Bayesian nonparametric methods for modelling ecological data and stochastic processes for modelling species interactions

Alex Diana

School of Mathematics, Statistics and Actuarial Science
University of Kent

This dissertation is submitted for the degree of
Doctor of Philosophy

June 2020

I would like to dedicate this dissertation to my supervisors, Eleni Matechou and Jim Griffin, for their excellent guidance, direction and wisdom. I could not have imagined having two better supervisors for my PhD study. This piece of work would have been impossible without their help.

I would also like to dedicate this work to my parents and my sister, who have always supported the choices I made throughout my life, and to whom I express much love.

Finally, I would like to thank my girlfriend and my closest friends in these three years, who have contributed to make this experience a wonderful journey.

Acknowledgements

I would like to thank my supervisors, Eleni Matechou and Jim Griffin, for their help during my study. I gratefully acknowledge also the funding received towards my PhD from the The University of Kent with a vice-chancellor's graduate teaching assistantship and also the School of Mathematics, Statistics and Actuarial Science.

Abstract

In this thesis, we present four manuscripts, described in the second to fifth chapter. Chapter 2 presents a Bayesian nonparametric model for capture-recapture (CR) data collected at different sites and for several years. To estimate arrival and departure patterns at the different sites and years, we build an extension of the Dirichlet process, the Hierarchical Dependent Dirichlet process, which allows us to perform density estimation jointly across different sites and in the presence of covariates. In this case, we use a year-specific covariate, and model the correlation structure of the covariate across years using a multivariate Gaussian process. In Chapter 3, we present a model for estimating entry and exit patterns, as well as the population size, using count data (CD), by employing a Polya Tree (PT) prior. In Chapter 4 we present several extensions of chapter 3. More specifically, we extend the model to CR and to ring-recovery data and develop a joint model for CR and CD. In addition, we consider the case when multiple data-sets are modelled at the same time, by defining a hierarchical extension of the PT, which we define as Hierarchical Logistic PT. Finally, we extend the model to the case of long time series, by borrowing ideas from the Optional PT. Chapter 5 presents a spatial model to estimate interactions between multiple species using CR data. The model uses a vector of interaction point process (IPP), which allows us to estimate interactions between and within species. The use of an IPP leads to an intractable ratio of normalising constants (RNC), and hence we use the Monte Carlo Metropolis Hastings algorithm to approximate the RNC with an importance sampling estimate. The supplementary material for each paper is presented in the appendix.

Table of contents

1 Introduction **1**

2 A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK **5**

2.1 Introduction . . . 7
2.2 The existing model . . . 9
2.3 Theory . . . 10
2.3.1 Hierarchical Dependent Dirichlet Process mixtures . . . 10
2.3.2 Multivariate Gaussian Process (MGP) . . . 12
2.4 Bayesian nonparametric model for CR data collected at multiple sites and multiple years 14
2.4.1 Sampling scheme . . . 14
2.4.2 Arrival and Departure Process . . . 14
2.5 Application . . . 16
2.5.1 Simulations . . . 16
2.5.2 BTO's Constant Effort Sampling Scheme Data . . . 19
2.6 Discussion . . . 24
2.6.1 MCMC Algorithm . . . 26

3 A Polya Tree Based Model for Unmarked Individuals in an Open Wildlife Population **31**

3.1 Introduction . . . 33
3.2 The Polya Tree Prior . . . 34
3.3 Model . . . 35
3.4 Computational Notes . . . 37
3.5 Application . . . 38
3.6 Conclusion . . . 40

4 A general modelling framework for ecological data based on the Polya Tree prior **41**

4.1 Introduction . . . 43
4.2 Background . . . 45
4.2.1 Sampling schemes of population surveys . . . 45

4.2.2 The Polya Tree prior . . . 46
4.3 Univariate PT models . . . 47
4.3.1 Example: Cormack-Jolly-Seber . . . 48
4.4 Replicate PT . . . 51
4.5 Bivariate PT models . . . 52
4.5.1 Bivariate partitions . . . 52
4.5.2 Joint analysis of capture-recapture and count-data . . . 54
4.5.3 Ring-Recovery . . . 58
4.6 The hierarchical Logistic PT . . . 60
4.6.1 A hierarchical model across different data-sets . . . 61
4.6.2 Modelling long time-series of ecological data . . . 62
4.7 Conclusion . . . 66

5 A vector of softcore processes for modelling interactions between species using capture-recapture data. 67
5.1 Introduction . . . 69
5.2 Interaction point processes . . . 71
5.2.1 Vectors of point processes . . . 72
5.2.2 Bayesian inference for interaction point processes . . . 75
5.3 Spatial capture-recapture model . . . 77
5.4 Simulation study . . . 78
5.5 Application: CR data on leopards and tigers . . . 79
5.6 Discussion . . . 79

6 Discussion 81

References 85

Appendix A Supplementary material for A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK 91
A.1 Posterior distributions . . . 91
A.1.1 Posterior distributions for μ and Σ . . . 91
A.1.2 Posterior for b . . . 92
A.2 Posterior update of α and γ . . . 93
A.3 Results on Multivariate Gaussian process . . . 95
A.4 Prior distributions . . . 96
A.5 Results . . . 96

Appendix B Supplementary material for A general modelling framework for ecological data based on the Polya Tree prior 101
B.1 Partitions . . . 101
B.2 Proof of Lemma 1 . . . 102
B.3 Likelihoods for capture-recapture data . . . 103
B.3.1 Likelihood of the CJS . . . 103
B.3.2 Likelihood of open-population CR . . . 103
B.4 Case studies . . . 104
B.4.1 Spoonbills . . . 104
B.4.2 Ring-Recovery . . . 104
B.4.3 Newts . . . 104
B.4.4 Moths . . . 105
B.5 Posterior inference . . . 106
B.5.1 Cormack-Jolly-Seber model . . . 106
B.5.2 Joint model for capture-recapture and count data . . . 106
B.5.3 Ring Recovery model . . . 107
B.5.4 Hierarchical model . . . 108
B.5.5 Long count data model . . . 110

Appendix C Supplementary material for A vector of softcore processes for modelling interactions between species using capture-recapture data. 113
C.1 Simulation algorithm for interaction point processes . . . 113
C.2 MCMC Algorithm . . . 114

Chapter 1

Introduction

This thesis considers the results from four papers considering novel Bayesian methods and their applications to statistical ecology, more specifically, the study of wildlife populations. The study of wildlife populations is becoming increasingly important as in this period of rapid environmental change it is paramount to be able to monitor the changes in populations using statistical models that utilise the data arising from different sampling protocols.

When modelling wildlife populations, several quantities are of interest to conservation scientists and researchers. One of the main quantities of interest is usually population size. In the case of closed populations, that is, when individuals are assumed to remain at the study area for the entire duration of the study, population size corresponds to the number of individuals present at the site during the study. For open populations, that is, when individuals are allowed to enter and leave the study area, population size is interpreted as the number of individuals that become available for detection at least once. In addition, when considering open populations, researchers may also be interested in estimating the arrival/departure or birth/death patterns of the population, that is, when individuals are expected to enter or exit the study area and their corresponding length of stay or lifespan, depending on the application. In the four manuscripts, the terms arrival/departure and entry/exit or birth/death are used interchangeably, as according to the data considered births (deaths) are indistinguishable from entry (exit). Throughout this introductory chapter we use the terms entry and exit, which is also the terminology used in the fourth chapter. We always assume throughout the thesis that individuals are only allowed to enter and exit the study area once so that emigration is permanent. Also, when sampling is performed in more than one locations, the data can be used to estimate the spatial patterns of individuals. The term spatial pattern encompasses more generally the density of the individuals in the site area, or how the density changes according to the characteristics of the site, which can be expressed by covariates. However, spatial data have also been used recently to estimate the interactions between individuals [74].

The type of sampling protocol used depends on the time and funding constraints and on the characteristics of the population. One of the most used sampling protocols, which we consider in three of the manuscripts, is capture-recapture (CR). CR data are collected by visiting a site on

several occasions, called sampling occasions, and each time capturing a number of individuals and marking the individuals never caught before. However, as physically marking individuals is often difficult, in the case where the individuals can be individually identified using natural marking, the CR protocol can be performed via camera-traps, by taking pictures of individuals. The other two sampling protocols considered in our work are count data (CD) and ring-recovery (RR). The CD protocol is similar to CR, but in each sampling occasion, a number of individuals is counted without individually identifying the individuals. The RR protocol relies on marking a number of individuals in each sampling occasion and recovering them only when they are dead without attempting to recapture them alone. The advantage of the CR protocol over CD and RR is the information available at the individual level, which allows for more precise estimates of population size, capture probability and arrival and departure patterns. In fact, as opposed to CR, it is more difficult to separately estimate all three quantities in the case of CD and RR [2]. In fact, in the third chapter we use CD to estimate arrival and departure patterns by relying on informative prior distributions, while in the fourth chapter we use RR to estimate age-specific survival patterns by using two data-sets, where for one of the two data-sets the individuals have known age at the time of marking.

The first three papers focus on the study of entry and exit patterns. In particular, we assume a density for the individual arrival and departure times and we use a Bayesian nonparametric (BNP) approach for modelling this density. BNP is an increasingly popular field in Bayesian statistics and a vast literature is present especially for BNP approaches to density estimation. According to the BNP approach, a prior is defined on an infinite dimensional space, as the number of parameters is infinite compared to the parametric case, where this number is finite. The benefit of the BNP approach is its flexibility, as the complexity of BNP grows with the number of observations.

The second chapter relies on the popular Dirichlet process (DP) prior [28], which can be seen as the start of the nonparametric literature. Although this prior is of limited use on its own because its draws are discrete, it is common to use the DP as the mixing measure of the parameters of a continuous kernel to define a nonparametric prior on continuous densities. The DP has been extended in many directions. Two of the most notable extensions are the Dependent Dirichlet process (DDP), which allows us to introduce covariates in the process [51], and the Hierarchical Dirichlet process (HDP) [84], which allows us to model different data-sets hierarchically. In the second chapter, we are interested in performing density estimation at different sites and in presence of covariates, and hence to achieve this we combine the idea of the DDP and the HDP and define the Hierarchical Dependent Dirichlet process (HDDP).

The third and the fourth chapter use the Polya tree (PT) prior [29, 48], which is another common prior for distributions in the nonparametric literature. The PT prior is an extension of the DP and the advantage over the DP is that it allows us to model continuous distributions. The PT works by defining a prior distribution on the masses of sets of nested partitions on the sample space. Inference with the PT prior is usually performed by truncating the process using only a finite number of sequences of these sets, until a given level of resolution is required. This is especially suited in statistical ecology, as it is not possible to obtain information at resolutions finer than the one defined by the sampling

occasions, and hence the sampling occasions provide a natural way to define the resolution of the PT. The third chapter is concerned with the estimation of arrival and departure patterns using CD, while the fourth chapter extends these ideas to the case of CR, RR and to data collected at different sites or different years. In this paper, we also define two extensions of the PT: the replicate PT (RPT), which allows us to set constraints on the distributions, and the Hierarchical Logistic PT (HLPT), which allows us to model different distributions hierarchically, in a similar fashion to the HDP.

In the fifth chapter, we are concerned with estimating interactions between different wildlife populations. To model the interactions between individuals, we rely on the theory of spatial point processes (PP) and consider the latent locations of the individuals as realizations from a PP. PPs are a vast area of research in statistics and they are a modelling tool when the objects of interest are the positions of points in space [75, 16]. The starting point for modelling point patterns is the Poisson process, which is simply characterised by an intensity function that determines the number of points in each subset of the space. The Poisson process serves as a model class for the no interaction case in spatial PPs. In fact, even if more flexible approaches need to be considered, the Poisson process is a useful reference process to rely on to build other processes. One extension of the Poisson process is the class of Cox processes, which are processes for clustered or aggregated point patterns and are obtained by taking the realization of a random field as the intensity function of a Poisson process. Examples of Cox processes are Neyman-Scott processes [63], shot noise Cox processes [7] and Log Gaussian Cox processes [60]. Another extension of point processes is the class of interaction PPs (IPP), and are a useful modelling tool when interactions between points are considered. The study of IPPs started with Markov PPs in statistical physics for studying particle interaction systems [80, 70]. Markov PPs are a special case of IPPs, which are defined by constructing a process with a density function with respect to a Poisson process and letting the density function determine the interaction between points. Inference for IPPs is usually difficult as the normalising constant of the likelihood is unknown except for trivial cases (such as the Poisson processes). In frequentist settings, the most common approach is to use a maximum pseudo-likelihood estimator (MPLE) in the place of the classic maximum-likelihood estimator (MLE). This procedure is justified by convergence theorems for MPLE estimators. In Bayesian settings, the use of IPPs leads to inference in the presence of an intractable ratio of normalising constants. A vast literature is present on inference procedures that allow to deal with intractable ratios of normalising constants in the Bayesian setting, as this problem arises also when working with exponential random graph models. The most notable method is the Exchange sampler algorithm [62], while other popular methods are path sampling [64] and the Monte Carlo Metropolis-Hastings algorithm [50].

Chapter 2

A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK

Abstract

Environmental changes in recent years have been linked to phenological shifts, which in turn are linked to the survival of species. The work in this paper is motivated by capture-recapture data on blackcaps collected by the British Trust for Ornithology as part of the Constant Effort Sites monitoring scheme. Blackcaps overwinter abroad and migrate to the UK annually for breeding purposes. We propose a novel Bayesian nonparametric approach for expressing the bivariate density of individual arrival and departure times at different sites across a number of years as a mixture model. The new model combines the ideas of the hierarchical and the dependent Dirichlet process, allowing the estimation of site-specific weights and year-specific mixture locations, which are modelled as functions of environmental covariates using a multivariate extension of the Gaussian process. The proposed modelling framework is extremely general and can be used in any context where multivariate density estimation is performed jointly across different groups and in the presence of a continuous covariate.

2.1 Introduction

Describing abundance, distribution and phenology of wild animals is key to understanding the drivers of populations and therefore to designing effective conservation strategies. During this period of rapid environmental change and degradation of the natural world, it is important to develop statistical methods that utilise currently available data to provide increased understanding of species dynamics and the impact of climate change on species. The annual cycle of migratory species makes them particularly sensitive to impacts of climate change, but also makes them challenging to study. In this paper we study the phenology and abundance of migratory birds in Great Britain, in order to better understand their populations and the impacts of climate. Phenology has been linked to the survival of species, with populations that did not show a phenological response to climate change declining, as birds fail to breed at the time of maximal food abundance [6, 58].

Capture-recapture (CR) is one of the most commonly employed protocols in ecology to estimate the main demographic parameters of a wildlife population. CR is performed by visiting a site several times and capturing and marking a subset of the individuals before releasing them back into the population. The work in this paper is motivated by CR data on birds, collected by the British Trust for Ornithology (BTO) at different sites since 1983 as part of the Constant Effort Sites (CES) monitoring scheme, described in Peach et al. [66]. Specifically, we consider data on blackcaps, that are known to

breed in the UK but overwinter in Africa. The CES scheme has been already adopted across Europe. For instance, [26] used data from constant effort ringing protocols in Western Europe to assess the productivity of several bird species, while [43] estimated annual survival from similar data.

Individuals of the same species are expected to share many of their migratory behaviours even if breeding at different sites. This led us to adopt a joint modelling approach for their migration pattern across different sites. Such a modelling approach is also motivated from the fact that fewer than 15 birds were captured at least once and fewer than 5 were captured more than once in 80% of the sites. Such small sample sizes prohibit us from studying phenology or estimating population sizes at these sites when modelling data at each site separately, as for example using the approach of [55] (MC17). Instead, a joint modelling approach enables us to study migration patterns across the UK without being limited to only using sites where large numbers of individuals are caught. There is also considerable interest in determining the effect of changes in environmental conditions due to climate change on the migration patterns of animals, including birds. In order to link phenological changes to environmental conditions, we introduce a year-specific weather covariate, specifically the average North-Atlantic Oscillation (NAO), in modelling phenology, expressed through the arrival and departure density of individuals at the different sites.

In ecological applications, parametric models often entail assumptions on the population studied that are difficult to assess in practice. In particular, as wildlife populations typically present considerable heterogeneity, the use of parametric models in ecology can be prone to model misspecification. As a result, Bayesian nonparametric models have recently been more frequently adopted. The most popular nonparametric prior employed in these applications is the Dirichlet Process (DP) prior of [28]. The DP is a prior for densities that can be centered around any continuous distribution. However, as samples from the DP are always discrete distributions, the DP is often convolved with a continuous kernel when used as a prior for continuous distributions. The result of this convolution is called a DP mixture and gives rise to a mixture distribution with an *a-priori* infinite number of mixture components. Thanks to this flexibility, this prior has been adopted in several ecological applications. First, Dorazio et al. [22] extended the N-mixture model of [77] with the DP mixture of normals to allow for a variable number of mixture components in the prior distribution of population sizes. Ford et al. [30] used a DP mixture to model heterogeneity in capture and survival probabilities in a closed population of whales. Manrique-Vallier [54] used a DP mixture of product-Bernoulli distributions to estimate the size of a closed population in multiple CR data. Finally, MC17 used the Gamma process [47], which can be expressed in terms of the DP, to model the arrival intensity of a population given a CR dataset.

