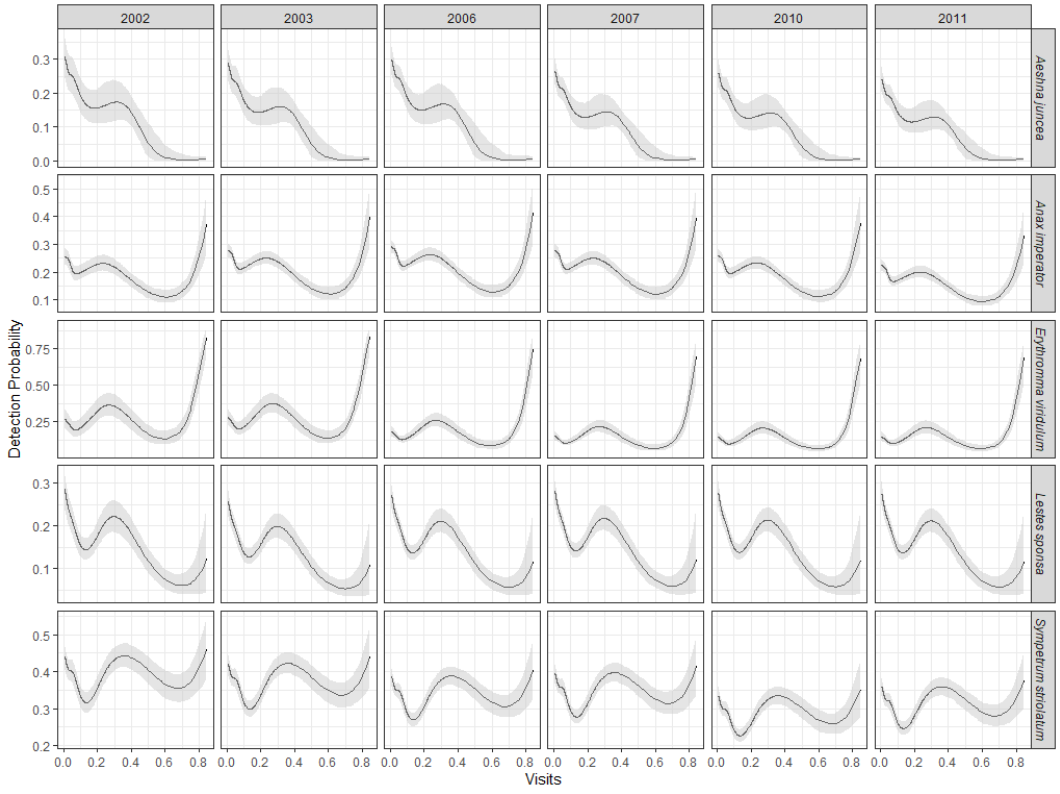


Figure 5.6: Relationship between detection probabilities and human activities index for Odonata species. Shaded gray area represents 95% credible intervals.

Estimated species richness provides evidence that species composition has remained relatively constant from 2002 to 2012 and also confirms the well-known fact that Odonata species richness is higher at lower latitudes (Fig. 5.7). The mean inter-annual changes in species occupancy, shown in Figure 5.8, suggests that the Odonata community has not experienced any major changes from year to year. However, the total growth rate representing the net change in the total number of sites where each individual species occur vary widely across species. More than half of the species did not show any evidence of a significant change in their total growth rate (24 out of the 41 species), which was determined if their credible interval contained zero (Table 5.1 shows a sample of 18 individual Odonata species' total growth rate). For those species that showed a significant change in their total growth rate, there were some species with a decreasing growth rate i.e. $\lambda_{tot} < 1$ (e.g. *A. juncea* and *S. sanguineum* in Table 5.1), and some other species with a very high value for this parameter (such as the invasive *C. viridis* and *E. viridulum*). Thus, these results can be used to identify the potential threat of invasive species to the whole community (A complete list of the 41 species' total growth rate can be found in the appendix D.4.1).

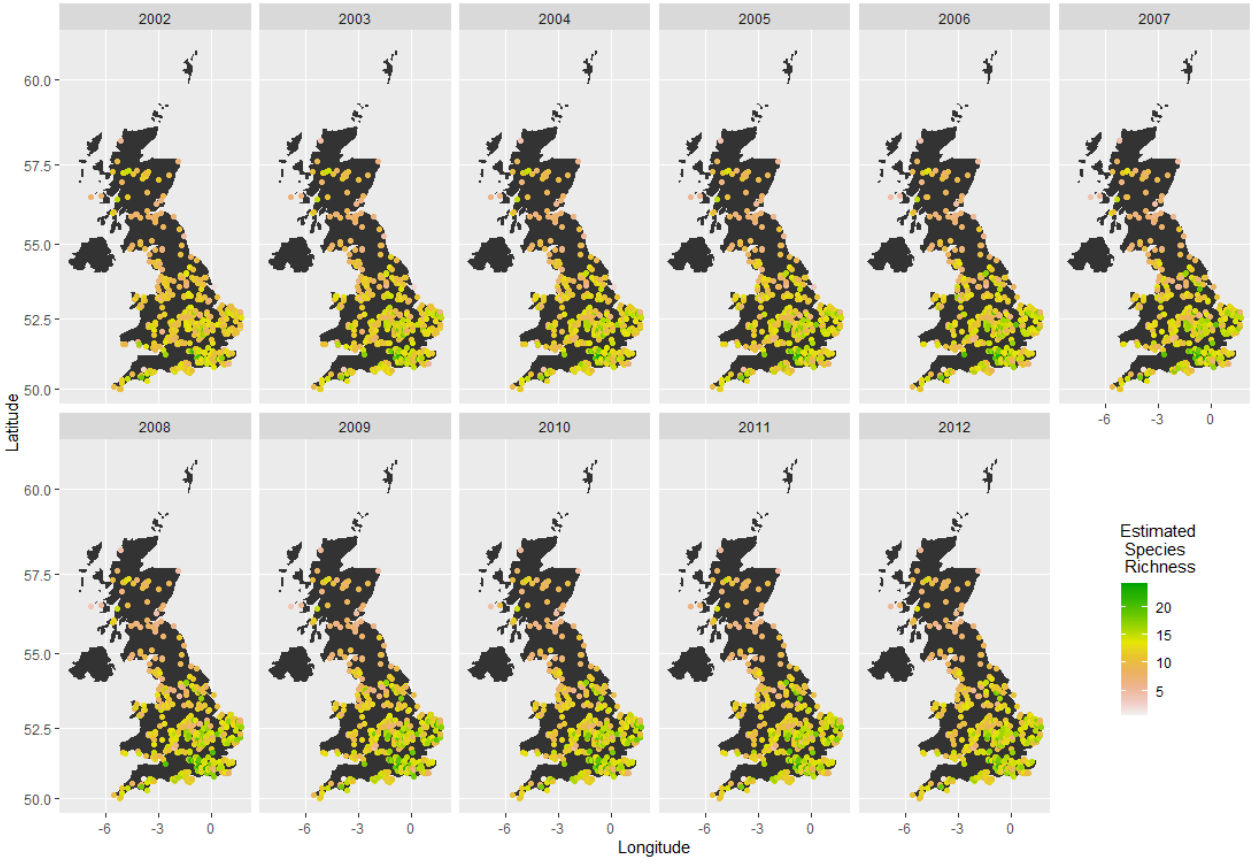


Figure 5.7: Odonata estimated species richness maps from 2002 to 2012.

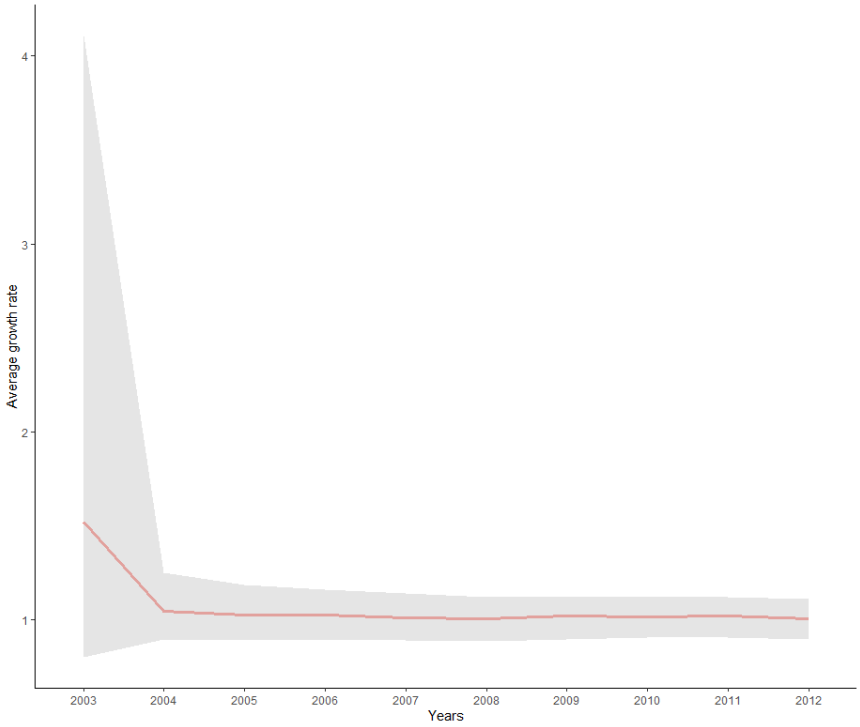


Figure 5.8: Odonata community UK's average growth rate estimated from 2002 to 2012. Shaded gray area represents 95% credible intervals.

Table 5.1: Estimated total growth rate for 18 Dragonflies and Damselflies species. Red rows correspond to species with a decreasing growth rate, blue rows are species showing increasing rates and gray rows are species with no evidence of a change in their total growth rate.

Species	$\hat{\lambda_{tot}}$	95% Credible interval
Aeshna cyanea	0.97	[0.83 , 1.14]
Aeshna grandis	1.26	[1.09 , 1.45]
Aeshna juncea	0.74	[0.58 , 0.93]
Aeshna mixta	1.16	[1.01 , 1.32]
Anax imperator	0.97	[0.84 , 1.13]
Brachytron pratense	1.60	[1.23 , 2.10]
Calopteryx virgo	1.47	[1.02 , 2.15]
Ceriagrion tenellum	1.02	[0.60 , 1.69]
Chalcolestes viridis	85.90	[3.45 , 503.06]
Coenagrion hastulatum	0.90	[0.39 , 2.11]
Coenagrion mercuriale	0.60	[0.29 , 1.12]
Enallagma cyathigerum	1.06	[0.98 , 1.16]
Erythromma najas	1.44	[1.19 , 1.74]
Erythromma viridulum	4.00	[2.22 , 7.10]
Ischnura pumilio	0.81	[0.27 , 1.96]
Libellula quadrimaculata	1.29	[1.11 , 1.51]
Sympetrum sanguineum	0.78	[0.65 , 0.93]
Sympetrum striolatum	1.03	[0.94 , 1.14]

Another way to visualize the community's temporal occupancy pattern is presented in Figure 5.9 with the yearly proportion of occupied sites in the study. Similarly to the previous results, this confirms that the proportion of occupied sites by the whole Odonata community has presented very little changes over the years. However, the individual species heterogeneity in the proportion of occupied sites shown in Figure 5.10 can be used to identify species populations occupancy status. For instance, *A. juncea* and *S. sanguineum* populations which were described previously by their decreasing total growth rate, showed a gradual decline in the proportion of occupied sites. Other species such as *C. virgo* and *E. najas* show an increasing trend on the first years (2002-2005) that reaches a stability point afterwards. Similarly, invasive species *C. viridis* and *E. viridulum* proportion of occupied sites have raised from < 1 and 5% to 5% and 25% respectively, with *C. viridis* showing a much steeper increment (see appendix for a large sample of species-specific estimated proportion of occupied sites).

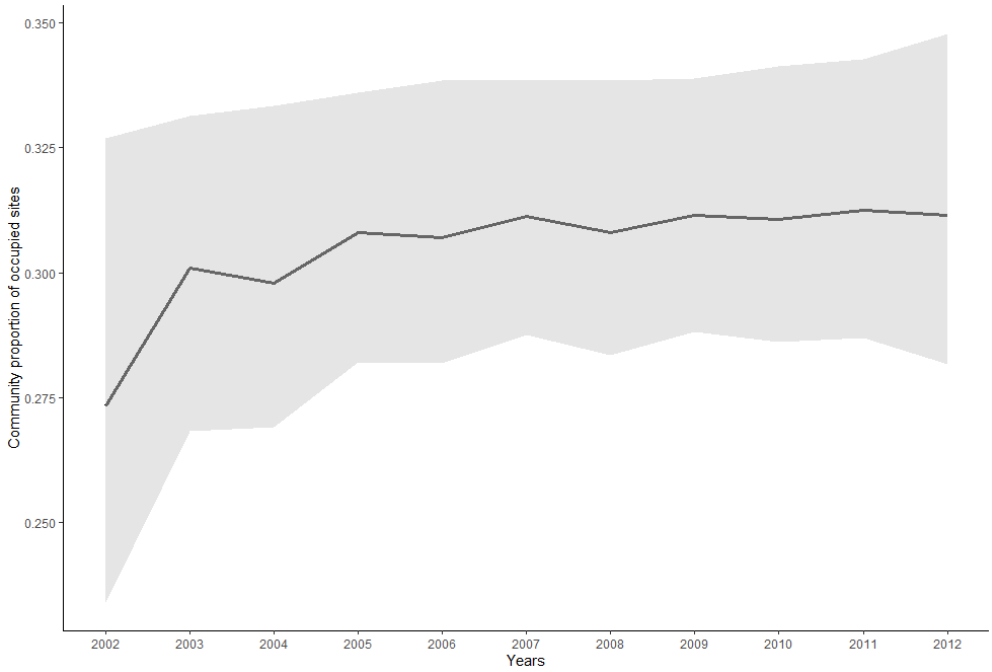


Figure 5.9: Odonata community mean proportion of occupied sites from 2002 to 2012. Shaded gray area represents 95% credible intervals.

A sensitivity analysis was performed by comparing the model results under different priors (see discussion section of chapter 4). Across all priors, model convergence was better when either logistic(0,1) or zero centered t-student ($\sigma = 1.566$; $\nu = 7.7663$) distributed priors where specified for the community mean parameters. Similarly, Uniform (0,5) and Half-Cauchy distributed priors for the community variance parameters led to better results than inverse-gamma conjugate priors. Figure 5.11 displays the traceplots for the species traits hyperparameters drawn from logistic (0,1) priors. Overall good mixing is highlighted by the traceplots, low posterior autocorrelation (Fig. 5.12) and Gelman–Rubin diagnostic < 1.1 indicates good convergence for model 5.5 (see appendix D.4.1 for colonization and survival hyperparameter convergence diagnostics).

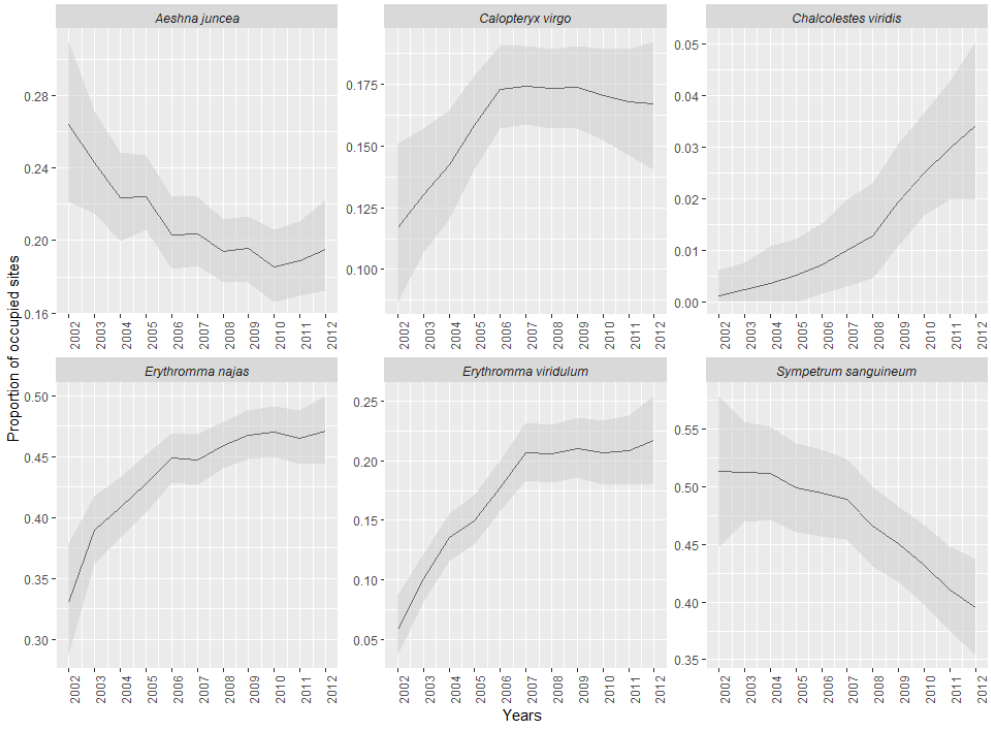


Figure 5.10: Odonata species estimated proportion of occupied sites from 2002-2012. Shaded gray area represents 95% credible intervals.

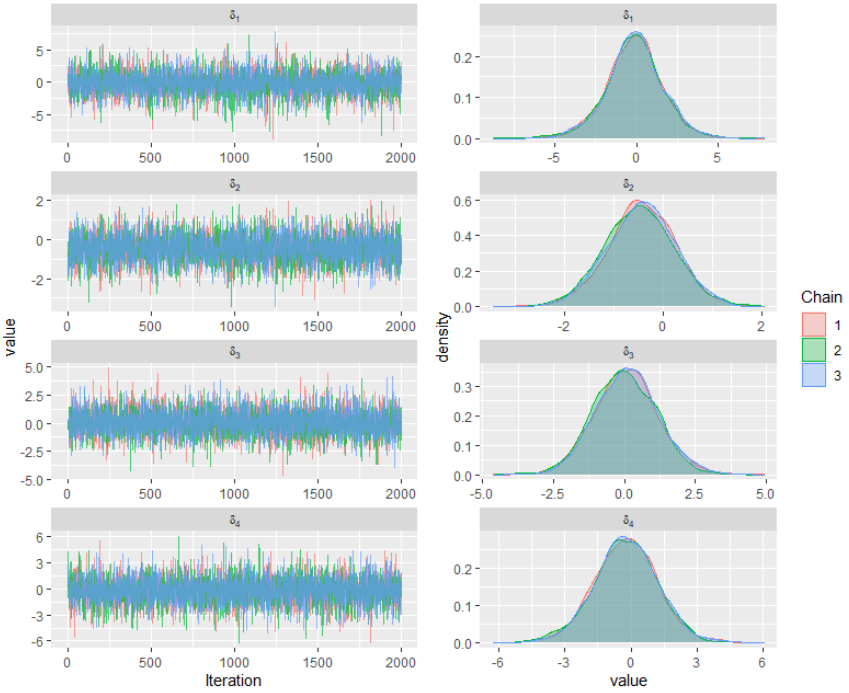


Figure 5.11: Odonata species traits community mean parameters traceplots for three independent chains.

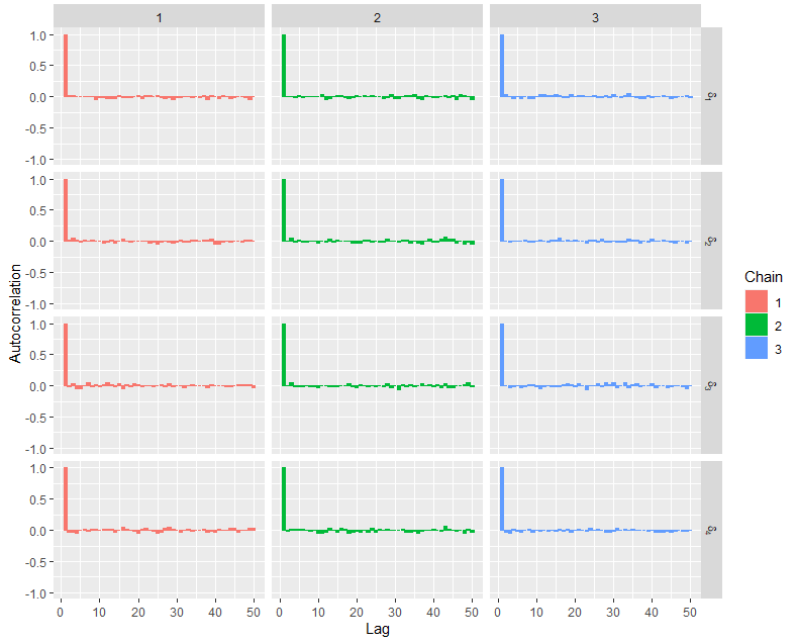
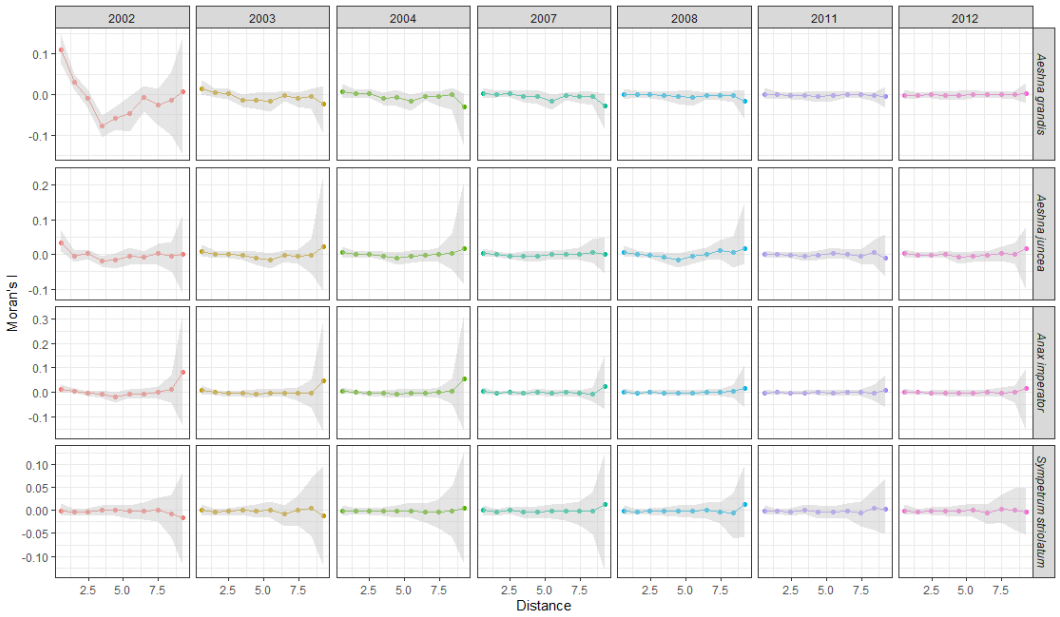
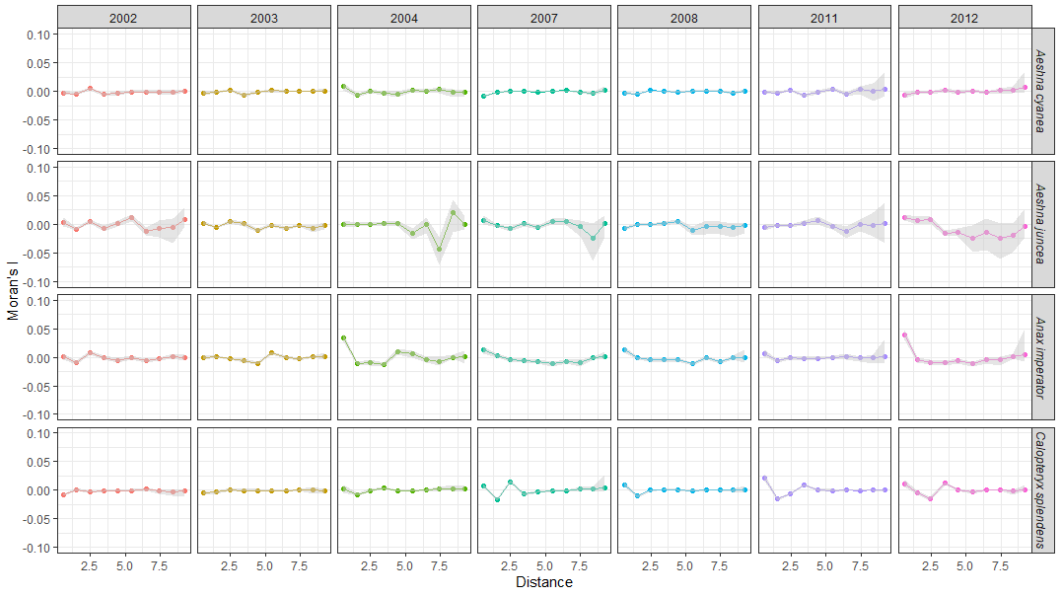


Figure 5.12: Odonata species traits community mean parameters posterior autocorrelation for three independent chains.

Spatial autocorrelation calculated for the state process residuals (Eqn. 5.7) using Moran's I (Eqn. 5.10) at increasing distances of 1 km shows little evidence of remaining spatial autocorrelation (Fig. 5.13 (a)). Note that Morans' I indicates that occupancy residual spatial autocorrelation is larger during the first year for a few species (e.g. *A. grandis* Fig. 5.13 (b)). This could be due to initial occupancy not being defined by any smooth terms, suggesting that the smooth terms in colonization and survival processes decrease the unaccounted spatial autocorrelation in the state process. On the other hand, the spatial autocorrelation observed in the observational process residuals correlogram (Fig. 5.13) even though significant, was very close to zero with no strong patterns (the largest observed departures from zero were less than 0.05 for *A. juncea*). Thus, this residual assessments did not indicate any major unexplained spatial correlation, suggesting that, in addition to species specific traits, human activity index is able to capture most of the spatial autocorrelation in the observational process. This would be expected as the index is computed based on a series of human activities centers and, since the volunteer work in citizen science projects is usually undertaken in sites with easy access, this index will reflect those locations humans frequently visit where species are more likely to be detected. However, it is important to notice that Moran's I shows a small but significant spatial autocorrelation over specific time periods for species like *A. cyanea* and *A. juncea* in Figure 5.13. Several occupancy models have been developed to account for such spatial autocorrelation in the observational process by specifying an explicit spatial occupancy model as described in Wright et al. (2019). However, in this chapter, the observational model described in equation 5.6 is proposed as an alternative formulation by using the Julian date to account for population size changes over time (along with the number of visits and the species' list length as a proxy for sampling effort).



(a)



(b)

Figure 5.13: Morans' I statistics calculated for occupancy model state process (a) and observational process (b) residuals at increasing distances of 1 km. Shaded gray area represents 95% credible intervals.

The results from fitting the observational model 5.6 for presence/inferred-absences are presented below. Estimated detection probabilities show a clear quadratic relationship with the Julian date (Fig. 5.14). The highest probability of detecting a species occurs during late spring and early Autumn which is consistent with the frequency of sampling records reported in Figure 5.2. This seasonal pattern corresponds to the time of year where the majority of Odonata adults are developed and thus, more likely to be detected (Braune et al., 2008) (see appendix D.4.2 for a larger subset of species detection probabilities estimates). The correlograms for the observational model residuals (equations 5.6 and 5.8) show little evidence of remaining spatial autocorrelation (Fig. 5.15). The amount of unexplained spatial autocor-

relation observed in the empirical correlograms based on the model that account for the date when each site was visited, the number of visits per site and the species list, is similar to the model that incorporates the human activities index to correct for unequal observation effort. However, the algorithm did not converge for several parameters in spite of the already large number of iterations (possible identifiability issues due to small sample sizes and computation memory limits). For instance, mean and variance community parameters from which species-specific baseline initial occupancy, colonization and survival effects are drawn showed a good mixing across three different independent chains (Fig. 5.16). Similarly, the parameters measuring the effect that temperature has on survival and colonization (and their corresponding variances) showed good convergence as indicated by the traceplots in Figure 5.17, with the exception of τ_{b_2} which is the precision parameter controlling the smoothness of the colonization term. Poor mixing was also observed for the observational model parameters (Fig.5.18). Specifically for the species-specific effects which had previously shown good convergence under the aggregated detection parametrization (Figures 5.11 5.12) and the precision parameter τ_{b_3} controlling the smoothness of the sampling dates effect (a sample of these parameter dignostics plots are presented in the appendix). Thus, the large number of parameters that need to be estimated from the presence/absence observational model formulation and the reduced number of sites not only makes the algorithm's running time slow, but also makes the parameter convergence challenging.

In summary, the proposed flexible dynamic occupancy model that considers the aggregated number of detections is computationally more stable and efficient compared to the model with presence-absence observational structure that uses dates and species length to correct for unequal observational effort. The issues with the latter formulation could be due to an overparametized model where there is not enough data to estimate the large number of parameters. A possible alternative would be to simplify such model. Based on the results the number of parameters in the model could be reduced by (1) using a simple quadratic term for modelling the day each site was visited (see discussion about species voltinism in the next section) while specifying a common intercept for all species as originally proposed by (Termaat et al., 2019) and (2) simplifying the latent model structure by removing the smooth term describing the relationship between survival and temperature which has proven to be smoother than the relationship between temperature and colonization. The following section discusses the ecological results and methodological remarks and caveats derived from this work, some of which are revisited in-depth in chapter 6.

‘

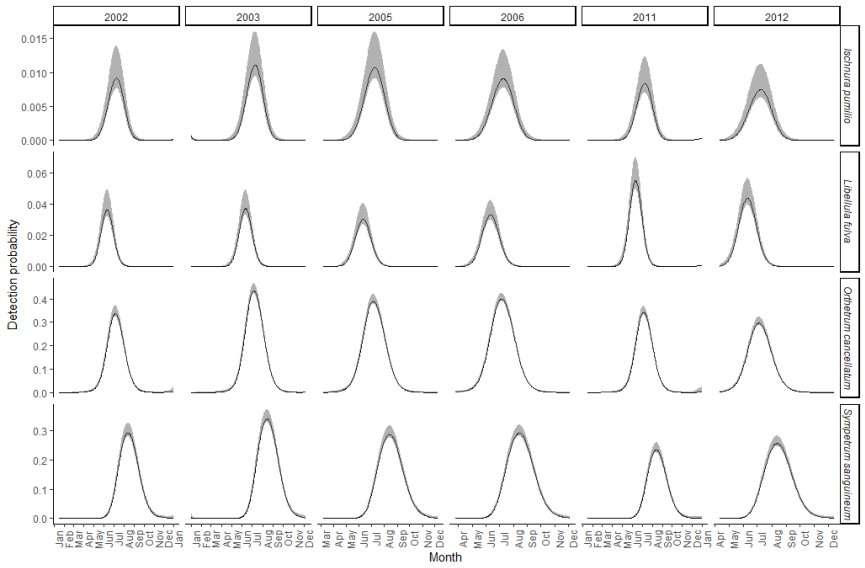


Figure 5.14: Relationship between detection probabilities and Julian date for Odonata species. Shaded gray area represents 95% credible intervals.

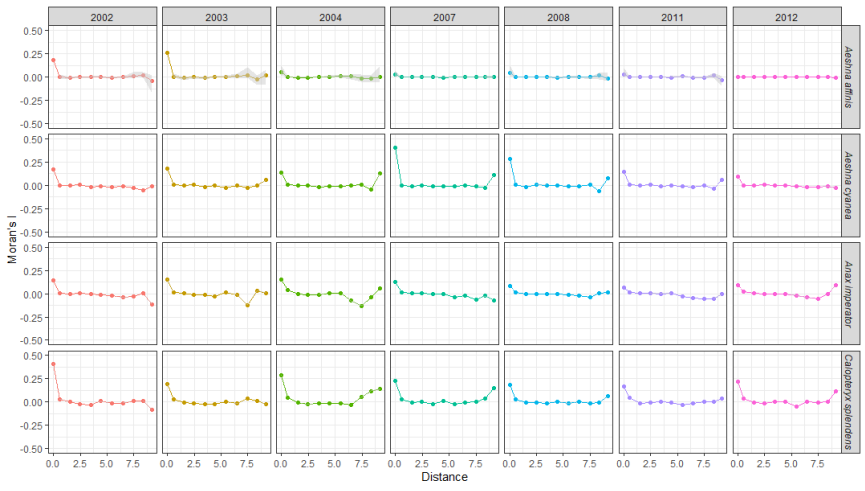


Figure 5.15: Morans' I statistics calculated for presence/absence observational process residuals at increasing distances of 1 km.

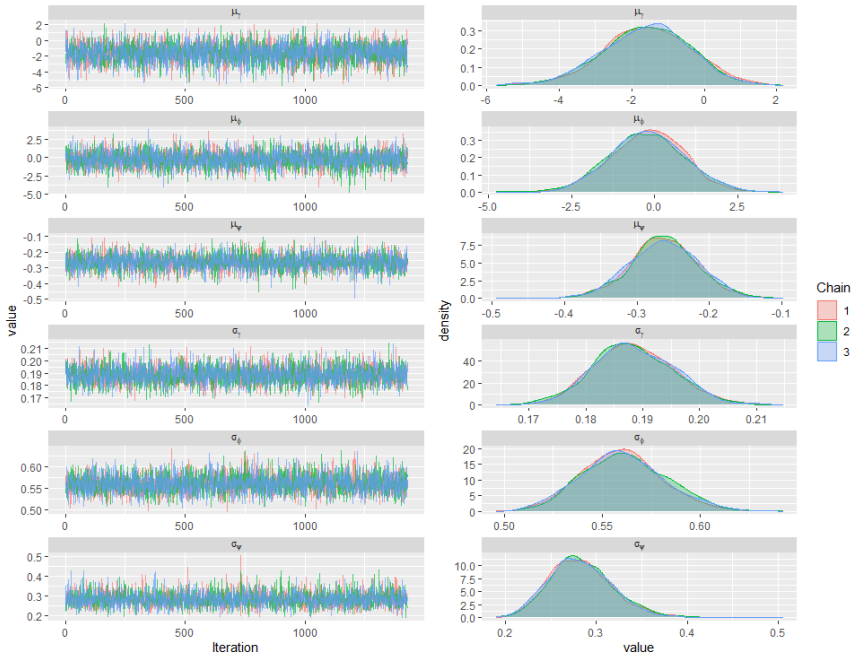


Figure 5.16: Density and traceplots for community baseline occupancy parameters drawn from the Odonata flexible occupancy model with a presence/absence observational model structure from three independent chains.