Our model extends MC17 by borrowing ideas from two other popular nonparametric priors, the Hierarchical Dirichlet process (HDP) of Teh et al. [84] and the single-p Dependent Dirichlet process (DDP) of MacEachern [51]. The former is an extension of the DP for data collected in several groups, while the latter is an extension of the DP that allows the introduction of covariates. Combining these two models, we define the Hierarchical Dependent Dirichlet process (HDDP), which can be used

as the mixing measure of a continuous kernel to estimate densities as functions of continuous and categorical covariates.

As a result, our model is completely flexible in the sense that it assumes a mixture distribution with an *a-priori* infinite number of mixture components for the arrival and departure distribution at each specific site and in each year. Moreover, as a result of the clustering properties of the model, these mixture components can be shared across different sites. The ecological interpretation is that birds at different sites can belong to the same cohort, sharing similar migration behaviour, which in the model equates to one of the mixture components. Thus, even if there is no information available on the number of cohorts of birds with similar migratory behaviour, the model can naturally adapt to any number of cohorts, by varying the number of mixture components in each site-specific density.

The paper is organized as follows. In Section 2.2 we describe the existing model of MC17. In Section 2.3 we introduce the mathematical concepts necessary to define the model presented in this paper. In Section 2.4 we define the new model proposed. The results of fitting the model to simulated data and to the BTO data are presented in Section 2.5. Section 2.6 concludes the paper and introduces some potential future directions. The details of the sampler are presented in the appendix.

2.2 The existing model

The model of MC17 performs inference from a single CR dataset. As mentioned in the introduction, CR data are collected by capturing individuals present at the site during K repeated sampling occasions. The data can be summarised in the form of a matrix H, with individual capture histories of the D caught individuals represented in the rows and the K capture occasions represented in the columns of the matrix. The capture history of individual i, H_i, corresponding to the i-th row of H, has k-th element equal to 1 if the individual was caught at the k-th sampling occasion, and equal to 0 otherwise.

The probability of capturing an individual that is present, p, is assumed to be constant across sampling occasions and common between individuals. The population size, which corresponds to the overall number of individuals that visited the site, is denoted by N.

Moreover, the model assumes that birds can enter the site at any continuous time, ζ, called the arrival time, and stay for a time δ, referred to as length of stay. The arrival time of each individual is sampled from a Poisson process with intensity $\nu(\zeta)$, which is taken to be a mixture of normal distributions $\nu(\zeta) = \int_{-\infty}^{\infty}\int_{0}^{\infty} \mathrm{N}(\zeta|\mu,\sigma^2)\, G(d\mu, d\sigma^2)$, where G is a Gamma process with shape αG_0 and scale τ, where $\alpha, \tau > 0$ and G_0 is a distribution function. The Gamma process is a completely random measure [46], whose Levy intensity is given by $\nu(ds, dx) = \exp(-\frac{s}{\tau})s^{-1}ds\ \alpha G_0(dx)$. It is closely related to the more popular DP, as the latter arises as normalisation of the Gamma process [28, 47], since the normalised random measure $P(\cdot) = \frac{G(\cdot)}{G(\Omega)}$, where Ω is the sample space, is distributed as a DP. Thanks to this property, the Gamma process can be decomposed as $G = \omega P$, where P is distributed as a DP with concentration parameter α and corresponds to the normalized density of the process and $\omega \sim \mathrm{Gamma}(\alpha, \tau)$ is the overall intensity of the process. The intensity $\nu(\zeta)$ can be

expressed as

$$\nu(\zeta) = \omega \underbrace{\int_{-\infty}^{\infty}\int_{0}^{\infty} \mathrm{N}(\zeta|\mu,\sigma^2)\, P(d\mu, d\sigma^2)}_{f_X} \tag{2.2.1}$$

Given G, the sample size N is distributed as a Poisson(ω) and the arrival times $\zeta_1,\ldots,\zeta_N$ are i.i.d. from f_X. The previous representation motivates the use of the intensity function, as it allows us to sample the population size and the arrival times conditionally independent on each other, as used for example in Wolpert and Ickstadt [88].

The length of stay is modelled by a survival function with piecewise constant hazard rate f_Y. The model can be expressed through latent variables in a hierarchical form as

$$\begin{cases} H_{ik}|\zeta_i,\delta_i,p \sim \text{Bernoulli}(pz_{ik}) & i=1,\ldots,N \quad k=1,\ldots,K \\ \zeta_i \overset{i.i.d.}{\sim} f_X & i=1,\ldots,N \\ \delta_i \overset{i.i.d.}{\sim} f_Y & i=1,\ldots,N \\ N|\omega \sim \text{Poisson}(\omega) \\ \omega|\alpha,\tau \sim \text{Gamma}(\alpha,\tau) \end{cases} \tag{2.2.2}$$

where z_{ik} is 1 if individual i is available at sampling occasion k (if $\zeta_i < t_k < \zeta_i + \delta_i$) and 0 otherwise.

In this paper, we jointly model arrival and lengths of stay non-parametrically and extend the work of MC17 by defining the Hierarchical Dependent Dirichlet process, which allows us to jointly model data collected

- at different sites, while sharing information between sites, using the properties of the HDP, and
- across different years, accounting for the effect of a continuous covariate on migration patterns, with correlation over time modelled using a multivariate Gaussian process.

2.3 Theory

2.3.1 Hierarchical Dependent Dirichlet Process mixtures

Before introducing the Hierarchical Dependent Dirichlet Process (HDDP) we present some standard models from the Bayesian nonparametrics literature.

The Dirichlet Process (DP), already mentioned in the introduction, is a random measure F with two parameters: a distribution G_0, called the base measure, and a positive real number α, called the concentration parameter, which tunes the variability of F around the base measure. It is denoted by DP(α, G_0) and it can be represented as $\sum_{i=1}^{\infty} \phi_i \delta_{\theta_i}$, with $\theta_i \sim G_0$ and the ϕ_is generated according to the stick-breaking process [82]. According to this process, given a sequence of variables $v_i \sim \text{Beta}(1,\alpha)$, the weights are generated as $\phi_i = \left(\prod_{j=1}^{i-1} v_j\right) v_i$. The θ_i are often referred to as cluster locations, while the ϕ_i are called weights.

A popular extension of the DP, designed to work with data collected in different groups, is the Hierarchical Dirichlet process of [84]. In order to model data from different groups, the HDP assumes a random measure, F_j, for the j-th group, and a global random probability measure F_0. The global measure is assumed to have a DP prior $F_0 \sim \mathrm{DP}(\gamma, G_0)$, while the group-specific random measures have independent DP prior $F_j \sim \mathrm{DP}(\alpha, F_0)$. Parameter γ tunes the variability of F_0 around G_0 and α tunes the variability of F_j around F_0. According to the stick-breaking representation, $F_0 = \sum_{i=1}^{\infty} \phi_i \delta_{\theta_i}$ and $F_j = \sum_{i=1}^{\infty} \pi_{ij} \delta_{\theta_i}$, and the distribution of the weights $\pi_{\cdot j}$ can be obtained in closed form as $\pi_{kj} = \left(\prod_{i=1}^{k-1} v_{ij}\right) v_{kj}$ where $v_{kj} \mid (\alpha, \phi_1, \ldots, \phi_k) \sim \mathrm{Beta}\left(\alpha\phi_k, \alpha\left(1 - \sum_{l=1}^{k} \phi_l\right)\right)$. Hence, every F_j is essentially obtained by keeping the same atoms of F_0 but redistributing the weights. No variation is induced in the cluster locations of the group-specific DPs.

The HDP is often conveniently described via the Chinese restaurant franchise (CRF) representation. According to the CRF representation, every observation in a group corresponds to a customer in a restaurant. In addition, the cluster locations of F_0 $\theta_1, \ldots, \theta_K \overset{i.i.d.}{\sim} G_0$, represent the dishes that can be served in the restaurant. To link the customers to the dishes, customer i in restaurant j is assigned to a table t_{ij}, while table t in restaurant j is assigned to dish k_{jt}. As a consequence, the dish served to customer i in restaurant j is $k_{jt_{ij}}$, which we define as c_{ij}. In addition, following the notation established in the literature, n_{jt} denotes the number of customers sitting at table t in restaurant j, m_k the number of tables serving dish k and M the total number of tables.

Thanks to the CRF representation, we can express the distribution of the allocations c_{ij} of customers to dishes by first defining the distribution of allocations t_{ij} of customers to tables and then the distribution of the allocations k_{jt} of tables to dishes. We can generate a sample from the CRF by sampling iteratively according to the following scheme. A new customer is assigned to an

$$\begin{cases} \text{existing table } t & \text{with probability } \frac{n_{jt}}{n_{jt}+\alpha} \\ \text{new table } t^\star \text{ serving existing dish } \theta_k & \text{with probability } \frac{\alpha}{n_{jt}+\alpha}\,\frac{m_k}{M+\gamma} \\ \text{new table } t^\star \text{ with new dish } \theta_{k^\star} \sim G_0 & \text{with probability } \frac{\alpha}{n_{jt}+\alpha}\,\frac{\gamma}{M+\gamma} \end{cases}$$

Likewise, a table is assigned to an

$$\begin{cases} \text{existing dish } \theta_k & \text{with probability } \frac{m_k}{M+\gamma} \\ \text{new dish } \theta_{k^\star} \sim G_0 & \text{with probability } \frac{\gamma}{M+\gamma} \end{cases}$$

The implied distribution on the c_{ij} is defined as $\mathrm{CRF}(\alpha, \gamma)$.

In the application to the BTO dataset, the birds are represented by the customers and the dishes correspond to the same migratory behaviour. Thanks to the CRF, groups of birds belonging to different sites can still share the same migratory behaviour if they are assigned to tables serving the same dish.

Another extension of the DP, designed to work with general covariates, is the Dependent Dirichlet Process (DDP) of MacEachern [51]. The DDP is a random measure F_x that can be written as

$$F_x = \sum_{k=1}^{\infty} \phi_k \delta_{\theta_i(x)}$$

where the cluster locations $\theta_i(x)$ are drawn independently from a stochastic process G_x, allowing F_x to depend on continuous covariates, if a continuous process, such as a GP, is assumed for G_x. The weights ϕ_i are drawn from the stick-breaking process as in the standard DP. More information on other nonparametric priors can be found in Hjort et al. [37].

In this paper, we perform density estimation conditionally on general covariates in a context where we have several groups. To achieve this, we combine the idea of the HDP and the DDP defining the Hierarchical Dependent Dirichlet Process (HDDP) as a HDP where the DP F_0 in the top level is replaced by a DDP.

Definition 2.3.0.1. *Let G_x be a stochastic process. The measures F_{jx} are said to follow a HDDP prior if, for each group j and each value x of the covariate*

$$\begin{cases} F_x = \sum_{k=1}^{\infty} \phi_k \delta_{\theta_k(x)} & \theta_k(x) \sim G_x \\ F_{jx} = \sum_{k=1}^{\infty} \pi_{kj} \delta_{\theta_k(x)} \end{cases} \tag{2.3.1}$$

where the weights ϕ_k and π_{kj} follow the same distribution as the weights of the HDP.

As we can see, the covariate, x, is introduced in the top-level and not in the group-specific DPs, which implies that the effect of the covariate is assumed to be the same across groups. However, as the DDP is assumed as a prior distribution for the group-specific measures, the weights are constant for each value of the covariate.

As opposed to a standard dataset analysed in [84], our data have an additional third dimension, given by the covariate x. However, as mentioned above, the covariate only affects the cluster locations. As a result, the CRF representation of the HDP can be used to describe the HDDP, since the covariate does not play a role when assigning the observations to clusters.

To conclude, we term as Hierarchical Dependent Dirichlet process mixtures the process obtained when the HDDP is used as the mixing measure of the parameters of a continuous kernel.

2.3.2 Multivariate Gaussian Process (MGP)

Before introducing the MGP we start by describing the univariate version. A GP is a prior distribution on a function $f : \mathbb{R}^q \to \mathbb{R}$, defined by the distribution of f evaluated on any finite collection of points $(x_1, \ldots, x_n)$. Specifically, we write $f \sim \mathrm{GP}(0, k)$ if, for any $(x_1, \ldots, x_n) : x_i \in \mathbb{R}^q$

$$(f(x_1), \ldots, f(x_n)) \sim \mathrm{N}(0, K((x_1, \ldots, x_n), (x_1, \ldots, x_n))$$

where $\{K((x_1, \ldots, x_n), (x_1, \ldots, x_n))\}_{ij} = \sigma^2 k(x_i, x_j)$ and k is a correlation function. In our case, we consider the Gaussian radial basis function $k(x, x') = \exp\left(-\frac{|x-x'|^2}{l^2}\right)$, with $l > 0$. For more information on Gaussian processes, see Rasmussen [73].

In the case of multivariate data, that is, if f is a function from $\mathbb{R}^q$ to $\mathbb{R}^p$, the MGP prior is defined based on the matrix normal distribution. A variable X is said to follow a matrix normal distribution $\mathrm{MN}(M,U,V)$ if $vec(X) \sim \mathrm{N}(vec(M), V \otimes U)$, where U is called the among row covariance matrix, V is called the among column covariance matrix and $\otimes$ is the Kronecker product.

The MPG prior on f is defined in the following way.

Definition 2.3.0.2. *Let Σ be a $p \times p$ positive definite matrix and μ an $n \times p$ matrix. We say that $f = (f_1, \ldots, f_p) \sim MGP(\mu, K, \Sigma)$ if*

$$((f_1(x_1), \ldots, f_p(x_1)), \ldots, (f_1(x_n), \ldots, f_p(x_n))) \sim MN(\mu, K((x_1, \ldots, x_n), (x_1, \ldots, x_n)), \Sigma).$$

This construction of the multivariate Gaussian process is also presented in [12]. By defining the MGP in terms of the matrix normal distribution, we have implicitly assumed that the cross-covariance matrix of the vector $((f_1(x_1), \ldots, f_p(x_1)), \ldots, (f_1(x_n), \ldots, f_p(x_n)))$ is separable, that is, it can be factorised as $\Sigma \otimes K((x_1, \ldots, x_n), (x_1, \ldots, x_n))$.

The advantage of this construction is that if we assume that the observations $y_i = (y_{i1}, \ldots, y_{ip})$ are generated according to

$$y_i \sim \mathrm{N}((f_1(x_i), \ldots, f_p(x_i))^T, \Sigma)$$

the posterior predictive distribution of new observations is available in closed form. Assuming we have available observations (x, y), the posterior predictive distribution for new observations with covariates $x^\star$ is:

$$y^\star \sim \mathrm{N}(vec(\mu_2) + (K_\star(K+I)^{-1} \otimes I)(y - vec(\mu_1)), ((K(x^\star, x^\star) + I) - K_\star(K+I)^{-1} {K_\star}^T), \Sigma)$$

where $K := K(x,x)$, $K_\star := K(x, x^\star)$.

In addition, we can account for the effect of covariates on the mean. If we have R covariates arranged in an $n \times R$ matrix X and coefficients $\beta = \begin{bmatrix} \beta_1^1 & \ldots & \beta_1^p \\ \vdots & & \vdots \\ \beta_R^1 & \ldots & \beta_R^p \end{bmatrix}$, we define the MGP as:

$$((f_1(x_1), \ldots, f_p(x_1)), \ldots, (f_1(x_n), \ldots, f_p(x_n))) \sim \mathrm{MN}(X\beta, K((x_1, \ldots, x_n), (x_1, \ldots, x_n)), \Sigma)$$

A useful property of this construction is that, if a prior distribution $\mathrm{MN}(b, B, \Sigma)$ is assumed for β, the marginal distribution of f is still a MGP prior of the form $\mathrm{MGP}(K^\star(K^{-1}XB^\star)B^{-1}b, K^\star, \Sigma)$, with $K^\star = (K + XBX^T)^{-1}$ and $B^\star = K^{-1}X(X^TK^{-1}X + B^{-1})^{-1}$. The calculations can be found in the supplementary material [21].

2.4 Bayesian nonparametric model for CR data collected at multiple sites and multiple years

The data can be expressed in the form H_{ijy}, where H_{ijy} is the capture history, defined in Section 2.2, of individual i at site j in year y, and we perform sampling at J sites in Y different years. At site j and year y, captures take place on C_{jy} sampling occasions at times $t_1^{jy}, \ldots, t_{C_{jy}}^{jy}$. Sampling times and the number of sampling occasions may differ across sites and years. We denote by x_y the value of the year-specific environmental covariate associated with year y. The site and year specific covariate associated with capture probability at site j and year y is denoted by λ_{jy}.

2.4.1 Sampling scheme

Capture probabilities are modelled using a logistic mixed effects model, where the site-specific intercept is assumed to be constant across years in the same group and all intercepts share a common prior distribution. The model for capture probability at site j in year y can be written as

$$\begin{cases} \text{logit}(p_{jy}) = \alpha_j^p + \lambda_{jy}\beta^p \\ \beta^p \sim \text{N}(0, B^p) \\ \alpha_j^p \sim \text{N}(a_0^p, A_0^p) \end{cases}$$

where B^p is the prior variance of β^p and a_0^p, A_0^p are chosen according to expert knowledge.

The choice of a mixed effects model is motivated by the study design of the CES scheme, according to which, sampling at the different sites is performed with the same effort. However, additional site characteristics, such as habitat and structure of the site, present an additional source of variation affecting capture probability that is not explained by the covariate, but instead modelled by the site-varying intercepts.