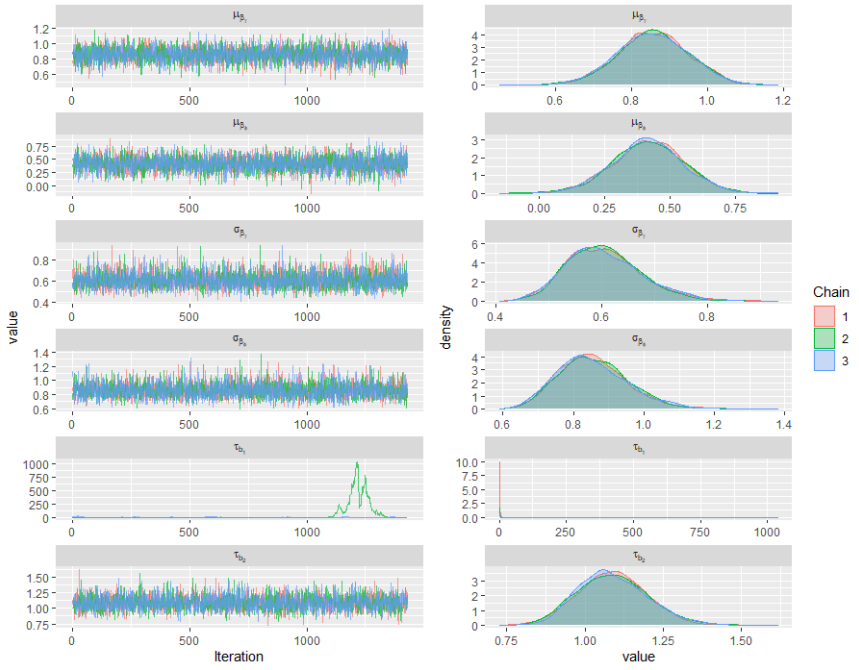


Figure 5.17: Density and traceplots for community detection parameters drawn from the Odonata flexible occupancy model with a presence/absence observational model structure from three independent chains.

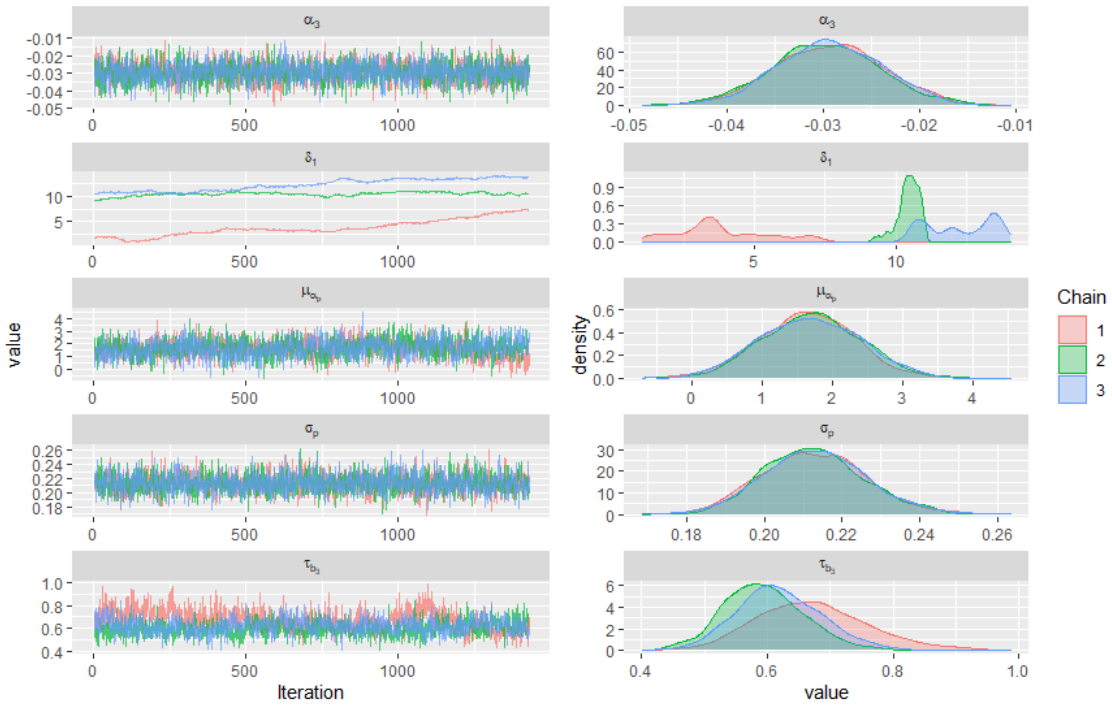


Figure 5.18: Density and traceplots for community temperature effect parameters drawn from the Odonata flexible occupancy model with a presence/absence observational model structure from three independent chains.

5.3 Model remarks and caveats

5.3.1 Modelling considerations

GAM-based species' distribution models have become an important tool for analyzing biodiversity data because of their flexibility to incorporate non-linear effects in the model structure (Elith and Leathwick, 2009). In this sense, the proposed model is conceptually similar but also enables non-linear relationships for the latent variables to be described while accounting for the observational process. Accounting for sampling effort and detection probabilities in the observation process is of major importance when working with large citizen data where observations are collected without a standardized field sampling protocol for a frequently arbitrary selection of sampling sites (van Strien et al., 2010, 2013). This variation in observation effort causes detection probabilities to vary over time and space (Kéry et al., 2010). Thus, to correct for unequal observation effort in this type of study Kéry et al. (2010) and van Strien et al. (2013) introduced daily species lists in combination with occupancy models to obtain estimates that are equivalent to those produced with a standardized field sampling scheme. In this chapter, the Chapman et al. (2020) human activity index was also proposed as an alternative predictor to account for uneven sampling effort.

Occupancy model residual definitions from Wright et al. (2019) were used to verify and identify residual spatial correlation among occupancy and detection. Moran's I residual diagnostics empirical correlograms assessing the flexible dynamic occupancy model for Odonata communities in the UK only provided evidence of residual spatial correlation among detections. However, the detected spatial autocorrelation was extremely low for both observational model parametrizations. Hence, introducing the human activity index of Chapman et al. (2020) as a predictor for the observational process proved to be as efficient as the daily species list proposed by van Strien et al. (2010) to correct for unequal sampling effort bias associated with these type of citizen-science data studies.

It is important to notice that some model inadequacies when using these types of residuals have been reported in the simulation study of Wright et al. (2019). For instance, the effectiveness of using occupancy residuals to identify spatial correlation among occupancy probabilities is reduced when a fully explicit spatial term is introduced on the detection probabilities. This might be caused by the additional source of uncertainty due to spatial dependency among occupancy probabilities. Hence, occupancy estimates with very large errors (as reported for some of the most rare or elusive Odonata species in this work) could mask a possible underlying spatial structure that is not been captured by the occupancy residuals.

While this approach evaluates the detection and occupancy components separately, both components are linked with each other because the occupancy posterior distribution depends on both occupancy and detection probabilities. Thus, the effectiveness of a separate residual assessment depends on how sparse the data are, and how much information available there is. If daily species lists are short, the number of visits is small or species occurrences are scarce, the process of identifying separate spatial structures for the two model components will be difficult. Nevertheless, conditioning residuals on the latent occupancy state provides a practical way to derive each residual component directly from the posterior without any additional transformation. For example, Warton et al. (2017) proposed to use "Dunn–Smyth" residuals based on an integral transformation to make residuals follow a standard normal distribution. However, such transformation introduces random noise that reduces the effectiveness to identify unaccounted spatial correlation (Wright et al., 2019).

Moreover, posterior predictive probability (PPP) can also be used to identify unexplained spatial structures in occupancy models (Broms et al., 2016; Wright et al., 2019). Wright et al. (2019) uses this approach based on the discrepancy between Moran's I calculated from the observation/latent variable residuals (e.g. $I^{[s]}(z,\psi) = z^{[s]} - \psi^{[s]}$ for occupancy residuals) and the residuals computed for a new data set simulated from the current parameter values at each iteration ($I^{[s]}(z_{sim},\psi) = z_{sim}^{[s]} - \psi^{[s]}$). Then, the Moran's I PPP $Pr(I^{[s]}(z_{sim},\psi) > I^{[s]}(z,\psi))$ is the proportion of posterior draws where the observed value is larger than the predictive value, a PPP value close to 0.5 indicates that predictions from the model are consistent with the observed data set (thus, a large deviation from this value provides evidence for unaccounted residual spatial correlation). However, this approach focuses on one distance class only and depends on (1) the neighbourhood definition (e.g. queens neighbours) and (2) the PPP cutoff that indicates spatial correlation (how large the PPA value must be for the spatial correlation to be considered meaningful).

Authors proposed that any PPP > 0.90 provides evidence of unexplained spatial correlation, but this could depend on the spatial scale and the species distribution range. For instance, it would be interesting to assess if PPP values less than 0.90 can provide evidence of moderate - to strong spatial correlation for short-ranged species on a regional scale. Moreover, although this approach can be potentially useful for a regular lattice where sites' adjacency can be easily determined, the sites on the Odonata case study are not within a regular grid due to the locations being determined by the presence of a waterbody. Thus, Moran's I definition based on the existing distance between sites and visualized with empirical correlograms provides a more sensible description that allows for the spatial correlation across multiple distance classes to be displayed simultaneously.

The results also indicate that by including smooth terms, the spatial autocorrelation in the model residuals is reduced for several Odonata species highlighting the importance of considering non-linear relationships to describe the diverse collection of species detection and occupancy responses. However, there are important caveats regarding this case study that need to be discussed and that can be potentially useful for practitioners interested in fitting these types of models to similar scenarios.

First, the model fitting has proven to be a computationally demanding process compared to the occupancy model applications presented in previous chapters. While previous applications of the occupancy model for the Odonata data set ranged between 3 to 8 hours, the flexible dynamic occupancy model approximate run time was one and a half weeks for the model with the aggregated number of detections and over 2 weeks for the model with presence/inferred-absence binary detections (run on a PC with i7-8086K CPU 4.00GHz with 48.0 GB RAM). Moreover, several runs were needed to compile the different sets of parameters due to large number of parameters and computation storage limits and convergence was not satisfied for several parameters in the binary detections sub-model. State process parameters, observational model parameters, and the model residuals were each obtained on separate runs resulting in an extremely long running time period. Hence, some authors have proposed running this type of models on external computer clusters such as LISA (van Strien et al., 2013) (a further discussion on computational efficiency of occupancy models and possible alternatives approaches will be reviewed in the following chapter).

These long run times can be caused by the complexity of the model where a large number of unknown parameters imposes a very large number of MCMC samples needed for the algorithm to converge. Running time is also enlarged when there are many rarely detected species (Broms et al., 2016). In chapter 2 half of the species in this study were identified as rare. Thus, estimating parameters for these species can be difficult due to the lack of information, especially at those sites with no detections for any species.

The lack of a consistent selection of sites to be monitored over time and the unbalanced number of surveys within each time period represents a major drawback for citizen data projects that yields to an uneven species lists for each site across the years. This reduces the total number of sites from which species absences can be inferred (i.e. sites where at least one species has been observed), making the task of fitting a temporal model slow and difficult and thus, limiting the number of terms (either smooth terms or more complex spatial effects) that can be included.

Hence, the choice of priors is important to ensure the algorithm's convergence for the largest possible collection of species, sites and years (although in practice this will be a time consuming process that does not guarantee the convergence of all the parameters in the model, specially for rare species at those sites without any detection records). Several studies have discussed the impact that different priors have on occupancy model estimates (e.g. Broms et al. (2016), Northrup and Gerber (2018) and Outhwaite et al. (2018)).

Sensitivity analysis to the choice of priors in this work showed consistent results with these previous studies and with the simulation analysis presented in chapter 4. For example, standard deviation parameters' weakly informative half-Cauchy distributed priors, which have been recommend by Gelman et al. (2006) in small sample size situations, helped with convergence in some situations where commonly-used Uniform (0,5) priors failed. The drawback of Uniform (0,5) priors is that at its upper limit, the precision parameter takes a minimum value of $\tau = 1/5^2$. Thus, if the data provide evidence that the true precision is < 0.04, the posterior mass would be dragged towards the upper limit set by the prior and would not be able to fully incorporate the information from the data. Similarly, too vague normal priors for logit-scaled mean parameters that should have low probabilities outside the (-5,5) range can show a U-shaped distribution with a high density close to 0 or 1 (see discussion in chapter 4). Hence, Normal ($0,2.25^2$) priors which are relatively flat over the (0,1) range (Broms et al., 2016), logistic (0,1) (Northrup and Gerber, 2018), or Student-t (0, 1.566, 7.763) distributed priors (Dorazio et al., 2011) have often been used in occupancy models as they produce more numerically stable algorithms. Nonetheless, there were no obvious differences when comparing the Odonata flexible dynamic occupancy model results with such priors.

Another aspect that needs to be considered before fitting the proposed dynamic flexible occupancy model is the definition of the primary sampling period where occupancy among visits is assumed to be constant. This largely depends on the species ecological responses. For instance, dragonflies and damselflies' primary sampling periods can be strongly linked with species voltinism, i.e. the number of generations that complete their life cycle within a year period (Braune et al., 2008). Once nymphs and juveniles complete their development, it is reasonable to assume that the adults distribution will remain relatively constant within their flight duration period until the next generation of individuals emerge van Strien et al. (2010). Once the primary sampling periods have been determined, a thoughtful examination of the potential covariates affecting both species occupancy and detection must take place. Due to the citizen data aforementioned limitations, only a few potential predictors might be introduced in the model and thus it is important to consider the ecological importance these covariates have (variable selection in occupancy models is an interesting area for future development that will be discussed in the next chapter). In this case study, only a single state-level covariate was used, thus the trade-off between fitting a non-parametric model and the number of terms in the model need further exploration. For instance, a pragmatic approach would be to assume a linear relationship between multiple covariates and the response, and carefully select those effects that are introduced as nonlinear terms in the model (for example by using binned-residual plots; see discussion in chapter 6).

5.3.2 Ecological insights

In this chapter, the proposed flexible dynamic occupancy model has been applied to describe the relationship between temperature and Odonata occupancy patterns in the UK. The model results suggest that higher temperature levels significantly increase the survival probabilities of Odonata species, and have a non-linear relationship with their colonization probabilities (where species-specific responses vary widely but generally show an increasing trend with temperature). These results are consistent with the observed range expansion response of Continental Europe' Odonata species towards increasing temperatures as reported by Termaat et al. (2019).

Several studies have demonstrated the role that temperature plays in dragonflies and damselflies life histories and population dynamics. For instance, high temperatures have proven to lead to earlier emergence of adults, longer flight seasons and enhanced voltinism, which could lead to higher colonization rates (Braune et al., 2008). However, the underlying mechanisms from which temperature modulates species responses are varied and complex. For example, Corbet et al. (2006) discussed how photoperiod changes regulate larval temperature-development relationships. Thus, the photoperiodic cues to which species larvae are exposed along a latitudinal gradient might also play a crucial role in early-adult development which could facilitate site's colonization at lower latitudes (Corbet et al., 2006; Hassall and Thompson, 2008).

In addition to photoperiodic modulation, colonization and survival can also be affected by the species-specific adaptation to varying hydroperiods which yields to enhanced voltinism as reported for *Lestes* damselflies species (De Block et al., 2008). Low temperatures have also been reported to affect species-specific responses. For example, damselflies wings' shape can change along a latitudinal gradient. At low temperatures, the wing's shape departs from an effective shape that can lead to a decline in the population fitness and potentially affect the probability of colonizing new environments and thereby increasing the risk of local extinctions within northern range margins (Hassall et al., 2008).

Temperature is also linked with other species morphological traits such as body size. Although this is a less studied subject, studies have shown that larger body-sized species are more likely to occur at lower temperatures (Hassall and Thompson, 2008). However, Johansson (2003) reported a non-linear U-shaped latitudinal body size pattern for *Enallagma cyathigerum*. This spatial variation in body size patterns could explain the results found in chapter 2 where the detection estimates' uncertainty is greater for species with smaller body sizes than for species with larger body sizes. On average, species with small body sizes will be less likely to occur (and thus to be detected) at the sampled sites and thus, there will be less information from which to estimate their detection probabilities, resulting in wider uncertainty regions. However, more studies are needed to have a better understanding of the role that latitudinal body size patterns have on life histories of Odonata species and how this impact the occupancy and detection of a species.

A previous study by Corbet et al. (2006) showed how the number of Odonata generations per year (for over 250 species) is negatively correlated with latitude. Therefore, the majority of the UK species exhibit a univoltine life cycle (completing one generation within a year), which is consistent with the observed

quadratic relationship between the estimated detection probabilities and sampling date. Given the type of habitat each species are known to occupy (temporary waters, lentic perennial waters, or lotic waters) and the latitude where these species are present, it is expected for most of the UK species to complete at least one generation per year where the adults are more likely to be detected in late spring and early Autumn.

However, other species of insect (including some Odonata species) might produce more than one offspring per year. Thus, it is expected for detections to vary based on the different occasions when adults emerge. Defining a flexible term in this sense allows for different types of voltinism responses to be captured while accounting for yearly changes in the population sizes. In small sample sizes situations researchers might need to choose between including a flexible terms to describe the effect that species voltinism has on detection, and including flexible terms to capture the effect that environmental covariates have on population dynamics.

Although species-specific responses vary widely, species richness, average and total growth rates, and proportion of occupied sites estimates suggest that the overall community structure has not experienced any major temporal changes. A recent study by Termaat et al. (2019) investigating the effect that climate change has had on Odonata species communities in Europe found similar results for the period 1990–2015. From the 50 species in this study, authors reported 26 species to have an increasing trend, 12 to have a stable population with a non-significant trend, 2 species with negative trends, and 10 species with no significant change and standard errors too large to detect a 5% trend. Authors were able to identify a general increasing trend for the UK species by classifying them into cold-dwelling and warm-dwelling species based on a species-specific temperature index computed with the mean annual temperatures at each of the occupied sites. However, the overall trend among both cold-dwelling and warm-dwelling species follows the same pattern found for the 40 species analyzed through the dynamic flexible occupancy model in this chapter (specifically for the time period 2002-2012). Similarly, a recent study by Outhwaite et al. (2020) analyzing UK invertebrates (including dragonflies) temporal distribution trends from 1970 to 2015 showed no evidence of any significant increase for the dragonflies mean occupancy for the 2000-2015 period.

The novel contribution of the proposed model in this chapter with respect to these previous studies is the flexible framework for simulating and analyzing complex relationships between the population latent structure and site-level metrics while providing flexibility in the observational process definition to characterize changes in species distributions over time. The results from this model can be used to identify vulnerable species and the potential threat of invasive species to the whole community. This enables the formulation of new research questions to be explored in future work relating occupancy models such as computational efficient methods to incorporate multiple flexible terms, producing model diagnostics and validation metrics, variable selection (dimensionality reduction approaches), incorporate species interactions, and optimization of sampling protocols and practices. Hence, the next chapter will address the general insights gained throughout this research and will discuss future areas of development.

Chapter 6

Final discussion and conclusions

Modern methods for quantifying, predicting and mapping species distribution have played a crucial part in biodiversity's conservation and management. These methods have helped ecologists to understand different aspects of biodiversity conservation such as biological communities assemblage, trophic interactions, population dynamics, conservation of threatened and endangered species, and the effect of management in terrestrial, freshwater, and marine environments (Elith and Leathwick, 2009). To study these ecological processes, linkages between modelling techniques, ecological theory, and data collection must be developed. Such linkages should make a distinction between the sources of uncertainty associated with the ecological processes of interest and those observational errors that, within the context of species distribution modelling, are frequently induced by the species imperfect detection (Cressie et al., 2009). Over the last decade, the rapid development of statistical methods to estimate species distribution has been facilitated by new data collection schemes and information systems that enable large-scale biodiversity metrics to be obtained by integrating information from different sources such as museum collections, planned *in situ* systematic surveys, radiotelemetry studies, and citizen-science projects where people contribute with their personal observations (Kellner and Swihart, 2014; Koshkina et al., 2017; Altwegg and Nichols, 2019; Devarajan et al., 2020). Unfortunately, analyzing such data is challenging due to imperfect detection caused by the observer error, species rarity, or environmental conditions that make the task of detecting the species difficult (Kellner and Swihart, 2014; Isaac et al., 2014).

Despite numerous studies demonstrating that detection is rarely perfect or constant among species and that ignoring it can cause an underestimation of the true species distribution, there is still a reportedly important gap in ecological studies addressing detectability, specially for small invertebrates such as Odonata (Kellner and Swihart, 2014; Devarajan et al., 2020). Over the last decade, very few studies have addressed Odonata large-scale distribution patterns under detectability bias despite their well-known importance as bioindicators of aquatic ecosystems well-being (van Strien et al., 2010, 2013). Motivated by this, this research has explored and developed novel statistical methods to estimate species distribution under imperfect detection that enables for different attributes in biological communities and their relationship with the environment to be described.

Specifically, the research questions addressed by these methods are: *How important is it to account for imperfect detection?*, *How can rare species in a community be identified?*, *what environmental conditions (e.g. connectivity and stressors) drive species distributions?*, and *How can population dynamics changes over time be modeled in a flexible framework?* Such methods were applied to characterize Odonata species diversity across UK freshwaters, since their species presence records are only partially observed due to imperfect detection.

In this work, a comparative analysis between Bayesian and Classical methods has been presented to estimate species occupancy under imperfect detection using different programming languages to assess computational efficiency and the effect of sampling design on model performance has been discussed. Then, different models were fitted and compared to estimate Odonata distribution in sites defined by the presence of a waterbody. Based on these results, two different approaches to quantify rare species in a community were developed, discussed and applied to identify Odonata rare species. A two-stage modelling approach was developed to evaluate how a large suite of environmental metrics available over nested spatial scales shape rare species distributions or any other quantity derived from an occurrence-based occupancy model. To identify temporal changes in a species population dynamics a flexible model was developed and analyzed using computer simulation. Then, this method was applied to describe Odonata population dynamics in the UK across a 10 year time period. This model was used to describe the effect that temperature has on shaping the species distribution while fitting two different observational sub-model structures to account for unequal observation effort which is commonly associated with citizen science data. The next section will review some of these findings and discuss the contribution of this research to how species distributions are analyzed under imperfect detection. Then, some of the challenges and practices involving how data are collected and analyzed will be discussed based on the results derived from this work and, finally, the perspectives and future steps this research has led into will be discussed.

6.1 Research outcomes

6.1.1 Accounting for imperfect detection

The comparative simulation-based analysis between Bayesian and Classical inference for a single season occupancy model showed that Bayesian occupancy estimate errors are lower than the errors associated with the likelihood-based estimates when species occupancy/detection probabilities are low. It has also been shown that while Bayesian analysis can be computationally intensive, systems such as Nimble (de Valpine et al., 2017) can help to reduce the algorithm's running time (see discussion on computational efficiency in the next section). It is important to notice that the efficiency of these methods depends on (1) how elusive the target species are, and (2) the sampling design. For instance, if the probability of detecting a species is high (e.g above 0.8), the results from fitting an occupancy model will be similar to those obtained through a standard logistic regression.

Moreover, from a classical-inference point of view, sparse data induced by low occupancy/detection probabilities will result in similar estimation bias regardless of whether imperfect detection is accounted for or not (Welsh et al., 2013). However, as mentioned by Guillera-Arroita (2017), accounting for imperfect detection allows having an honest representation of the uncertainty associated with the data collection process. Even if the sampling design minimizes the observational error (e.g. Figures 2.4 and 2.5 in chapter 2), the estimates obtained through methods that account for imperfect detection will match those from methods that ignore it. So it is a better practice to account for detection bias rather than ignore it and thus potentially produce severely biased estimators.

An extensive revision by Kellner and Swihart (2014) of a collection of studies addressing imperfect detection, found that 70% of such studies estimated detection probabilities to be less than 0.5, and 50% of the articles reported species detection probabilities to be less than 0.3, highlighting the importance of accounting for detection bias in species distribution modelling. Having information about species detection will enable us to make a distinction between an otherwise confounded occupancy and detection processes since each observation would portray the combined effect of these two processes. Thus, a formal assessment of whether an important bias is introduced by imperfect detection can be only made by making a clear distinction between occupancy and detection. To do so, the data collection schemes and the subsequent analysis must be informative about the detection process (Bailey et al., 2014) and thus, imperfect detection should be considered across the different stages of a study, i.e, during the design, collection and analysis of species occurrence data (Guillera-Arroita, 2017).

The importance of accounting for imperfect detection also applies for mutiple-species studies where occupancy models proved to be an efficient tool for making inference about the whole biological community structure and composition (Dorazio and Royle, 2005). For instance, chapter 2 highlights the advantages of using a Bayesian hierarchical structure to estimate individual species responses by assuming each species arise from a common community with shared attributes. A multi-species occupancy model was then fitted to estimate Odonata distribution in UK freshwaters and community/diversity metrics were derived to characterizes species richness. The results indicated that the estimated occupancy probabilities varied widely within the Odonata community but detection probabilities were estimated to be less than 0.5 for all the species. Additionally, species richness proved to be higher at lower latitudes and species-specific traits were found to have a significant effect on detection probabilities. These results illustrate the base model that was further developed in the next chapters to address the aforementioned research questions.

6.1.2 Modelling rare species distribution

To identify rare species composition in the biological communities two occurrence-based methods were explored. First, Leroy et al. (2012) index of relative rarity (IRR) was computed as a derived quantity of the aforementioned multi-species occupancy model to account for imperfect detection. As an alternative approach, a mixture occupancy model was developed to classify rare species based on their relative occupancy probabilities.

The IRR efficiency to characterize species rarity depended on the occupancy threshold at which a species are considered rare. The choice of this threshold can be subjective if there is no prior information about the species rarity (Leroy et al., 2012). Moreover, different rarity cut-off values can yield very different results, and thus it is difficult to make a fair comparison against other methods. On the other hand, the proposed mixture occupancy model proved to be an efficient method to quantify the number of rare species in a community. By using computer simulations, the model results were compared under different scenarios obtaining a high classification performance across all of them. Both methods were applied to the Odonata occurrence records to quantify the proportion of rare species on a national scale across waterbodies in the UK. Rare species composition across the UK was found to be scarce and homogeneous across space, possibly due to the scale of the study. Species rarity is a scale-dependent process (Hartley and Kunin, 2003) and thus integrating the proposed approach with multiscale sampling designs (Nichols et al., 2008) could provide an accurate description of how rare species are assembled at different spatial scales. Thus, the proposed mixture occupancy model represents a new approach to identify and quantify the composition of rare species in a community while accounting for detectability bias. Using this modelling framework opens lines of research and future developments for the understanding of how species rarity can be measured in a wide range of scenarios some of which have already been discussed in chapter 3 and will be reviewed in the final section of this chapter.

Then, a two-stage statistical modelling framework was proposed for analyzing how environmental metrics describing freshwater connectivity interacted with land-use change to affect rare species distributions. While the focus of these methods was the composition of rare species (estimated with the mixture occupancy model and the Bayesian IRR) it could be also applied to any other community/diversity metric derived from an occupancy model such as species richness or the Jaccard dissimilarity index discussed in chapter 2. This 2-stage approach estimates detectability bias in stage 1 and incorporates the effects of the covariates on the adjusted species distribution in stage 2. The second stage evaluates the effect of site-level covariates on the composition of rare species richness by using Random forests to identify the explanatory variables that are “important" to the response and selecting a reduced set using prediction MSE as criteria. Then, the reduced set of potential explanatory variables is considered in a generalised additive model (GAM), allowing for smooth, nonlinear relationships, with interactions to be modeled using tensor products (Stage 1 uncertainties are included through inverse-variance weighting in this second step).

Strong relationships between the composition of rare species and connectivity/stressors effect could not be determined possibly due to the homogeneity found in rare species distributions across sites which is influenced by the spatial scale at which rarity is defined (constrained by both the extent and grain of the study, see discussion on the spatial scale in section 6.2), and because of individual species responses varying widely to environmental conditions. Nevertheless, by taking this two-step analysis researchers can synthesize a collection of estimates into a single estimate while properly propagating the uncertainty associated with the first estimation process. This is particularly useful in a community occupancy framework because it allows estimates from one analysis based on partially observed occurrences to be used in

a second analysis to relate them to environmental variables. This approach can also be used as an initial exploratory step in hierarchical modelling (Kéry and Royle, 2015), or to perform variable selection in a much more computationally efficient manner compared to fitting a fully Bayesian Hierarchical model and perform variable selection all at once, which has proven to be problematic and computationally intensive (Tenan et al., 2014), especially with correlated variables (this is a less studied subject within occupancy modelling framework that will be discussed in the last section of this chapter). For instance, the large number of parameters in such a Bayesian analysis can be so large that estimation would not be able to be obtained all at once as in the proposed flexible Odonata dynamic occupancy model in chapter 5, making the analysis difficult and hard to communicate with a broader class of audience. Thus, the proposed two stage-approach enables an otherwise too complex hierarchical model to be simplified and analyzed in different stages to provide a pragmatic computationally efficient method for choosing the most relevant predictors affecting an ecological response of interest while propagating the uncertainty associated with this quantity on a second analysis.

6.1.3 Developing a flexible modelling framework to understand population dynamics

Understanding how sites become occupied or unoccupied by a collection of species over time, and how these colonization and extinction dynamics depend on environmental predictors is of major interest for ecologists (MacKenzie et al., 2003). While ignoring imperfect detection leads to bias in colonization and survival estimates (Kéry et al., 2013) a review by Tingley and Beissinger (2009) found that only a third of studies assessing population dynamics considered imperfect detection. On the other hand, studies that did consider detection bias have defined colonization and survival as linear functions of environmental covariates (e.g. Green et al. (2019) defined survival and colonization probabilities as a function of grassland cover) and have not yet defined these relationships in a flexible framework. Thus, a non-parametric Bayesian dynamic occupancy model using penalized thin-plate splines was developed here to account for the non-linear relationships between occupancy probabilities (colonization and survival) and site-level covariates. This model also allows for smooth terms to be included in the observational model to account for unequal observational effort across surveys.

The proposed model captured, with a relatively low error, the different individual species responses across different simulated scenarios (even for sites that were visited only once) and proved to be a very efficient method to describe the general community/diversity metrics such as species richness, proportion of occupied sites, and population growth rates. This model was further developed and applied to estimate Odonata population dynamics and to investigate the effect that temperature has on species colonization and survival probabilities. Species-specific response probabilities vary widely in the community. Colonization probabilities showed a more complex non-linear relationship with temperature than survival. Moreover, the model results suggested that both of these probabilities increase with medium-high temperature levels. The community metrics derived from this model indicate that Odonata communities have

not experienced any major changes over the time period of the study. This can be the result that more than half of the species in this study did not show evidence of significant temporal changes in their occupancy state. However, by assessing species' individual responses researchers can identify potentially vulnerable species, for example by looking at the individual growth rates. Developing a flexible dynamic occupancy model that also allows for smooth terms to be introduced in the observational sub-model allows for heterogeneity in species detection to be captured. In this work, the Chapman et al. (2020) human activity index was proposed as a predictor to account for uneven sampling effort. Unfortunately, the information required to compute this index is not always available, specially at the same grid level at which other environmental covariates and occurrences records are measured. Thus, a frequent approach to deal with sampling bias is to introduce the date when each site was visited, the number of visits, and the species list length as an alternative formulation of the observational model (van Strien et al., 2010). Incorporating a smooth term to describe the relationship between the human activity index and species detection probability proved to be computationally more stable than using a flexible term for the date when each site was visited. However, making an ecological interpretation of this relationship is difficult due to the index complexity in measuring the proximity-weighted contributions from different sources of human influence such as local land cover, population density, and transport infrastructure. On the other hand, while incorporating the sampling record date proved to be computationally demanding, describing the non-linear relationship between detection probabilities and the date when each site was visited has a potentially novel ecological application to study the distributions of species with different classes of voltinism.

In summary, occupancy models have proved to be a powerful tool that provides a flexible framework to incorporate complex structures that can be used to address different ecological problems. In this work, Occupancy models were developed to describe Dragonflies and Damselflies distribution in the UK and to identify rare species in a community. These models were also used in the aforementioned two-stage approach to describe how environmental conditions shape species distribution, and finally a flexible modelling framework was proposed to fit these class of models to characterize species occupancy changes over time. There are, however, some important modelling features that need to be considered in order for any conclusions drawn from studies involving species distribution and detectability to be valid. Some of these considerations are challenging and have not been fully explored. Thus, the following section will address this and discuss some of the challenges and practices identified from working with species distribution modelling under imperfect detection.