2.4.2 Arrival and Departure Process

We denote by ζ_{ijy} and δ_{ijy} respectively the arrival time and length of stay of individual i at site j in year y. We do not work directly with arrival and departure times because these two quantities do not lie in $\mathbb{R}^2$ (departure is obviously always later than arrival) and this would imply the need to work with a bivariate truncated normal, for which conjugate schemes are not available, resulting in computationally intensive inference. Instead, we choose to work with arrival times and a transformation of the length of stay, $\eta := h(\delta)$, in order to make the latter lie in $\mathbb{R}$. Although the logarithm is the common choice, it would lead to a lognormal behaviour in the right tail once we assign a normal prior distribution to $h(\delta)$, as the tails of the DP mixture behave approximately as the tails of the kernel. In order to have a normal behaviour also in the right tail, we choose $h(x) = \begin{cases} \log(x) & x \leq 1 \\ x - 1 & x > 1 \end{cases}$.

Borrowing ideas from MC17, we assume that for each site, arrival times and transformed lengths of stay are drawn from a Poisson process with non-homogeneous intensity ν_{jy}, modelled as

$$\nu_{jy}(\zeta,\eta) = \omega_j \int \mathrm{N}(\zeta,\eta|\mu_{x_y},\Sigma)dP_{jy}(\mu_{x_y},\Sigma) \tag{2.4.1}$$

where P_{jy} is the year and site-specific mixing measure of the parameters μ_{x_y} and Σ of the normal distribution, and ω_j is the site-specific intensity. The link with MC17 is clear if we compare (2.4.1) with (2.2.1). The bivariate density ν_{jy} of arrival times and lengths of stay is allowed to be site and year dependent, by replacing the DP with a HDDP, unlike MC17, who use a univariate DP mixture.

To achieve this, we define $\theta = (\mu,\Sigma,\beta)$, where μ is the $Y \times 2$ matrix of all the means μ_{x_y} of arrival and departure times for each covariate value, Σ is the 2×2 covariance matrix, β is an $R \times 2$ matrix expressing the trend of the means across the years and R is the dimension of the year-specific covariate (including the intercept). The prior distributions for these quantities are

$$\begin{cases} \mu \sim \mathrm{MGP}(X\beta, K(y,y), \Sigma) \\ \Sigma \sim \mathrm{IW}(\nu_0, \Sigma_0) \\ \beta \sim \mathrm{MN}(b, B, \Sigma) \end{cases} \tag{2.4.2}$$

where IW is the inverse-Wishart distribution, ν_0 is the number of degrees of freedom, $\mathbb{E}[\mathrm{IW}(\nu_0,\Sigma_0)] = \frac{1}{\nu_0-3}\Sigma_0$, $X = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_Y \end{bmatrix}^T$, $y = (1,\dots,Y)$, b is an $R \times 2$ matrix and B is an $R \times R$ matrix.

The measure $P_{jy}(\theta)$ is allowed to be year and site dependent by assuming the HDDP prior defined in (2.3.1), where P_{jy} has the same prior as the F_{jx}. As shown in the appendix, the choice of such prior distribution for θ will allow us to make straightforward posterior inference when the measure F_{jx} is convolved with a bivariate Gaussian kernel, as in our case. Keeping in mind the explicit expression of the DP, the resulting model for a specific year and site can be written as

$$f_{jy}(\zeta,\eta) = \sum_{k=1}^{\infty} \pi_{kj} N(\zeta,\eta|(\mu_k)_y, \Sigma_k)$$

where the (μ_k, Σ_k) are shared between groups.

Every cluster has its own regression coefficient β with a common prior distribution $\mathrm{MN}(b,B,\Sigma)$. However, in order to estimate the overall trend across all clusters, we assign an additional hyperprior distribution $b \sim \mathrm{MN}(b_0, B_0, \Sigma_b)$. The posterior distribution for b will give the overall trend of arrival and length of stay for all groups across the years.

For the overall intensity of the process, ω_j, we keep the same prior distribution as in the Gamma process case but in order to share information between sites and years, we assume that intensities now have a prior distribution $\omega_j|\alpha,\tau_j \sim \mathrm{Gamma}(\alpha, \frac{\alpha}{\tau_j})$ where α is the standard shape and τ_j is the mean of the Gamma distribution. The parameters τ_j, α and γ are assumed to have Gamma prior distributions, which is a standard choice.

The model can be summarised with the introduction of latent variables.

$$
\begin{cases}
H_{ijyl} \mid \zeta_{ijy}, \eta_{ijy}, p_{ij} \sim \text{Bernoulli}(p_{ij} z_{ijyl}) & i = 1,\dots,N_{jy} \quad j = 1,\dots,J \quad y = 1,\dots,Y \quad l = 1,\dots,C_{jy} \\
z_{ijyl} = \begin{cases} 1 & \text{if } \zeta_{ijy} < t_l^{jy} < \zeta_{ijy} + \delta_{ijy} \\ 0 & \text{otherwise} \end{cases} & i = 1,\dots,N_{jy} \quad j = 1,\dots,J \quad y = 1,\dots,Y \quad l = 1,\dots,C_{jy} \\
(\zeta_{ijy}, \eta_{ijy}) \mid c_{ijy}, \{\mu_k\}, \{\Sigma_k\} \sim \text{N}((\mu_{c_{ijy}})_{x_y}, \Sigma_{c_{ijy}}) & i = 1,\dots,N_{jy} \quad j = 1,\dots,J \quad y = 1,\dots,Y \\
\mu_k \sim \text{MGP}(X\beta_k, K(x,x), \Sigma_k) & k = 1,\dots,K \\
\Sigma_k \sim \text{IW}(\nu_0, \Sigma_0) & k = 1,\dots,K \\
\beta_k \sim \text{MN}(b, B, \Sigma_k) & k = 1,\dots,K \\
b \sim \text{MN}(b_0, B_0, \Sigma_b) & \\
c_{ijy} \sim \text{CRF}(\alpha, \gamma) & i = 1,\dots,\sum_{y=1}^{Y} N_{jy} \quad j = 1,\dots,J \quad y = 1,\dots,Y \\
N_{jy} | \omega_j \sim \text{Poisson}(\omega_j) & j = 1,\dots,J \qquad y = 1,\dots,Y \\
\omega_j | \alpha, \tau_j \sim \text{Gamma}\left(\alpha, \frac{\alpha}{\tau_j}\right) & j = 1,\dots,J
\end{cases}
$$

where K is the total number of clusters and in the CRF assignments of c_{ijy} the variable j indexes the groups and i and y index the observations.

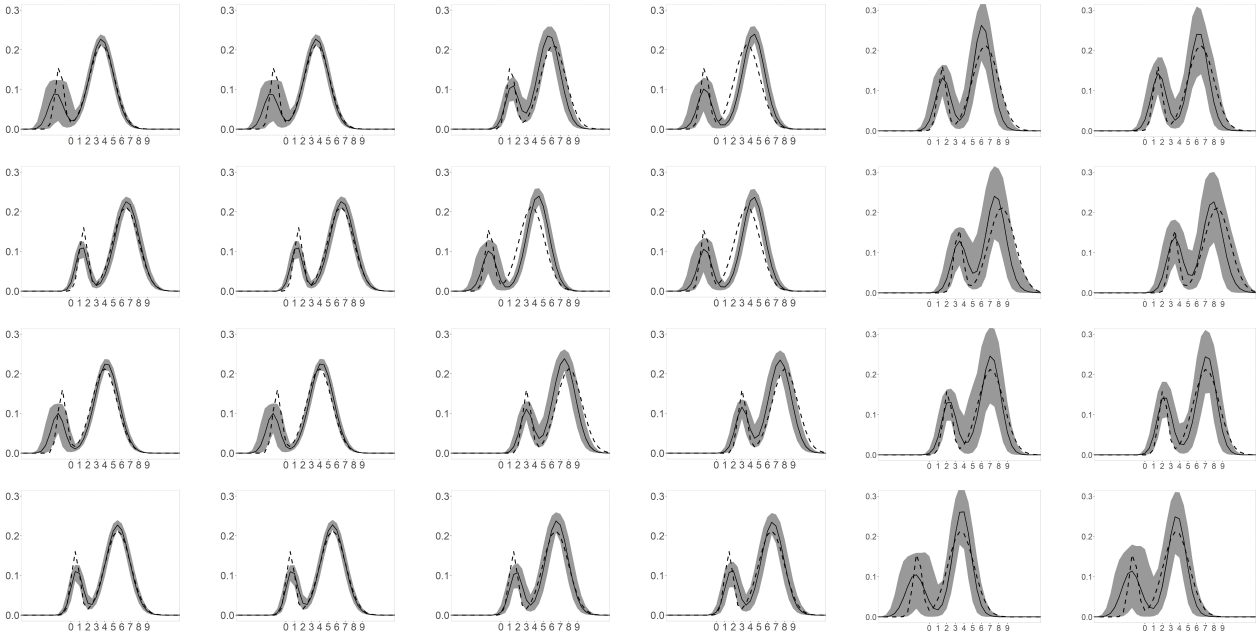
2.5 Application

2.5.1 Simulations

In order to assess the performance of the model, we have simulated several sets of data and compared the posterior distributions of the main quantities of interest with the true values used to simulate the data. The simulated data consist of 2 sites and 16 years, with 10 sampling occasions in each year. In order to have population sizes similar to the ones in the CES data, the site-specific intensities ω_j of the prior distribution of the population sizes are sampled from a Gamma distribution with mean 60 and variance 200, population sizes for each year are then sampled from a Poisson with the intensity ω_j sampled above. Arrival times and lengths of stay are sampled keeping in mind the CES data, which consist of a mixture of individuals with different patterns of arrival and stay. In particular, it is known [43] that there are two groups of birds that use the sites; "residents" that breed at the sites and may return in subsequent years, and "transients" that pass through the site on the way to breeding grounds further north, or wintering grounds further south. To model this behaviour, we sample from the following mixture distribution

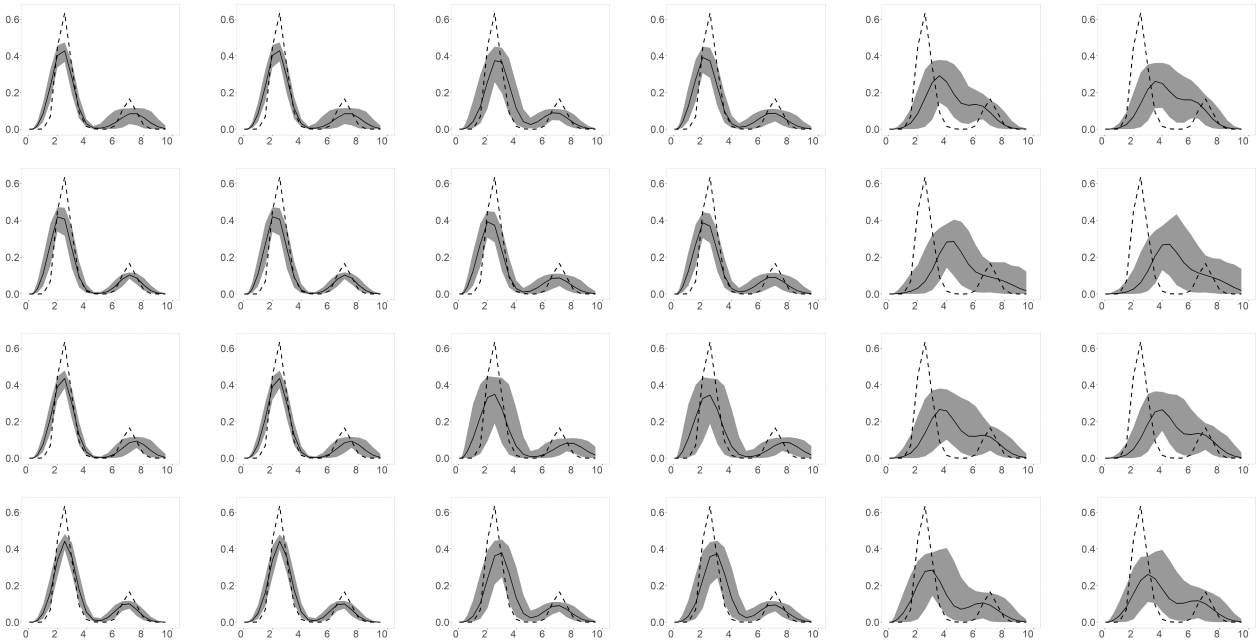
$$
\begin{bmatrix} \zeta_{ijy} \\ \delta_{ijy} \end{bmatrix} \sim 0.8 \begin{bmatrix} \text{N}(6 + 1\, x_y, 1.5) \\ \text{Gamma}(25, 10) \end{bmatrix} + 0.2 \begin{bmatrix} \text{N}(1 + 1\, x_y, .5) \\ \text{Gamma}(210, 30) \end{bmatrix}
$$

where y indexes the year. The values x_y of the covariate are sampled from a $\text{N}(0,1)$.



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig. 2.1 Arrival times - Posterior densities for the three sets of simulations, shown for 2 sites (columns) and a subset of 4 years (rows). The solid line represents the posterior mean, the dashed line represents the true distribution used to simulate the data and the grey area represents the 95% posterior credible interval (PCI).



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig. 2.2 Lengths of stay - Posterior densities for the three sets of simulations, shown for 2 sites (columns) and a subset of 4 years (rows). The solid line represents the posterior mean, the dashed line represents the true distribution used to simulate the data and the grey area represents the 95% PCI.

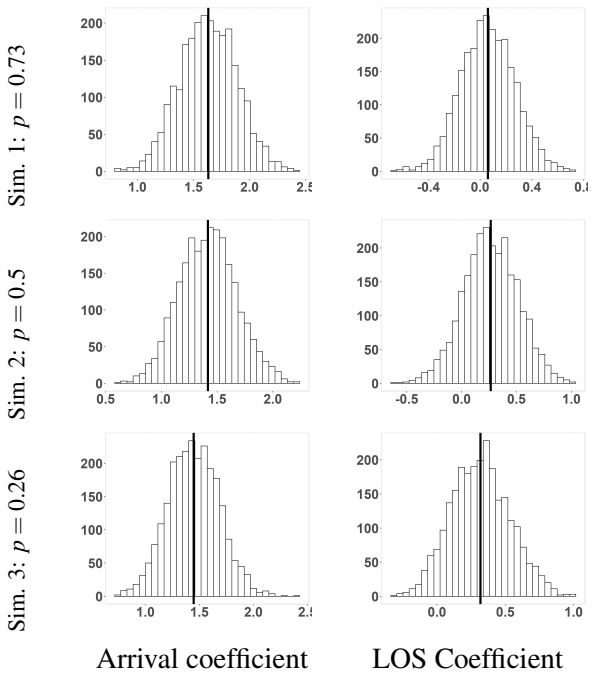


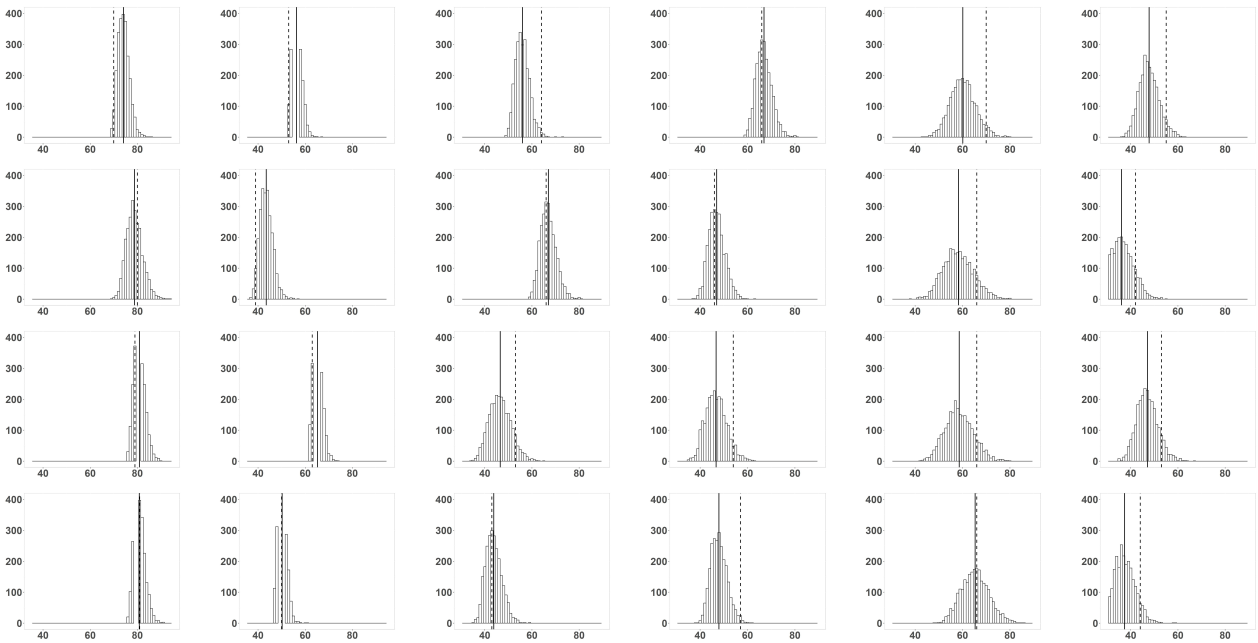
Fig. 2.3 Posterior distributions of the regression coefficients of arrival times, b_{21}, and length of stay, b_{22}, for the three sets of simulations, with the solid line representing the posterior mean. The true value is fixed at 1.5 for the arrival times and 0 for the lengths of stay.

We performed three sets of simulations, each of them with different values of the capture probabilities. We sample values from a logistic-normal with scale 0.1 and location equal to, respectively 1, 0 and -1 for the three sets of simulations, which corresponds to capture probabilities centred around, respectively, 0.73, 0.5 and 0.26.