6.2 Methodological considerations and challenges

6.2.1 Defining the scale of the study

The first chapter of this thesis described the statistical principles involved with species distribution modelling and the ecological context in which the proposed methods have been applied. This first chapter also introduced (1) the information exchange and (2) the scale dependency, two key concepts that need to be considered in order for any statistical inference to be valid. The information exchangeability defines the unidirectional structure in which species richness can be inferred from the species occurrence data which can itself be inferred from abundance data. This information structure is ultimately derived from an underlying spatial point process (see chapter 1 and 2) which depends on the spatial scale defined by the grain and extent of the study. As mentioned by Koshkina et al. (2017), the extent of the study is driven by the purpose of the analysis. For example, a national scale study like the Odonata citizen science data set has a macroecological scope that aims to determine the general occupancy pattern in Odonata communities occurring in freshwaters all across the UK. However, the patterns observed on this large scale might not reflect what happens on a local or regional extent. For example, the definition of species rarity in chapter 3 depends on the relative occupancy probabilities among all the species that occur on a region, thus species with local distributions that are considered rare on a national scale could be common on a regional scale and therefore might not have the same importance from a local conservation point of view. The spatial scale is also determined by the grain, i.e. the grid cell size of the spatial units in which species occurrence records and environmental covariates are measured. While modern information systems have proven to be an important tool for manipulating both species records and environmental data at different grain levels, very often the grid cell in which occurrence data is measured needs to be adjusted in order for it to be consistent with the resolution in which environmental covariates are obtained (Fithian and Hastie, 2013). This constrains the model's parameter interpretation to the available resolution and narrows the conclusions drawn from these models to the spatial scale in which both species occurrences and environmental variables were recorded. For instance, the models and applications derived from this research were constrained to the 1 km grid cell defined by the Ordnance Survey National Grid. This grain size has been reported for recent studies involving Dragonflies and Damselflies distribution (e.g. van Strien et al. (2010, 2013); Outhwaite et al. (2020)) and seems to be appropriate given the Odonata species flight ranges (from discussion with British Dragonfly Society experts, 2020). The methods derived from this research can be applied to different taxonomic groups but will be determined by the extent and grain of the study which should be defined based on the ecological context of the species.

It is important to mention a special class of abundance-based occupancy model that have been developed to produce estimates that are invariant to the choice of spatial scale for counts data (Fithian and Hastie (2013); Dorazio (2014); Koshkina et al. (2017); Martino et al. (2021)). While this is beyond the scope of this research, this class of models represents an interesting area for future research to integrate data from different sources such as opportunistic and planned surveys while accounting for observer error and site selection bias.

This approach is based on estimating the species occupancy rates, i.e. expected number of individuals per unit area, to provide a sensible metric to quantify the occupancy probability ψ simultaneously for all sites $j = 1, \ldots, M$ by parametrizing the standard occupancy model in terms of an inhomogeneous Poisson point process:

$$\psi_j = \Pr(N(A_j) > 0) = 1 - \exp\left(- \int_{A_j} \lambda(\mathbf{s}) ds \right). \tag{6.1}$$

Where $N(A)$ denotes the number of individuals in region A that depends on the mean intensity $\lambda(\mathbf{s})$ of the process over the region.

This occupancy model likelihood can be rewritten in terms of this abundance-based occupancy model parametrization:

$$\begin{aligned} L(\psi_j, p_{jk}) = & \left[\prod_j^J \left(1 - \exp\left\{ - \int_{A_j} \lambda_N(\mathbf{s}) ds \right\} \right) \prod_j^K p_{ij}^{y_{jk}} (1 - p_{jk})^{1 - y_{jk}} \right] \quad \times \\ & \prod_{J+1}^M \left[\left(1 - \exp\left\{ - \int_{A_j} \lambda_N(\mathbf{s}) ds \right\} \right) \prod_j^K (1 - p_{jk}) + \exp\left\{ - \int_{A_j} \lambda(\mathbf{s}) ds \right\} \right]. \end{aligned} \tag{6.2}$$

Then, conditional on an individual being present, the probability of a species being detected in an opportunistic survey ($p(s)$) at locations $s_1, \ldots, s_J$ ($J < M$) is determined by a thinned Poisson process where the expected number of detections is given by $\nu(A) = \int_A \lambda(\mathbf{s}) p(s) ds$.

Koshkina et al. (2017) showed how the parameter resulting from combining these different sources of data are invariant to the choice of spatial scale and enables detection probabilities to account for both site selection bias and observer error simultaneously. This approach can be implemented in a Bayesian inference framework using the integrated nested Laplace approximation (INLA) (Martino et al., 2021) and is a promising area for modelling species distribution under imperfect detection. Unfortunately, the extent to which Poisson processes can be inferred from occurrence data can be limited due to the information exchangeability constrain that depends on the grain of the study and the intensity of the underlying point process. As mentioned in chapter 1, the mean abundance and occupancy are proportional to each other when the grain of the study is small and the intensity function produces a small number of observations per cell, this relationship has been exploited in binomial-mixture models to make inference about abundance based on the information provided by occurrence data sets (Royle and Nichols, 2003). However, as the grain and intensity increase, the proportion of occupied sites increases and the relationship between abundance and occupancy departs from linearity, making occupancy a non-informative measure for either the underlying point process or the abundance (Kéry and Royle, 2015).

6.2.2 Sampling bias in citizen science data

Freshwater monitoring programs have become essential to describe, predict and map species whose occupancy patterns depend on different freshwater attributes (e.g. temperature,alkalinity , oxygen levels and macroecological processes such as connectivity) across large geographic and temporal scales (Altwegg and Nichols, 2019). Unfortunately, collecting data from such studies can be costly due to the amount of effort and experience needed for collecting and identifying samples from different species, limiting the spatial and temporal resolution of the study (Dennis et al., 2017). Thus, citizen science projects, involving volunteers who help to collect data and monitor sites, offer a cost-effective solution to address research questions at spatial and temporal scales that would be otherwise difficult to achieve (Aceves-Bueno et al., 2017). However, the analysis of these data is challenging because of variable sampling effort, species imperfect detection, and the number of specialists capable of identifying certain species (Altwegg and Nichols, 2019). While the number of specialists studying specific organisms might be limited in certain regions, species such as Odonata characterized by large colorful bodies, can be easily spotted and identified by participants that have received a relatively low amount of training (van Strien et al., 2010, 2013). Despite this advantage, these opportunistic recording schemes usually lack of a standardized protocol, resulting in uneven sampling effort (e.g. participants will have different recording skills and will tend to visit sites that are easier to access) (Tulloch et al., 2013). Moreover, very often protocols for citizen-science projects only ask volunteers to record species detections resulting in presence only data where there is no information about the species absence (Altwegg and Nichols, 2019), which is the case for the Odonata occurrences record analyzed in this work. Thus, in this research Kéry et al. (2010) and van Strien et al. (2010) approach was followed to generate species absences based on the occurrences of other species recorded at each site.

This has proven to be an efficient approach that produces estimates that are similar to those obtained through monitoring schemes with standardized sampling protocols (Kéry et al., 2010). Species non-detection records generated by this method should reflect as much a possible the biological similarity between the target species for which non-detections are inferred and the background species used to infer that target species “observed" absence (Kéry et al., 2010). In this work, it has been assumed that the complete list of recorded Odonata species provides enough evidence to infer the absence of a target species confirmed by the presence of any species in the list other than the target species. Conditioning the non-detection data to certain species with similar biological behaviors (e.g. species of the same guild or with similar habitat selection), as suggested by Kéry et al. (2010), yields a more sensible ecological interpretation of the model parameters. However, by doing so there is a potential risk of having shorter species lists and fewer or non-existing absence records for the species belonging to a specific group with shared biological attributes. For instance, van Strien et al. (2010) reported that occupancy trends estimated from single records and short daily lists were difficult to detect due to the large standard errors caused by the uncertainty in the detection probabilities estimates. Moreover, missing non-detection records are also problematic as they tend to show convergence issues which become more evident when working over large time periods (see discussion in chapter 5). Under a temporal framework, missing non-detection

could be produced either by a species' population decline or by inconsistent sampling intensity, and thus introduces bias to the observational process (Kéry et al., 2010).

To correct for this bias, the number of visits at each site was incorporated into the single-season occupancy models developed in chapters 2 and 3. Conditioned on the number of visits, it was assumed that detection probability was the same across sites and only depended on the species-specific traits. To make parameters identifiable, sites with a single visit only were removed. Despite this, the number of sites was large enough to estimate the occupancy parameters with a reasonably low error. For the more complex temporal occupancy models presented in chapter 5, the human activity index, species list length, number of visits, and the date each species was observed were incorporated to correct for unequal sampling effort and population size changes over time (note that the number of sites was greatly reduced due to the lack of repeated visits and large amount of missing non-detection records, resulting in larger standard errors and convergence issues for some species).

This approach to account for unequal sampling effort has become an increasing choice in recent studies involving citizen science data (Outhwaite et al., 2020). However, van Strien et al. (2013) reported that occupancy estimates based on "opportunistic" surveys were often higher than those obtained through standardized monitoring protocol, possibly due to the difference in the sampling area coverage each sampling scheme has. For instance, while standardized monitoring schemes usually sample a specific location characterized by a unique water-body type, observers in "opportunistic" surveys visit several watertypes and collect information of species occurring in different habitats, leading to higher occupancy rates (van Strien et al., 2013). Unfortunately, observers in "opportunistics" surveys usually fail to report the sites where species have not been seen. Thus, encouraging participants involved with citizen science data projects to report the sites they visit even though none of the species on the checklist are detected would result in a better practice that could reduce the bias induced by the lack of presence/absence observations.

6.2.3 Model fitting: advances and challenges

During the past decade, a wide range of occupancy models have been developed and applied to a variety of scenarios. From a single species occupancy model (MacKenzie et al., 2002) to complex spatio-temporal models (Rushing et al., 2019), occupancy models have proven to be a flexible tool that ecologists and conservationist use for studying different aspects of species distributions. As occupancy models began to grow in complexity, the approaches and software used to fit such models changed accordingly. In this work, the current available approaches to fit the standard occupancy model were compared. Classical approaches to fit such models were a popular choice among ecologists during the first decade after MacKenzie et al. (2002) occupancy model was first introduced (Kellner and Swihart, 2014) (programs such as PRESENCE (Hines, 2006) or `unmarked` (Fiske and Chandler, 2011) library in R were a common choice for modelling species distribution under imperfect detection (Guillera-Arroita, 2017)).

While classical methods proved to be the most computationally efficient method to estimate occupancy parameters, their lack of flexibility to model complex scenarios and the numerical instability

caused by sparse data made Bayesian methods the most frequent approach to address detection bias in species distribution modelling over the past few years (Devarajan et al., 2020). As discussed in chapter 2, Bayesian modelling hierarchical structures enables species information to be exchanged across species. Thus, species with sparse occurrence data can "borrow" information from other data-rich species which facilitates the estimation process for even the most rare and elusive species (for which estimation through classical methods usually failed). The Bayesian hierarchical formulation of Occupancy models assume that species are conditionally independent given a common/shared community parameter that represents the parameter's average value among species in the study. Thus, the Bayesian formulation of Occupancy models not only facilitates the otherwise difficult analysis of rare species distributions but also allows inference about the whole community to be made. By taking a Bayesian framework, different aspects need to be considered such as, the choice of priors, the model structure and assumptions, the method to estimate the posterior distribution, and finally how to assess the model convergence and fit. While several of these criteria are common for any Bayesian fitted model and have already been discussed along this research, there are certain specific features of Occupancy models that need further discussion.

6.2.3.1 Computational efficiency

In chapter 2, different programming languages to sample from the posterior were compared. A recent review by Devarajan et al. (2020) showed a strong preference for fitting Bayesian multi-species occupancy using BUGS, JAGS or Stan languages. While different R-packages have been developed to fit such models (e.g `hSDM` library encompassing a collection of occupancy species distribution models based on C_{++} (Vieilledent, 2019); `stocc` aimed to fit spatial occupancy models using probit-based Gibbs sampling algorithm (Johnson, 2021) ; `ubms` implements an user-friendly based collection of occupancy models using in Stan (Kellner, 2021)) the flexibility enabled by JAGS or Stan to define different modelling structures has made them a popular tool among researchers. Unfortunately, Bayesian inference can be computationally intensive and languages such as BUGS or JAGS haven proven to be computationally less efficient compared to samplers written directly in R or C_{++} (see Table 2.2 in chapter 2). As discussed in chapter 5, the time to run and compile such models increases as model structure grows in complexity. This has also been reported by Clark and Altwegg (2019) where a written C_{++} Gibbs algorithm to obtain posterior samples of a spatial occupancy model was 12 times faster than a JAGS or Stan written samplers. Unfortunately, coding such samplers in C_{++} can be a challenging task that often results in large complex pieces of code that are hard to read and reproduce, especially for ecologists or conservationists with little programming experience. Moreover, because of the wide range of scenarios that occupancy models can address, producing a single package that encompasses these diverse and complex model structures is difficult. Thus, the `nimble` library provides a powerful system that improves computational efficiency by enabling models written in BUGS to be compiled in C_{++}. Through this work, the `nimble` package was used to implement the different proposed models. The intuitive construction along with a variety of built-in sampling MCMC algorithms that can be customized makes Nimble a compelling tool for both ecologists and statisticians for fitting occupancy models. Thus, future species distribution studies seeking

a less demanding computational process will more likely be developed with this system. For instance, the recent `nimbleEcology` extension grants users the access to a collection of Nimble functions and algorithms specialized for building ecological models such as occupancy models (Ponisio et al., 2020). Nevertheless, complex models such as spatially explicit multispecies occupancy model (Clark and Altwegg, 2019) or the flexible dynamic multi-species occupancy model developed in chapter 4 will still be computationally challenging because of the nature of Bayesian estimation through MCMC methods. An alternative approach that has not yet been fully explored within the occupancy models scope is to approximate Bayesian inference using integrated nested Laplace approximation (INLA) (Meehan et al., 2017). A brief discussion of this approach will be described in the following section.

6.2.3.2 Modelling assumptions

Developments of occupancy models have great flexibility to incorporate different sources of variation in species detectability and occurrence (Devarajan et al., 2020). Current extensions for this class of models are widely applicable in different areas of conservation and management (Guillera-Arroita, 2017) and thus, it is important to assess whether the model assumptions are consistent with the ecological data and inference objectives (Devarajan et al., 2020). However, occupancy model assessment is difficult because of the two model components for detection and occurrence, both of which have assumptions that should be checked (Warton et al., 2017). Hence, occupancy model validation and checking are still areas under development that have been receiving attention in recent years (Broms et al., 2016; Wright et al., 2019). An adequate model assessment will allow for potential inadequacies regarding the model's underlying assumptions to be identified (Wright et al., 2019). For instance, all the proposed models in this research have assumed that observations across sites are independent and that the observed variation in species occurrences is determined entirely by the environmental condition at each site. However, incorporating several environmental predictors into a single hierarchical model can be difficult, especially for very complex models like the flexible occupancy model proposed in chapter 5. Thus, the most relevant variables for predicting species distribution can be selected based on the expertise of specialists on the target species ecology. Alternatively, the two-step modelling framework proposed in chapter 3 can also be used to find out the most relevant covariates affecting species distributions (a new approach for variable selection will be discussed in the next section). Regardless of the approach, a proper assessment of whether the independence assumption is satisfied should be made. If the species distributions is determined by a set of environmental drivers, then a correct model specification using a meaningful set of predictors should produce residuals that display a minimal spatial autocorrelation (Warton et al., 2017; Wright et al., 2019). The residual analysis discussed in chapter 5 did not show any evidence of strong residual autocorrelation for any of the two model components. However, it is important to notice that several spatially explicit models have been developed to correct for unaccounted spatial correlation by either introducing conditionally autoregressive structures (CAR) on site-level random effects (see Johnson et al. (2013) and Wright et al. (2019) for more details) or by including a spatial smoothing function for the coordinates of each site (Rushing et al., 2019). Violation of this assumption can overestimate the true species richness, occupancy and detection (Devarajan et al., 2020).

Occupancy models have also been developed under the assumption that target species arise from closed populations where there are no colonization or extinction process occurring during the span of the survey. In chapters 4 and 5 a dynamic occupancy model was developed to relax this assumption by specifying within-season (years) survey replicates. Kéry et al. (2010) showed this to be a robust method when applied to citizen science data that enables population trends to be estimated while providing enough information about species detection probability. However, as pointed out in chapter 5, determining the primary sampling periods were the occupancy state is allowed to change is not an easy task as it depends on the species ecology.

In this work, it has also been assumed that species occurrences are independent of each other, and that detection bias is only defined by a species false absence and not by the false-positive errors induced by individuals misidentification (i.e. $\Pr(y_{ij\cdot} = 1|z_{ij} = 0) = 0$). While there is little literature on methods addressing these assumptions, there are a few studies that are worthwhile mentioning. For instance, Rota et al. (2016) proposed to model the latent occupancy state using a multivariate Bernoulli random variable to relax the assumption of independence between species which allows describing situations in which species interact with each other. This could be of potential interest for modelling the interaction between invasive and native species where the latter are usually displaced as a result of inter-specific competition (MacKenzie et al., 2004). However, a major drawback of this approach is that the number of parameters for estimating the probability of all possible presence/absence combinations grows exponentially as the number of species increases, limiting the number of species that can be modeled by this approach (authors suggests a maximum of 10 species).

On the other hand, to account for false-positive errors, Royle and Link (2006) proposed using a finite mixture approach to model the detection miss-classification probabilities, i.e. $p_{ml} = \Pr(y_{jk} = m|z_j = l)$ for sites $j = 1,\ldots,M$ where a species is detected ($m = 1$) or not ($m = 0$) on each k visit given the occupancy state of whether the site is being truly occupied ($l = 1$) or not ($l = 0$). By modelling the number of times a species was detected at each site ($y_{j\cdot}$) the likelihood becomes:

$$L(\mathbf{p}, \psi, |\mathbf{y}) \propto \prod^{M} \left\{ \psi \left[p_{11}^{y_{j\cdot}} (1 - p_{11})^{K - y_{j\cdot}} \right] + (1 - \psi) \left[p_{10}^{y_{j\cdot}} (1 - p_{10})^{K - y_{j\cdot}} \right] \right\}. \tag{6.3}$$

While misidentification is less frequent than false absences and might be only relevant in certain scenarios (e.g. birds identification by sound), accounting for false-positive errors should receive more attention. Specifically for studies of opportunistic data records since volunteer observers might have varying recording skill levels making species misidentification more likely to occur (Rota et al., 2016), yet there are still no studies addressing how false-positive error rates affect the quality of the estimators in opportunistic data records.

Given all the different assumptions that occupancy models have, producing a unique diagnostic tool to assess if the model structure is appropriate for the data is difficult. This has been an area of active research, yet a common background on how occupancy model goodness-of-fit should be assessed has not been established (Broms et al., 2016). In consequence, different tools for diagnosing occupancy models fit have been proposed, some of which will be discussed in the following section.

6.2.3.3 Assessing the model goodness-of-fit

The general idea is to produce an overall metric to assess how well the models fit the data (Guillera-Arroita, 2017). A popular choice for assessing occupancy models goodness of fit is to compute a Bayesian p-value to summarize the posterior predictive probability (see discussion in chapter 5). The Bayesian p-value is calculated by comparing the proportion of posterior draws where a fit statistic calculated with simulated data from the fitted model, is larger than that same statistic but computed with the observed data instead (Gelman et al., 1996). Thus a value close to 0.5 indicates that observed data are consistent with the model predictions. The fit statistic is usually a discrepancy measure between the observed and expected values of the model. In occupancy models specifically, Freeman-Tukey statistic ($\sqrt{\mathbf{y}} - \sqrt{\mathbb{E}(\mathbf{y})^2}$) and, Pearson and Deviance residuals have often been used as discrepancy measures between the observed (and simulated) and predicted detections (Kéry and Royle, 2015; Broms et al., 2016; Rushing et al., 2019). However, Bayesian p-values calculated from these metrics are only informative about the observational process. Thus, in order to provide a separate diagnosis for each model component, Bayesian p-values can be computed based on the model's residuals definition introduced in chapter 5 by following Wright et al. (2019) or by using the "Dunn-Smyth" residuals proposed by Warton et al. (2017), (see discussion in chapter 5 about potential disadvantages that the latter formulation can have when assessing spatial autocorrelation).

Regardless of the discrepancy measure, the decision of whether a Bayesian p-value close to 0 or 1 indicates a lack of fit is rather subjective, and varies between authors (Kéry and Royle, 2015). Furthermore, computing Bayesian p-values is a computer-intensive process that does not provide information about the potential causes of lack of fit (Warton et al., 2017). For example, computational times for calculating Bayesian p-values for the simulation study in chapter 3 were 8 times slower despite the relatively small number of simulated species and sites. These computational constraints prevented the calculation of such values for the Odonata occupancy models developed in this work. Outhwaite et al. (2020) reported computational limitations for similar Odonata occurrence data in which posterior predictive checks could not be obtained.

Posterior predictive checks and Bayesian p-values can be recommended as a general approach to detect possible inadequacies in either of the two occupancy model components. However, these metrics should be presented along with other diagnostic tools that target specific model assumptions such as the correlograms in chapter 5 for addressing spatial correlation or binned residual plots to identify potentially important predictors that have been overlooked (Warton et al., 2017). These goodness-of-fit metrics can also be potentially used in simulation studies to validate different occupancy model structures and extensions. Hence, the ability of these metrics to detect violations of occupancy model assumptions under different scenarios needs further testing and is an undergoing area of research.

Assessing goodness-of-fit is related to model selection since selecting the best model from a set of candidate models is based on the relative comparison of how well different models fit the data (Warton et al., 2017). This is a very interesting and relevant area of research that has not been fully explored within the context of occupancy. Specifically for the Odonata case study analyzed in this research because of the

large number of potential predictors from which different candidates model can be generated. Selecting the “best" model among all those candidates is not by any means an easy task. Thus, a new approach for addressing this will be discussed in the next section.

Occupancy models are a powerful tool for estimating species occupancy while accounting for detection uncertainty. However, the assumptions and limitations that the single-season occupancy model has (MacKenzie et al., 2002) can be propagated to the different model extensions such as multispecies and dynamic models. If such model assumptions are not met, occupancy estimates will be biased leading to inaccurate interpretation of the true species distributions (Miller et al., 2015). Thus, a critical assessment of these assumptions is very important to identify the potential inadequacies when studying biological communities. These modelling inadequacies should not only be considered during the analysis stage but also during data collection and processing in order to reduce as much as possible the bias introduced when the model assumptions are not met. Contrasting existing approaches to properly assess modelling assumptions and testing them under different circumstances to detect violations of such assumptions is a promising area of research within species distribution modelling.

6.3 Future work and final comments

Over the last decade, the increasing awareness of how important accounting for the observational error in species distribution models is, has led to the development of different occupancy models (Guillera-Arroita, 2017). Such models have increasingly been used to monitor the species distribution and abundance in large-scale studies. This has been possible due to new information systems and collecting data schemes (such as citizen science data) that compile spatio-temporal data on a large collection of species and environmental predictors (Dennis et al., 2017).

Technological and methodological advances in monitoring wildlife populations have allowed researchers not only to answer but to formulate new increasingly complex questions about the role that imperfect detection plays in the study of species distributions (Devarajan et al., 2020). Throughout this research, several areas of future research within the context of species distribution modelling have been identified. Such areas reflect some of the challenges found among the different methods presented in this work. Thus, the following section addresses some of these new research questions that have been derived from the outcomes of this research.

6.3.1 Assessing species rarity across multiple spatial scales

The methods proposed in chapter 3 to identify rare species can be developed within a multiscale occupancy modelling framework (Nichols et al., 2008) to estimate rare species distributions across different nested spatial scales. This is equivalent to analyzing hierarchical split-plot designs where presence/absence occurrences are recorded in different spatial sub-units nested within a larger primary sampling unit (Kery et al., 2017). For instance, Aing et al. (2011) proposed introducing a unit-specific random effect to account for the dependency among sub-units within the same unit. In this approach, the occupancy state at those i primary units is given by $z_i \sim \text{Bernoulli}(\psi_i)$, where ψ_i is the larger-scale occupancy probability. Then, the j-th subunit presence/absence hierarchy is given by $a_{ij}|z_i \sim \text{Bernoulli}(z_i \times \theta_{ij})$, such that θ_{ij} is the subunit-level occupancy probability which is conditioned on the ith unit being occupied. Finally, the observed presence/absence occurrences across k repeated visits is modeled as $y_{ijk}|a_{ij} \sim \text{Bernoulli}(a_{ij} \times p_{ijk})$, where p_{ijk} is the occupancy probability. This approach can be integrated with the proposed mixture model to describe the rare species composition at different scales, assuming data are collected at each additional scale and that nested replicates are available at each level (e.g. replicates spatial sub-units for each site-unit). Moreover, the mixture occupancy models developed in chapter 3 assume occupancy and detection probabilities are constant across sites, however, Aing et al. (2011) multi-scale modelling framework require these probabilites to be defined in terms of site-level covariates or random effects to make the model parameters ψ_i and θ_{ij} identifiable (these parameters would be otherwise confounded as with the occupancy and detection parameters discussed in chapter 2 when replicated surveys are missing Eqn. 2.6).

6.3.2 Model selection and multicollinearity

With the growing concern over the effects that environmental changes have on species distributions and community composition, monitoring programs have now been collecting information on a large number of environmental predictors (Elith and Leathwick, 2009). A primary goal in Ecology is to understand how these site-specific covariates relate to species distribution by choosing the model that better fits the data from a collection of candidates models. However, very few studies have addressed modelling selection and contrasted different selection criterion for Bayesian occupancy models (Broms et al., 2016; Tenan et al., 2014). Models or variable selection in hierarchical Bayesian modelling has proven to be a challenging task that often involves computationally intensive methods that operate under certain regularity conditions that are not always met (Hobbs and Hooten, 2015; Tenan et al., 2014). A study by Warton et al. (2017) reported that from a total of 28 papers applying some class of occupancy model, more than half used AIC or AICc as a selection criteria. Unfortunately, these criteria are not ideal for Bayesian hierarchical occupancy models because the penalty function relies on the number of “free" parameters some of which will be constrained due to the hierarchical structure. In occupancy models, the latent state variables are constrained by both the prior and the likelihood (Hobbs and Hooten, 2015). Similarly, DIC has proven not to be appropriate for mixture models because the penalty term cannot be calculated

when the posterior mean is calculated for categorical parameters which unequivocally is what the latent state variables in occupancy models are (see Plummer (2008) and Lunn et al. (2012) for more details). Recall that the marginal distribution of Occupancy models (Eqn .2.9) can be expressed as a zero-inflated binomial distribution which is a mixture of a binomial distribution and a point-mass and therefore DIC is not suitable for comparing these models (Hobbs and Hooten, 2015).

WAIC is another information criterion that has proven to be suitable for Bayesian hierarchical models because its fully Bayesian implementation is based on the posterior distribution rather than the point-wise posterior mean for calculating the effective number of parameters (Watanabe and Opper, 2010; Gelman et al., 2014). Unfortunately, WAIC relies on the assumption that observations are conditionally independent, which can be an issue for spatio-temporal models. Furthermore, the practical implementation of WAIC requires all parameters directly involved in the likelihood (i.e. all the stochastic nodes in the model) to be monitored (de Valpine et al., 2017). This includes all the latent state variables and coefficients that, as highlighted in chapter 5 for the flexible model, cannot always be monitored all at once due to computational limitations.

An alternative approach discussed in chapter 2 is model averaging by the inclusion of an indicator variable. Instead of fitting separate models, a single large model with all possible predictors is fitted including the indicator latent variables that select which of the possible sets of covariates should be included in the model (Tenan et al., 2014). Thus, the posterior probability that a covariate is selected can be calculated as the posterior mean of the indicator variable. This method allows averaged predictions to be produced for latent variables since each variable represents a posterior sample obtained from embedding all possible models in one large model according to the values that each indicator variable takes (MacKenzie et al., 2017).

However, convergence issues for this approach are not uncommon when too vague priors for the coefficients are specified (Hobbs and Hooten, 2015) because when the indicator variable takes a value of zero, the corresponding coefficient will be sampled from its prior resulting in indicator values of 1 less likely to occur in later samples because the coefficient will be far from the posterior mass. Additionally, this method can be a computationally inefficient process because all the coefficients will be sampled even if the corresponding indicator is 0 (de Valpine et al., 2017). Thus, different extensions to this approach have been proposed such as Gibbs Variable Selection (Dellaportas et al., 2000) and Stochastic Search Variable Selection (George and Foster, 2000) that use a joint prior for indicators and coefficients, or Reversible Jump MCMC that only sample those coefficients for which indicator variable is $\neq 0$ (Green, 1995) but none have been explored within the context of occupancy model and thus, represent an interesting area for future research that can be used to determine the effect that the connectivity and stressors (described in chapter 1) have on species Odonata distribution. However, this represents another challenge because of site-specific covariates that have been recorded in different spatial scales exhibiting high levels of correltion between them. Thus, a model selection approach that accounts for correlation among predictors needs to be considered.

Curtis and Ghosh (2011) proposed a Bayesian Approach that deals with multicollinearity while simul-

taneous selecting and clustering predictors in Linear Regression. Their approach is based on specifying a Dirichlet process prior in conjunction with a variable selection prior to identify and remove redundant predictors. This method has shown competing results to those obtained with standard penalized least squares methods such as ridge regression and The Least Absolute Shrinkage and Selection Operator (LASSO).

For a set of p centered and scaled predictors $x_{i1},\ldots,x_{ip}$ and coefficients $\beta_1,\ldots,\beta_p$, authors proposed a prior for regression coefficients β_j to be defined by a set of random draws $\theta_1,\ldots\theta_p$ from a Dirichlet process random distribution $\mathscr{P}(\cdot)$, and random draws $\gamma_1,\ldots,\gamma_p$ from a Bernoulli distribution. Authors proposed a normal distribution as a base distribution for the Dirichlet process, i.e. $\mathscr{P} \sim \text{DP}(\alpha,\Phi_{0,\eta^2}(\cdot))$. This is equivalent to setting a prior for the regression coefficients while allowing clustering of the predictors (see Curtis and Ghosh (2011) for more details). Similar to the mixture model developed in chapter 3, the Dirichlet process can be approximated by drawing $\zeta_1,\ldots,\zeta_H$ random variables from a Multinomial distribution with probabilities $\pi_1,\ldots,\pi_H$ (drawn from a Dirichlet distribution with parameters $\alpha/H,\ldots$) for a suitable large and fixed number of H classes.

Then, the prior on regression coefficients is defined as $\beta_j = \gamma_j\xi_{\zeta_j}$, such that γ_j and ξ_{ζ_j} are drawn from Bernoulli(δ) and a Normal $(0,\eta^2)$ distributions respectively. Thus, hierarchical Bayesian linear model under this parametrization can be written as:

$$
\begin{aligned}
Y_i\boldsymbol{\beta},\sigma^2 &\sim \text{Normal}(\mathbf{x}_i^T\boldsymbol{\beta},\sigma^2) \quad i=1,\ldots,n\\
\beta_j &= \gamma_j\xi_{\zeta_j} \quad j=1,\ldots,p\\
\gamma_j &\sim \text{Bernoulli}(\delta) \quad j=1,\ldots,p\\
\zeta_j &\sim \text{Multinomial}(\pi_1,\ldots,\pi_H) \quad j=1,\ldots,p\\
\xi_j|\eta^2 &\sim \text{Normal}(0,\eta^2) \quad j=1,\ldots,H\\
(\pi_1,\ldots,\pi_H) &\sim \text{Dirichlet}(\alpha/H,\ldots,\alpha/H)\\
\delta &\sim \text{Uniform}(0,1)\\
\sigma^2 &\sim \text{InvGamma}(a_\gamma,b_\gamma)\\
\eta^2 &\sim \text{InvGamma}(a_\eta,b_\eta).
\end{aligned}
\tag{6.4}
$$

Thus, future work aims to explore this method within occupancy models and to test different prior parametrizations and indicator-variable selection methods. To do so, the effect that multicollinearity has on the state and observational processes of a standard occupancy model and the degree to which the Curtis and Ghosh (2011) method improves the estimates' precision are going to be assessed.

6.3.3 INLA-based approach

It is important to notice that the large number of parameters involved with these methods makes the estimation process via traditional MCMC methods difficult and computationally expensive. Bayesian inference approaches using MCMC methods approximates the posterior distribution by constructing a Markov chains that converges to the target distribution after a reasonably large number of iterations. Unfortunately, the large number of iterations needed for the chains to converge usually results in a computationally demanding process (where a lot of samples are usually discarded as a result of the burnin or thinning processes). Thus, Rue et al. (2009) developed the integrated nested Laplace approximation (INLA) as an alternative computationaly efficient method for Bayesian inference.

The details for this method can be found in Blangiardo et al. (2013). In this section, only a brief description of INLA and its recent applications in species distribution modelling will be described to illustrate the potential this method has to provide a computationally efficient framework to address complex ecological questions similar to those investigated in chapters 4 and 5.

INLA is an approximation method that enables Bayesian inference to be made for latent Gaussian models (LGMs) exclusively. The general structure of a LGM is given by the likelihood, a latent field and the hyperparameters of the latent field, i.e.

$$\begin{aligned} \mathbf{y}|\mathbf{x},\theta_2 &\sim \prod_i p(y_i|\eta_i,\theta_2) \\ \mathbf{x}|\theta_1 &\sim p(\mathbf{x}|\theta_1) = \text{Normal}(0,\Sigma) \\ \theta = [\theta_1,\theta_2]' &\sim p(\theta), \end{aligned} \tag{6.5}$$

such that $\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m z_{im} + \sum_{l=1}^{L} f_l(w_{il})$. Here, $\mathbf{y}$ represents the observed data and $\mathbf{x}$ is the vector of latent effects $\mathbf{x} = [\boldsymbol{\eta}, \beta_0, \boldsymbol{\beta}, \mathbf{f}(\cdot)]$. Where $\boldsymbol{\beta}$ are the regression coefficients for a set of covariates $\mathbf{z}$ and $\mathbf{f}(\cdot)$ are a set of functions on some covariates $\mathbf{w}$ that can allocate different terms such as covariates nonlinear effects, time trends and seasonal effects, random effects (intercepts and slopes), and spatial or temporal random effects.

The structure in equation 6.5 can be generalized to non-Gaussian responses when the y_i's follow a distribution from an exponential family such that its conditional mean μ_i is defined as a function of the linear predictor η_i via a link function $g(\cdot)$, i.e. $g(\mu_i) = \eta_i$ (thus, the likelihood does not has to be normal even if the latent effects are).

Additionally, it is assumed that (1) each variable $\mathbf{y}$ is conditionally independent given the latent effects $\mathbf{x}$ and the hyperparameters θ_2, and (2) the random field defined by the latent structure $\mathbf{x}$ is a Gaussian Markov Random Field (GMRF)(achieved by placing Gaussian priors on each parameter) with zero mean and precision matrix $Q(\theta_1) = \Sigma^{-1}$ where $Q(\theta_1)$ is a semi-positive definite matrix that depends on the vector of hyperparameters θ_1 (Wang et al., 2018).

The latent GRMF structure of the model assumes conditional independence in $\mathbf{x}$ ($x_i \perp\!\!\!\perp x_j|\mathbf{x}_{-ij} \Leftrightarrow Q_{ij}(\theta_1) = 0$) yielding to a sparse precision Q that is easier to invert, improving computation efficiency. Then, by letting $\theta = (\theta_1, \theta_2)$ the joint posterior distribution of $\mathbf{x}$ and θ is:

$$\begin{aligned} p(\mathbf{x}, \theta|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)p(\theta) \\ &\propto \prod_i p(y_i|x_i, \theta) \times |Q(\theta)|^{1/2} \times exp\left\{-\frac{1}{2}x'Q(\theta)x\right\} p(\theta) \\ &= |Q(\theta)|^{1/2} \times \exp\left\{-\frac{1}{2}x'Q(\theta)x + \sum_i \log(p(y_i|x_i, \theta))\right\} p(\theta). \end{aligned} \tag{6.6}$$

Instead of estimating the joint posterior, INLA focuses on the individual posterior marginals of the elements of the latent field and hyperprior distributions:

- $p(x_j|\mathbf{y}) = \int p(x_j|\theta, \mathbf{y})p(\theta|\mathbf{y})d\theta$
- $p(\theta_k|\mathbf{y}) = \int p(\theta_k|\mathbf{y})d\boldsymbol{\theta_{-k}}$.

To compute such marginals, the conditionals $p(\theta|\mathbf{y})$ and $p(x_j|\theta, \mathbf{y})$ are approximated using Laplace approximation (see Rue et al. (2009) for details):

$$\tilde{p}(\theta|\mathbf{y}) \propto \left.\frac{\theta, \mathbf{x}|\mathbf{y}}{\tilde{p}(\mathbf{x}|\theta, \mathbf{y})}\right|_{\mathbf{x}=\mathbf{x}^*(\theta)}, \tag{6.7}$$

where $\tilde{p}(\mathbf{x}|\theta, \mathbf{y})$ is a Gaussian approximation to the full conditional of the latent effects with $\mathbf{x}^*(\theta)$ being the mode for a given θ. Then, the marginal posterior $p(x_j|\theta, \mathbf{y})$ can be approximated by integrating the hyperparameters out and marginalizing over the latent effects. Three different approximations to do so have been proposed by Rue et al. (2009). The first one uses the normal approximation in (6.7) which is computationally efficient but according to Blangiardo et al. (2013) can have poor results. The second one, which is computationally more expensive, uses the Laplace approximation again while partitioning $\mathbf{x} = [x_j, \mathbf{x}_{-\mathbf{j}}]$ resulting in:

$$p_{LA}(x_j|\theta, \mathbf{y}) \propto \left.\frac{p(\mathbf{x}, \theta|\mathbf{y})}{\tilde{p}(\mathbf{x}_{-\mathbf{j}}|x_j, \theta, \mathbf{y})}\right|_{\mathbf{x}_{-\mathbf{j}}=\mathbf{x}_{-\mathbf{j}}{}^*(x_i,\theta)}, \tag{6.8}$$

where $\tilde{p}(\mathbf{x}_{-\mathbf{j}}|x_j, \theta, \mathbf{y})$ is the Laplace normal approximation with mode $\mathbf{x}_{-\mathbf{j}}{}^*(x_i, \theta)$. This is computationally intensive because $\tilde{p}(\mathbf{x}_{-\mathbf{j}}|x_j, \theta, \mathbf{y})$ must be computed for each x_i. Thus, the third option denoted *simplified Laplace approximation*, provides an efficient yet reliable approximation to $p(x_j|\theta, \mathbf{y})$ by using Taylor series expansion up to the third order on the numerator and denominator of equation 6.8 that corrects the normal approximation for location and skewness by including a cubic spline term (see Rue et al. (2009) for details). Then, the marginal posterior distributions are approximated by $\tilde{p}(x_j|\mathbf{y}) \approx \int \tilde{p}(x_j|\theta, \mathbf{y})\tilde{p}(\theta|\mathbf{y})d\theta$.

The estimation procedures takes several steps that are not going to be included here, but involves Newton-methods to find the mode of $\tilde{p}(\theta|\mathbf{y})$ and then a grid-search strategy (although other approaches such as central composite design are used to reduce the computational costs) to find suitable points for the numerical integration.

The `INLA` library in R has been developed to approximate Bayesian inference for Latent Gaussian Models using INLA (Rue et al., 2009). This approach has recently been explored to address species distribution modelling under imperfect detection. For instance, Meehan et al. (2017) analyzed a N-mixture occupancy model using R-INLA and compared its implementation with commonly used JAGS and likelihood-based method.

The N-mixture occupancy model is a class of abundance-based occupancy models that describes a mixture between the latent abundance state at site $N_i \sim \text{Pois}(\lambda_i)$ with the expected abundance λ_i commonly modeled as a log-linear function of site-level covariates (e.g. $\log(\lambda_i) = \beta_0 + \beta_1 x_i$) , and binomial distribution describing the observed individual species counts across repeated surveys, i.e. $y_{ij}|N_i \sim \text{Binomial}(N_i, p_{ij})$, where detection probabilities are defined by a set of site-survey covariates (e.g. $\text{logit}(p_{ij} = \alpha_0 + \alpha_1 g_{ij})$. Meehan et al. (2017) showed how parameter estimates for both latent and observational processes obtained with R-INLA were practically identical with those obtained with traditional likelihood-based methods or JAGS, but proved to be computationally more efficient (10 times faster than traditional likelihood estimation, and 500 faster than JAGS implementation). Unfortunately, some issues were raised by the authors. For instance, currently, INLA does not support site-survey level covariates to be specified and therefore, such variables need to be averaged in order for INLA to estimate site-varying detection probabilities. This may raise an issue when site-survey covariates are important for the data generating process or when a covariates is used to account for uneven sampling effort such as date of the visit. Additionally, random effects can only be specified in the observational component and thus, spatial effects or flexible terms cannot yet be developed for the latent process (which is the main point of interest in species distribution modelling). Current developments of R-INLA employs only Poisson-binomial and negative binomial-binomial mixtures, and therefore occupancy-based models like the ones developed in this research cannot be fitted yet. However, Bachl et al. (2019) recently developed the `inlabru` package that uses a stochastic partial differential approach to INLA (Lindgren et al., 2011) that approximates the Gaussian random fields to account for spatial dependence in ecological data. This approach enables point data and count data to be modeled as a spatial Poisson process and has recently been applied by Martino et al. (2021) to study the spatial distribution of dolphins while estimating detection probabilities in a distance sampling scheme. Bayesian spatial modelling developments using the INLA approach are still in the early stages and extensions are currently being developed (Bachl et al., 2019). Thus, the variety of species distribution models that can be fitted is somewhat restricted but future research will certainly enlarge their flexibility to add more complex structures.

Bibliography

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., and Anderson, S. E. (2017). The accuracy of citizen science data: a quantitative review. *Bulletin of the Ecological Society of America*, 98(4):278–290.

Adrian, R., O'Reilly, C. M., Zagarese, H., Baines, S. B., Hessen, D. O., Keller, W., Livingstone, D. M., Sommaruga, R., Straile, D., Van Donk, E., et al. (2009). Lakes as sentinels of climate change. *Limnology and oceanography*, 54(6part2):2283–2297.

Aing, C., Halls, S., Oken, K., Dobrow, R., and Fieberg, J. (2011). A bayesian hierarchical occupancy model for track surveys conducted in a series of linear, spatially correlated, sites. *Journal of Applied Ecology*, 48(6):1508–1517.

Altwegg, R. and Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, 10(1):8–21.

Amoros, C. and Bornette, G. (2002). Connectivity and biocomplexity in waterbodies of riverine floodplains. *Freshwater biology*, 47(4):761–776.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an r package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.

Bailey, L. L., MacKenzie, D. I., and Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, 5(12):1269–1279.

Banner, K. M., Irvine, K. M., Rodhouse, T. J., Donner, D., and Litt, A. R. (2019). Statistical power of dynamic occupancy models to identify temporal change: Informing the north american bat monitoring program. *Ecological Indicators*, 105:166–176.

Begon, M., Harper, J. L., Townsend, C. R., et al. (1986). *Ecology. Individuals, populations and communities.* Blackwell scientific publications.

Blackburn, T. M. and Gaston, K. J. (1997). Who is rare? artefacts and complexities of rarity determination. In *The biology of rarity*, pages 48–60. Springer.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with r-inla. *Spatial and spatio-temporal epidemiology*, 4:33–49.

Bled, F., Nichols, J. D., and Altwegg, R. (2013). Dynamic occupancy models for analyzing species' range dynamics across large geographic scales. *Ecology and evolution*, 3(15):4896–4909.

Bomphrey, R. J., Nakata, T., Henningsson, P., and Lin, H.-T. (2016). Flight of the dragonflies and damselflies. *Phil. Trans. R. Soc. B*, 371(1704):20150389.

Braune, E., Richter, O., Söndgerath, D., and Suhling, F. (2008). Voltinism flexibility of a riverine dragonfly along thermal gradients. *Global Change Biology*, 14(3):470–482.

Broennimann, O., Vittoz, P., Moser, D., and Guisan, A. (2005). Rarity types among plant species with high conservation priority in switzerland. *Botanica Helvetica*, 115(2):95–108.

Broms, K. M., Hooten, M. B., and Fitzpatrick, R. M. (2016). Model selection and assessment for multi-species occupancy models. *Ecology*, 97(7):1759–1770.

Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Chandler, R. B., Muths, E., Sigafus, B. H., Schwalbe, C. R., Jarchow, C. J., and Hossack, B. R. (2015). Spatial occupancy models for predicting metapopulation dynamics and viability following reintroduction. *Journal of Applied Ecology*, 52(5):1325–1333.

Chapman, D. S., Gunn, I. D., Pringle, H. E., Siriwardena, G. M., Taylor, P., Thackeray, S. J., Willby, N. J., and Carvalho, L. (2020). Invasion of freshwater ecosystems is promoted by network connectivity to hotspots of human activity. *Global Ecology and Biogeography*, 29(4):645–655.

Clark, A. E. and Altwegg, R. (2019). Efficient bayesian analysis of occupancy models with logit link functions. *Ecology and evolution*, 9(2):756–768.

Collier, B. A., Groce, J. E., Morrison, M. L., Newnam, J. C., Campomizzi, A. J., Farrell, S. L., Mathewson, H. A., Snelgrove, R. T., Carroll, R. J., and Wilkins, R. N. (2012). Predicting patch occupancy in fragmented landscapes at the rangewide scale for an endangered species: an example of an american warbler. *Diversity and Distributions*, 18(2):158–167.

Collins, S. D. and McIntyre, N. E. (2015). Modeling the distribution of odonates: a review. *Freshwater Science*, 34(3):1144–1158.

Contreras-Balderas, S. (1984). Environmental impacts in cuatro ciénegas, coahuila, méxico: a commentary. *Journal of the Arizona-Nevada Academy of Science*, pages 85–88.

Corbet, P. and Brooks, S. (2011). *Dragonflies (Collins New Naturalist Library, Book 106)*, volume 106. HarperCollins UK.

Corbet, P. S., Suhling, F., and Soendgerath, D. (2006). Voltinism of odonata: a review. *International Journal of Odonatology*, 9(1):1–44.

Crainiceanu, C., Ruppert, D., and Wand, M. (2005). Bayesian analysis for penalized spline regression using winbugs. *Journal of Statistical Software, Articles*, 14(14):1–24.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3):553–570.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Crooks, K. R. and Sanjayan, M. (2006). Connectivity conservation: maintaining connections for nature. In Crooks, K. R. and Sanjayan, M., editors, *Conservation Biology*, pages 1–20. Cambridge University Press, Cambridge.

Curtis, S. M. and Ghosh, S. K. (2011). A bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression. *Journal of statistical theory and practice*, 5(4):715–735.

De Block, M., McPeek, M. A., and Stoks, R. (2008). Life-history evolution when lestes damselflies invaded vernal ponds. *Evolution: International Journal of Organic Evolution*, 62(2):485–493.

De Chazal, J. and Rounsevell, M. D. (2009). Land-use and climate change within assessments of biodiversity change: a review. *Global Environmental Change*, 19(2):306–315.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). Bayesian variable selection using the gibbs sampler. *BIOSTATISTICS-BASEL-*, 5:273–286.

Denes, F. V., Silveira, L. F., and Beissinger, S. R. (2015). Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation. *Methods in Ecology and Evolution*, 6(5):543–556.

Dennis, E. B., Morgan, B. J., Freeman, S. N., Ridout, M. S., Brereton, T. M., Fox, R., Powney, G. D., and Roy, D. B. (2017). Efficient occupancy model-fitting for extensive citizen-science data. *PloS one*, 12(3):e0174433.

Devarajan, K., Morelli, T. L., and Tenan, S. (2020). Multi-species occupancy models: review, roadmap, and recommendations. *Ecography*.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, 23(12):1472–1484.

Dorazio, R. M. (2016). Bayesian data analysis in population ecology: motivations, methods, and benefits. *Population ecology*, 58(1):31–44.

Dorazio, R. M., Gotelli, N. J., and Ellison, A. M. (2011). Modern methods of estimating biodiversity from presence-absence surveys. In *Biodiversity loss in a changing planet*. InTech.

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

Dorazio, R. M., Royle, J. A., Söderström, B., and Glimskär, A. (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, 87(4):842–854.

Dunstan, P. K., Foster, S. D., and Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, 222(4):955–963.

Eaton, M. J., Hughes, P. T., Hines, J. E., and Nichols, J. D. (2014). Testing metapopulation concepts: effects of patch characteristics and neighborhood occupancy on the dynamics of an endangered lagomorph. *Oikos*, 123(6):662–676.

Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. *Methods in ecology and evolution*, 1(4):330–342.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697.

Ellis, M. M., Ivan, J. S., Tucker, J. M., and Schwartz, M. K. (2015). rspace: Spatially based power analysis for conservation and ecology. *Methods in Ecology and Evolution*, 6(5):621–625.

Ellstrand, N. C. and Elam, D. R. (1993). Population genetic consequences of small population size: implications for plant conservation. *Annual review of Ecology and Systematics*, 24(1):217–242.

Fagan, W. F., Aumann, C., Kennedy, C. M., and Unmack, P. J. (2005). Rarity, fragmentation, and the scale dependence of extinction risk in desert fishes. *Ecology*, 86(1):34–41.

Fattorini, S. (2009). Assessing priority areas by imperilled species: insights from the european butterflies. *Animal conservation*, 12(4):313–320.

Fergus, C. E., Lapierre, J.-F., Oliver, S. K., Skaff, N. K., Cheruvelil, K. S., Webster, K., Scott, C., and Soranno, P. (2017). The freshwater landscape: lake, wetland, and stream abundance and connectivity at macroscales. *Ecosphere*, 8(8):e01911.

Fernández-i Marín, X. (2016). ggmcmc: Analysis of MCMC samples and Bayesian inference. *Journal of Statistical Software*, 70(9):1–20.

Fiske, I. and Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, 43(10):1–23.

Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, 7(4):1917.

Flather, C. H. and Sieg, C. H. (2007). Species rarity: definition, causes and classification. *Conservation of rare or little-known species: Biological, social, and economic considerations*, pages 40–66.

Freeman, M. C., Pringle, C. M., Greathouse, E. A., and Freeman, B. J. (2003). Ecosystem-level consequences of migratory faunal depletion caused by dams. In *American Fisheries Society Symposium*, volume 35, pages 255–266.

Geary, D. (1989). Mixture models: Inference and applications to clustering. *Journal of the Royal Statistical Society Series A*, 152(1):126–127.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4):1360–1383.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

George, E. and Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747.

Golfieri, B., Hardersen, S., Maiolini, B., and Surian, N. (2016). Odonates as indicators of the ecological integrity of the river corridor: Development and application of the odonate river index (ori) in northern italy. *Ecological indicators*, 61:234–247.

Green, A. W., Pavlacky Jr, D. C., and George, T. L. (2019). A dynamic multi-scale occupancy model to estimate temporal dynamics and hierarchical habitat use for nomadic species. *Ecology and evolution*, 9(2):793–803.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, 40(2):281–295.

Guillera-Arroita, G. and Lahoz-Monfort, J. J. (2012). Designing studies to detect differences in species occupancy: power analysis under imperfect detection. *Methods in Ecology and Evolution*, 3(5):860–869.

Guillera-Arroita, G., Morgan, B. J., Ridout, M. S., and Linkie, M. (2011). Species occupancy modeling for detection data collected along a transect. *Journal of agricultural, biological, and environmental statistics*, 16(3):301–317.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009.

Hadley, W. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039.

Hartley, S. and Kunin, W. E. (2003). Scale dependency of rarity, extinction risk, and conservation priority. *Conservation biology*, 17(6):1559–1570.

Hassall, C. and Thompson, D. J. (2008). The effects of environmental warming on odonata: a review. *International Journal of Odonatology*, 11(2):131–153.

Hassall, C., Thompson, D. J., and Harvey, I. F. (2008). Wings of coenagrion puella vary in shape at the northern range margin (odonata: Coenagrionidae). *International Journal of Odonatology*, 11(1):35–41.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., and Hooten, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3):632–646.

Hefley, T. J. and Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1(2):87–97.

Hines, J. (2006). Presence: Software to estimate patch occupancy and related parameters, version 11.5.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press.

Hobbs, N. T. and Hooten, M. B. (2015). *Bayesian models: a statistical primer for ecologists*. Princeton University Press.

Holmes, E. E., Lewis, M. A., Banks, J., and Veit, R. (1994). Partial differential equations in ecology: spatial interactions and population dynamics. *Ecology*, 75(1):17–29.

Isaac, J. L., Vanderwal, J., Johnson, C. N., and Williams, S. E. (2009). Resistance and resilience: quantifying relative extinction risk in a diverse assemblage of australian tropical rainforest vertebrates. *Diversity and Distributions*, 15(2):280–288.

Isaac, N. J., van Strien, A. J., August, T. A., de Zeeuw, M. P., and Roy, D. B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5(10):1052–1060.

Jeltsch, F., Bonte, D., Pe'er, G., Reineking, B., Leimgruber, P., Balkenhol, N., Schröder, B., Buchmann, C. M., Mueller, T., Blaum, N., et al. (2013). Integrating movement ecology with biodiversity research-exploring new avenues to address spatiotemporal biodiversity dynamics. *Movement Ecology*, 1(1):6.

Johansson, F. (2003). Latitudinal shifts in body size of enallagma cyathigerum (odonata). *Journal of Biogeography*, 30(1):29–34.

Johnson, D. S. (2021). *stocc: Fit a Spatial Occupancy Model via Gibbs Sampling*. R package version 1.31.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., and Pond, B. A. (2013). Spatial occupancy models for large data sets. *Ecology*, 94(4):801–808.

Karron, J. D. (1997). Genetic consequences of different patterns of distribution and abundance. In *The biology of rarity*, pages 174–189. Springer.

Kellner, K. (2017). *jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses*. R package version 1.4.9.

Kellner, K. (2021). *ubms: Bayesian Models for Data from Unmarked Animals using 'Stan'*. R package version 1.0.2.

Kellner, K. F. and Swihart, R. K. (2014). Accounting for imperfect detection in ecology: a quantitative review. *PLoS One*, 9(10):e111436.

Kéry, M., Guillera-Arroita, G., and Lahoz-Monfort, J. J. (2013). Analysing and mapping species range dynamics using occupancy models. *Journal of Biogeography*, 40(8):1463–1474.

Kery, M., Royle, A., and Meredith, M. (2017). *AHMbook: Functions and Data for the Book 'Applied Hierarchical Modeling in Ecology'*. R package version 0.1.4.

Kéry, M., Royle, J., Schmid, H., Schaub, M., Volet, B., Häfliger, G., and Zbinden, N. (2010). Correcting population trend estimates from opportunistic observations for observation effort using siteoccupancy modeling. *Conserv Biol*, 24:1388–1397.

Kéry, M. and Royle, J. A. (2008). Hierarchical bayes estimation of species richness and occupancy in spatially replicated surveys. *Journal of Applied Ecology*, 45(2):589–598.

Kéry, M. and Royle, J. A. (2009). Inference about species richness and community structure using species-specific occupancy models in the national swiss breeding bird survey mhb. In *Modeling demographic processes in marked populations*, pages 639–656. Springer.

Kéry, M. and Royle, J. A. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., and Stone, L. (2017). Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 8(4):420–430.

Krebs, C. J. (1972). The experimental analysis of distribution and abundance. *Ecology. New York: Harper and Row*.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., and Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research synthesis methods*, 10(1):83–98.

LaPoint, S., Balkenhol, N., Hale, J., Sadler, J., and Ree, R. (2015). Ecological connectivity research in urban areas. *Functional Ecology*, 29(7):868–878.

Lele, S. R., Moreno, M., and Bayne, E. (2012). Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology*, 5(1):22–31.

Leroy, B., Canard, A., and Ysnel, F. (2013). Integrating multiple scales in rarity assessments of invertebrate taxa. *Diversity and Distributions*, 19(7):794–803.

Leroy, B., Petillon, J., Gallon, R., Canard, A., and Ysnel, F. (2012). Improving occurrence-based rarity metrics in conservation studies by including multiple rarity cut-off points. *Insect Conservation and Diversity*, 5(2):159–168.

Li, Y., Lord-Bessen, J., Shiyko, M., and Loeb, R. (2018). Bayesian latent class analysis tutorial. *Multivariate behavioral research*, 53(3):430–451.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Liittschwager, D. (1994). *Witness: Endangered Species of North America*. Chronicle Books.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC press.

MacKenzie, D. and Hines, J. (2017). *RPresence: R Interface for Program PRESENCE*. R package version 2.12.10.

MacKenzie, D., Nichols, J. D., Hines, J. E., Knutson, M. G., and Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, 84(8):2200–2207.

MacKenzie, D., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

MacKenzie, D., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., and Hines, J. E. (2017). *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier.

MacKenzie, D. I., Bailey, L. L., and Nichols, J. D. (2004). Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology*, 73(3):546–555.

MacKenzie, D. I., Nichols, J. D., Sutton, N., Kawanishi, K., and Bailey, L. L. (2005). Improving inferences in population studies of rare species that are detected imperfectly. *Ecology*, 86(5):1101–1113.

MacKenzie, D. I. and Royle, J. A. (2005). Designing occupancy studies: general advice and allocating survey effort. *Journal of applied Ecology*, 42(6):1105–1114.

Makambi, K. H. (2004). The effect of the heterogeneity variance estimator on some tests of treatment efficacy. *Journal of biopharmaceutical statistics*, 14(2):439–449.

Martino, S., Pace, D. S., Moro, S., Casoli, E., Ventura, D., Frachea, A., Silvestri, M., Arcangeli, A., Giacomini, G., Ardizzone, G., et al. (2021). Integration of presence-only data from several sources. a case study on dolphins' spatial distribution. *arXiv preprint arXiv:2103.16125*.

McCarthy, M. A. and Masters, P. (2005). Profiting from prior information in bayesian analyses of ecological data. *Journal of Applied Ecology*, 42(6):1012–1019.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Meehan, T. D., Michel, N. L., and Rue, H. (2017). Estimating animal abundance with n-mixture models using the r-inla package for r. *arXiv preprint arXiv:1705.01581*.

Miller, D. A., Bailey, L. L., Grant, E. H. C., McClintock, B. T., Weir, L. A., and Simons, T. R. (2015). Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, 6(5):557–565.

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., and Hines, J. E. (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology*, 45(5):1321–1329.

Noon, B. R., Bailey, L. L., Sisk, T. D., and McKelvey, K. S. (2012). Efficient species-level monitoring at the landscape scale. *Conservation Biology*, 26(3):432–441.

Northrup, J. M. and Gerber, B. D. (2018). A comment on priors for bayesian occupancy models. *PloS one*, 13(2):e0192819.

Oesterwind, D., Rau, A., and Zaiko, A. (2016). Drivers and pressures–untangling the terms commonly used in marine science and policy. *Journal of Environmental Management*, 181:8–15.

Outhwaite, C. L., Chandler, R. E., Powney, G. D., Collen, B., Gregory, R. D., and Isaac, N. J. (2018). Prior specification in bayesian occupancy modelling improves analysis of species occurrence data. *Ecological Indicators*, 93:333–343.

Outhwaite, C. L., Gregory, R. D., Chandler, R. E., Collen, B., and Isaac, N. J. (2020). Complex long-term biodiversity change among invertebrates, bryophytes and lichens. *Nature ecology & evolution*, 4(3):384–392.

Peach, M. A., Cohen, J. B., and Frair, J. L. (2017). Single-visit dynamic occupancy models: an approach to account for imperfect detection with atlas data. *Journal of applied ecology*, 54(6):2033–2042.

Pedersen, E. J., Miller, D. L., Simpson, G. L., and Ross, N. (2019). Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ*, 7:e6876.

Pedersen, O., Andersen, T., Ikejima, K., Zakir Hossain, M., and Andersen, F. Ø. (2006). A multidisciplinary approach to understanding the recent and historical occurrence of the freshwater plant, littorella uniflora. *Freshwater Biology*, 51(5):865–877.

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics*, 9(3):523–539.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Pollock, K. H. (1982). A capture-recapture design robust to unequal probability of capture. *The Journal of Wildlife Management*, 46(3):752–757.

Pollock, L. J., Thuiller, W., and Jetz, W. (2017). Large conservation gains possible for global biodiversity facets. *Nature*, 546(7656):141–144.

Ponisio, L. C., de Valpine, P., Michaud, N., and Turek, D. (2020). One size does not fit all: Customizing mcmc methods for hierarchical models using nimble. *Ecology and evolution*, 10(5):2385–2416.

Powney, G. D., Brooks, S. J., Barwell, L. J., Bowles, P., Fitt, R. N. L., Pavitt, A., Spriggs, R. A., and Isaac, N. J. B. (2014). Morphological and geographical traits of the british odonata. *Biodiversity Data Journal*, 2:e1041.

Pringle, C. (1997). Exploring how disturbance is transmitted upstream: going against the flow. *Journal of the north american Benthological society*, 16(2):425–438.

Pringle, C. (2001). Hydrologic connectivity and the management of biological reserves: a global perspective. *Ecological Applications*, 11(4):981–998.

Pringle, C. (2003). The need for a more predictive understanding of hydrologic connectivity. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 13(6):467–471.

Pringle, C. (2006). Hydrologic connectivity: a neglected dimension of conservation biology. In Crooks, K. R. and Sanjayan, M., editors, *Conservation Biology*, pages 233–254. Cambridge University Press, Cambridge.

Pringle, C. and Triska, F. J. (2000). Emergent biological patterns and surface-subsurface interactions at landscape scales. In Jones, J. B. and Mulholland, P. J., editors, *Streams and Ground Waters*, Aquatic Ecology, pages 167 – 193. Academic Press, San Diego.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabinowitz, D. (1981). Seven forms of rarity. in 'the biological aspects of rare plant conservation'.(ed. h synge) pp. 205–217.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Ritz, C. and Streibig, J. C. (2008). *Nonlinear regression with R*. Springer Science & Business Media.

Rosenberg, D. M., McCully, P., and Pringle, C. M. (2000). Global-scale environmental effects of hydrological alterations: introduction. *BioScience*, 50(9):746–751.

Rota, C. T., Ferreira, M. A., Kays, R. W., Forrester, T. D., Kalies, E. L., McShea, W. J., Parsons, A. W., and Millspaugh, J. J. (2016). A multispecies occupancy model for two or more interacting species. *Methods in Ecology and Evolution*, 7(10):1164–1173.

Royle, J. A. (2006). Site occupancy models with heterogeneous detection probabilities. *Biometrics*, 62(1):97–102.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Elsevier.

Royle, J. A. and Kéry, M. (2007). A bayesian state-space formulation of dynamic occupancy models. *Ecology*, 88(7):1813–1823.

Royle, J. A. and Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841.

Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence–absence data or point counts. *Ecology*, 84(3):777–790.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Rushing, C. S., Royle, J. A., Ziolkowski, D. J., and Pardieck, K. L. (2019). Modeling spatially and temporally complex range dynamics when detection is imperfect. *Scientific reports*, 9(1):1–9.

Sadoti, G., Zuckerberg, B., Jarzyna, M. A., and Porter, W. F. (2013). Applying occupancy estimation and modelling to the analysis of atlas data. *Diversity and Distributions*, 19(7):804–814.

Schroeder, B. (2008). Challenges of species distribution modeling belowground. *Journal of Plant Nutrition and Soil Science*, 171(3):325–337.

Smallshire, D. and Swash, A. (2018). *Britain's Dragonflies: A Field Guide to the Damselflies and Dragonflies of Great Britain and Ireland-Fully Revised and Updated Fourth Edition*, volume 12. Princeton University Press.

Sollmann, R., Eaton, M. J., Link, W. A., Mulondo, P., Ayebare, S., Prinsloo, S., Plumptre, A. J., and Johnson, D. S. (2021). A bayesian dirichlet process community occupancy model to estimate community structure and species similarity. *Ecological Applications*, 31(2):e02249.

Souza, V., Espinosa-Asuar, L., Escalante, A. E., Eguiarte, L. E., Farmer, J., Forney, L., Lloret, L., Rodríguez-Martínez, J. M., Soberón, X., Dirzo, R., et al. (2006). An endangered oasis of aquatic microbial biodiversity in the chihuahuan desert. *Proceedings of the National Academy of Sciences*, 103(17):6565–6570.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Steffan-Dewenter, I., Münzenberg, U., Bürger, C., Thies, C., and Tscharntke, T. (2002). Scale-dependent effects of landscape context on three pollinator guilds. *Ecology*, 83(5):1421–1432.

Stoks, R. and McPeek, M. A. (2003). Predators and life histories shape lestes damselfly assemblages along a freshwater habitat gradient. *Ecology*, 84(6):1576–1587.

Sutherland, C., Elston, D., and Lambin, X. (2014). A demographic, spatially explicit patch occupancy model of metapopulation dynamics and persistence. *Ecology*, 95(11):3149–3160.

Taylor, P. D., Fahrig, L., and With, K. A. (2006). Landscape connectivity: a return to the basics. In Crooks, K. R. and Sanjayan, M., editors, *Conservation Biology*, pages 29–43. Cambridge University Press, Cambridge.

Tenan, S., O'Hara, R. B., Hendriks, I., and Tavecchia, G. (2014). Bayesian model selection: the steepest mountain to climb. *Ecological Modelling*, 283:62–69.

Termaat, T., van Strien, A. J., van Grunsven, R. H., De Knijf, G., Bjelke, U., Burbach, K., Conze, K.-J., Goffart, P., Hepper, D., Kalkman, V. J., et al. (2019). Distribution trends of european dragonflies under climate change. *Diversity and Distributions*, 25(6):936–950.

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine*, 18(20):2693–2708.

Tingley, M. W. and Beissinger, S. R. (2009). Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in ecology & evolution*, 24(11):625–633.

Tulloch, A. I., Mustin, K., Possingham, H. P., Szabo, J. K., and Wilson, K. A. (2013). To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions*, 19(4):465–480.

Utzeri, C. and Raffi, R. (1983). Observations on the behaviour of aeshna affinis (vander linden) at a dried-up pond (anisoptera: Aeshnidae). *Odonatologica*, 12(2):141–151.

van Strien, A. J., Termaat, T., Groenendijk, D., Mensing, V., and Kery, M. (2010). Site-occupancy models may offer new opportunities for dragonfly monitoring based on daily species lists. *Basic and Applied Ecology*, 11(6):495–503.

van Strien, A. J., van Swaay, C. A., and Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, 50(6):1450–1458.

Vere-Jones, D. (1988). An introduction to the theory of point processes. *Springer Ser. Statist., Springer, New York*.

Vieilledent, G. (2019). *hSDM: Hierarchical Bayesian Species Distribution Models*. R package version 1.4.1.

Wan, H., Cushman, S., and Ganey, J. (2018). Habitat fragmentation reduces genetic diversity and connectivity of the mexican spotted owl: a simulation study using empirical resistance models. *Genes*, 9(8):403.

Wang, X., Yue, Y., and Faraway, J. J. (2018). *Bayesian regression modeling with INLA*. Chapman and Hall/CRC.

Warton, D. I., Stoklosa, J., Guillera-Arroita, G., MacKenzie, D. I., and Welsh, A. H. (2017). Graphical diagnostics for occupancy models with imperfect detection. *Methods in Ecology and Evolution*, 8(4):408–419.

Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12).

Welsh, A. H., Lindenmayer, D. B., and Donnelly, C. F. (2013). Fitting and interpreting occupancy models. *PLoS One*, 8(1):e52015.

Wikle, C. K. (2003). Hierarchical bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394.

Wikle, C. K. and Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test*, 19(3):417–451.