In order to choose the value of the length scale parameter l of the MGP, we have performed a sensitivity analysis considering the values 0.1, 0.3 and 0.5, obtaining practically identical results. Thus we fixed the value to 0.3, as values outside the range considered would give a correlation between close points which is either too large or too small for our application.

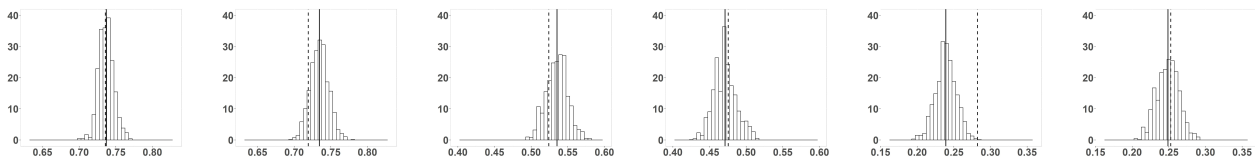
The posterior distributions of the arrival densities and lengths of stay for the three sets of simulations are shown (for a subset of 4 years), respectively, in Fig. A.2 and A.3. In the case of the arrival densities, the posterior mean densities closely resemble the true densities. As capture probability decreases, the estimates present, as expected, more variance and the model splits one of the modes in two separate clusters. In the case of the lengths of stay, for all simulated data the posterior mean density is smoother than the true distribution, a fact that becomes progressively more evident as capture probability decreases.

The posterior distributions of the regression coefficients b_{21} and b_{22} are shown in Fig. 2.3. The estimates of the posterior means are similar and close to the true values, but the cases with lower capture probabilities exhibit more variance in the estimates. The posterior distributions of the population sizes are shown in Fig. 2.4, where it can be seen that, aside from the case with



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig. 2.4 Population sizes - Posterior densities for the three sets of simulations, shown for 2 sites (columns) and a subset of 4 years (rows). The solid line represents the posterior mean, the dashed line represents the true population size.



Simulation 1: $p = 0.73$ Simulation 2: $p = 0.5$ Simulation 3: $p = 0.26$

Fig. 2.5 Capture probabilities - Posterior densities for the three sets of simulations. The solid line represents the posterior mean, the dashed line represents the true capture probability.

lowest capture probability, the posterior mean is generally close to the true value, which is always included in the corresponding 95% PCI. Clearly, population size is consistently either over-estimated or under-estimated at some site. This is due to the model assuming that mean population size at each site is constant over time. The posterior distributions of the capture probabilities are shown in Fig. 2.5. As was the case when inferring population size, the posterior variance increases as capture probability decreases.

2.5.2 BTO's Constant Effort Sampling Scheme Data

We apply the model to CR data of blackcaps collected by the BTO at several breeding and stopover sites across the UK. We discarded all the juvenile birds as, being born at the site in the same year they are captured, they do not provide any information on the arrival density. Even though the complete

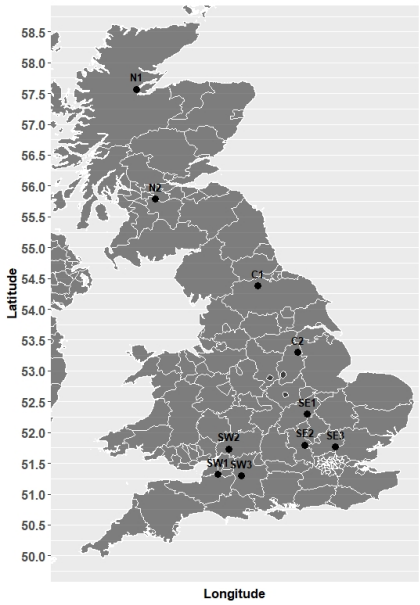


Fig. 2.6 Map of the CES sites used in the analysis, with site ID shown above the sites.

data consist of more than 100 sites for more than 20 years, we work on a subset of 10 sites across 16 consecutive years, from 1998 to 2013, with a total of 3401 birds caught, as working with the entire data would not be feasible in terms of computational time. We selected these 10 sites by choosing the subset where sampling occurred for the highest number of consecutive years, because we are interested in estimating the regression coefficient for the year-continuous covariate. The locations of the sites are indicated on the map shown in Fig. 2.6.

The prior specification is based on previous studies [66, 43]. Arrival times and lengths of stay are modelled in weeks, and their prior distribution is chosen to have 95% of the mass of the arrival distribution from three weeks before the start of the sampling period up to the end of it, and 95% of the mass of the departure distribution from the start of the sampling period, up to three weeks after the end. The prior on the capture probability and the prior on the mean τ_j of the intensity of the population size are shown in the supplementary material [21].

As a year-specific covariate, we use the average North-Atlantic Oscillation (NAO) in the months from January to April, as these are the months preceding the sampling period. This choice is motivated by the fact that the NAO is thought to represent the overall trend of global temperatures. The covariate λ_{jy} used to model the capture probability is the length of the net placed at each site.

We present results for 5 sites, out of the 10 shown in Fig. 2.6, for years 2003, 2005, 2007, 2009 and 2011. Additional plots can be found in the supplementary material [21]. Between these sites, we chose two sites in the South, two in the center and one in the North, in order to highlight differences in the densities for sites at several latitudes. We first focus on the arrival distributions, shown in Fig. 2.7. All of the distributions present a mode before the first sampling occasion, which can be interpreted as the result of the many individuals arriving before sampling has begun. In fact, all of the data show a

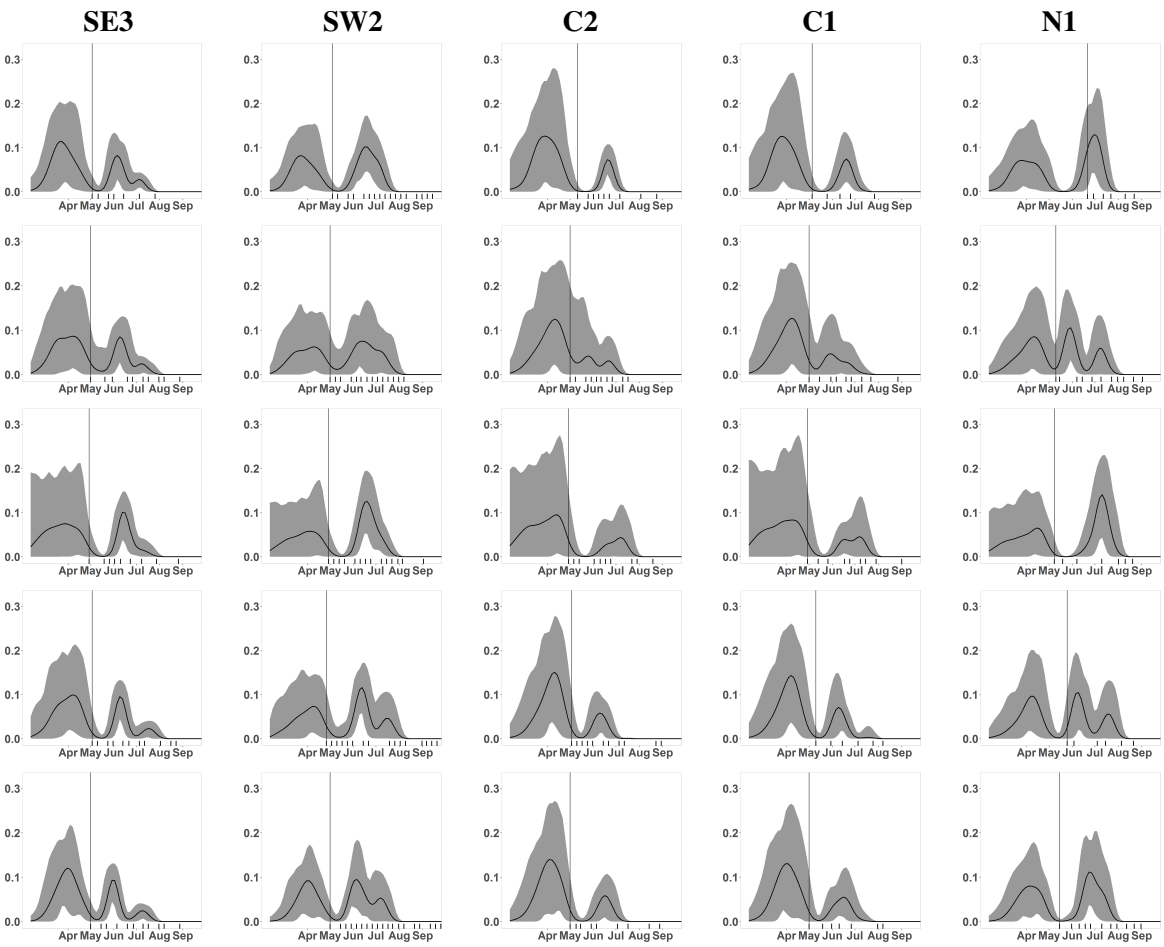


Fig. 2.7 Arrival times - Posterior distribution for a subset of 5 sites, with site names given on top of each column, and 5 years. The black line shows the posterior mean density and the grey area shows the 95% PCI. The sampling occasions are shown in bold on the x-axis and the black line shows the first sampling occasion.

high number of captures in the first and second sampling occasions, while the number decreases in the middle of the sampling period. The remainder of the peaks are likely to correspond to the transient birds arriving at the sites later in the season only for feeding. It can also be noticed that northern sites (e.g. $N1$ and $N2$) present a higher number of birds arriving later in the season, suggesting that the birds arriving in the UK stop first at the southern sites before reaching the sites in the north. The length of stay densities, presented in Fig. 2.8, also exhibit several peaks because of the presence of the breeding birds and transient birds. However, due to the lack of data in some of the sites, the two modes are likely to merge in some cases.

Population sizes for the same sites and years as those considered in Fig. 2.7 are presented in Fig. 2.9. Comparison with the posterior densities of capture probabilities in Fig. 2.10 shows that, as expected, smaller estimates of the capture probability are generally associated with greater variance

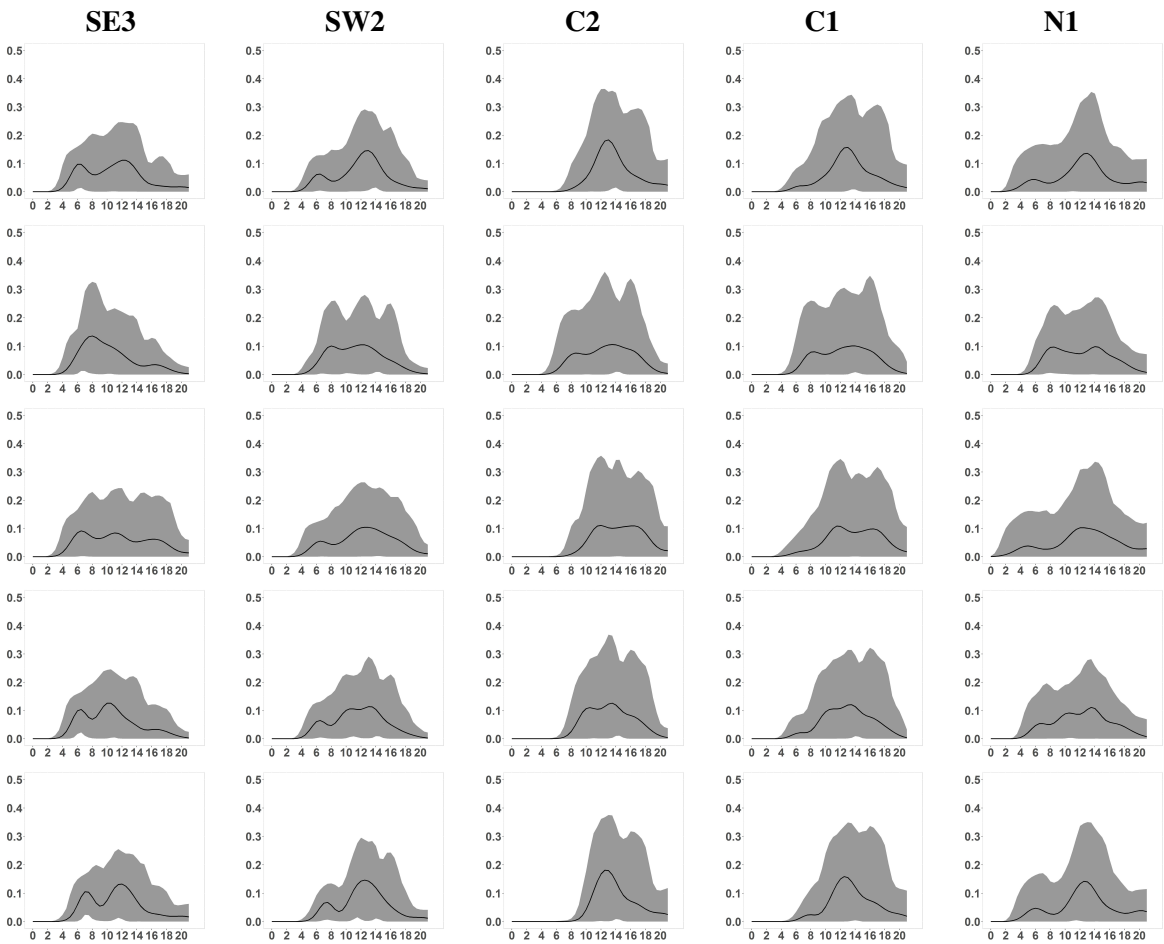


Fig. 2.8 Lengths of stay - Posterior distribution for a subset of 5 sites, with site names given on top of each column, and 5 years. The black line shows the posterior mean density and the grey area shows the 95% PCI.

in population size estimates. Moreover, northern sites present overall lower population sizes than southern sites.

The posterior distributions of the coefficients are shown in Fig. 2.11. The 95% PCIs of the arrival time and length of stay components of the regression coefficient b include 0, suggesting that the NAO has no effect on the patterns of arrival and length of stay, which agrees with previous findings [76, 34]. However, this does not necessarily imply that the arrival and length of stay distributions in the clusters do not exhibit trends across the years, but it might be that some clusters have positive shifting trends while others have negative shifting trends, which would imply an overall posterior close to 0.

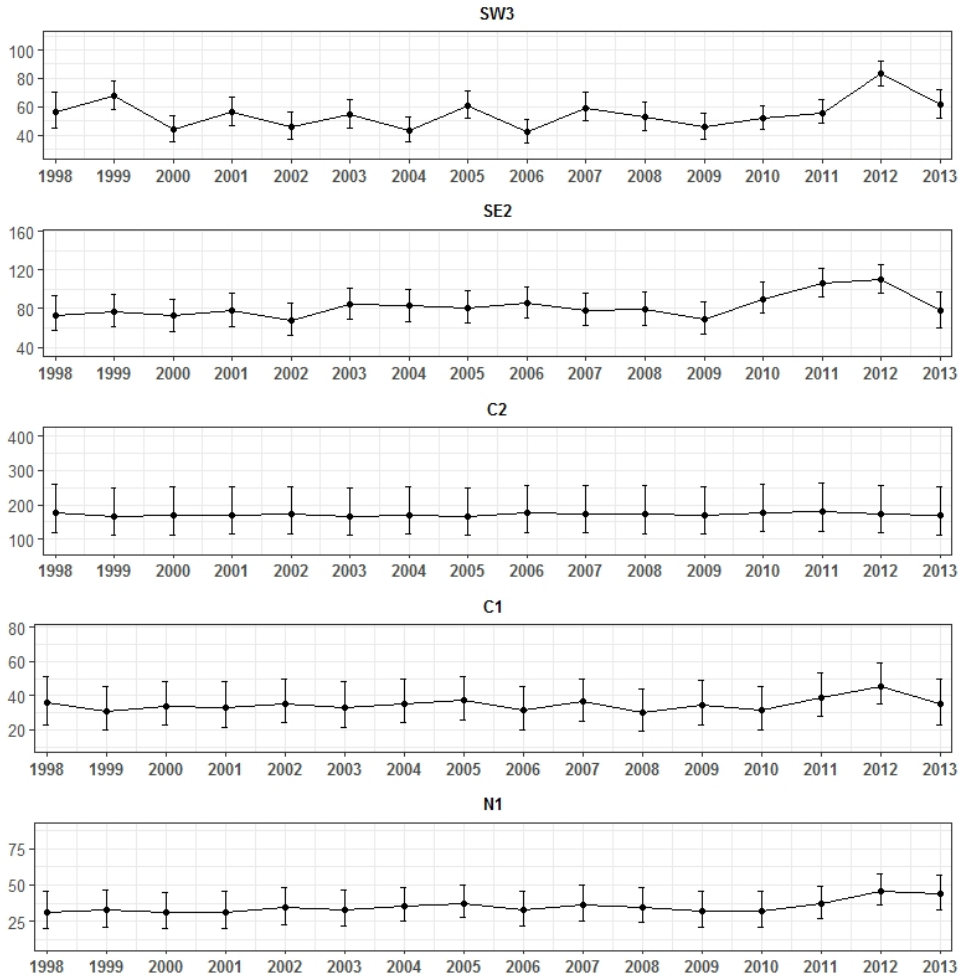


Fig. 2.9 Posterior densities of the population sizes for a subset of 5 sites, with site names given on top of each plot. The bars show the 95% credible intervals, while the dots show the posterior means.