Winston, M. R., Taylor, C. M., and Pigg, J. (1991). Upstream extirpation of four minnow species due to damming of a prairie stream. *Transactions of the American Fisheries Society*, 120(1):98–105.

Wintle, B. and Bardos, D. (2006). Modeling species–habitat relationships with spatially autocorrelated observation data. *Ecological Applications*, 16(5):1945–1958.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Wright, W. J., Irvine, K. M., and Higgs, M. D. (2019). Identifying occupancy model inadequacies: can residuals separately assess detection and presence? *Ecology*, 100(6):e02703.

Yamaura, Y., Royle, J. A., Kuboi, K., Tada, T., Ikeno, S., and Makino, S. (2011). Modelling community dynamics based on species-level abundance models from detection/nondetection data. *Journal of Applied Ecology*, 48(1):67–75.

Yu, J. and Dobson, F. S. (2000). Seven forms of rarity in mammals. *Journal of Biogeography*, 27(1):131–139.

Glossary

Alluvial floodplain landform next to a river or a stream where sediment has been deposited by running water over long periods of time.

Biological community assemblages of a group of species occurring in a particular area or habitat.

Catchment areas of land where water from various sources (e.g. groundwater, precipitation and aquifers) is collected.

Colonization an ecological process by which a species spread to new sites that were previously unoccupied.

Ecological connectivity landscape physical structures or species behavioural strategies that either limit or facilitates organisms movement across space.

Elusive species species that are difficult to detect during a survey due to multiple biological and environmental factors (e.g. species rarity, areas that are difficult to survey, species traits such as camouflage, daytime/temporal activity patterns or body sizes that make the tasks of detecting them difficult).

Environmental stressor physical, chemical, or biological conditions that limit species productivity an distributions.

Extinction an ecological process by which a previously occupied site becomes unoccupied by one or more species, resulting in the dissolution of local populations .

IUCN Red list The International Union for Conservation of Nature's Red List of Threatened Species encompasses an comprehensive inventory of the global conservation status of animal, fungi and plant species by assessing the risk of extinction of each species.

Population dynamics The study of how population size and density for one or more species change over time and the biological and physical processes that affect those changes. In this research, population dynamics refer to those colonization and extinction processes that determine species distributions over time and space which are driven by underlying demographic emigration/migration and mortality/natality mechanisms.

Rare species are species with very scarce distributions that are very uncommon and infrequently encountered in nature. In this research, rare species are defined based on their relative occurrence probabilities with respect to the occurrences of other species on a national scale using a 1 km grid.

Riparian connectivity lands along watercourses and water bodies such as streams, rivers, lakes and ponds that facilitates wildlife and plants movement across habitats.

Species list number of different species that are recorded during a survey.

Species richness also denoted as α diversity, is a biological diversity metric defined by the total number of species occurring in a particular area or habitat.

Species abundance biological diversity metric defined by the number of individuals of a particular species.

Stralher number an integer that indicates the complexity in a river network based on the level of branching determined by each of the segments of a stream or river.

Voltinism the number of generations or broods an organism produces in a year.

Appendices

A.1 Chapter 2 appendix

A.1.1 Metropolis-Gibbs sampler algorithm for a two-stage occupancy model

Description of the Gibbs/ MH algorithm used to fit the single species two-stages occupancy model presented in chapter 2. A description of the conditional distributions necessary for implementing a Gibbs sampler for a constant detection occupancy model with site varying occupancy probabilities are described in section 2.2.2.

Suppose the following occupancy model:

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\psi_j) \\ \text{logit}(\psi_j) &= \beta_0 + \beta_1 \times x_j \\ y_{j\cdot} &\sim \text{Binomial}(K, p) \end{aligned} \tag{9}$$

Where the likelihood can be expressed as a zero-inflated binomial:

$$[y_{j\cdot}|\beta_0,\beta_1,z,p,K,x] = \prod_j^M \binom{K}{y_{j\cdot}} p^{y_{j\cdot}}(1-p)^{K-y_{j\cdot}}\psi_j + \mathbb{I}(y_{j\cdot}=0)(1-\psi_j). \tag{10}$$

A Gibbs/ MH algorithm implies sampling from $[\beta|z]$, $[p|z,y]$, and $[z_j = 1|p,\beta,y]$ for all sites where the species was not detected. The algorithm goes as folllows:

1. Set initial values $\beta = [\beta_0^{(0)},\beta_1^{(0)}], p^{(0)}$ and $z^{(0)}$, e.g. initial values for the occupancy state can be defined as $z_j^{(0)}\mathbb{I}(y_{j\cdot} > 0)$.

2. For each iteration $\mathbf{s} \in 1,\ldots,S$:

   2.1. Draw candidate values of β from the symmetric proposal distribution, centered at the current sampled value, e.g. starting with β_0 this is:

   $$\beta_0^{(s)} \sim \text{Normal}(\beta_0^{(s-1)}, \sigma^2),$$

   where σ^2 is the tuning parameter.

   2.2. Evaluate the marginal likelihood (equation 10) and the prior at the candidate value, i.e. $[y_{j\cdot}|\beta_0^{(s)},\beta_1^{(s-1)},z^{(s-1)},p^{(s-1)},K,x]$ and $\text{Normal}(\beta^{(s)},\tau^2)$ respectively.

   2.3. Compute Metropolis acceptance probability (r)

$$r = \frac{[y_{j\cdot}|\beta_0^{(s)},\beta_1^{(s-1)},z^{(s-1)},p^{(s-1)},K,x] \times \text{Normal}(\beta^{(s)},\tau^2)}{[y_{j\cdot}|\beta_0^{(s-1)},\beta_1^{(s-1)},z^{(s-1)},p^{(s-1)},K,x] \times \text{Normal}(\beta^{(s-1)},\tau^2)} \tag{11}$$

2.4. Accept the proposed move $\beta_0^{(s)}$ with probability

$$\alpha(\beta_0^{(s-1)}, \beta^{(s)})_0 = \min(1, r(\beta_0^{(s-1)}, \beta_0^{(s)}) \tag{12}$$

Otherwise reject the proposed move, and the chain stays at the same position $\beta_0^{(s)} = \beta_0^{(s-1)}$

3. Repeat steps 2.1 through 2.4 to update β_1.

4. Update z_j only for the sites where the species was not detected (i.e. $y_{j\cdot} = 0$)

4.1. Define $\psi_j^{(s)} = \text{logit}^{-1}(\beta_0^{(s)} + \beta_1^{(s)} x_j)$

4.2. Update $[z_j^{(s)} = 1 | y_{j\cdot} = 0, \psi^{(s)}, p^{(s-1)}]$

4.3. Update likelihood $[y_{j\cdot} | \beta_0^{(s)}, \beta_1^{(s)}, z^{(s)}, p^{(s-1)}, K, x]$

5. Update detection probabilities directly from its full conditional (MH/Gibbs algorithm can be used if detection probabilities vary across sites due to site-level covariates).

$$[p^{(s)} | y, z^{(s)}, \psi_j^{(s)}] = \text{Beta}(1 + \sum_j y_{j\cdot}, 1 + \sum_j z_j^{(s)} \times K - \sum_j y_{j\cdot}) \tag{13}$$

A.1.2 Simulation study: Occupancy model with covariates

The following section explores the singles species occupancy model when occupancy and detection probabilities vary with site-level covariates. Let the following single season occupancy model:

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\psi_j) \\ \text{logit}(\psi_j) &= \beta_0 + \beta_1 x_j \\ y_{jk} &\sim \text{Bernoulli}(z_j p_{jk}) \\ \text{logit}(p_{jk}) &= \alpha_0 + \alpha_1 g_{jk} \end{aligned} \tag{14}$$

Where z_j is the latent occupancy state of whether site j is occupied with an occupancy probability denoted by ψ_j defined as a function of the species baseline occupancy (β_0) and the effect covariate x_j on the species occurrences. The observational model is defined by the observed species detection y_{jk} during visit k and the detection probability p_{jk} defined by the baseline detection (α_0) and the effect of covariate g_{jk}.

A total of 500 sites visited on three sampling occasions were specified to simulate $\mathbf{x}$ and $\mathbf{g}_k$ covariates that were randomly drawn from a Uniform (-1,1) distribution. Slopes $\beta_1 = 3$ and $\alpha_1 = -3$ were specified to simulate a positive and negative relationship between site-level covariates and occupancy and detection probabilities respectively. The Gibbs sampler algorithm described in section A.1.1 was implemented in R for a total of 10,000 iteration, a thinning of 10 and a burnin period of 2000. Normal(0,0.01) priors

were specified for each parameter. Figures 1 through 3 show the results when the baseline occupancy and detection probabilities are both 0.5. Then, figures 4 through 6 show the results when the baseline occupancy is 0.5 and baseline detection probability is 0.2.

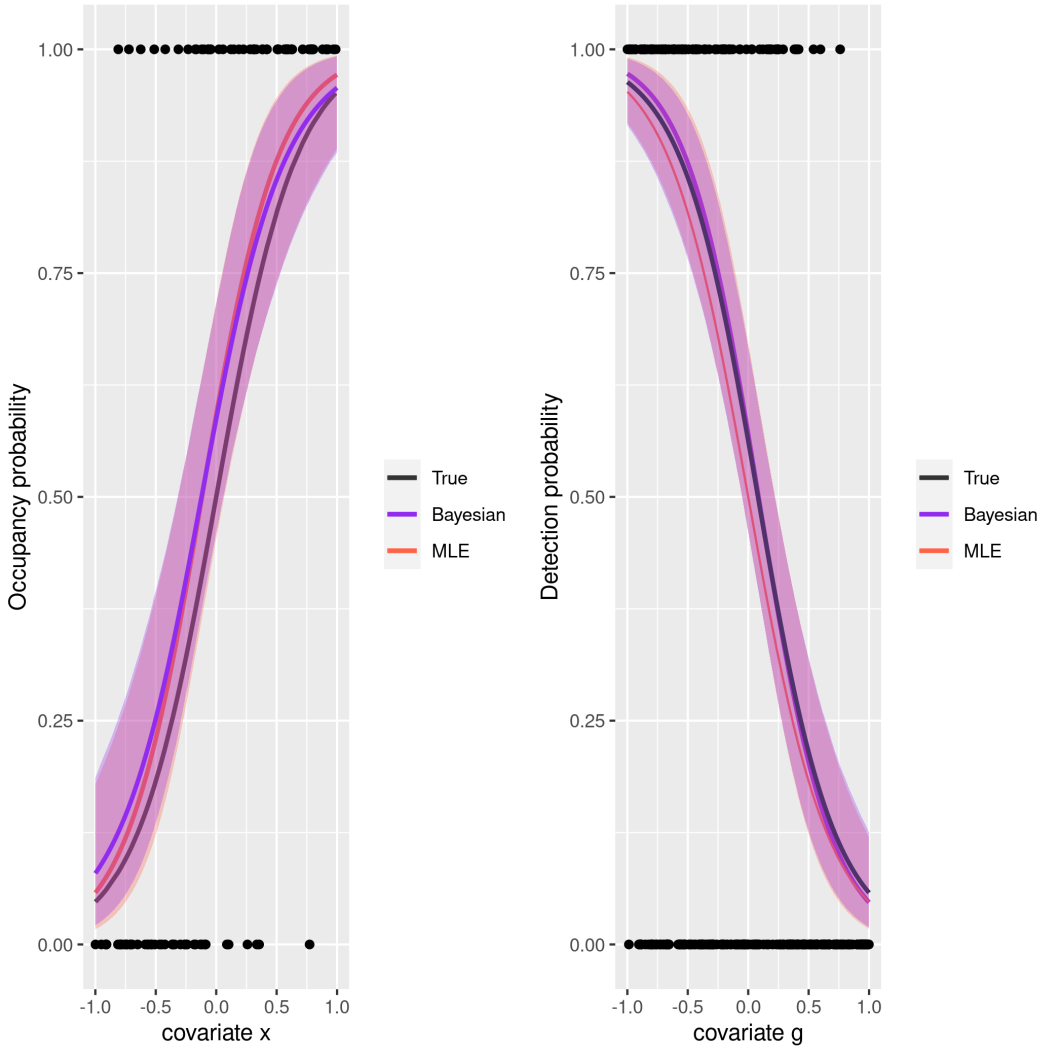


Figure 1: Estimated relationships between occupancy probability and simulated covariate x (left) and between detection probability and covariate g (right) from an occupancy model fit to the simulated data set with a baseline occupancy and detection probabilities of 0.5. Black lines represent the true relationship between covariates and the response with points showing the true presence/absence. Red lines indicate maximum likelihood estimates and 95% CIs (red shaded region). Purple lines indicate bayesian estimates with 95% Credible intervals (purple shaded region).

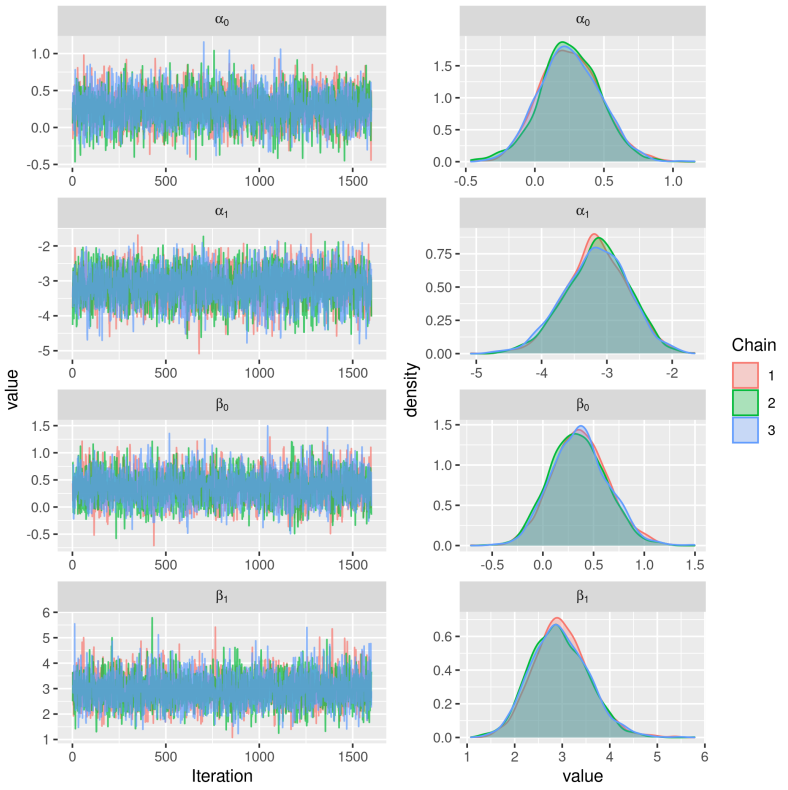
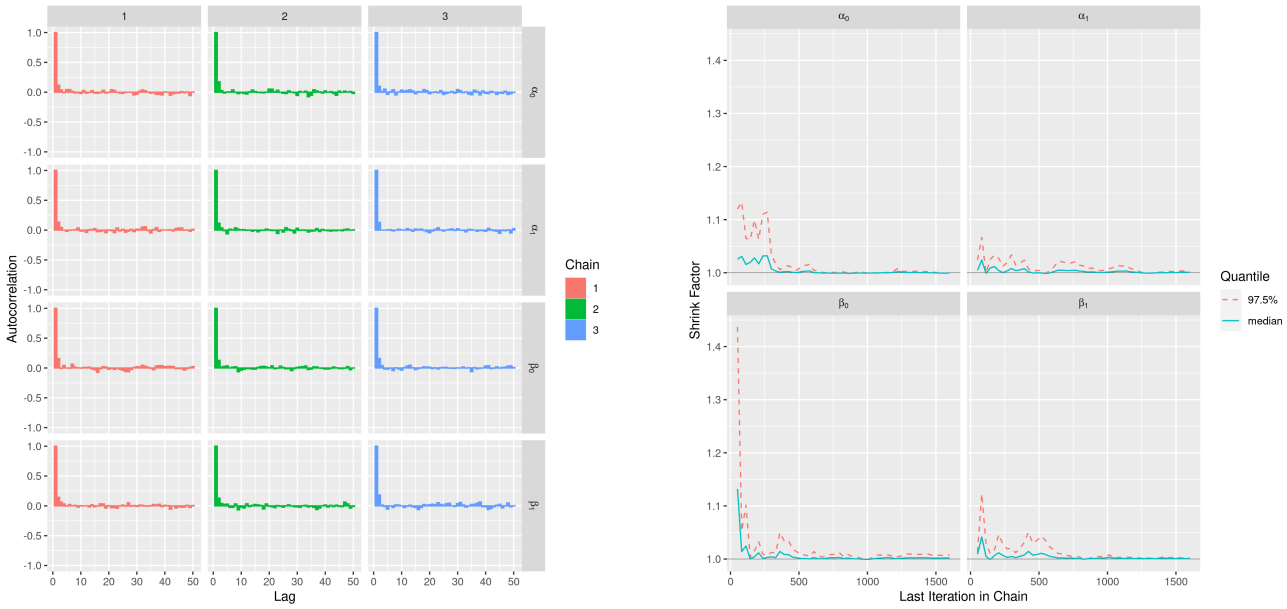


Figure 2: Trace and density plots from an occupancy model fit to the simulated data set with a baseline occupancy and detection probabilities of 0.5.



(a) (b)

Figure 3: Autocorrelation and Gelman-Rubin diagnostic plots from an occupancy model fit to the simulated data set with a baseline occupancy and detection probabilities of 0.5

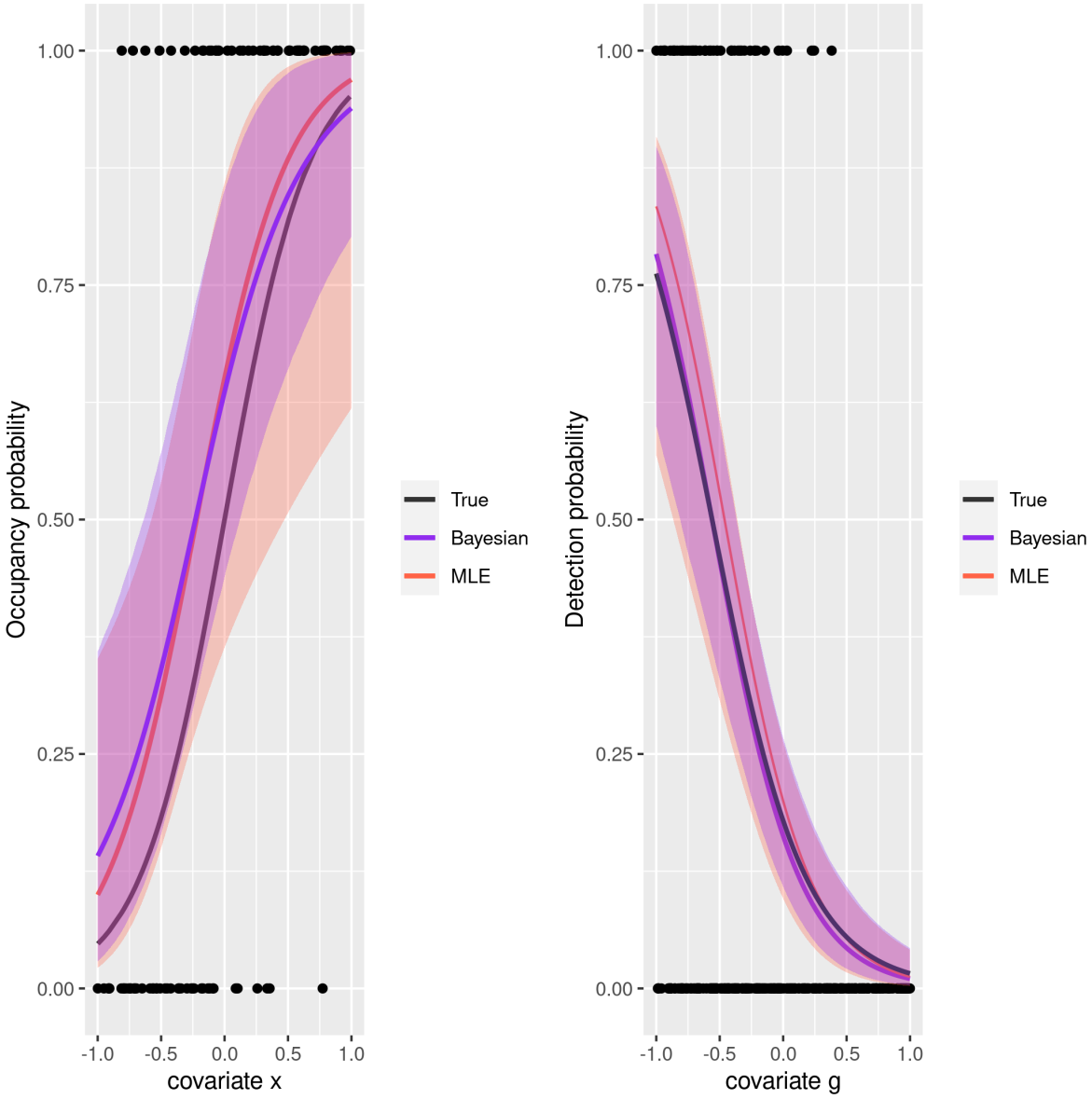


Figure 4: Estimated relationships between occupancy probability and simulated covariate x (left) and between detection probability and covariate g (right) from an occupancy model fit to the simulated data set with a baseline occupancy of 0.5 and detection probabilities of 0.2. Black lines represent the true relationship between covariates and the response with points showing the true presence/absence. Red lines indicate maximum likelihood estimates and 95% CIs (red shaded region). Purple lines indicate bayesian estimates with 95% Credible intervals (purple shaded region).

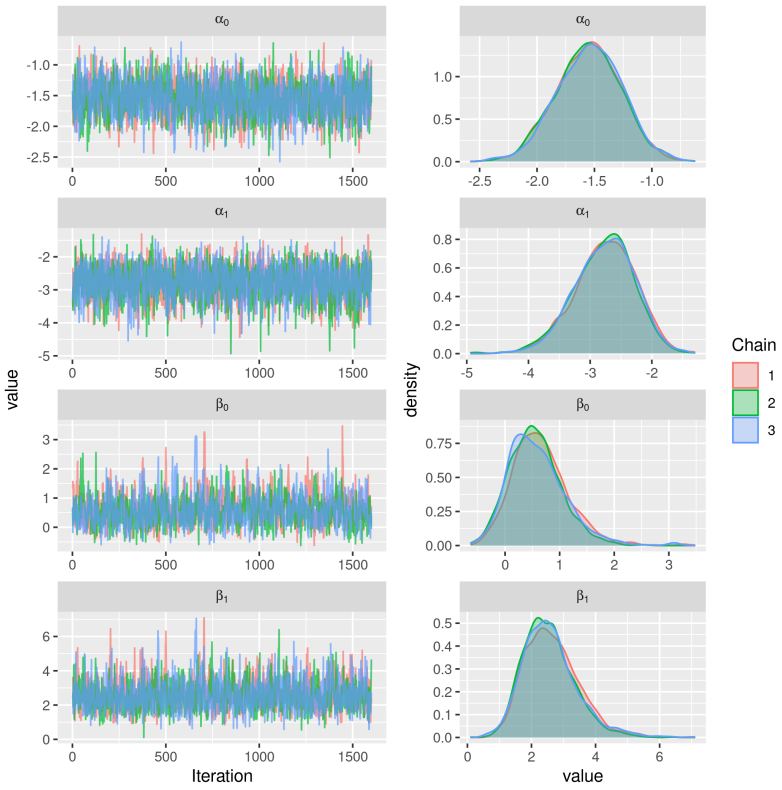
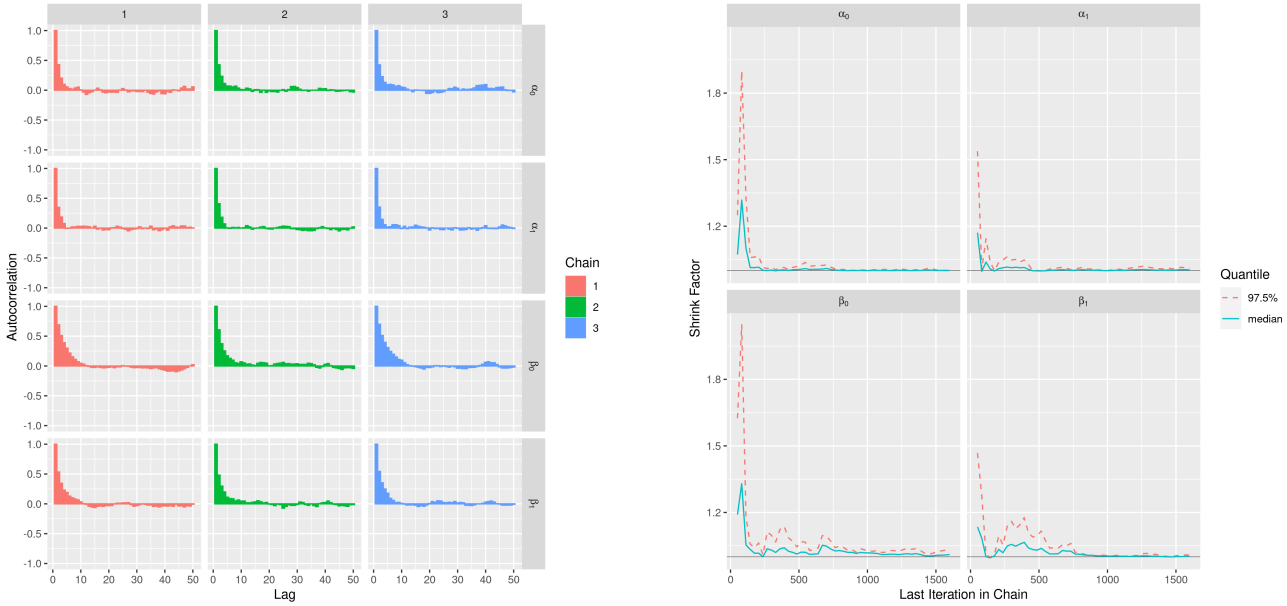


Figure 5: Trace and density plots from an occupancy model fit to the simulated data set with a baseline occupancy of 0.5 and detection probabilities of 0.2



(a) (b)

Figure 6: Autocorrelation and Gelman-Rubin diagnostic plots from an occupancy model fit to the simulated data set with a baseline occupancy and detection probabilities of 0.5

A.1.3 Odonata multiple species occupancy model

A.1.3.1 Fitting separate models to each species

The following figures show model (2.38) convergence diagnostics. This model is equivalent to fitting individual occupancy model to each of the species in the community. A sample of the species specific occupancy/detection probabilities parameters (ψ and p respectively) convergence diagnostics are shown for the following species: (1) *Aeshna affinis* (2) *Aeshna grandis* (3) *Coenagrion mercuriale* (4) *Enallagma cyathigerum* (5) *Gomphus vulgatissimus* and (6) *Lestes dryas*.

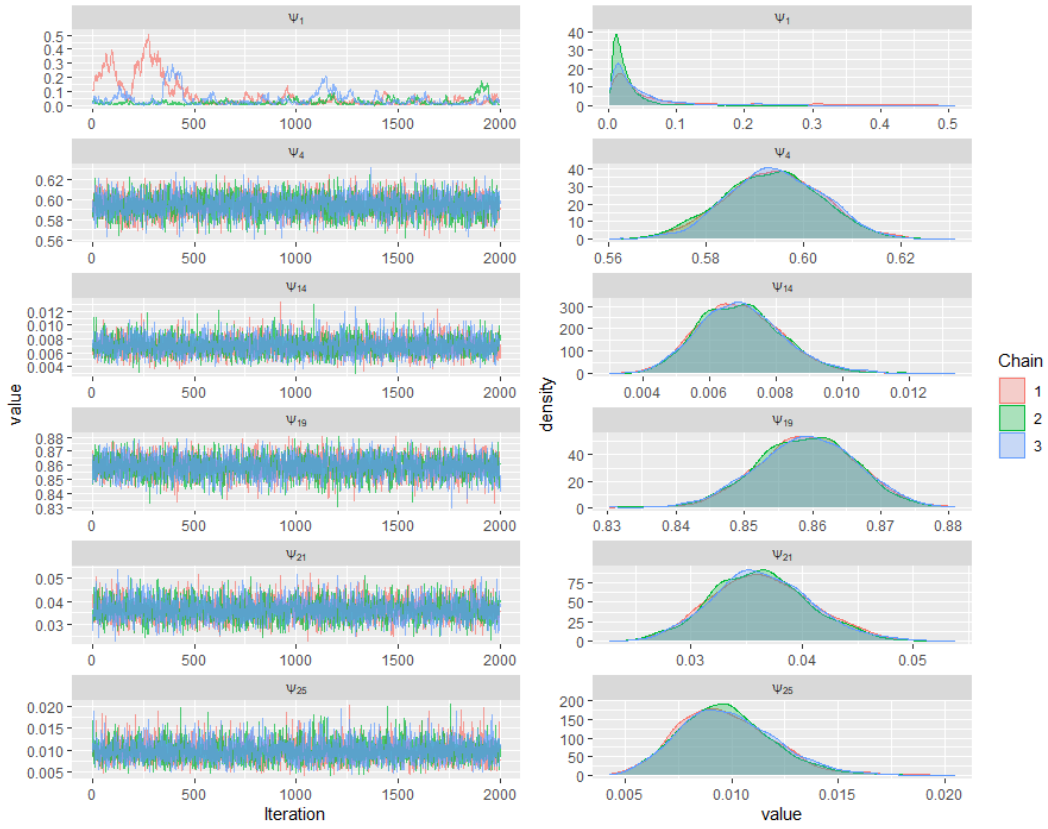


Figure 7: Species-specific occupancy probabilities trace and density plots from a multiplespecies occupancy model fit separately to each Odonata species.

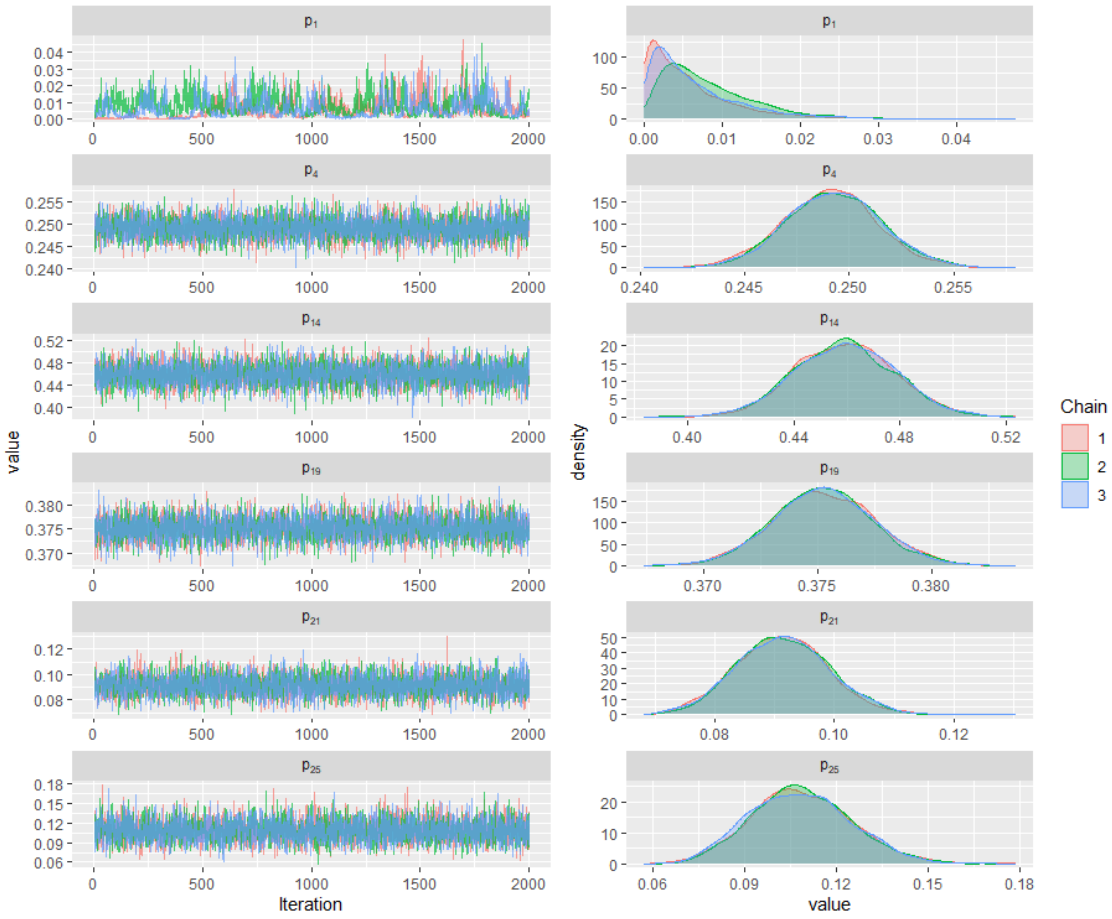
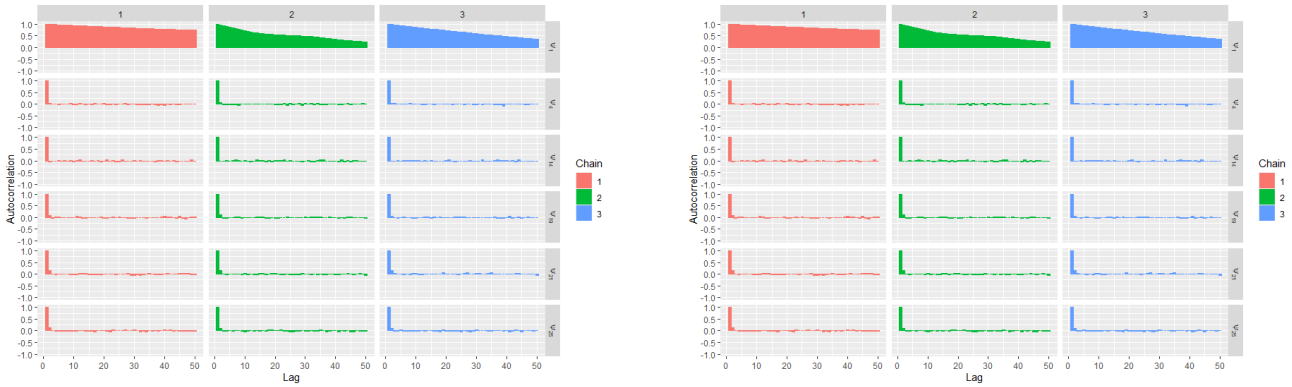


Figure 8: Species-specific detection probabilities trace and density plots from a multiplespecies occupancy model fit separately to each Odonata species.



(a) (b)

Figure 9: Species-specific occupancy (left) and detection (right) probabilities autocorrelation plots from a multiplespecies occupancy model fit separately to each Odonata.

A.1.3.2 Species random effect occupancy model convergence diagnostics

The following plots show species-specific model (2.39) parameters convergence diagnostics for the same aforementioned subset of species where species parameters are assumed to arise form the same multivariate normal prior distribution.

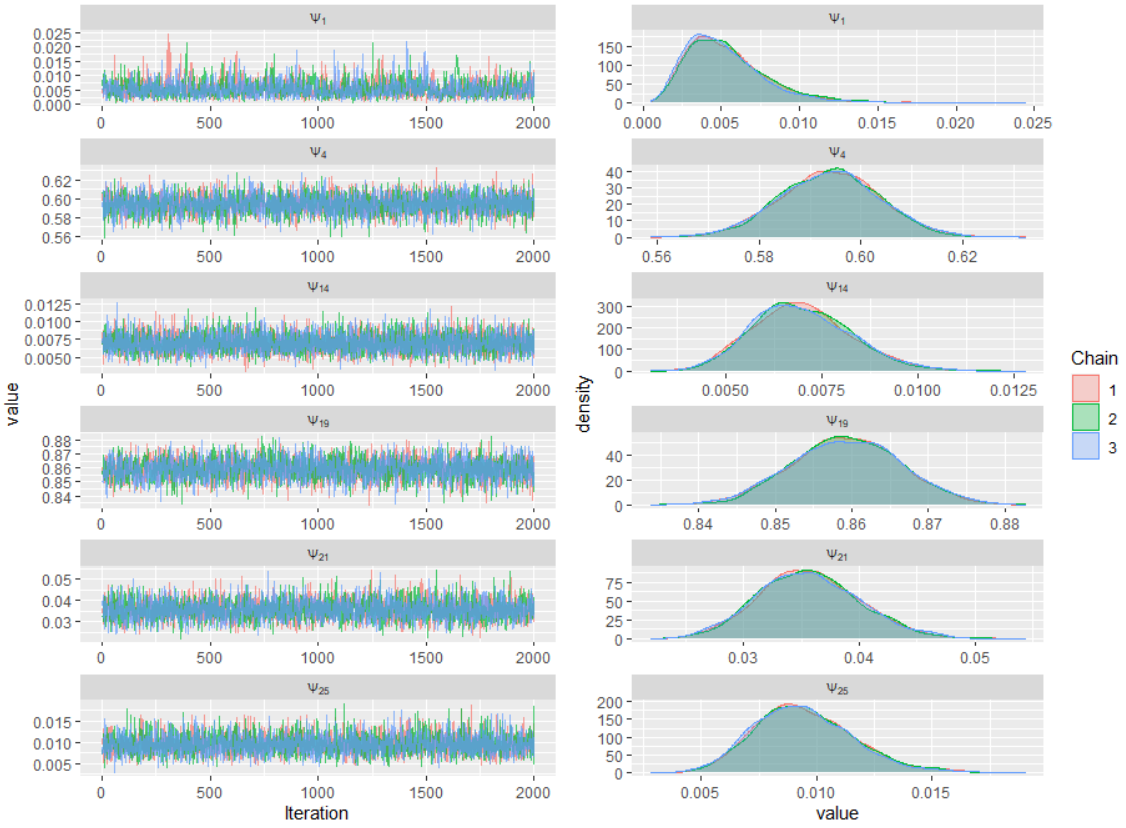


Figure 10: Species-specific occupancy probabilities trace and density plots from a multiplespecies occupancy model assuming the same prior distribution for the species-specific effects.

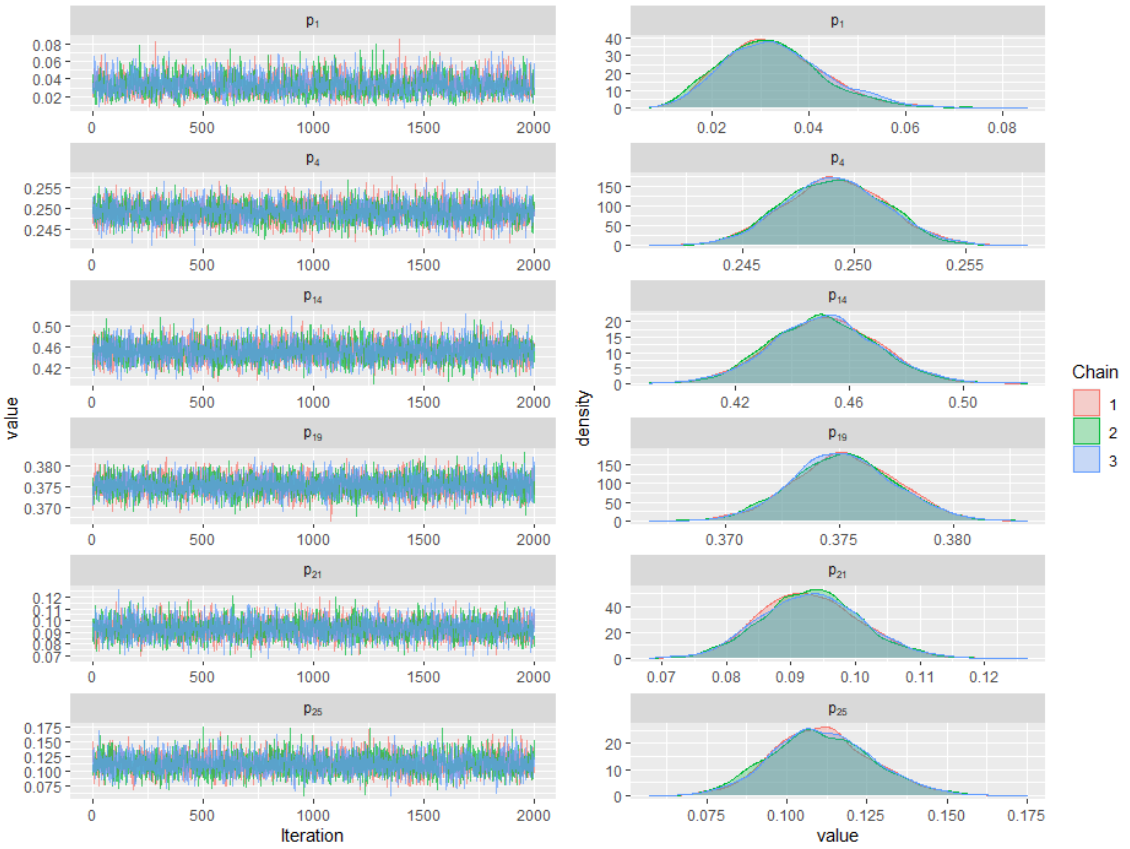
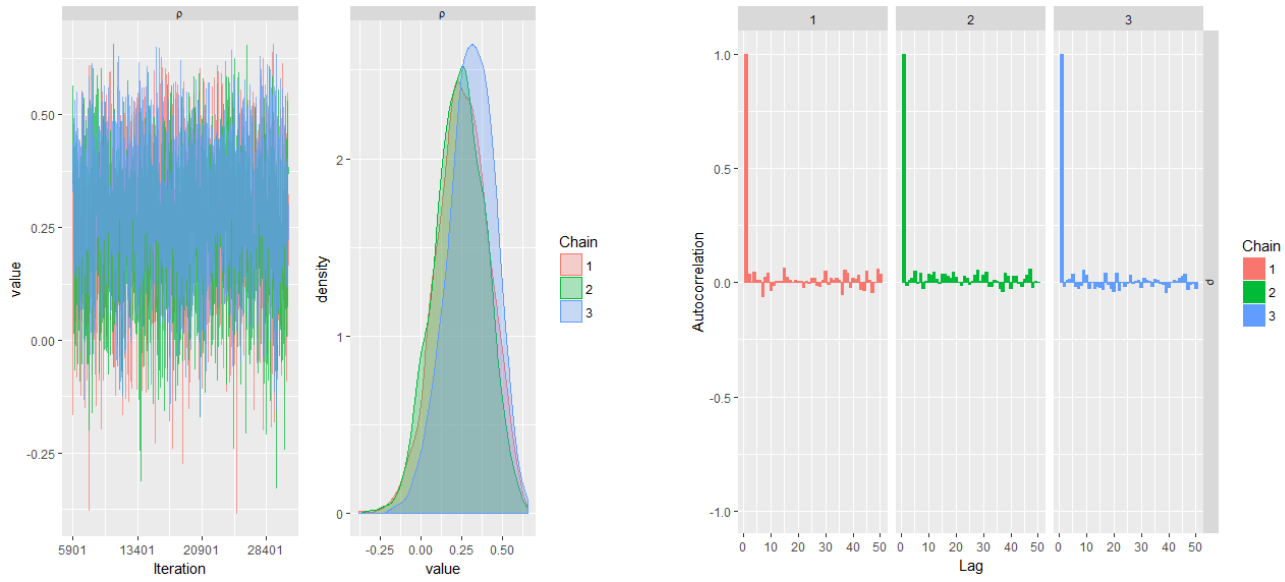


Figure 11: Species-specific detection probabilities trace and density plots from a multiplespecies occupancy model assuming the same prior distribution for the species-specific effects.



(a) ρ trace plots

(b) autocorrelation plot for ρ posterior samples

Figure 12: Diagnostics plots for ρ posterior samples from an multiplespecies occupancy model assuming the same prior distribution for the species-specific effects.

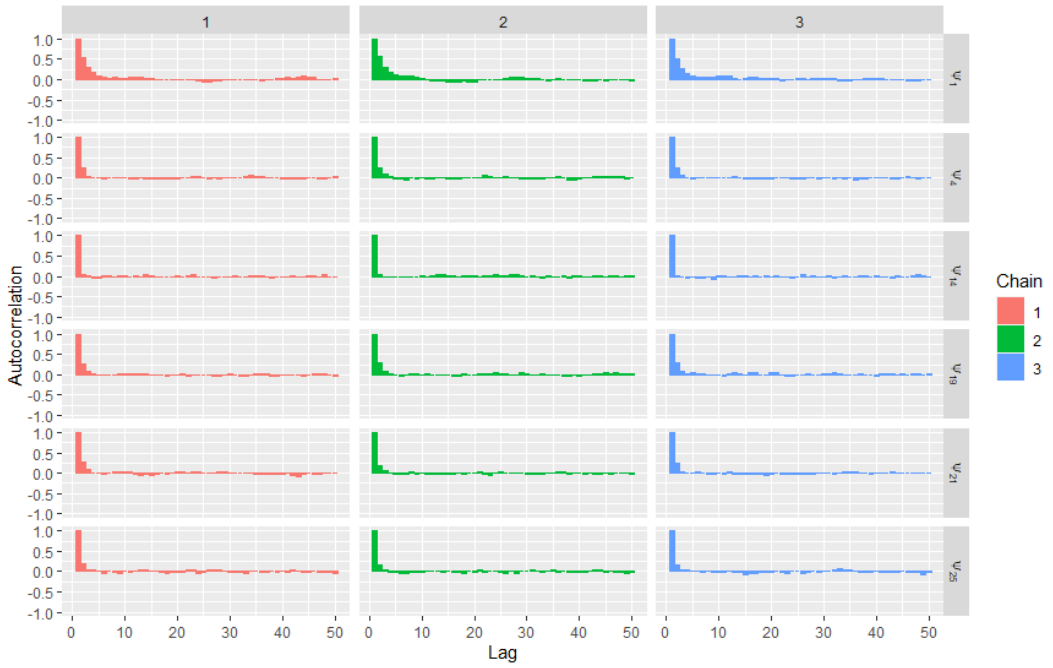


Figure 13: Species-specific occupancy probabilities autocorrelation plots from a multiplespecies occupancy model assuming the same prior distribution for the species-specific effects.

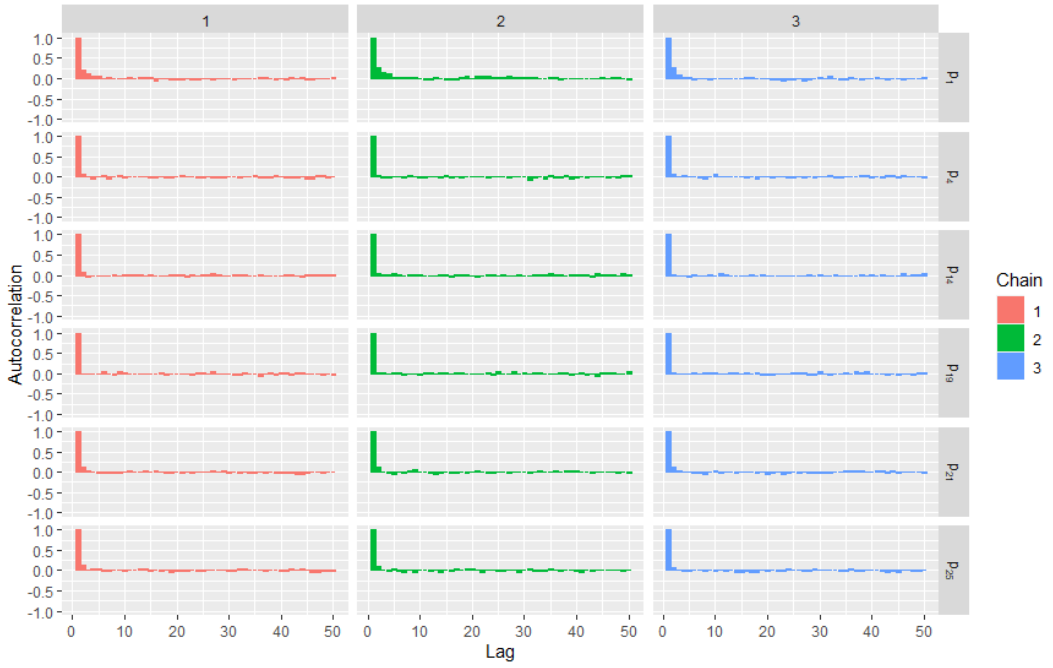


Figure 14: Species-specific detection probabilities autocorrelation plots from a multiplespecies occupancy model assuming the same prior distribution for the species-specific effects.

B.2 Chapter 3 appendix

B.2.1 Bayesian Index of realtive rarity result

The following table shows the rarity weights computed for each Odonata species based on a 25% quantile threshold.

Table 1: Species Estimated rarity weights for computing Leroy's Relative Rarity Index

Species	weights
Aeshna affinis	0.2429
Aeshna caerulea	0.0905
Aeshna cyanea	0.0000
Aeshna grandis	0.0000
Aeshna juncea	0.0000
Aeshna mixta	0.0000
Anaciaeschna isoceles	0.0293
Anax imperator	0.0000
Brachytron pratense	0.0000
Calopteryx splendens	0.0000
Calopteryx virgo	0.0000
Ceriagrion tenellum	0.0095
Coenagrion hastulatum	0.2056
Coenagrion mercuriale	0.2156
Coenagrion puella	0.0000
Coenagrion pulchellum	0.0001
Cordulegaster boltonii	0.0000
Cordulia aenea	0.0000
Enallagma cyathigerum	0.0000
Erythromma najas	0.0000
Gomphus vulgatissimus	0.0074
Ischnura elegans	0.0000
Ischnura pumilio	0.0367
Lestes barbarus	0.3320
Lestes dryas	0.1733
Lestes sponsa	0.0000
Leucorrhinia dubia	0.0670
Libellula depressa	0.0000
Libellula fulva	0.0000
Libellula quadrimaculata	0.0000
Orthetrum cancellatum	0.0000
Orthetrum coerulescens	0.0000
Platycnemis pennipes	0.0000
Pyrrhosoma nymphula	0.0000
Somatochlora arctica	0.0765
Somatochlora metallica	0.0015
Sympetrum danae	0.0000
Sympetrum sanguineum	0.0000
Sympetrum striolatum	0.0000

B.2.2 Bayesian Occupancy Mixture Model convergence diagnostics

This section includes the convergence conventional graphical diagnostics for the proposed Occupancy Mixture Model fitted to the simulated scenario 1 and 2 (Figures 15 - 18), and to the Odonata data set (Figures 19 - 20).

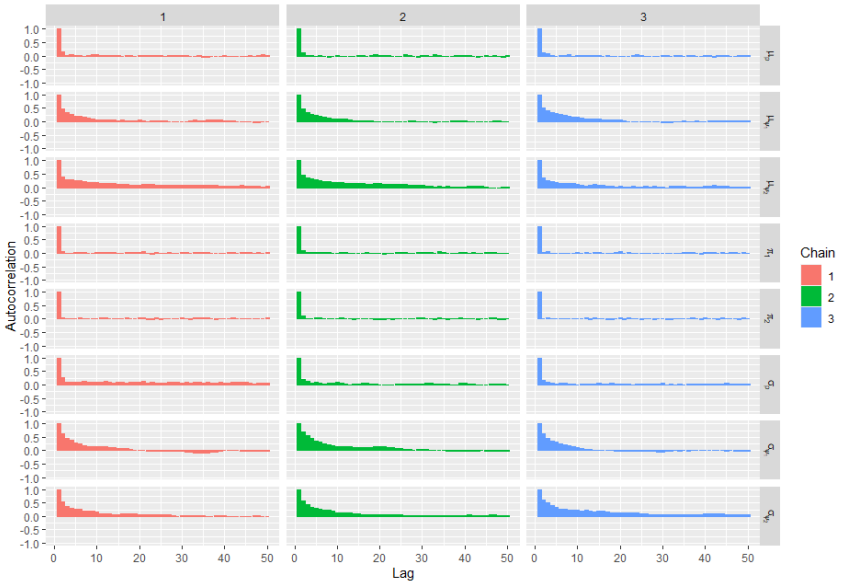


Figure 15: Autocorrelation plot for mixture occupancy model fitted to scenario 1 with proportional allocation.

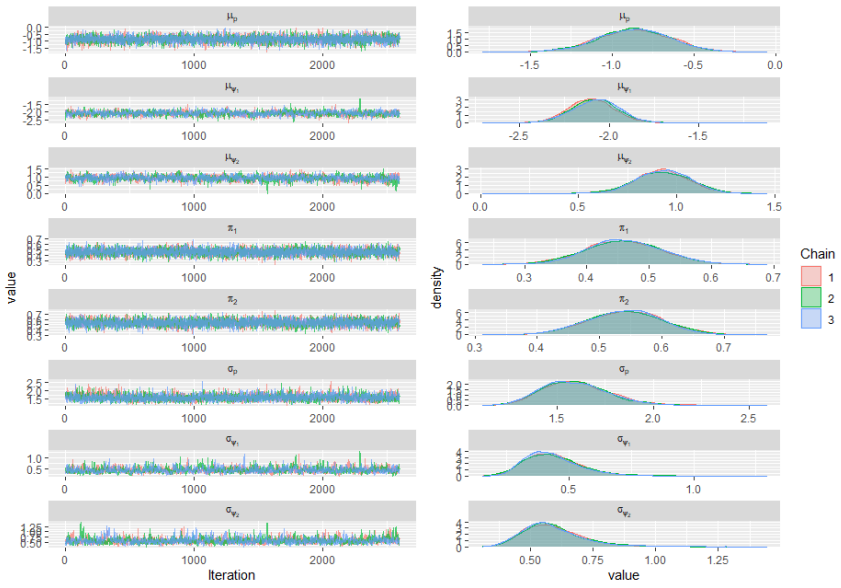


Figure 16: Density and traceplots for mixture occupancy model fitted to scenario 1 with proportional allocation.

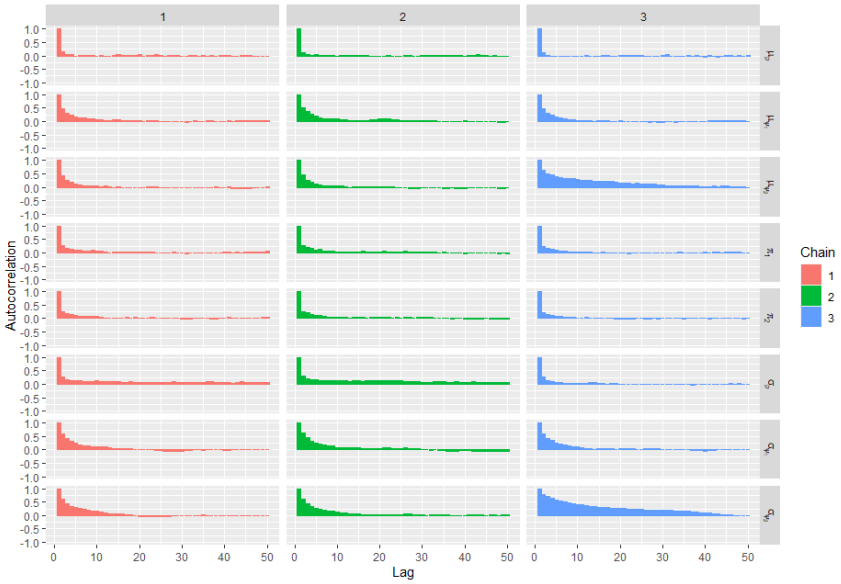


Figure 17: Autocorrelation plot for mixture occupancy model fitted to scenario 2 with a greater proportion of commons species.

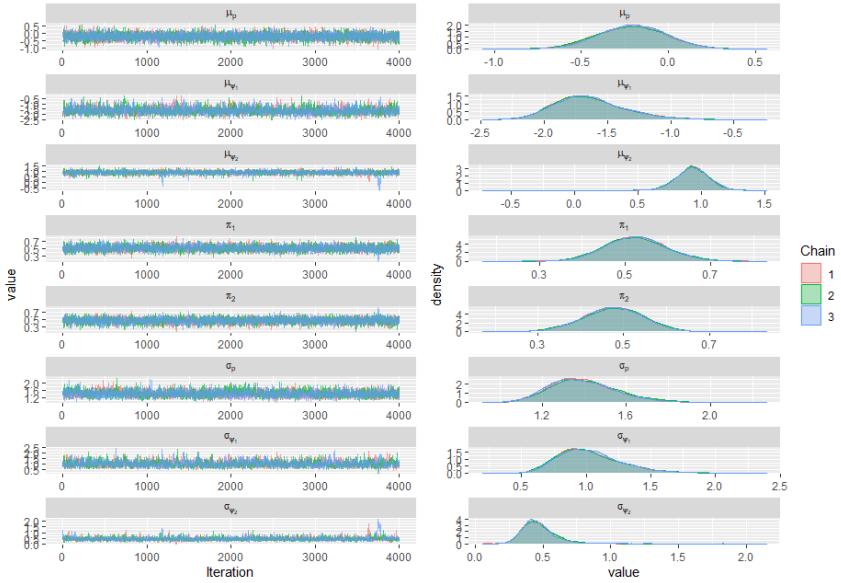


Figure 18: Density and traceplots for mixture occupancy model fitted to scenario 2 with a greater proportion of commons species.

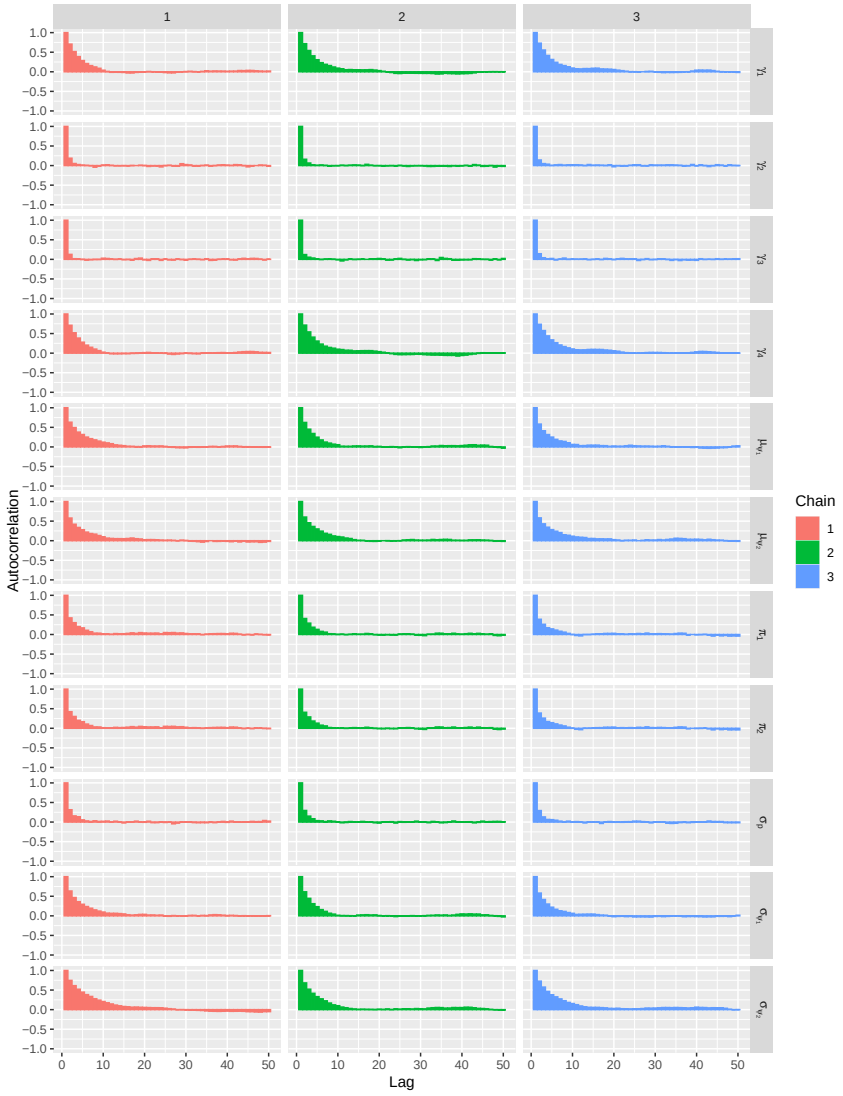


Figure 19: Autocorrelation plot for odonata mixture occupancy model.

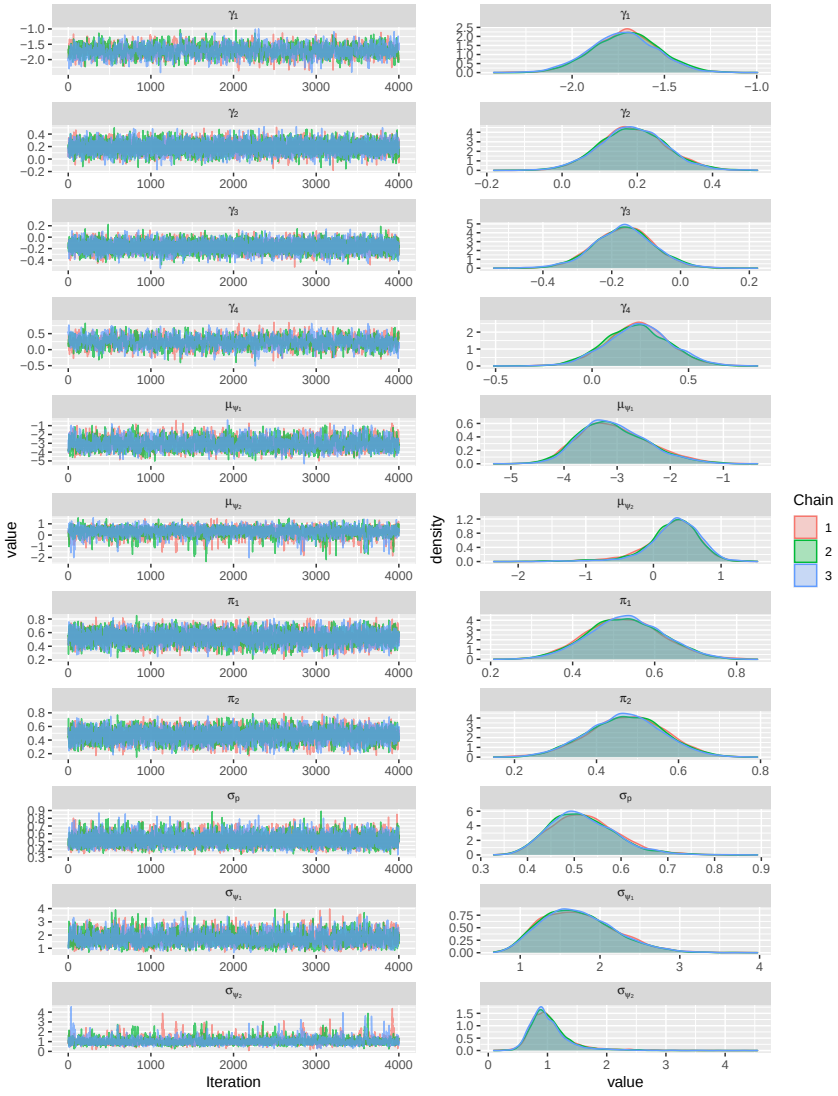


Figure 20: Density and traceplots for model odonata mixture occupancy model.

B.2.3 Two-step GAM diagnostic check

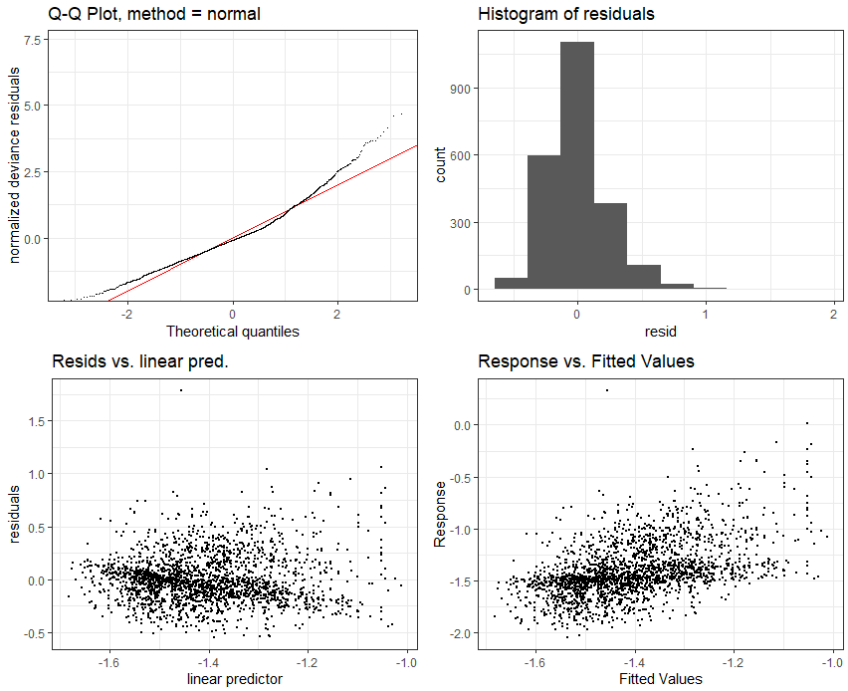


Figure 21: Diagnostics plot for proportion of rare species GAM using an additive variance structure.

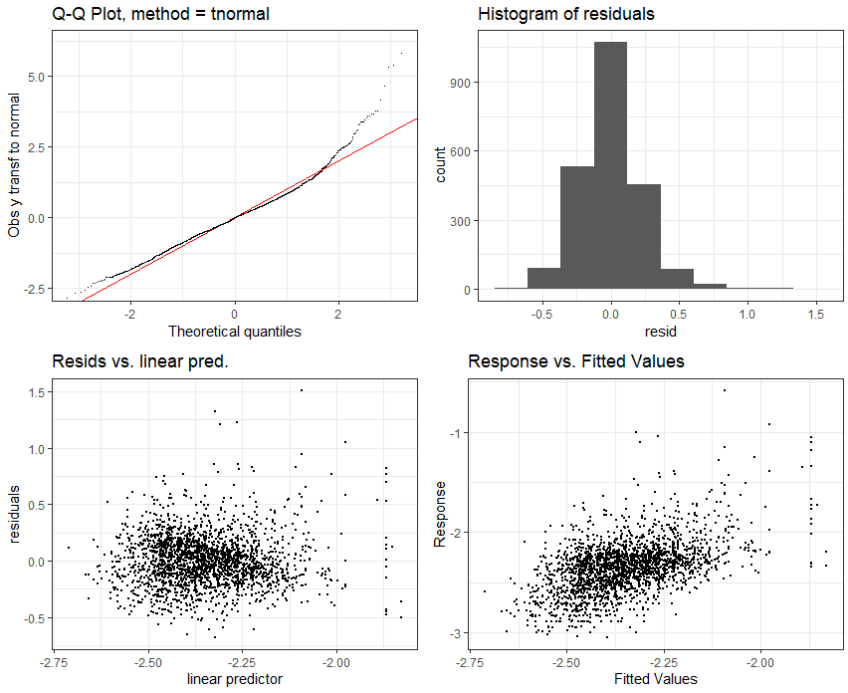


Figure 22: Diagnostics plot for IRR GAM with no variance structure.