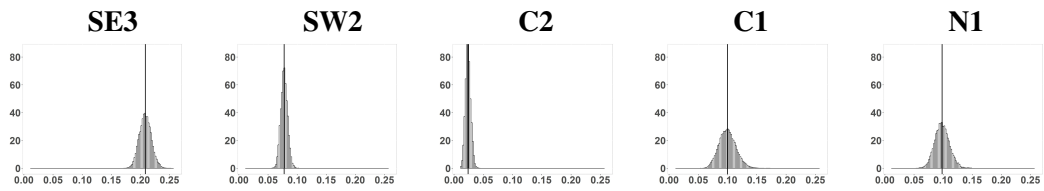


Fig. 2.10 Posterior densities of the capture probabilities for a subset of 5 sites, with site names given on top of each column. The black vertical line shows the posterior mean.

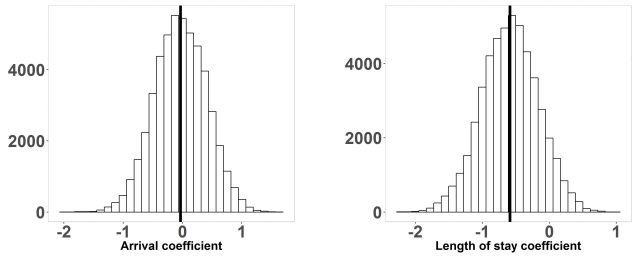


Fig. 2.11 Posterior distribution of the arrival time and length of stay components of the regression coefficient b. The black line represents the posterior mean.

2.6 Discussion

In this paper we have developed a model to estimate arrival and departure distributions in a multi-site and multi-year capture-recapture data set with annual environmental covariates and site-specific variation, and applied this model to real data. Moreover, we have performed a simulation study to assess the validity of the model on simulated data with similar features, obtaining encouraging results even when capture probability is low.

The dataset used in our application consists of a mixture of breeding and transient birds. Although breeding birds tend to return to the same site in different years, transient birds change the site they visit from year to year. As a result, changes in population sizes at each given site across the years do not reflect an actual change in the number of birds of the population. For this reason, in section 2.4.2, we chose not to adopt a model for the evolution of population sizes over time but we only assume that population sizes are sampled from the same common distribution. Because of the lack of site-fidelity of blackcaps, changes in the populations' behaviour are not evaluated by analyzing the evolution of population sizes but instead by observing the changes in phenology, summarised in the arrival and departure distribution for each site and year, in relation to an indicator of global temperature, as the NAO in our case.

In this model, we did not track the same individuals across the years. The choice is motivated by the fact that the number of recaptures of the same individuals in different years is too low to motivate such a modelling approach. However, the model could be further extended in the cases of species exhibiting higher longevity and site-fidelity than the blackcaps.

We have followed the approach of MC17, using a Bayesian nonparametric approach to estimate the arrival and departure densities. However, MC17 only allow the estimation of the arrival density for a single site and year. In our work, we first added an additional level of complexity modelling nonparametrically both the arrival and departure density. Then, as our goal was to perform density estimation in several sites conditional on year-specific covariates, we extended their model to account for these additional effects. The starting point to achieve a joint modelling of data collected at different sites is the use of the Hierarchical Dirichlet Process (HDP) of Teh et al. [84] in place of the DP. However, since this model does not allow the introduction of continuous covariates, we

further extend the HDP defining the Hierarchical Dependent Dirichet Process. Lastly, to introduce a correlation structure over time, we started from the idea of the Gaussian process (GP) and, as we are modelling arrival and length of stay jointly and the GP can only give univariate outputs, we adopted an extension of the GP to the case of p-dimensional outputs. Another interesting definition of the GP with multiple outputs can be found in Álvarez and Lawrence [1]. However, the advantage of our construction is that we still maintain the useful conjugacy properties of the GP, which allows us to straightforwardly use the sampling schemes available for the HDP.

The Bayesian nonparametric model defined here is extremely general and can be used in a generic context where multivariate density estimation is to be performed jointly across different groups and in the presence of a continuous covariate, which extends the model presented in De Iorio et al. [17].

As it is clear from equation (2.3.1), the model can account for covariates only in the cluster locations, however Griffin and Leisen [35] have defined a nonparametric model, known as compound random measures, which can account for covariates in the cluster weights. However, in the case of compound random measures, inference is more difficult as the sampling scheme based on the CRF cannot be used anymore. Moreover, our model allows the introduction of covariates only across time, while in some scenarios it could be useful to adopt spatial covariates, for example the latitude of the site, in order to account for differences in arrival patterns according to site-specific covariates. Even if our model accounts for additional random effects from site to site, explaining the variation through covariates would require a change to the structure of the model.

In Section 2.5.2, we mention that we choose a subset of the data in order to be able to run the algorithm in a feasible computational time. In fact, given the large number of observations, one of the challenges of the model is the computational complexity, which scales linearly with the number of observations. This is a common drawback of all algorithms based on the Chinese Restaurant representation, as the sampler requires to update the cluster allocations of each individual by computing the probability of belonging to each cluster, which is a computationally expensive operation. Sampling from the posterior of DP mixtures without having to update the cluster allocations as in Escobar and West [27] is still an open problem, and it goes beyond the scope of this paper. A potentially faster algorithm to sample from a DP mixture model, based on parallel computation, has been proposed by Ge et al. [32]. Moreover, inference for the HDDP mixtures is performed on the space of the latent arrival times and lengths of stay, which further slows down the mixing, making necessary to run the MCMC for more iterations. An alternative algorithm to speed up the mixing has been proposed by Jain and Neal [39].

We note that we have not used any spatial information on the sites and, as a result, sites are assumed to be exchangeable, in the sense that permuting the site labels has no effect on our inference. This is generally the case when data are collected at a number of sites but the models employed are not spatially-explicit. See for example the occupancy model by MacKenzie et al. [52] and extensions as well as the N-mixture model by Royle [77] and extensions. Since there is only a small number of sites, which are not in close proximity to one another, any spatial autocorrelation in our application is expected to be low.

The code used to generate results has been written in R [71], while the most computationally expensive part of the algorithm, such as the Gibbs sampler for the clusters allocation, has been written integrating C++ and R using the Rcpp package of Eddelbuettel et al. [23]. The code is available upon request.

Appendix

2.6.1 MCMC Algorithm

The vector of unknown parameters is:

$$(\{\zeta_{ijy}\},\{\delta_{ijy}\},\{c_{ijy}\},\mu_k,\Sigma_k,b,\{N_{jy}\},\{p_{jy}\},\alpha,\gamma,\{\tau_j\},\{\omega_j\}).$$

Clearly the posterior distribution is intractable and we obtain samples from it using the following steps in a Gibbs sampler: cluster allocations $\{c_{ijy}\}$ are sampled using the update in [84] that makes use of the CRF representation, while cluster locations $\{\mu_k,\Sigma_k\}$ are updated conditional on the allocations. The population sizes N_{jy} are updated using the rejection algorithm employed in MC17. The arrival times and lengths of stay $(\zeta_{ijy},\delta_{ijy})$ are sampled jointly using a simple MH update. To update b, first we sample the β_k and then we sample b from its full conditional. Finally, capture probabilities are updated using a MH step. For the remaining hyperparameters α, γ, τ_j and ω_j, we can sample directly from the full conditional.

A detailed description of each Gibbs sampler can be found below.

1. **Sample the cluster means and covariance matrices (μ_j,Σ_j):**

   For each cluster $k=1,\dots,K$, we sample (μ_k,Σ_k) from the posterior distribution:

$$p(\mu_k,\Sigma_k|\{\zeta_{ijy}\},\{\delta_{ijy}\},\{c_{ijy}\},b,B,\nu_0,\Sigma_0)\propto$$
$$p(\mu_k|\{\zeta_{ijy}\},\{\delta_{ijy}\},\{c_{ijy}\},\Sigma_k,b,B)\,p(\Sigma_j|\{\zeta_{ijy}\},\{\delta_{ijy}\},\{c_{ijy}\},B,\nu_0,\Sigma_0)$$

   As shown in the supplementary material [21], the posterior distribution for μ_k is still a MN distribution, while the posterior distribution for Σ_k is still an inverse-Wishart. In our application, to efficiently compute the posterior distributions, we rely on the fact that the covariate, being year-specific, takes only as many values as the number of years, Y. Thus, instead of building the covariance matrix of the MGP using all the individuals in the cluster, we calculate the covariance computed using only the value of the covariates at the observed points. Moreover, as these points are fixed in the model, the inverse of the covariance matrix of the GP can be precomputed.

2. **Sample the allocation $\{c_{ijy}\}$ of individuals to the different clusters:**

Following Teh et al. [84] and the notation of section 2.3.1, the variables c_{ijy} are updated using the CRF representation defined in Section 2.3.1, by first sampling the allocations t_{ij} of the customers to the tables and then the allocations k_{jt} of the tables to the dishes.

We use the superscript $^{-ij}$ to indicate that the quantities are computed removing customer i from restaurant j and $^{-jt}$ when removing table t from restaurant j.

At each step of the Gibbs sampler, individual i in group j having covariate y is assigned to either an existing table in the current restaurant, a new table serving an existing dish or a new table serving a new dish

$$\begin{cases} \text{existing table } t & \text{with probability} \quad \frac{n_{jt}^{-ij}}{n_{jt}^{-ij}+\alpha}\, p((\zeta_{ijy},\delta_{ijy})|x_y,\mu_{k_{jt}},\Sigma_{k_{jt}}) \\ \text{new table } t^\star \text{ with existing dish } k & \text{with probability} \quad \frac{\alpha}{n_{jt}^{-ij}+\alpha}\frac{m_k^{-ij}}{M^{-ij}+\gamma}\, p((\zeta_{ijy},\delta_{ijy})|x_y,\mu_k,\Sigma_k) \\ \text{new table } t^\star \text{ with new dish } k^\star & \text{with probability} \quad \frac{\alpha}{n_{jt}^{-ij}+\alpha}\frac{\gamma}{M^{-ij}+\gamma}\, p((\zeta_{ijy},\delta_{ijy})|x_y,b,\Sigma_0,\nu_0) \end{cases}$$

Similarly, tables are allocated to

$$\begin{cases} \text{existing dish } k & \text{with probability} \quad \frac{m_k^{-jt}}{M^{-jt}+\gamma}\, p(\{(\zeta_{ijy},\delta_{ijy})\}_{t_{ij}=t}|\{x_y\},\mu_{k_{jt}},\Sigma_{k_{jt}}) \\ \text{new dish } k^\star & \text{with probability} \quad \frac{\gamma}{M^{-jt}+\gamma}\, p(\{(\zeta_{ijy},\delta_{ijy})\}_{t_{ij}=t}|x_y,b,\Sigma_0,\nu_0) \end{cases}$$

As opposed to the original algorithm of [84], instead of computing the posterior distribution of $(\zeta_{ijy},\delta_{ijy})$ conditional on the current elements in the cluster, which would excessively slow down the algorithm if repeated for each point, we compute the update conditional on the cluster locations (μ_k,Σ_k) computed in the previous step.

3. **Sample the population sizes $\{N_{jy}\}$:**

   Following MC17, for each site j and year y, conditional on the measure P_{jy}, the arrival times and length of stay

   $\zeta_{(D_{jy}+1):N_{jy},j,y}, \delta_{(D_{jy}+1):N_{jy},j,y}$ of the uncaptured birds are distributed from a non-homogeneous Poisson process with intensity

$$\nu_0(\zeta,\delta) = \nu(\zeta,\delta)\,\mathbb{P}(H=(0,\dots,0)|\zeta,\delta,p)$$

   It follows that samples from the posterior distribution of

$$\left(N_{jy}, c_{(D_{jy}+1):N_{jy},j,y}, \zeta_{(D_{jy}+1):N_{jy},j,y}, \delta_{(D_{jy}+1):N_{jy},j,y}\right)$$

   can be obtained with a rejection algorithm in the following way. First, sample $N_0 \sim \text{Poisson}(\omega_j)$, then, for $i = 1,\dots,N_0$ sample:

$$\{c_{ijy}\}_{i=1,\dots,N_0} \mid \mathrm{CRF}(\alpha,\gamma)$$

$$(\zeta_{ijy},\eta_{ijy}) \sim \mathrm{N}(\mu_{c_{ijy}},\Sigma_{c_{ijy}})$$

$$H_{ijy} \sim Pr(\zeta_{ijy},\eta_{ijy},p_{jy})$$

and accept the individual if capture history H_{ijy} has no non-zero entries. The new population size is given by $D_{jy}+\tilde{N}_0$ where D_{jy} is the number of captured individuals at site j in year y and $\tilde{N}_0$ is the number of accepted individuals.

4. **Sample the hyperparameters** $(\alpha,\gamma,\{\tau_j\},\{\omega_j\})$**:**

   τ_j and ω_j are updated as

$$\omega_j \sim \mathrm{Gamma}\left(\alpha+\sum_{i=1}^{Y} N_{ji}, \frac{\alpha}{\tau_j}+Y\right)$$

$$p(\tau_j|\alpha,\omega_j) \propto p(\tau_j)p(\omega_j|\alpha,\tau_j) \propto \mathrm{Gamma}(\alpha_\tau,\beta_\tau)\tau_j^{-\alpha}e^{-\alpha\frac{\omega_j}{\tau_j}}$$

   The posterior distributions for α and γ are found by adapting the update for the concentration parameter of the DP presented in Escobar and West [27]. Details are presented in the supplementary material [21]. An exact expression of the likelihood of α and γ given an allocation of individuals to the cluster can be found in Camerlenghi et al. [10].

5. **Sample the mean** b **of the prior distribution of the cluster-specific regression coefficients:**

   In order to improve the mixing for the posterior distribution of b, we introduce the variables $\delta_k := \beta_k - b$. After sampling the δ_k from their posterior distribution (which can be found in the supplementary material [21]), the posterior distribution of b given δ_k can be computed as

$$p(b|\{\beta_k\}_{k=1,\dots,K},B,b_0,B_0,\{\Sigma_k\}_{k=1,\dots,J}) \propto$$

$$p(\{\beta_k\}_{k=1,\dots,K}|b,B)p(b|b_0,B_0) \propto \prod_{k=1}^{K}\mathrm{MN}(\beta_k|b,B,\Sigma_k)\mathrm{MN}(b|b_0,B_0,\Sigma_0) \propto$$

   The complete formulas can be found in the supplementary material.

6. **Sample the latent arrival times and length of stay** $\{\zeta_{ijy}\},\{\delta_{ijy}\}$**:**

   Given the continuous arrival time and length of stay of each individual, if we consider the partition defined by the sampling occasions $t_1^{jy},\dots,t_{C_{jy}}^{jy}$, we can define as b_{ijy} and d_{ijy} the intervals where individual ijy respectively arrives and departs. Given these intervals, the posterior distribution for $(\zeta_{ijy},\delta_{ijy})$ is:

$$p(\zeta_{ijy},\delta_{ijy}|\mu_{c_{ijy}},\Sigma_{c_{ijy}},H_{jy},p_{jy}) \propto p(\zeta_{ijy},\delta_{ijy}|\mu_{c_{ijy}},\Sigma_{c_{ijy}})p(H_{jy}|p_{jy},\zeta_{ijy},\delta_{ijy}) =$$

$$\mathrm{N}(\zeta_{ijy},\eta_{ijy}|\mu_{c_{ijy}},\Sigma_{c_{ijy}})\ p_{jy}^{\sum_{k=1}^{C_{jy}} H_{ijyk}}(1-p_{jy})^{d_{ijy}-b_{ijy}-\sum_{k=1}^{C_{jy}} H_{ijyk}}$$

7. **Sample the coefficient $\{\alpha_{jy}^p\}$ and β^p of the capture probabilities $\{p_{jy}\}$:**

   Although not available in analytic form, the posterior distribution can be computed as

$$p(\{\alpha_j^p\},\beta^p|\{\zeta_{ijy}\},\{\delta_{ijy}\},\{x_{jy}\}) \propto p(\beta^p|b_0^p,B)p(\{\alpha_j^p\}|a_0^p,A_0^p)p(\{H_{jy}\}|\{\zeta_{ijy}\},\{\delta_{ijy}\},p_{jy}) =$$

$$\mathrm{N}(\beta^p|b_0^p,B)\prod_j \mathrm{N}(\alpha_j^p|a_0^p,A_0^p)\ \prod_{j,y}\prod_{i=1}^{N_{jy}} p_{jy}^{\sum_{k=1}^{C_{jy}} H_{ijyk}}(1-p_{jy})^{d_{ijy}-b_{ijy}-\sum_{k=1}^{C_{jy}} H_{ijyk}}$$

Chapter 3

A Polya Tree Based Model for Unmarked Individuals in an Open Wildlife Population

Abstract

Many ecological sampling schemes do not allow for unique marking of individuals. Instead, only counts of individuals detected on each sampling occasion are available. In this paper, we propose a novel approach for modelling count data in an open population where individuals can arrive and depart from the site during the sampling period. A Bayesian nonparametric prior, known as Polya Tree, is used for modelling the bivariate density of arrival and departure times. Thanks to this choice, we can easily incorporate prior information on arrival and departure density while still allowing the model to flexibly adjust the posterior inference according to the observed data. Moreover, the model provides great scalability as the complexity does not depend on the population size but just on the number of sampling occasions, making it particularly suitable for data-sets with high numbers of detections. We apply the new model to count data of newts collected by the Durrell Institute of Conservation and Ecology, University of Kent.