C.3 Chapter 4 appendix

C.3.1 Generalized penalized splines dynamic occupancy model diagnostics

The following figures show a sample of the species-specific parameters traceplots for the flexible dynamic occupancy model developed in chapter 4 and fitted to simulated data under a different visit-sampling schemes.

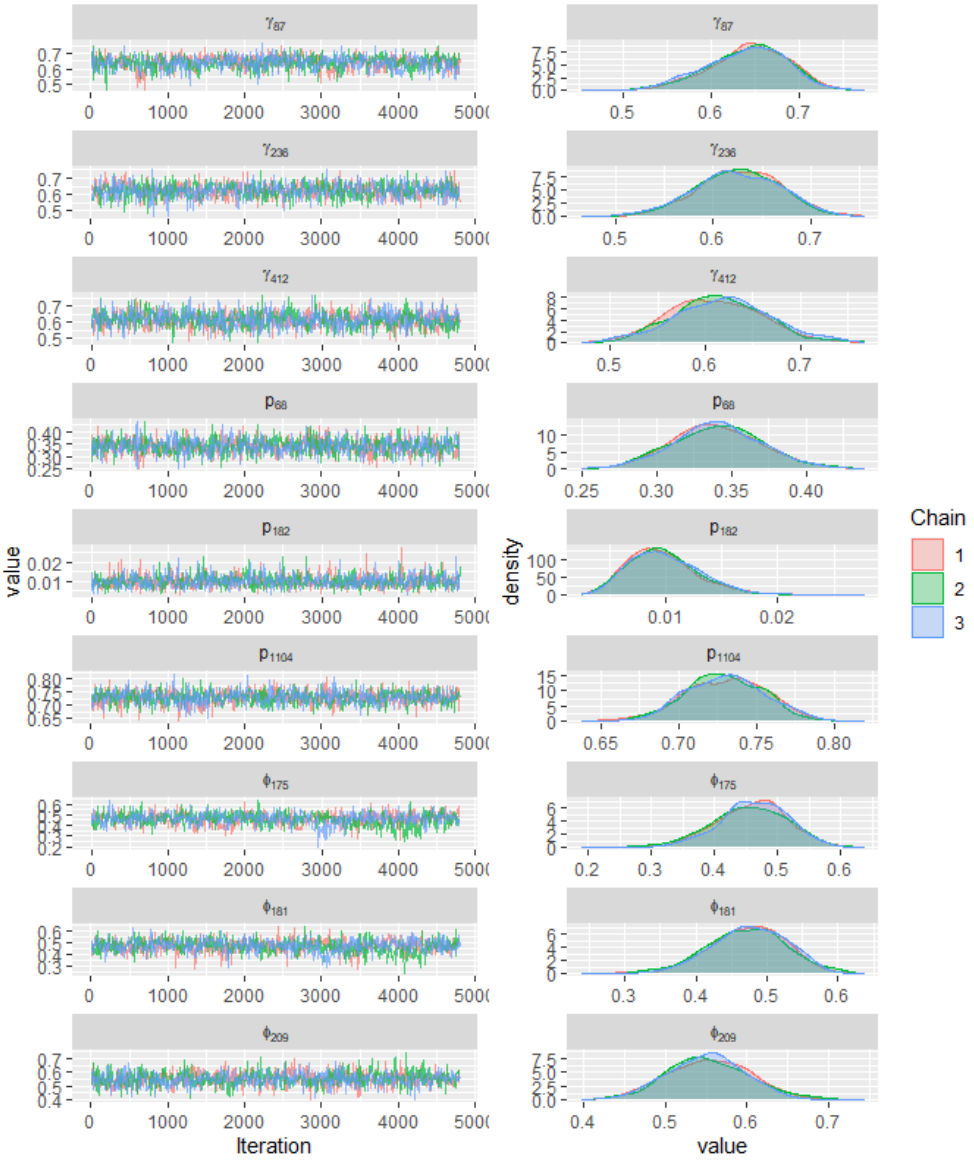


Figure 23: Trace plots and posterior densities for generalized penalized splines MSOM colonization, survival and detection parameters when data from three visits is used.

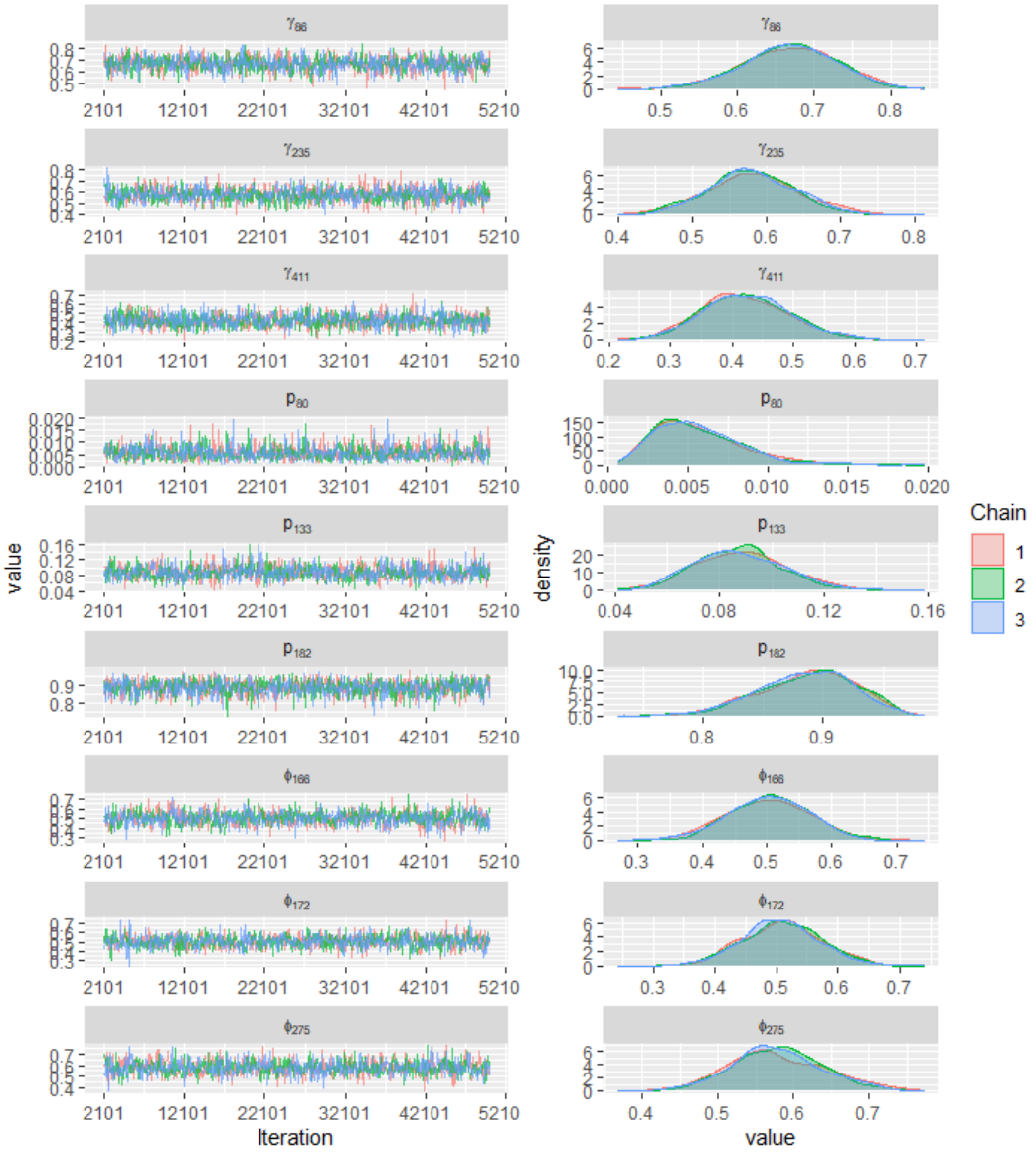
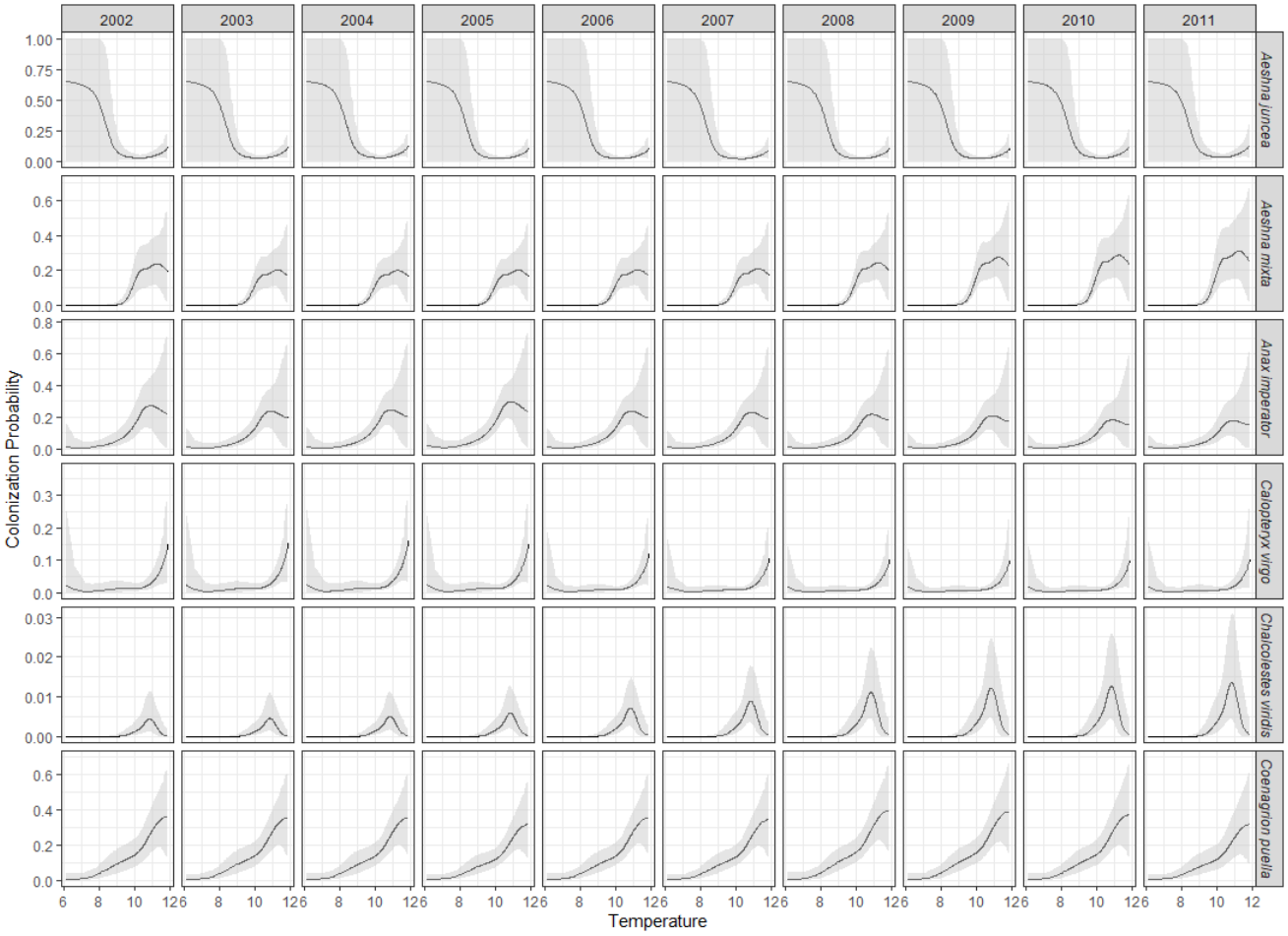


Figure 24: Trace plots and posterior densities for generalized penalized splines MSOM colonization, survival and detection parameters when data from a single visits is used.

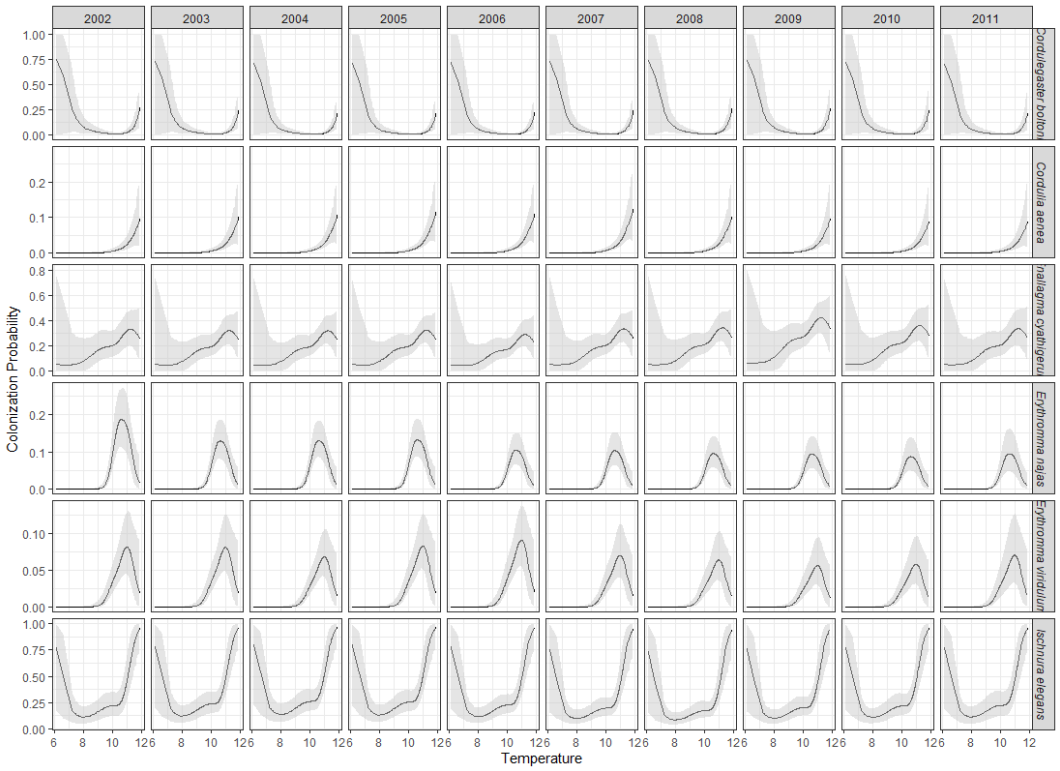
D.4 Chapter 5 appendix

D.4.1 Odonata flexible dynamic occupancy model aggregated number of species records

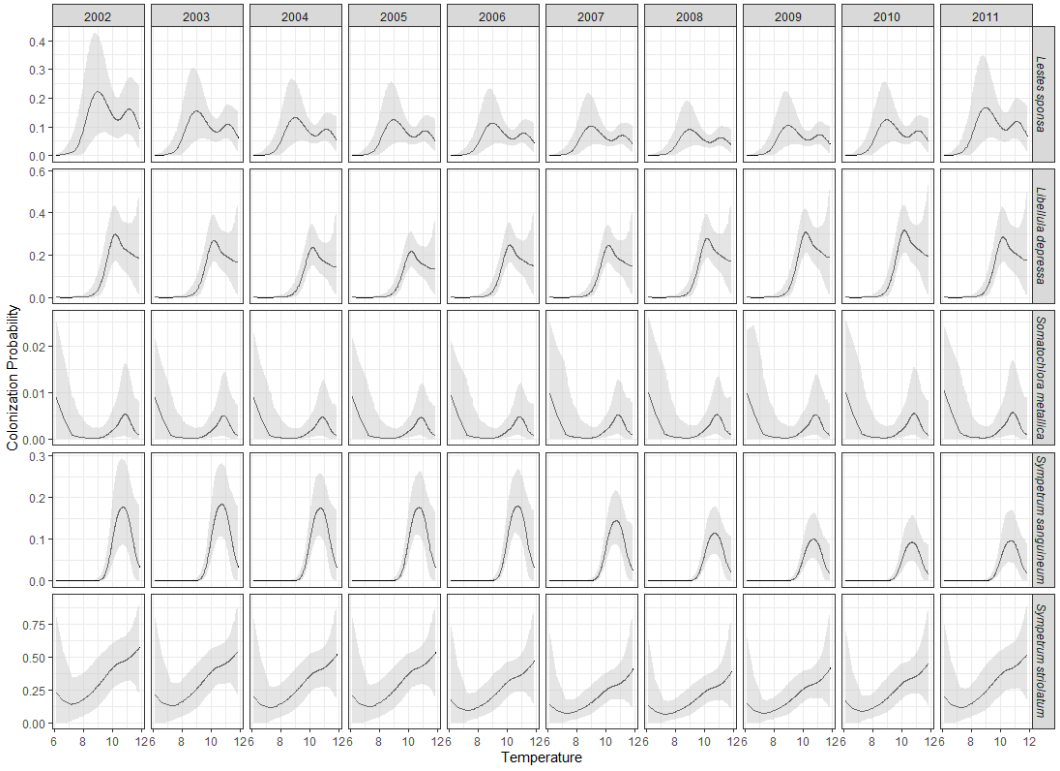
The following figures show the relationship between temperature and the population dynamics for 17 species as well as traceplots showing the mixing for a sample of parameters for certain species at selection of sites and years. In addition, the relationship between the human density index and detection probabilities for these species. This selection of species and years summarizes the main species-specific effects found among the 41 species analyzed with the flexible occupancy model.



(a)

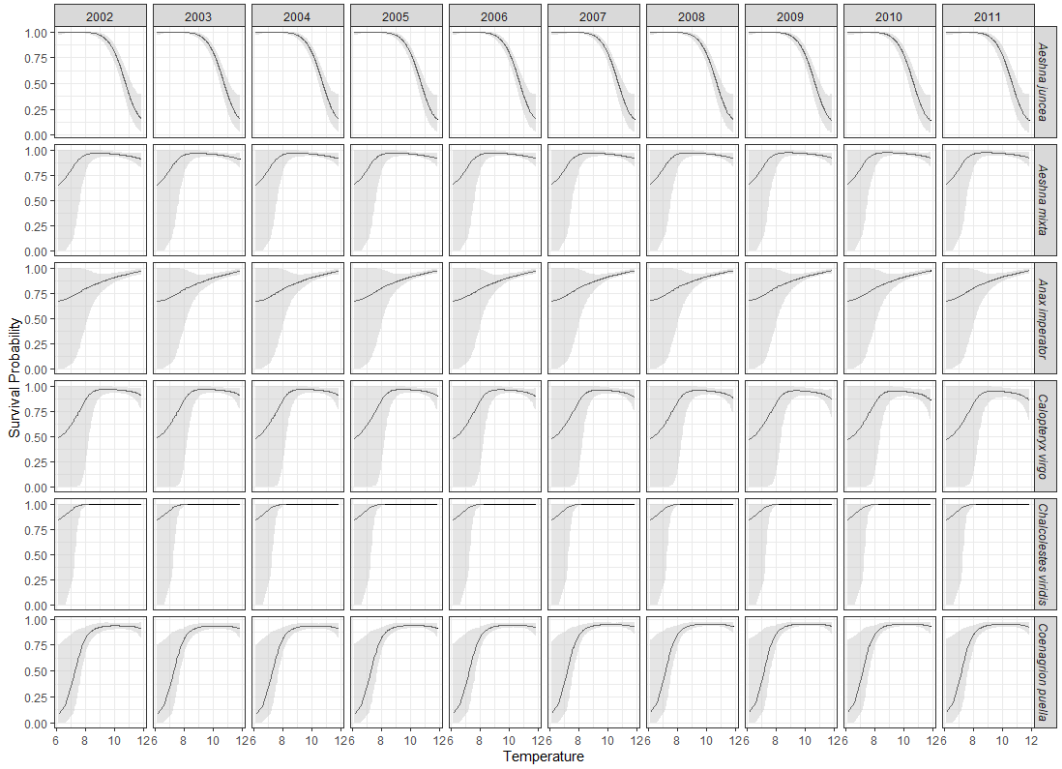


(a)

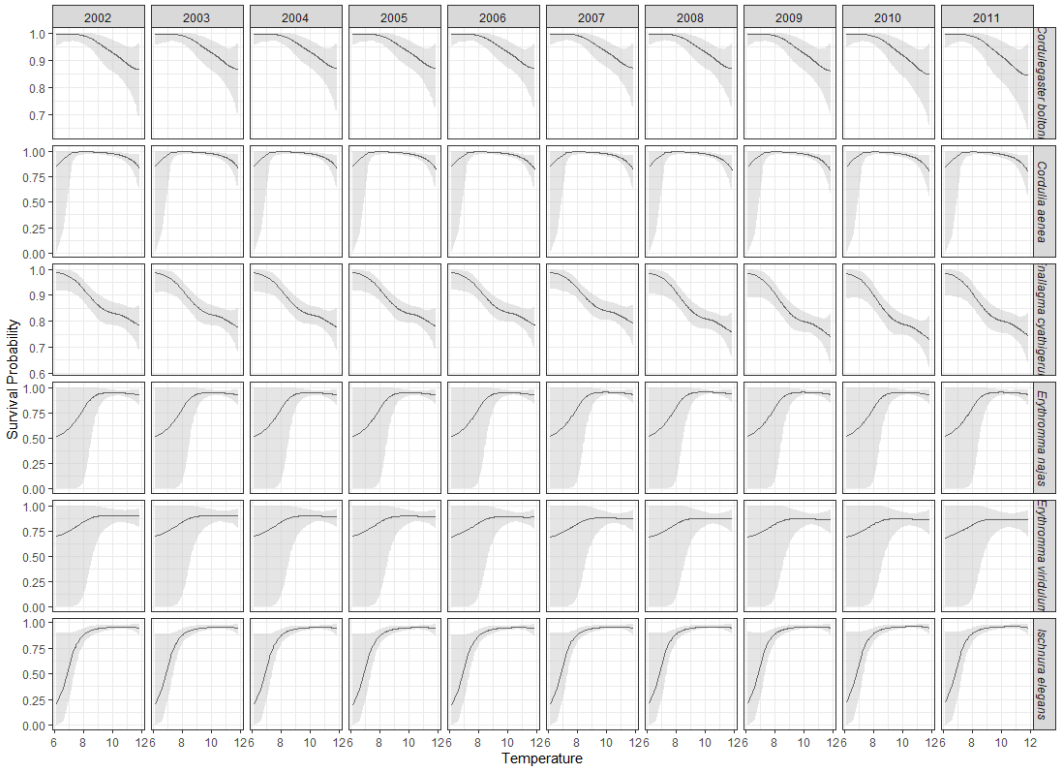


(b)

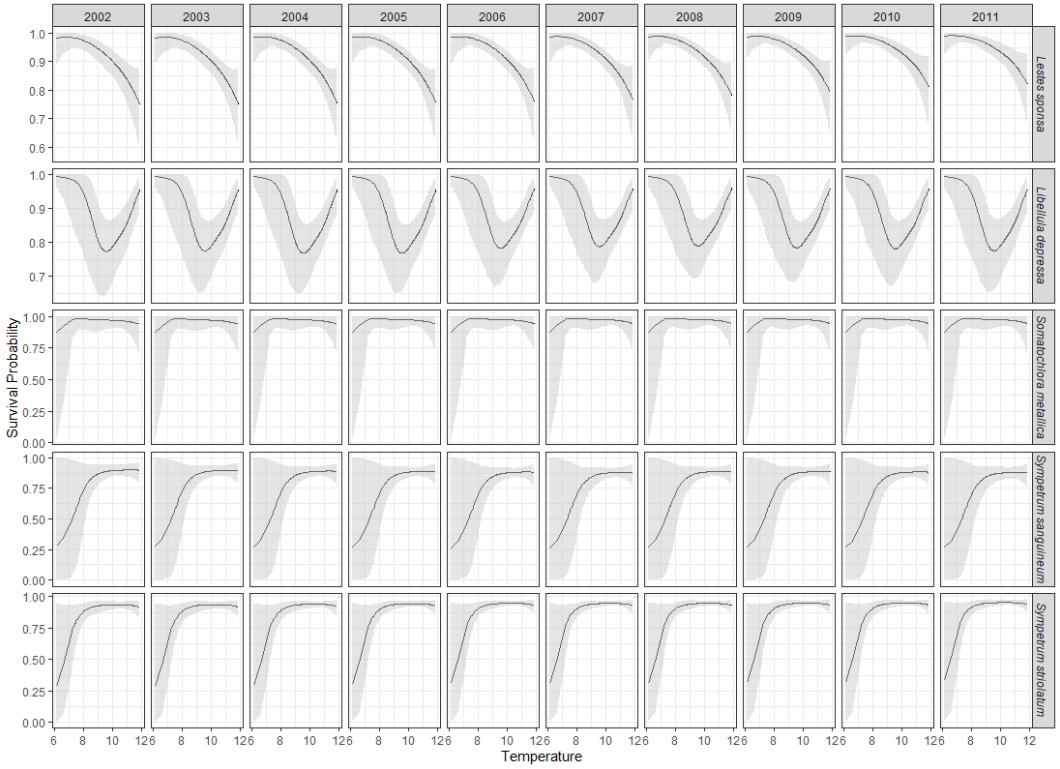
Figure 25: Relationship between colonization probabilities and temperature (° C) for 17 Odonata species from 2002 to 2010. Shaded gray area represents 95% credible intervals.


2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Aeshna juncea
Aeshna mixta
Anax imperator
Calopteryx virgo
Chalcolestes viridis
Coenagrion puella
Survival Probability
Temperature

(a)

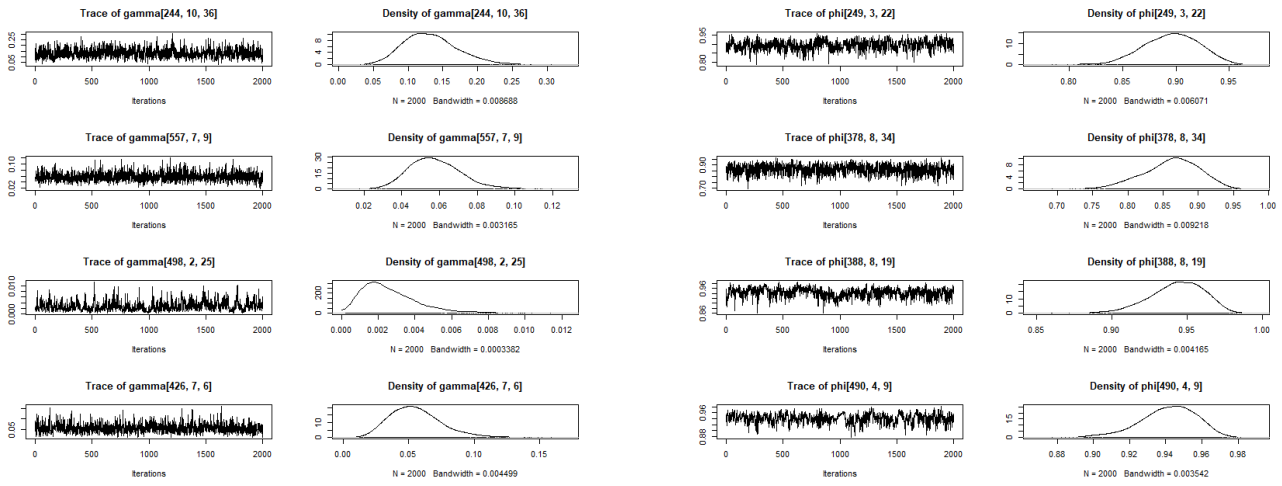

2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Cordulegaster boltoni
Cordulia aenea
Hallagma cyathigerum
Erythromma najas
Erythromma viridulum
Ischnura elegans
Survival Probability
Temperature

(b)



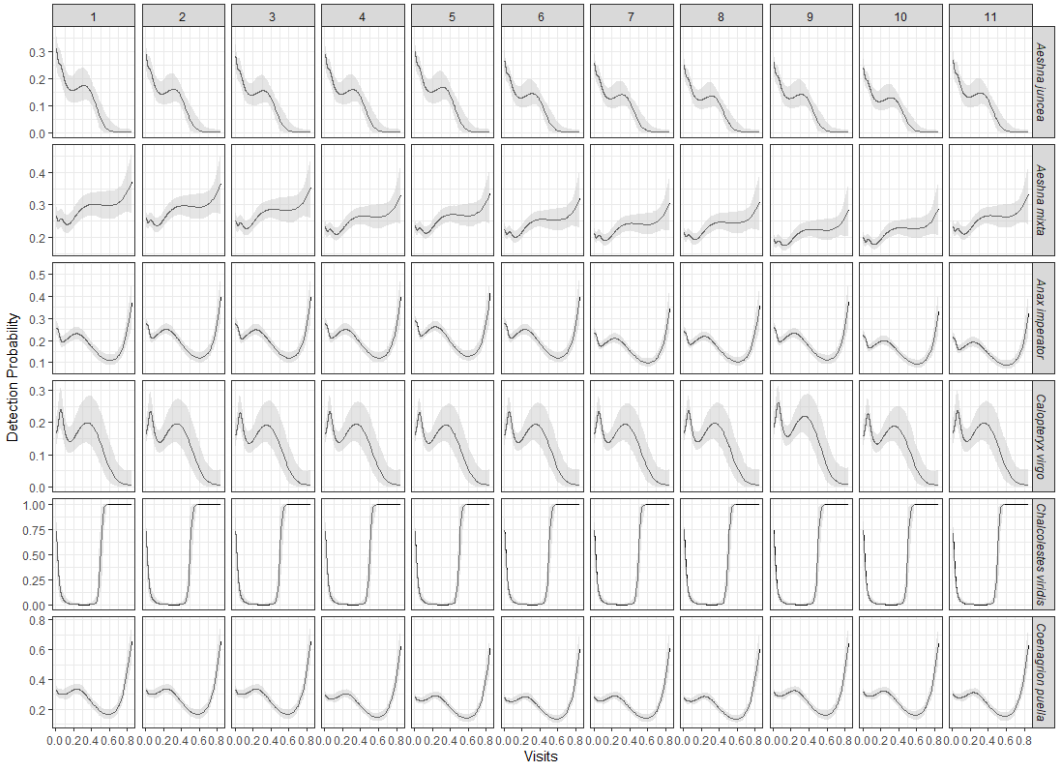
(a)

Figure 26: Relationship between survival probabilities and temperature (° C) for 17 Odonata species from 2002 to 2010. Shaded gray area represents 95% credible intervals.

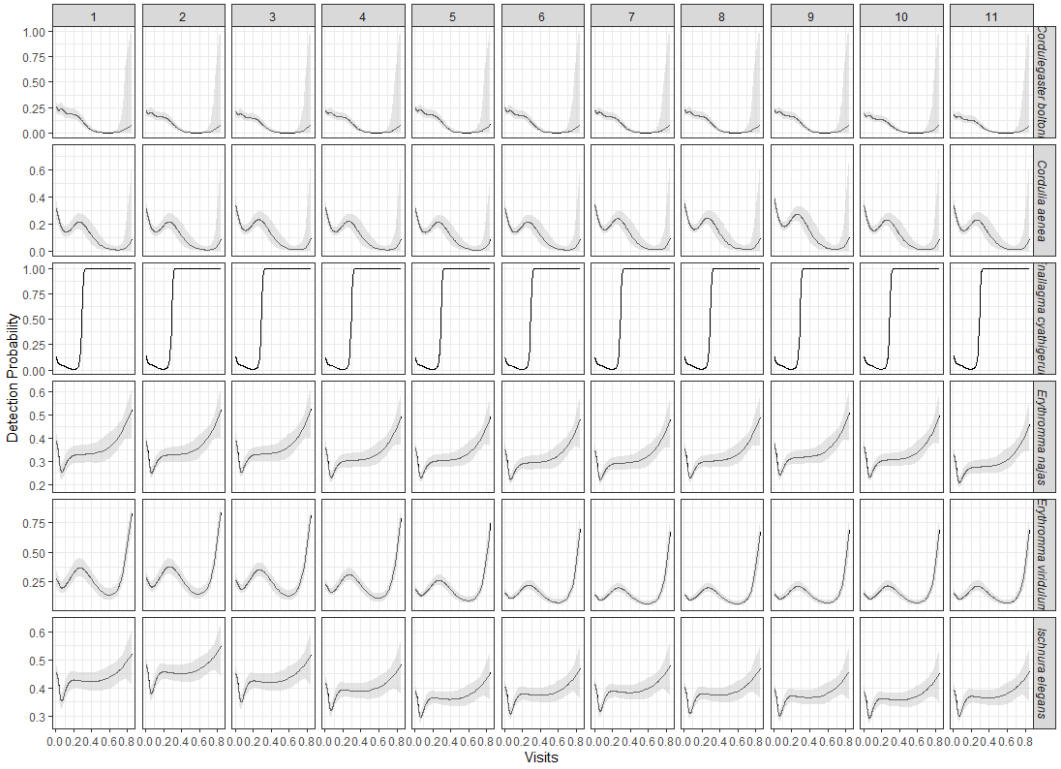


(a) (b)

Figure 27: Trace plots for a sample of the colonization (a) and survival (b) parameters for 8 different species at different sites during four years.