3.1 Introduction

Monitoring wildlife populations presents particular challenges. For example, it is typically not possible to perform a census of the population of interest by encountering all of the individuals. One of the most cost and time effective ways to monitor a wildlife population is to collect counts of the population on repeated sampling occasions (SO). This protocol is considerably easier to perform than a capture-recapture (CR) scheme as it does not require physical capture or unique identification of the individuals in the population. The most popular model for analysing count data in a frequentistic setting is the N-mixture model introduced in [77], which allows the estimation of population size and capture probability when the population is closed, that is the same individuals are present throughout the study period. However, when the data are sparse or detection probability is low, N-mixture models are known to suffer from parameter identifiability issues and may give rise to infinite estimates for population size [18]. In a Bayesian setting, the natural way to solve issues of parameter identifiability is to assume informative prior distributions on detection probability or on population size in order to obtain sensible posterior distributions.

In this paper we work in a Bayesian framework and we relax the assumption of population closure, allowing for individuals to enter and leave the site (and thus become available or unavailable for detection) at random times, but still assuming that emigration is permanent. These random arrival and departure times are sampled from a distribution with unknown parameters. However, the absence

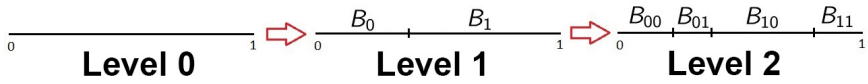


Fig. 3.1 Scheme of the partitions at the first three levels of the Polya tree.

of closure makes it more challenging to separately estimate capture probability, population size and density of arrival/departure times. Hence, it is of primary importance to assume informative prior distributions in order to obtain ecologically sensible posterior distributions.

In order to allow for the posterior distribution to correctly adjust to the data without relying on parametric assumptions, we use a Bayesian nonparametric (BNP) approach to choose the prior distribution of the bivariate distribution of arrival and departure times. In particular, we work with Polya Trees, which in the BNP framework are the main alternative to Dirichlet process mixture models for modelling continuous distributions. More information on other nonparametric priors can be found in [37].

3.2 The Polya Tree Prior

We model the joint density of arrival and departure times, using the Polya tree (PT) prior, defined in [48]. A PT has two parameters: the first is a sequence of nested partitions Π of the sample space Ω ($\Omega = \mathbb{R}^2$ in our case), while the second parameter, α, is a sequence of positive numbers associated with each set of each partition.

The partition at the first level, π_1, is obtained by splitting the sample space in two sets, B_0 and B_1. Then for the partition at the second level, π_2, we split each of the two sets in two additional sets B_{00}, B_{01} and B_{10}, B_{11}, respectively.

$$\pi_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}, \qquad B_{00} \cup B_{01} = B_0, \quad B_{10} \cup B_{11} = B_1.$$

The same process is repeated to generate the partitions at the remaining level. A visual representation of the scheme for $\Omega = [0,1]$ is given in Fig. 3.1.

The PT prior is defined in terms of the (random) mass associated with each set of each partition. By defining $\varepsilon_1 \dots \varepsilon_m$ as a generic sequence of 0s and 1s, $B_{\varepsilon_1 \dots \varepsilon_m}$ as a generic set of the partition and $\alpha_{\varepsilon_1 \dots \varepsilon_m}$ as the associated parameter, the mass associated to $B_{\varepsilon_1 \dots \varepsilon_m}$ by the Polya Tree is

$$G(B_{\varepsilon_1 \dots \varepsilon_m}) = \prod_{i=1}^{m} Y_{\varepsilon_1 \dots \varepsilon_i} \tag{3.2.1}$$

where $Y_{\varepsilon_1 \dots \varepsilon_{i-1} 0}$ is a $\text{Beta}(\alpha_{\varepsilon_1 \dots \varepsilon_{i-1} 0}, \alpha_{\varepsilon_1 \dots \varepsilon_{i-1} 1})$ random variable and $Y_{\varepsilon_1 \dots \varepsilon_{i-1} 1} = 1 - Y_{\varepsilon_1 \dots \varepsilon_{i-1} 0}$. For example, $G(B_{01}) = Y_0(1 - Y_{00})$ where $Y_0 \sim \text{Beta}(\alpha_0, \alpha_1)$ and $Y_{00} \sim \text{Beta}(\alpha_{00}, \alpha_{01})$.

A conjugate scheme for a PT can be constructed if we assume a PT prior for a distribution G, and we have observations $y_1, \dots, y_n \sim G$, since the posterior distribution $G \mid y_1, \dots, y_n$ is still a PT. The

parameters $\alpha_\varepsilon^\star$ of the posterior distribution can be computed as $\alpha_\varepsilon^\star = \alpha_\varepsilon + n_\varepsilon$ where n_ε is the number of observations falling into set B_ε.

A common choice is to center the PT on a pre-specified distribution G_0, which means that, for every set B of the partition, $\mathbb{E}[G(B)] = G_0(B)$. In this paper we will set the $\alpha_{\varepsilon_1\ldots\varepsilon_{j-1}0}$ and $\alpha_{\varepsilon_1\ldots\varepsilon_{j-1}1}$ associated with the sets $B_{\varepsilon_1\ldots\varepsilon_{j-1}0}$ and $B_{\varepsilon_1\ldots\varepsilon_{j-1}1}$ to be proportional to the mass assigned to these sets from G_0, that is $\alpha_\varepsilon = c_\varepsilon \times G_0(B_\varepsilon)$, where c_ε is a scaling parameter tuning the overall variance around the mean distribution. Finally, we assume that G_0 has random parameter η and we let this vary by placing an additional prior on it, leading to what is known as Mixtures of Polya Trees (MPT), as defined in [36].

3.3 Model

The data consist of the number of individuals, D_k, detected on SO k, with $k = 1,\ldots,K$. We denote by N_k the (latent) number of individuals available for detection at SO k and by p the detection probability, assumed to be constant for each individual and each SO. Clearly $D_k \sim \text{Binomial}(N_k, p)$.

We do not assume that individuals are present throughout the study period but we instead assume that their arrival and departure times are random. These times are assumed to be sampled from a Poisson process, with intensity that can be written as $\omega \times \tilde{\nu}$ where ω is the overall mass of the process and $\tilde{\nu}$ is a probability density function. The MPT is employed as a prior for $\tilde{\nu}$ and we call P_0 the prior distribution on the hyparameters of the centring distribution G_0, which we define later. As departure is always greater than arrival, $\tilde{\nu}$ is defined on $\{(x,y) \in \mathbb{R}^2 | x < y\}$.

Although the data depend only on the latent number of individuals N_k, we introduce additional latent variables to make the inference of the PT more efficient. Let $\{t_k\}_{k=1,\ldots,K}$ be the times when samples are collected and take by convention $t_0 = -\infty$ and $t_{K+1} = \infty$. Additionally, let n_{ij} be the number of individuals having arrival times between t_i and t_{i+1} and departure times between t_j and t_{j+1} (with $n_{ij} = 0$ for $i > j$). The N_k can easily be obtained from the n_{ij} as $N_k = \sum_{j=k}^{K}\sum_{i=0}^{k-1} n_{ij}$.

We make the standard choices of a Beta prior distribution for detection probability and a Gamma prior distribution for the overall intensity of the process. The hierarchical structure of the model is the following:

$$
\begin{aligned}
&D_k \sim \text{Binomial}(N_k, p), && N_k = \sum_{j=k}^{K}\sum_{i=0}^{k-1} n_{ij}, \qquad k = 1,\ldots,K,\\
&n_{ij} \sim \text{Poisson}(\omega \times \omega_{ij}), && i = 0,\ldots,K, \quad j = 0,\ldots,K \quad i \le j,\\
&\omega_{ij} = \int_{t_i}^{t_{i+1}}\int_{t_j}^{t_{j+1}} \tilde{\nu}(x,y)\,dx\,dy, && i = 0,\ldots,K, \quad j = 0,\ldots,K, \quad i \le j,\\
&\omega \sim \text{Gamma}(a_\omega, b_\omega), && p \sim \text{Beta}(a_0, b_0),\\
&\tilde{\nu} \sim \text{PT}(\Pi, \alpha^\eta), && \eta \sim P_0.
\end{aligned}
$$

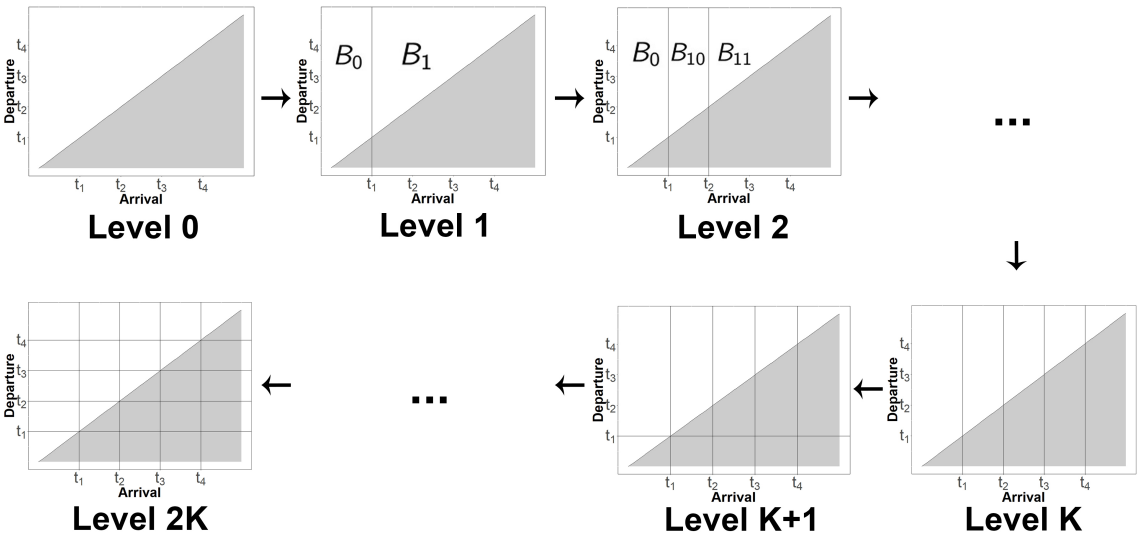


Fig. 3.2 Partitions of the Polya Tree. Observation occur only in the region above the diagonal. At step k, the set of individuals departing after the k-th sampling occasion is split into the individuals departing between the k-th and $(k+1)$-th sampling occasion and the individuals departing after the $(k+1)$-th sampling occasion. After having reached level K, each set is split according to the same procedure but with respect to the other dimension (arrival times).

In order to center the PT on a pre-specified distribution, we use the approach explained in Section 3.2 of using a fixed partition and choosing the α according to the value η of the parameters of the centring distribution. The dependence on η is thus only in the α.

The process used to create the partition is explained in Fig. 3.2. The last level corresponds to the partition defined by the sampling occasions. Since we use the latent variables n_{ij} and not the exact, on a continuous scale, individual arrival and departure times, it is not meaningful to build a finer level of the partitions, as no information is available about them.

We center the PT on a bivariate distribution with independent double exponential marginal distributions, with probability density function (pdf)

$$G_0(x_1,x_2;\mu_1,\mu_2,\lambda_1,\lambda_2) = \frac{1}{2\lambda_1}\exp\left(-\frac{|x_1-\mu_1|}{\lambda_1}\right)\frac{1}{2\lambda_2}\exp\left(-\frac{|x_2-\mu_2|}{\lambda_2}\right),$$

with the constraint that $G_0(x_1,x_2;\mu_1,\mu_2,\lambda_1,\lambda_2) = 0$ if $x_1 > x_2$.

The sets of the partition are squares and triangles, as shown in Fig. 3.2. The choice of the double exponential is motivated by the fact that integrals of this distribution on squares and triangles can be computed analytically, without resorting to numerical techniques. The hyperparameters (μ_1,μ_2) are given a bivariate normal prior distribution and λ_1 and λ_2 two independent Gamma prior distributions.

3.4 Computational Notes

Posterior inference is performed using a Gibbs sampler. While this is straightforward for some parameters, such as the detection probability p and the intensity ω of the Poisson process thanks to conjugacy, for other parameters posterior inference is not straightforward given that we are working with a PT.

The conditional distribution of the parameters $(\{n_{ij}\},\{\omega_{ij}\},\eta)$ given p and ω can be written as

$$p(\{n_{ij}\},\{\omega_{ij}\},\eta|\{D_k\},p,\omega) \propto p(\{D_k\}|\{N_k\},p)\ p(\{n_{ij}\}|\omega,\{\omega_{ij}\})\ p(\{\omega_{ij}\}|\Pi,\alpha_\eta)\ p(\eta)$$

$$\propto \prod_{k=1}^{K}\text{Binomial}(D_k|N_k,p)\ \prod_{i=1}^{K+1}\prod_{j=1}^{i}\text{Poisson}(n_{ij}|\omega\times\omega_{ij})\ p(\{\omega_{ij}\}|\Pi,\alpha_\eta)\ p(\eta)$$

where the distribution $p(\{\omega_{ij}\}|\Pi,\alpha_\eta)$ is given from the PT.

When writing the posterior distribution of the parameter η, we can integrate out the parameter ω_{ij}, by employing a different parametrisation of the PT. First, we define as n_ε the number of observations in set B_ε and as $q_{\varepsilon 0}$ the probability of assigning an observation in set $B_{\varepsilon 0}$ given that we are in B_ε which, according to the structure of the PT, has a $\text{Beta}(\alpha_{\varepsilon 0},\alpha_{\varepsilon 1})$ prior distribution. The marginal likelihood of the n_ε can be expressed as

$$p(\{n_\varepsilon\}|\{\alpha_\varepsilon\}) = \prod_\varepsilon \int \text{Bin}(n_{\varepsilon 0}|n_\varepsilon,q_{\varepsilon 0})\text{Beta}(q_{\varepsilon 0}|\alpha_{\varepsilon 0},\alpha_{\varepsilon 1})dq_{\varepsilon 0}.$$

Integrating out the probabilities $q_{\varepsilon 0}$ gives as a result a beta-binomial distribution. Since the probability mass function of the beta-binomial is

$$f(k|n,\alpha,\beta) = \binom{n}{k}\frac{B(k+\alpha,n-k+\beta)}{B(\alpha,\beta)},$$

the marginal likelihood of the latent variable n_ε given the hyperparameter η is

$$p(\{n_\varepsilon\}|\eta) \propto \prod_\varepsilon \frac{B(\alpha_{\varepsilon 0}+n_{\varepsilon 0},\alpha_{\varepsilon 1}+n_{\varepsilon 1})}{B(\alpha_{\varepsilon 0},\alpha_{\varepsilon 1})}.$$

The posterior distribution for the latent variable n_{ij} can be written as

$$p(n_{ij}|\{D_k\},p,\omega,\omega_{ij}) \propto \text{Poisson}(n_{ij}|\omega\times\omega_{ij})\prod_{k=1}^{K}\text{Binomial}(D_k|N_k,p).$$

The parameter is updated with a random walk with uniform proposal over $(n_{ij}-K_{ij},\ldots,n_{ij}+K_{ij})$, where K_{ij} is chosen according to the value of the n_{ij} chosen as a starting point for the MCMC. In our case, we choose K_{ij} to be $1/5$ of the starting point of n_{ij}.

The parameters ω_{ij} correspond to the masses assigned by the distribution $\tilde{\nu}$ to the sets in the partition of the last level of the Polya tree. Hence, they can be sampled as a product of Beta

distributions as in (3.2.1). The parameters of the PT are updated at each iteration conditional on the latent variable n_{ij}, using the standard update explained in Section 3.2.

3.5 Application

The data used in our application consist of weekly detections of great crested newts (GCN) (*Triturus cristatus*). This species generally start to migrate to ponds in late winter in order to breed. Subsequently, they leave the breeding site at the end of the summer and hibernate on land. Sampling took place in ponds located at the University of Kent with data collected between the end of February 2016 until the beginning of September of the same year, which covers a large part of the breeding period of the newts. Samples were collected on weeks $1-22$, $24-27$ and 29 of the season.

GCN are uniquely identifiable and hence individual capture histories of this population exist. A total of 69 individuals were captured during the study. However, in this case, the individual CR data have been collapsed to simple count data, which are obtained by recording the number of individual newts caught on each of the sampling weeks. It is believed that the population size is close to the sample size, and we choose a prior distribution for ω with mean 76 to represent our belief that around 90% of the individuals have been detected at least once. The variance has been chosen in order to have a relative weakly informative prior, as the 95% prior mass includes up to 135 individuals. Finally, 95% of the prior mass for detection probability is placed on the $(0.05, 0.35)$ interval, based on previous analyses of data on the same population.

Prior knowledge suggests that a considerable number of individuals tend to arrive at the site between the beginning of March and the end of April. Additionally, individuals depart between the end of May and the end of July. In order to translate this knowledge into our prior distributions, we choose hyperpriors for $\eta = (\mu, \lambda)$ such that 95% of the prior mass of the arrival and departure density is in the aforementioned ranges.

The posterior mean estimate of the population size is 89, while the posterior mean of the detection probability is 0.33. The two posterior distributions are shown in Fig. 3.3. The posterior distribution of the population size is different from the prior distribution, as more individuals are estimated to be at the site than expected by the ecologists. Moreover, in Fig. 3.4 we display the posterior mean of the latent number of individuals available at each sampling occasion together with the number of individuals counted. For some sampling occasions, the empirical estimated detection probability, estimated as the ratio between the estimated number of individuals available and the counted individuals, is outside the 95% posterior credible interval for detection probability. This suggests that detection probability is not constant across sampling occasions, as we have assumed in our model. According to expert knowledge, changes in detection probability might be due to differences in environmental conditions between sampling occasions, which affect behaviour of newts.