1
2
3
4
5
6
7
8
9
10
11
Aeshna juncea
Aeshna mixta
Anax imperator
Calopteryx virgo
Chalcolestes viridis
Coenagrion puella
Detection Probability
Visits

(a)


1
2
3
4
5
6
7
8
9
10
11
Cordulegaster boltoni
Cordulia aenea
Libellula cyathigeru
Erythromma najas
Erythromma viridulum
Ischnura elegans
Detection Probability
Visits

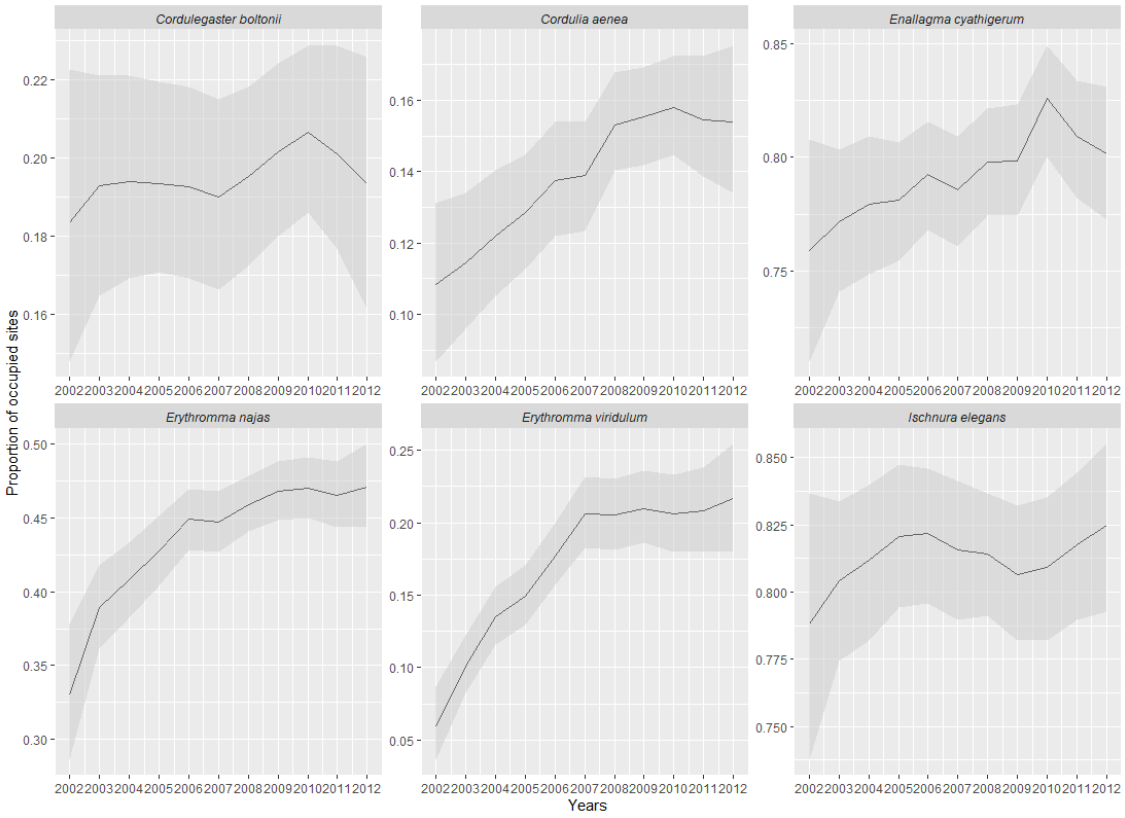
(b)

(a)

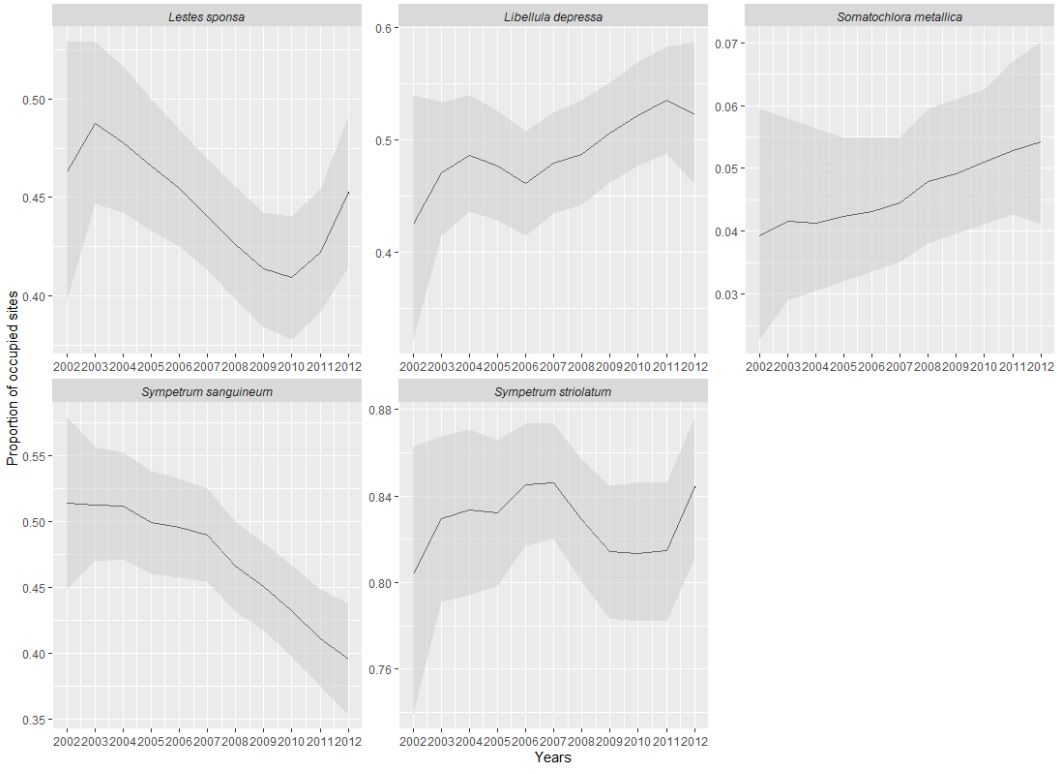
Figure 28: Relationship between detection probabilities and human activities index for Odonata species from 2002 to 2010. Shaded gray area represents 95% credible intervals.

Table 2: Estimated total growth rate for 41 Dragonflies and Damselflies species.

	Species	$\hat{\lambda}_{tot}$	95% Credible interval
1	Aeshna affinis	0.53	[0.03, 1.09]
2	Aeshna caerulea	1.46	[0.54 , 6.82]
3	Aeshna cyanea	0.97	[0.83 , 1.14]
4	Aeshna grandis	1.26	[1.09 , 1.45]
5	Aeshna juncea	0.74	[0.58 , 0.93]
6	Aeshna mixta	1.16	[1.01 , 1.32]
7	Anaciaeschna isoceles	1.11	[0.66, 1.92]
8	Anax imperator	0.97	[0.84, 1.13]
9	Brachytron pratense	1.60	[1.23, 2.10]
10	Calopteryx splendens	1.06	[0.87, 1.28]
11	Calopteryx virgo	1.47	[1.02, 2.15]
12	Ceriagrion tenellum	1.02	[0.60, 1.69]
13	Chalcolestes viridis	85.90	[3.45, 503.06]
14	Coenagrion hastulatum	0.90	[0.39, 2.11]
15	Coenagrion mercuriale	0.60	[0.29, 1.12]
16	Coenagrion puella	1.12	[1.00, 1.27]
17	Coenagrion pulchellum	1.50	[0.95, 2.40]
18	Cordulegaster boltonii	1.08	[0.77, 1.50]
19	Cordulia aenea	1.47	[1.06, 2.00]
20	Enallagma cyathigerum	1.06	[0.98, 1.16]
21	Erythromma najas	1.44	[1.19, 1.74]
22	Erythromma viridulum	4.00	[2.22, 7.10]
23	Gomphus vulgatissimus	38.04	[1.25, 201.61]
24	Ischnura elegans	1.05	[0.97, 1.14]
25	Ischnura pumilio	0.81	[0.27, 1.96]
26	Lestes barbarus	0.61	[0.00, 1.93]
27	Lestes dryas	1.62	[0.15, 7.48]
28	Lestes sponsa	0.98	[0.80, 1.18]
29	Leucorrhinia dubia	0.66	[0.38, 1.10]
30	Libellula depressa	1.25	[0.93, 1.70]
31	Libellula fulva	5.36	[2.41, 12.01]
32	Libellula quadrimaculata	1.29	[1.11, 1.51]
33	Orthetrum cancellatum	1.13	[0.96, 1.33]
34	Orthetrum coerulescens	1.12	[0.72, 1.73]
35	Platycnemis pennipes	0.96	[0.68, 1.32]
36	Pyrrhosoma nymphula	1.16	[1.02, 1.31]
37	Somatochlora arctica	0.65	[0.38, 1.09]
38	Somatochlora metallica	1.52	[0.80, 2.84]
39	Sympetrum danae	0.84	[0.66, 1.07]
40	Sympetrum sanguineum	0.78	[0.65, 0.93]
41	Sympetrum striolatum	1.03	[0.94, 1.14]



(a)



(b)

Figure 29: Odonata species estimated proportion of occupied sites from 2002-2012. Shaded gray area represents 95% credible intervals.

The following figures show the autocorrelation, density and trace plots for the Odonata flexible occupancy model under an aggregated number of records observational process.

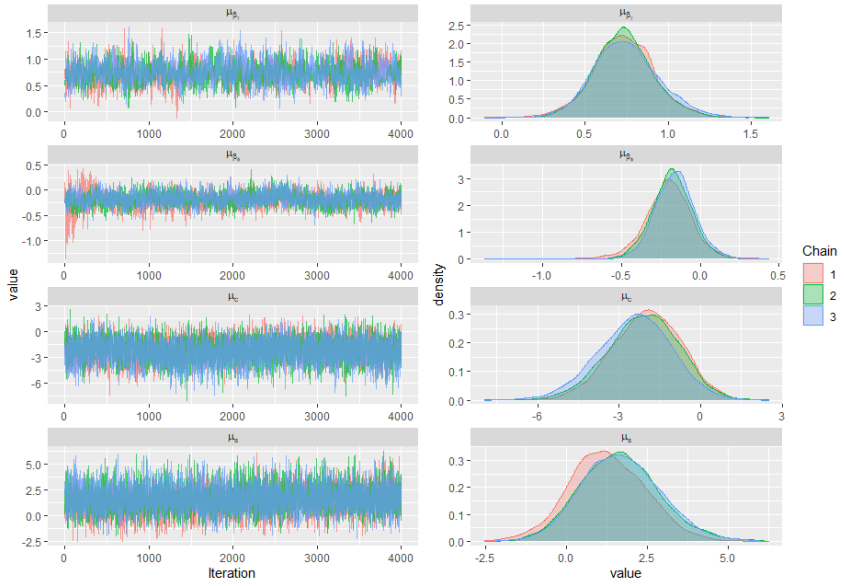


Figure 30: Density and traceplots for odonata flexible occupancy model.

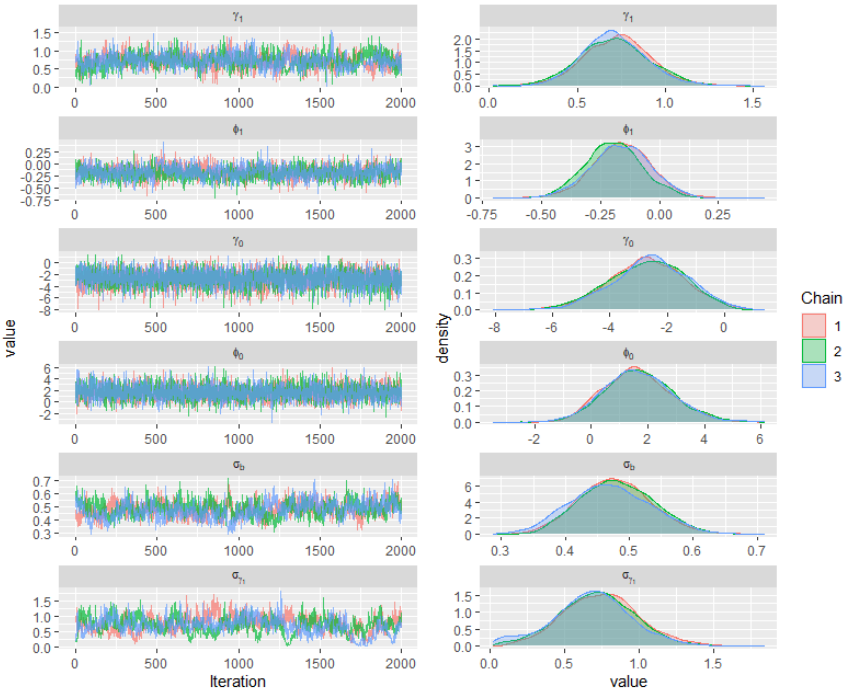


Figure 31: Density and traceplots for odonata flexible occupancy model.

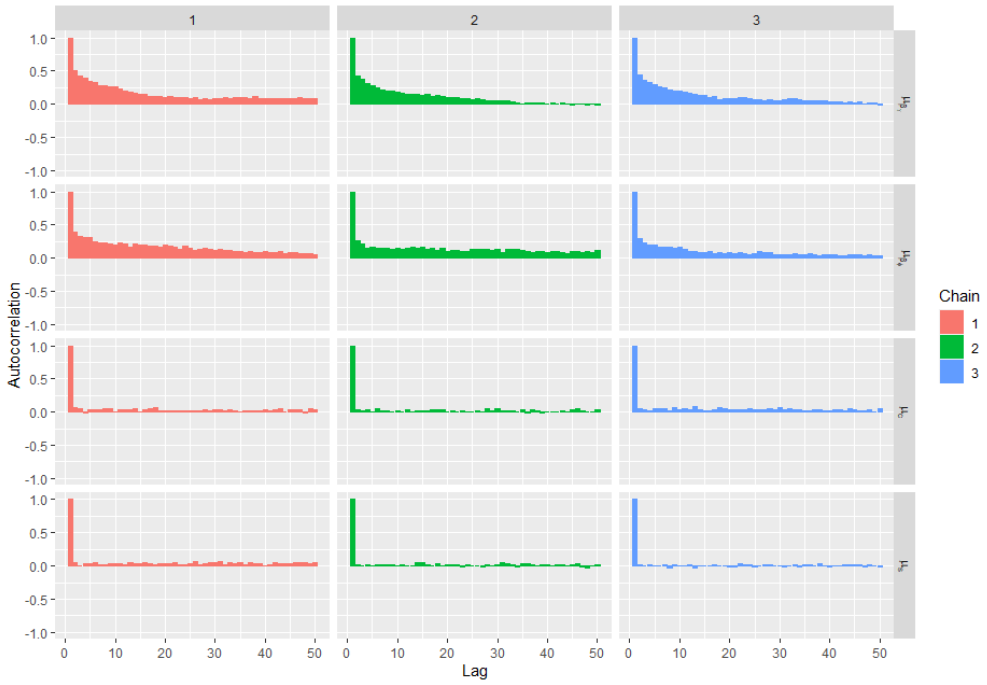


Figure 32: Autocorrelation plot for odonata flexible occupancy model.

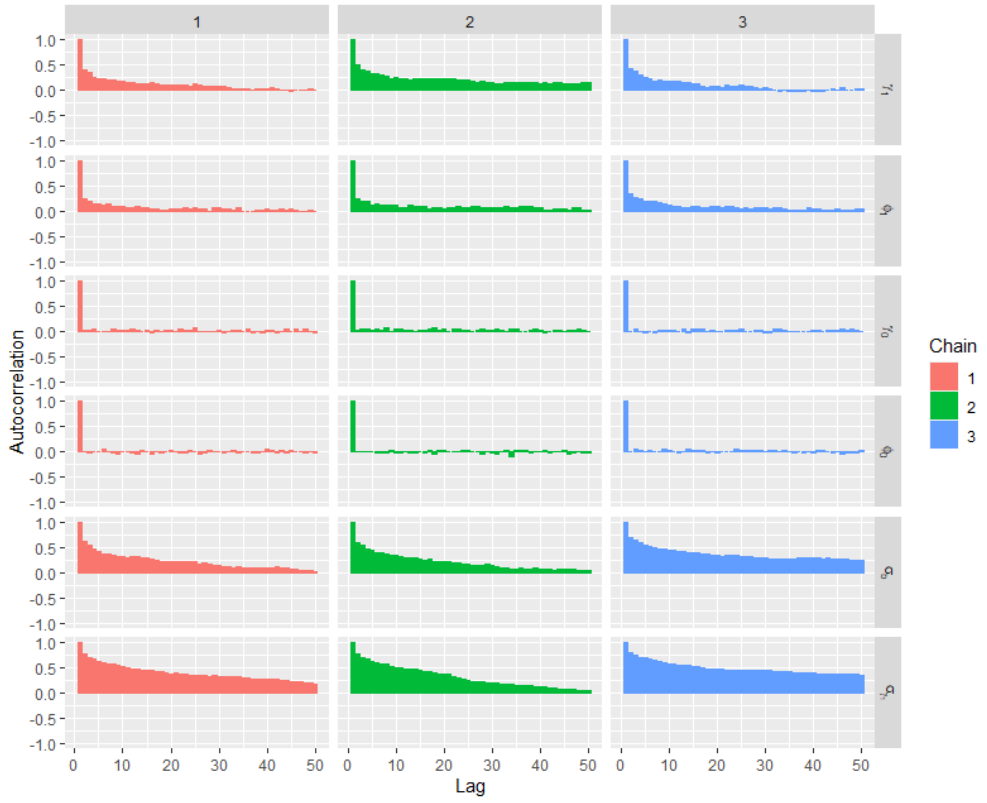
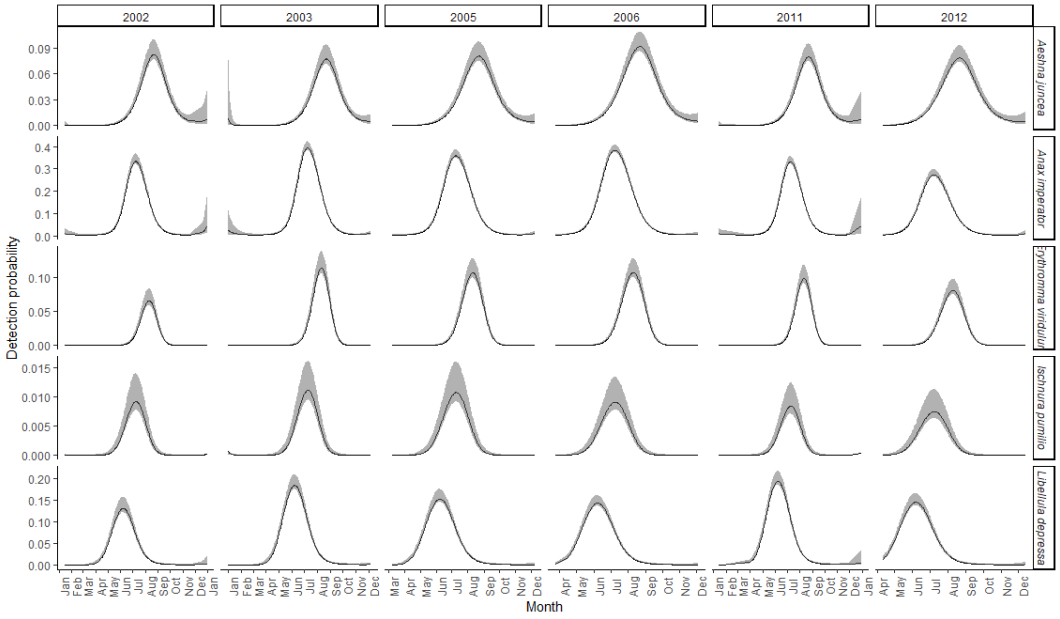


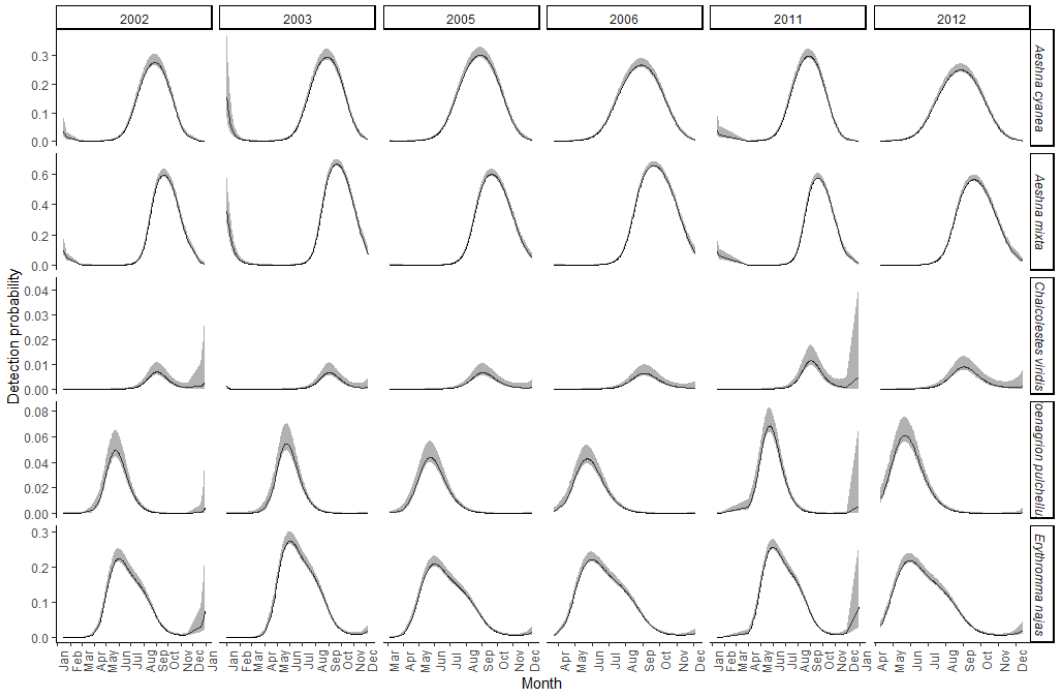
Figure 33: Autocorrelation plot or odonata flexible occupancy model.

D.4.2 Odonata flexible dynamic occupancy model presence/absence species records

The following figures show the results of fitting the proposed Odonata flexible dynamic occupancy model to the presence/absence occurrences data for a subset of species and years. This subset summarizes the main species-specific effects found among the 41 species analyzed with this model.

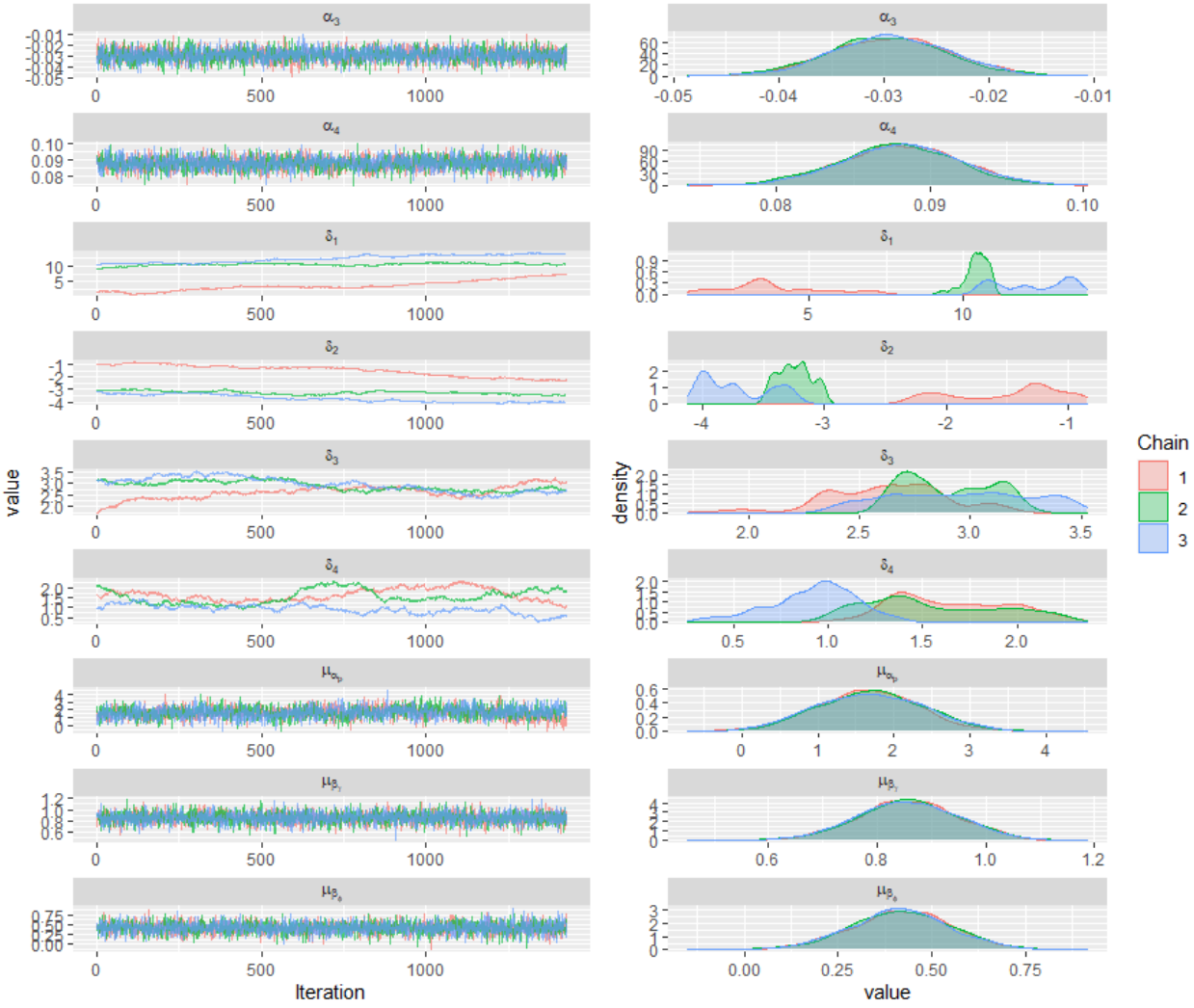


(a)

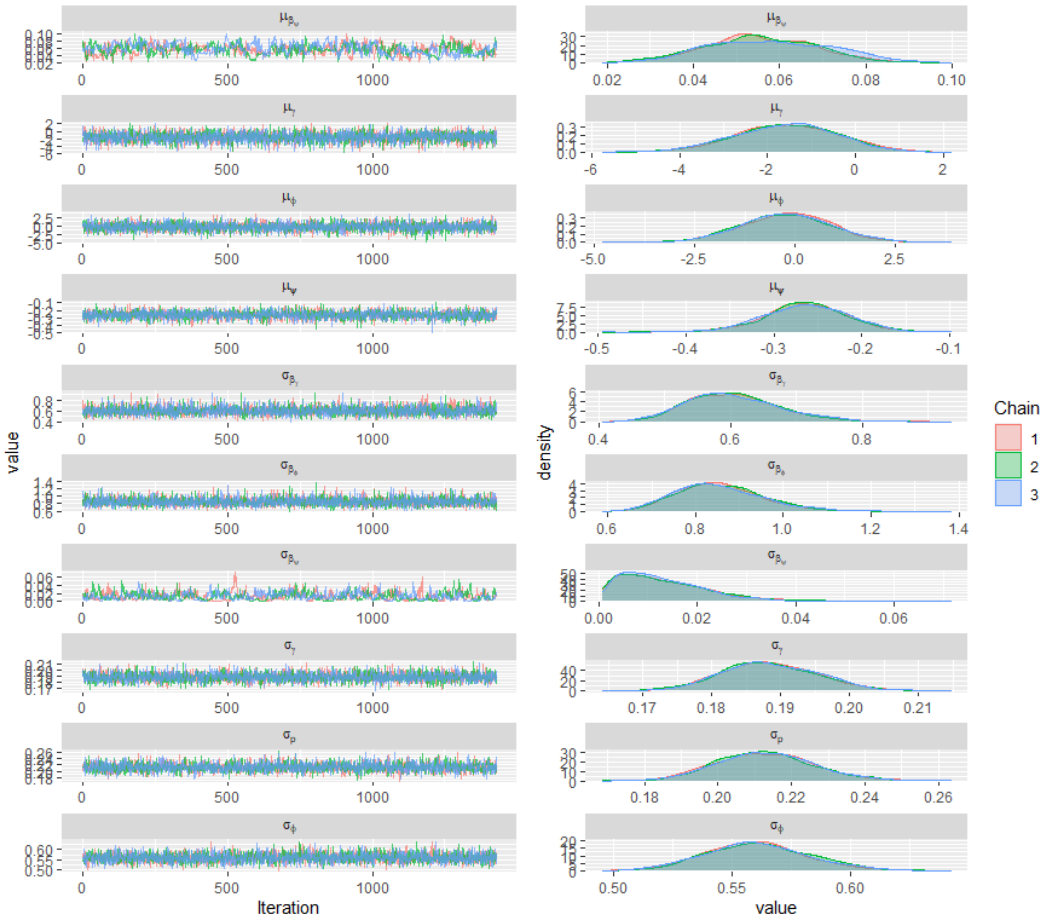


(b)

Figure 34: Relationship between detection probabilities and Julian date activities index for Odonata species. Shaded gray area represents 95% credible intervals.

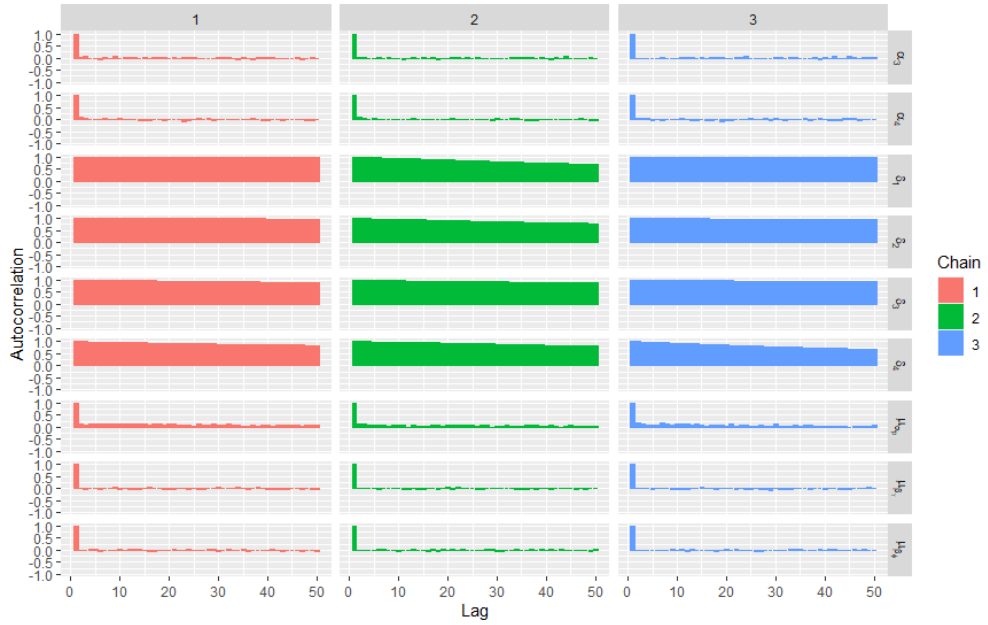

α3
α4
δ1
δ2
δ3
δ4
μα
μδ
μδ
value
density
Iteration
value
Chain
1
2
3

(a)

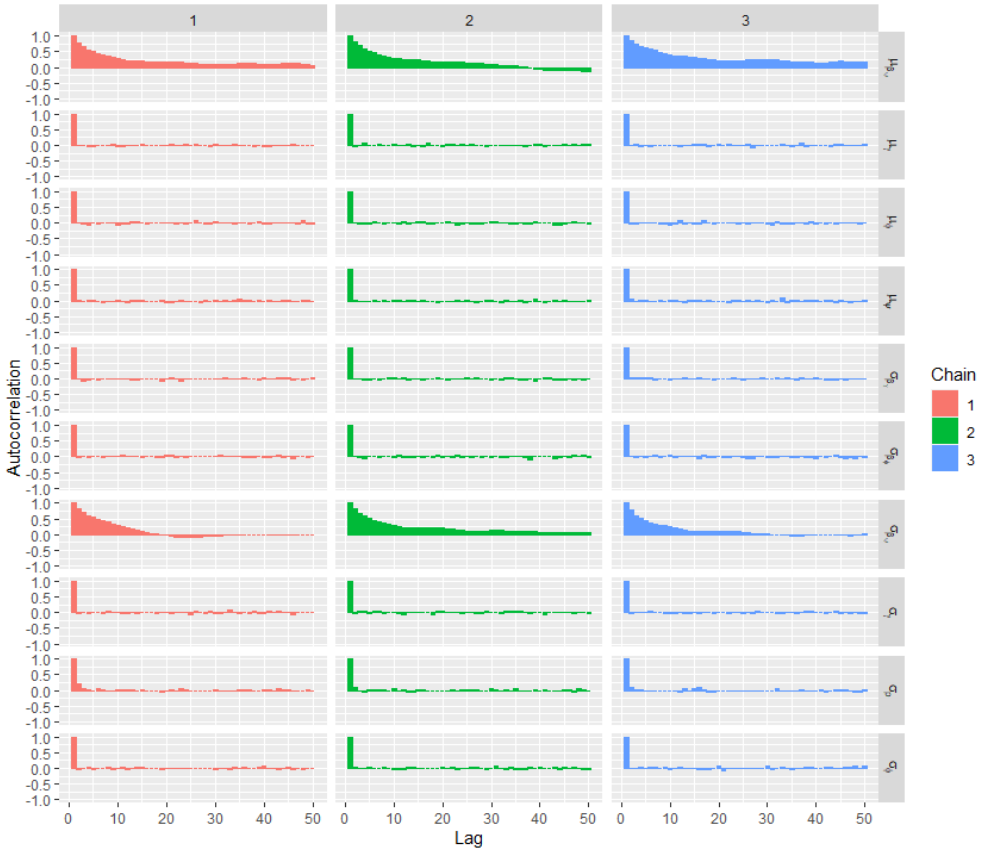


(a)

Figure 35: Density and traceplots for Odonata flexible occupancy model parameters with a presence/absence observational model structure from three independent chains.



(a)



(b)

Figure 36: Autocorrelation plots for Odonata flexible occupancy model parameters with a presence/absence observational model structure from three independent chains.