The posterior cumulative distribution functions of arrival and departure are also shown in Fig. 3.4. As 95% of the individuals are estimated to arrive before any individual has departed, the number of individuals is estimated to be fairly constant between sampling occasions 9 and 14.

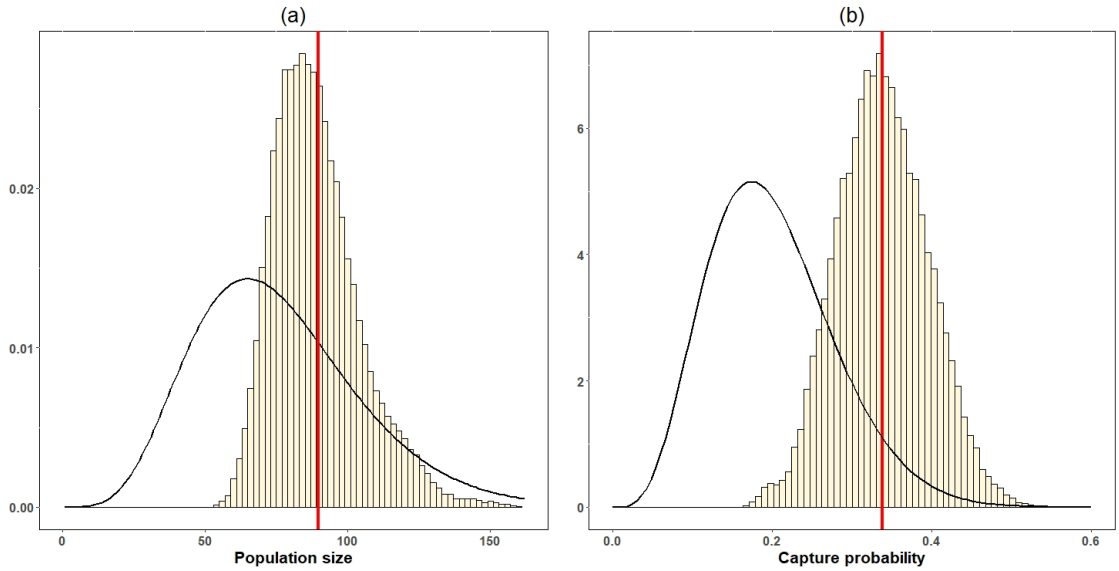


Fig. 3.3 Posterior distributions of population size (a) and detection probability (b), with the red vertical lines showing the posterior means and the black line showing the prior distribution in each case.

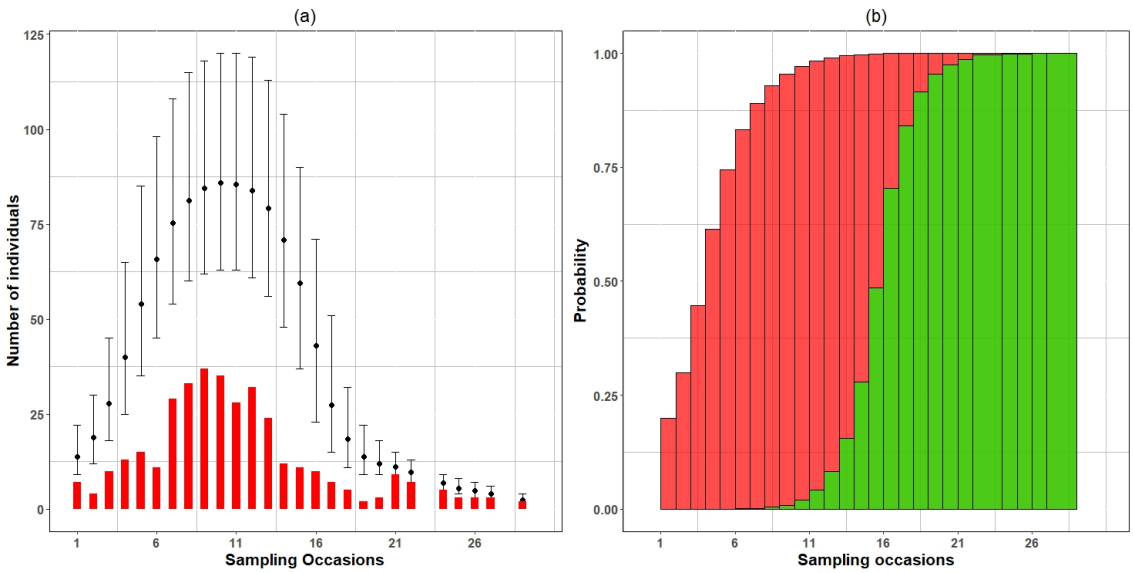


Fig. 3.4 (a) 95% posterior credible interval of the latent number of individuals available for detection each week, shown in black, and the number of individuals detected each week shown in red. (b) Posterior mean of the cumulative distribution function of arrival times (in red) and departure times (in green).

3.6 Conclusion

In this paper we have presented a BNP model for count data on an open wildlife population consisting of individuals entering and exiting the site at random times. By assuming a PT prior, we make no parametric assumptions on the shape of the arrival and departure distribution. Moreover, the implementation is fast as the computational complexity does not depend on the number of individuals but on the levels of the PT, which depends on the number of sampling occasions. However, given the small amount of information provided in count data, it is important to assume meaningful and informative prior distributions in order to have sensible posterior distributions. In this paper, we assume informative prior distributions for detection probability and for population size, available thanks to expert knowledge.

As we mentioned in the introduction, another common sampling protocol is CR which, as opposed to count data, provides individual information that can improve estimation. Hence, a possible extension is to model count data and CR data jointly. Another useful extension is to model data collected at different sites, by replacing the Polya tree prior with a hierarchical Polya tree prior, defined in [13].

Chapter 4

A general modelling framework for ecological data based on the Polya Tree prior

Abstract

Wildlife population monitoring can be performed using a number of different survey methods. Each survey method gives rise to a type of data. In the last five decades, a large number of associated statistical models have been developed for analysing these data. Although these models have been parameterised and fitted using different approaches, they have all been designed to model the pattern with which individuals enter and exit the population and to estimate population size. In this paper, we propose a novel Bayesian nonparametric framework for modelling ecological data based on the Polya Tree (PT) prior. We demonstrate how existing models for capture-recapture, count and ring-recovery data can be expressed within our proposed framework, and how different constraints on the parameters of these models can be imposed within our framework by introducing the replicated PT prior. Additionally, we introduce the Hierarchical Logistic PT prior for jointly modelling related data, for example collected at different sites or across different years and we consider the Optional PT prior for modelling long time series of data. Our work provides a unifying flexible framework for modelling the most commonly collected types of ecological data and we demonstrate our new approach using five different case studies on birds, amphibians and insects.

4.1 Introduction

In recent years, there has been an increased interest in monitoring wildlife populations due to the ongoing effects of climate change and habitat destruction [38]. However, monitoring wildlife populations is challenging because the observation process of animals in the wild is prone to errors. Therefore, statistical models need to be employed in order to infer population sizes [77], migration [68], phenological [19] or survival patterns [57] from ecological data at one or more sites.

Although these models are developed for data collected using different sampling protocols, they all focus on the estimation of the entry, exit and length of stay (LOS) patterns of populations, where entry can correspond to arrival/birth, exit to departure/death and LOS to retention/survival at the site. In this paper, we develop a general framework for building models to infer entry, exit and LOS patterns in a Bayesian setting and demonstrate that a number of existing statistical models for ecological data can be expressed within our new framework.

Several sampling schemes exist for monitoring wildlife populations. The choice of the sampling scheme to be employed in each case depends on the characteristics of the species of interest and of the study site, time and funding constraints and the research question(s) that each survey is looking

to answer. In this paper, we focus our attention on three of the most common sampling schemes, namely the collection of capture-recapture (CR) data, count data (CD) and ring-recovery (RR) data. Nevertheless, we note that our approach is not limited to these types of data and could easily be extended for example to removal data [65] or occupancy data [53]. Furthermore, we note that we only consider data collected on open populations, with individuals entering, by immigration or birth, and exiting, by emigration or death, the population during the study.

In all the aforementioned sampling protocols, data are collected at discrete times, typically referred to as sampling occasions. Every pair of consecutive sampling occasions defines an *interval*, during which individuals may enter or exit the study area. This suggests that an appropriate modelling approach for the individual *entry* and *exit* intervals should consider the *grid* defined by the sampling occasions in the bivariate space of entry and exit. This idea is exploited in Diana et al. [20] (D18), who defined a model for CD that allows the modelling of the entry and exit patterns. Their modelling approach is based on the Polya Tree (PT) prior [29], a popular nonparametric prior for densities, that is also a natural choice when modelling distributions defined on a grid.

The PT defines a prior on a distribution by defining an infinite sequence of nested partitions of the sample space, and recursively assigning the masses of the subsets of these partitions. Although the number of partitions is in principle infinite, the PT prior is usually defined only up to a specified level, which corresponds to defining a highly flexible parametric model instead of a nonparametric model. A good description of the PT and of related applications is given in [37].

Our modelling approach relies on defining a finite partition, where the last level corresponds to the grid defined by the sampling occasions, as explained above. In this way, we are defining a prior distribution for the probabilities of an individual to fall in each cell of the grid and there is no need to build a finer partition, as no information exists for partitions finer than the one defined by the grid. Once this grid is defined, we do not model individual entry and exit intervals but instead we work with the number of individuals in each cell of the grid. The advantage of the grid is that the computational complexity of our model depends on the number of sampling occasions instead of the number of individuals, which can lead to substantially faster inference when the number of individuals is much larger than the size of the grid.

We extend the approach of D18 by defining a model for CR, CD and RR, either in isolation or in combination by jointly modelling data collected using different sampling schemes. In addition, we develop the following two extensions of the PT: the replicate PT (RPT), which shares random variables (r.v.s) across different PTs, allowing us to impose constraints on model parameters, and the Hierarchical Logistic PT (HLPT), that allows us to model r.v.s across different distributions in a hierarchical fashion implying exchangeability across the different units. Finally, we consider a scenario when data are collected repeatedly and in high frequency. In this case, choosing the partition at which the data are analysed raises problems from a parsimony perspective, as choosing an excessively fine partition considerably increases the number of parameters. To overcome this problem, we use the Optional Polya Tree (OPT) of Wong et al. [89], which is an extension of the PT that allows us to infer the partition structure.

The paper is structured in the following way. In Section 4.2.1 we provide an overview of the sampling schemes considered as case studies in the paper. In Section 4.2.2 we describe the main features of the PT prior. In Section 4.3 we present and motivate the univariate partitions of the PT used in the paper and demonstrate the concept using CR data. The RPT prior is defined in Section 4.4. In Section 4.5.1 we define the bivariate partitions of the PT and define the joint model for CR data and CD and the model for RR data. The HLPT is defined in Section 4.6.1 and the use of the OPT is described in Section 4.6.2, with both sections presenting an application to CD collected across several sites or years. Section 4.7 concludes the paper and introduces some potential future directions.

4.2 Background

4.2.1 Sampling schemes of population surveys

In this paper, we consider models for data resulting from three types of ecological sampling schemes: CR, CD and RR. In all the three cases, we assume sampling is performed on K sampling occasions at times $t_1, \ldots, t_K$.

In the case of CR data, on each sampling occasion, a sample of the individuals is caught, and newly caught individuals are uniquely marked. If D is the total number of captured individuals, the captures can be summarised in a $D \times K$ matrix, H, where H_{ik} is 1 if the i-th individual was captured on sampling occasion k and 0 otherwise. Each row of the matrix is called a capture history. It is common to assume that there are no losses on capture so that all of the captured individuals are returned to the population after each sampling occasion. The starting point for modelling CR data is the Cormack-Jolly-Seber (CJS) model [14, 44, 81]. The CJS model is built conditional on the time of first capture of each individual. After the first capture, alive individuals of age a at time t_j are assumed to remain alive at the site until the following sampling occasion at time t_{j+1} with probability $\phi_{j,a}$ and, conditionally on being present at the site, they can be recaptured (or resighted) with probability p. In the following, we will always use the indexes f, k and l to indicate the times of first possible capture, first capture and last possible capture, respectively. As the model conditions on first capture, it does not allow the estimation of the population size but only of $\phi_{j,a}$ and p. We express the CJS model in a PT framework in Section 4.3.1.

In order to simultaneously model entry and exit from the population, Jolly-Seber [44, 81] type models, which estimate individual entry and exit intervals, have recently been developed [55]. These individual entry and exit intervals give rise to inferred individual *presence histories*, with entries equal to 1 on sampling occasions when the individual was available for capture and 0 otherwise. It is generally assumed that exit is permanent, that is, once an individual leaves the site it does not return.

CD are collected by visiting the site and simply detecting a number of individuals, without attempting to individually identify them. Similarly to CR data, detection probability in this case is less than 1 so that the counts do not constitute a complete census of the population. The data can be summarised in a vector $(C_1, \ldots, C_K)$, where C_j denotes the number of individuals detected on

sampling occasion j. This protocol is less costly and less disturbing for wildlife than collecting CR data, and hence CD have been widely used for monitoring purposes, and particularly for estimating the size of populations [77]. However, CD are not as informative as CR data because of the lack of information at the individual level.

Finally, according to the RR protocol, a subset of the population is captured and marked each year. In addition, individuals marked in previous years may be recovered after they die with probability λ. The data are usually summarised in an upper-triangular $K \times K$ matrix, R, where cell (i, j) denotes the number of individuals marked in year i and recovered dead in year j, and in a vector $(m_1, \ldots, m_K)$ where m_k denotes the total number of individuals marked in year k. RR data are typically used to estimate annual survival probabilities. [31] proposed a modelling strategy for birds of known age. Thereafter, several models have been proposed to circumvent the problem of unknown age [57, 42]. In this paper, we consider a model where individuals are of known age if caught at an age less than 1 year and jointly model data on individuals of known and unknown age, but greater than 1 year when unknown.

A more detailed description of these sampling schemes, as well as related sampling schemes, is provided in McCrea and Morgan [56].

4.2.2 The Polya Tree prior

The Polya tree (PT) prior [29, 48] is a nonparametric prior for probability distributions. It has two parameters: the first is a sequence of nested partitions Π of the sample space Ω, while the second parameter, α, is a sequence of positive numbers associated with each set of each partition. The PT is a more general case of the Dirichlet process (DP) [28], but as opposed to the DP the PT can induce positive correlation between adjacent sets [29], which instead always assigns negative correlation.

The partition at the first level, π_1, is obtained by splitting the sample space in K_Ω sets $\{B_0, \ldots, B_{K_\Omega - 1}\}$. Subsequently, to build the partition at the second level, π_2, each set B_i is split into K_i sets $\{B_{i0}, \ldots, B_{iK_i - 1}\}$, next each set B_{ij} is split into K_{ij} sets $\{B_{ij0}, \ldots, B_{ijK_{ij}-1}\}$ and so on. A visual representation of the scheme for dyadic splits ($K_{\varepsilon_1 \ldots \varepsilon_m} = 2\ \forall m$), is presented in Fig. 4.1.

By defining $\varepsilon_1 \ldots \varepsilon_m$ as a generic sequence of positive integers (such that $\varepsilon_i \in \{0, \ldots, K_{\varepsilon_1 \ldots \varepsilon_{i-1}} - 1\}$) and $B_{\varepsilon_1 \ldots \varepsilon_m}$ as the correspondent set of the partition, the random mass associated to $B_{\varepsilon_1 \ldots \varepsilon_m}$ by the PT is given by

$$G(B_{\varepsilon_1 \ldots \varepsilon_m}) = \prod_{i=1}^{m} V_{\varepsilon_1 \ldots \varepsilon_i} \tag{4.2.1}$$

where $(V_0, \ldots, V_{K_\Omega - 1}) \sim \text{Dirichlet}(\alpha_0, \ldots, \alpha_{K_\Omega - 1})$ and $(V_{\varepsilon_1 \ldots \varepsilon_{i-1} 0}, \ldots, V_{\varepsilon_1 \ldots \varepsilon_{i-1} K_{\varepsilon_1 \ldots \varepsilon_{i-1}} - 1}) \sim \text{Dirichlet}(\alpha_{\varepsilon_1 \ldots \varepsilon_{i-1} 0}, \ldots, \alpha_{\varepsilon_1 \ldots \varepsilon_{i-1} K_{\varepsilon_1 \ldots \varepsilon_{i-1}} - 1})$ for $i > 1$. For example, in the case where splits are dyadic, $G(B_{02}) = V_0 V_{02}$ where $(V_0, V_1, V_2) \sim \text{Dirichlet}(\alpha_0, \alpha_1, \alpha_2)$ and $(V_{00}, V_{01}, V_{02}) \sim \text{Dirichlet}(\alpha_{00}, \alpha_{01}, \alpha_{02})$. The V's can be interpreted as the conditional probabilities of falling in a particular set in the next level of the tree, that is, given $y \sim G$, where $G \sim \text{PT}$, $\mathbb{P}(y \in B_{\varepsilon_1 \ldots \varepsilon_m} | y \in B_{\varepsilon_1 \ldots \varepsilon_{m-1}}) = V_{\varepsilon_1 \ldots \varepsilon_m}$.

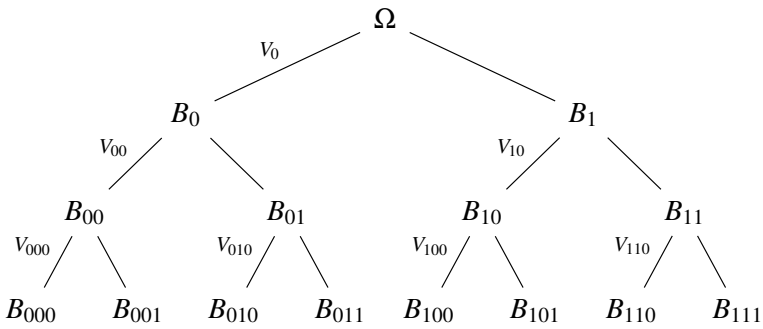


Fig. 4.1 Levels of partition of a PT with dyadic splits. Following Eq. 4.2.1, the mass in each set B is assigned by starting from Ω and multiplying the total mass by V_ε for a left split or $1-V_\varepsilon$ for a right split along the path until B is reached.

The PT is a conjugate prior for a distribution G if we have observations $y_1,\dots,y_n \overset{\text{i.i.d.}}{\sim} G$, since the posterior distribution $G \mid y_1,\dots,y_n$ is a PT with parameters Π and $\alpha^\star$, where $\alpha^\star_\varepsilon = \alpha_\varepsilon + n_\varepsilon$ and n_ε is the number of observations falling into set B_ε. Hence, the posterior distribution of a PT can be calculated without using MCMC techniques. In the following, we use the PT mainly with dyadic splits, which leads to a Beta distribution for the masses assigned in each split instead of a Dirichlet distribution.

The PT can be centered on any distribution G_0 in the sense that, for every set B of the partition, $\mathbb{E}[G(B)] = G_0(B)$. This can be achieved by choosing the α such that $\frac{\alpha_{\varepsilon_1\dots\varepsilon_m}}{\sum_{i=1}^{K_{\varepsilon_1\dots\varepsilon_{m-1}}} \alpha_{\varepsilon_1\dots\varepsilon_i}} = \frac{G_0(B_{\varepsilon_1\dots\varepsilon_m})}{G_0(B_{\varepsilon_1\dots\varepsilon_{m-1}})}$ where $\sum_{i=1}^{K_{\varepsilon_1\dots\varepsilon_{m-1}}} \alpha_{\varepsilon_1\dots\varepsilon_i}$ tunes the variance of the Vs.

In what follows, we will generally center the PT on a Laplace distribution, as the cumulative distribution function of the Laplace, required for computing the masses $G_0(B_{\varepsilon_1\dots\varepsilon_m})$, can be calculated analytically without having to perform numerical integration.

If we assume that the centering distribution G_0 has a parameter η on which an additional prior P_0 is assumed, we obtain what is known as Mixtures of Polya Trees (MPT), as defined in Hanson [36]. Hanson [36] originally used the MPT to obtain absolutely continuous draws from a PT with probability 1, which is achieved by obtaining a partition that varies according to the parameter η. In our case, we choose a fixed partition and hence we use the MPT only as a way to allow uncertainty in the value of the parameter η.

4.3 Univariate PT models

The partition of the PT can be chosen according to the specific application. To define a prior for the distribution of a time-to-event on the real line, Lavine et al. [49] (L94) proposed the following partition. Assume we have sampling times $t_1,\dots,t_K \in \mathbb{R}$ and we only observe the interval where each time-to-event x happens, that is, we observe whether $x < t_1$, $t_1 < x < t_2$ and so on. L94 proposed to

define Π such that $B_0 = (-\infty, t_1)$, $B_1 = (t_1, \infty)$, with B_1 further split into $B_{10} = (t_1, t_2)$, $B_{11} = (t_2, \infty)$ and so on. A representation is shown in Fig. 4.2. Following the notation of Section 4.2.2, this structure is obtained when $K_{B_0} = K_{B_{10}} = K_{B_{110}} = \cdots = 1$ and $K_\Omega = K_{B_1} = K_{B_{11}} = K_{B_{111}} = \cdots = 2$.

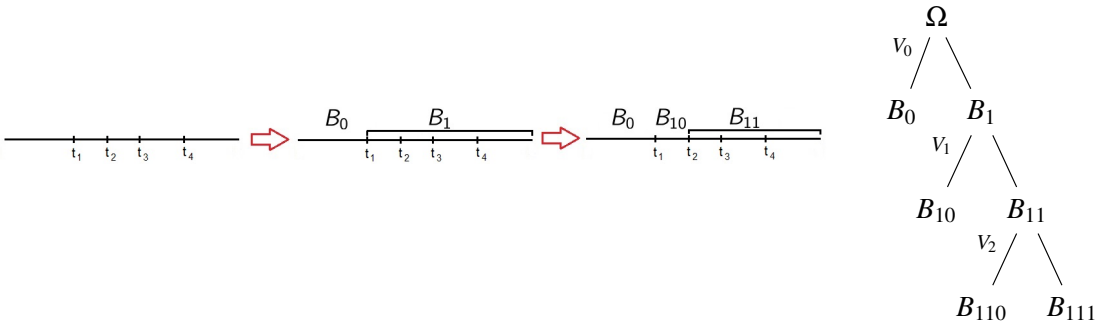


Fig. 4.2 Sequence of splits (left) and structure (right) of the L94 partition. The r.v. assigning the ratio of the masses of two sets for each branch is represented on the left branch. The mass assigned to the set in each right branch is always one minus the mass assigned to the set in the corresponding left branch.

Samples from a PT with the partition defined above have decreasing variance starting from the earlier sampling times. This follows easily from the following lemma, proven for simplicity for Beta r.v.s with mean $1/2$.

Lemma 1. *Given* $X_1 \sim \text{Beta}(\alpha, \alpha)$ *and* $X_2 \sim \text{Beta}(\beta, \beta)$*, if* $\beta \geq \alpha$ *then* $\text{Var}(\text{X}_1\text{X}_2) \leq \text{Var}(\text{X}_1)$.

The previous lemma shows that if the r.v.s $V_0, V_1, \ldots$, assigning the mass to the sets $B_0, B_1, \ldots$, have parameters whose sum is not decreasing, the variance of V_0, V_0V_1, $V_0V_1V_2$, ... is decreasing. Conversely, we can assume the variance is decreasing starting from the last sampling times by reversing the order of the split and splitting the partition using first t_K, then t_{K-1} and so on. We term the two partitions defined above as the forward LOS partition (FLP) and backward LOS partition (BLP), respectively. The idea of L94 is useful when modelling ecological data since, as mentioned above, sampling is typically performed only at discrete times, $t_1, \ldots, t_K$. An application of this partition is presented in Section 4.3.1.

In some cases, it is useful to build the partition in a single step by splitting using all the sampling times simultaneously. That is, we define $(B_0, \ldots, B_K)$ such that $B_i = (t_i, t_{i+1})$, where $t_0 = -\infty$, $t_{K+1} = +\infty$, and the masses are assigned according to a Dirichlet distribution. As mentioned before, this choice implies negative correlation between probabilities of any two disjoint sets. This partition is termed the uniform partition (UP).

4.3.1 Example: Cormack-Jolly-Seber

In this section we demonstrate how the CJS model can be expressed in a PT framework. As in the original CJS model, we define the model for the data conditioning on the time of first capture.

We first arrange all the marked individuals in a set of vectors $\mathbf{n}^k = (n_1^k, \ldots, n_{K-k+1}^k)$, where n_j^k indicates the number of individuals with first capture at time t_k and remaining at the site for j sampling occasions or, equivalently, with last *possible* capture at time t_{k+j-1}. Next, we assume a distribution G_k on the time of last possible capture for the individuals with first capture at time t_k or, equivalently, on the individuals belonging to vector $\mathbf{n}^k$. The sample space for G_k is $\Omega_k = (t_k, \infty)$, as exit time is left truncated by the time of first capture. Next, we assume a PT prior for each G_k, with partition taken to be the FLP. According to the FLP, the sample space Ω_k is split into $B_0^k = (t_k, t_{k+1})$ and $B_1^k = (t_{k+1}, \infty)$, next B_1^k is further split into $B_{10}^k = (t_{k+1}, t_{k+2})$ and $B_{11}^k = (t_{k+2}, \infty)$ and so on. The last level of the partition corresponds to the grid defined by the capture times, $(t_k, t_{k+1}, \ldots, t_K)$, and thus it uniquely determines the probabilities of each cell. A visual representation of the sets of the sample space is shown in Fig. 4.3. We note that the K partitions have different depths as for the individuals with first capture at time t_k, the partition is built only for $K-k+1$ levels.

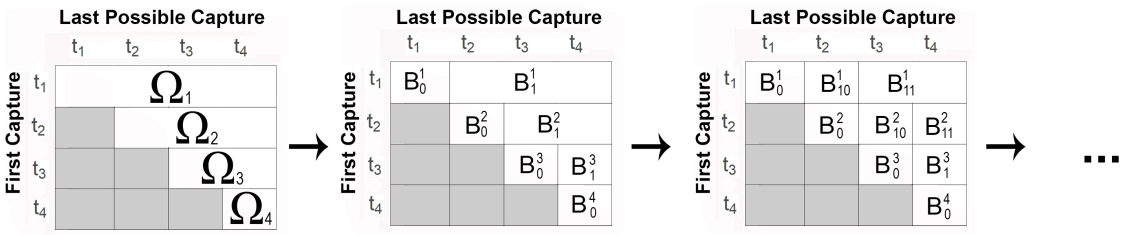


Fig. 4.3 Partitions of the PT in the CJS case. The first level corresponds to the sample spaces $\Omega_k = (t_k, \infty)$, consisting of the individuals with first capture at time t_k. In the second level, each sample space is split into $B_0^k = (t_k, t_{k+1})$ and $B_1^k = (t_{k+1}, \infty)$ and so on. The sets $B_{\mathbf{1}_j 0}^k$ and $B_{\mathbf{1}_j 1}^k$ correspond to the individuals first caught at time t_k that are present at the site at time t_{k+j} and exit and remain, respectively, at the site until the next sampling time, t_{k+j+1}.

For illustration purposes, we are going to assume that individuals are of age 1 when first caught at time t_k but will later show how to relax this assumption. We define V_j^k to be the probability of an individual aged 1 when first caught at time t_k exiting before time t_{k+j} conditional on being present and of age j at the site at time t_{k+j-1}. The V_j^k can also be seen as the ratio of masses $\frac{G_k(B_{\mathbf{1}_j 0})}{G_k(B_{\mathbf{1}_j})}$ (where $\mathbf{1}_j = (\underbrace{1, \ldots, 1}_{j \text{ times}})$) in the PT, as in Eq. 4.2.1, and hence they are assigned a Beta prior distribution, where the parameters are determined by centering the PT on a truncated Laplace distribution.

As mentioned in Section 4.2.1, in the standard CJS model, $\phi_{j,a}$ corresponds to the probability that an individual present and aged a at time t_j remains until time t_{j+1}. As individuals first caught at time t_k are of age j at time t_{k+j-1}, according to our model $V_j^k = 1 - \phi_{k+j-1,j}$.

The advantage of using a PT prior is that different assumptions regarding the model for parameter ϕ can be considered by sharing the r.v.s V_j^k across the different PTs. We list below a number of assumptions that can be made and demonstrate the concept by writing the probability of capture history 0101, $\mathbb{P}(0101)$, according to each assumption.

- *Constant case*: $\phi_{l,a} \equiv \phi \Rightarrow V_j^k \equiv V \ \forall k, j$. $\mathbb{P}(0101) = \phi^2 p(1-p)$.
- *Age-dependent case*: $\phi_{l,a} \equiv \phi_a \Rightarrow V_j^k \equiv V_j \ \forall k$, that is equivalent to assuming that $G_k \equiv G \ \forall k$. $\mathbb{P}(0101) = \phi_1 \phi_2 p(1-p)$.
- *Time-dependent case*: $\phi_{l,a} \equiv \phi_l \Rightarrow V_j^k \equiv V_{k+j}$. This is equivalent to assuming that the distribution of $G_k(x)$ is the same as the distribution of $G_{k-1}(x|x \in B_{k-1,1}) \ \forall k$. $\mathbb{P}(0101) = \phi_2 \phi_3 p(1-p)$.
- *Unconstrained case*: The most general case is obtained by assuming that all the V_j^k are different. Although this model is never fitted in practice because it is overparametrised, we can still write $\mathbb{P}(0101)$ as $\phi_{2,1}\phi_{3,2}p(1-p)$.

These assumptions are also depicted in Fig. 4.4.

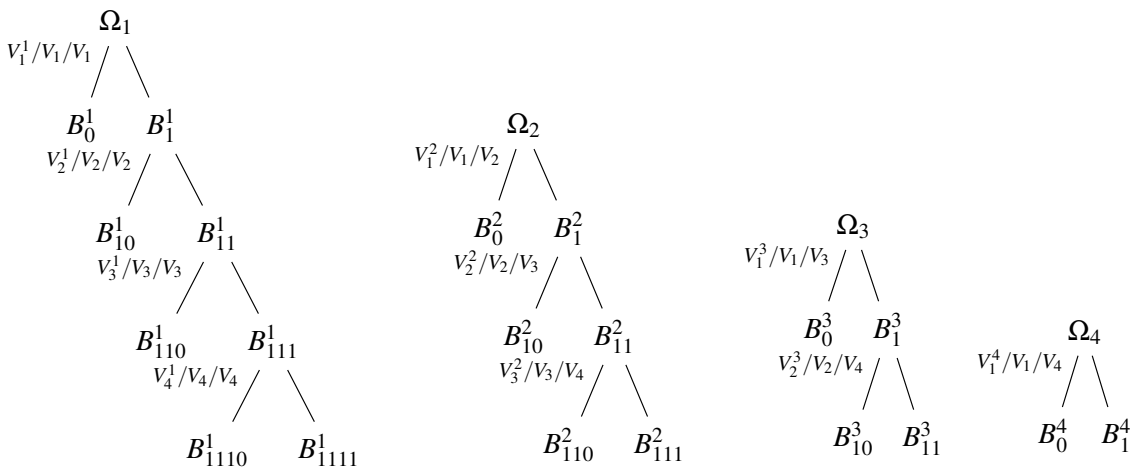


Fig. 4.4 Structure of the trees for the CJS model. The r.v. assigning the ratio of the masses of two sets for each branch is represented on the left branch. The r.v.s are in the order *unconstrained*/*age-dependent*/*time-dependent* case. The mass assigned to the set in each right branch is always one minus the mass assigned to the set in the corresponding left branch.

Assuming independence between sampling occasions and between individuals, the probability of observing the CR data, H, given the vectors $\mathbf{n}^k$ can be written explicitly. Although the $\mathbf{n}^k$ provide the frequencies of all possible presence histories, they do not give us a way to associate individuals and their capture histories to presence histories. For example, an individual with capture history 011 might have presence history 111 or 011. Therefore, we cannot simply express the probability of each observed individual capture history conditional on the corresponding inferred presence history to construct the likelihood function, but we need to compute the number of possible ways that the capture histories can be assigned to the inferred presence histories. The complete likelihood function is given in the supplementary material.

We note that, as opposed to the standard PT case where the observations, in this case the exit intervals, are observed, in our model we need to infer the exit intervals for each individual and therefore we need to employ an MCMC algorithm. The MCMC algorithm is used to sample from the posterior distributions of the latent vectors $\mathbf{n}^k$, the probability p of capturing an available individual and the r.v.s V. The details are given in Section 3.1 of the supplementary material.

The computational complexity of the algorithm is of order K^2, where K is the number of sampling occasions, as the dimension of the set of vectors $\{\mathbf{n}^k\}$ is K^2 (up to a proportionality constant). Alternatively, when the number of caught individuals D is smaller compared to K^2, an alternative scheme is to consider the individual exit intervals of each individual instead of the set of vectors $\{\mathbf{n}^k\}$. That is, we model directly the exit interval d_i of each marked individual, as for example in [55] and Diana et al. (in press).

4.4 Replicate PT

In the CJS model defined in the previous section, different models are obtained by assuming that parts of different trees are "replicated" across different branches. The same idea will also be used in the next sections, and hence we provide here a formal definition of the RPT. The definition we provide is for a single tree but it can easily be extended to the case of multiple trees.

Let Σ be a rule to generate a sequence of partitions from an initial seed set. For example, in the CJS example the rule Σ corresponds to FLP and the seed sets are $\Omega_1, \ldots, \Omega_K$. If G is a distribution having a PT prior with partition Π, we say that G has a replicate PT (RPT) structure across two sets B_ε and $B_{\varepsilon'}$ of the partition Π if the following constraints hold:

- The partitions of the trees starting from B_ε and $B_{\varepsilon'}$ are generated according to the same rule Σ, although the two partitions could be stopped after a different number of steps;
- For all $\varepsilon_1, \ldots, \varepsilon_m$, $V_{\varepsilon,\varepsilon_1,\ldots,\varepsilon_m} = V_{\varepsilon',\varepsilon_1,\ldots,\varepsilon_m}$.

The first condition states that the partitions of the trees starting from B_ε and $B_{\varepsilon'}$ are the same for a number of steps, while the second condition states that the r.v.s in the two trees are the same. The second condition is motivated by the CJS example, where the trees have different lengths, but it is also required in the models defined later in the paper. The definition is easily extended to the case where more than two parts of a tree are shared.

We can now express some of the assumptions of the CJS model in terms of the RPT just defined.

- The age-dependent case is obtained assuming that the trees have a replicate structure across the sets $\Omega_1, \ldots, \Omega_K$.
- The time-dependent case is obtained assuming a replicate structure across the pairs of trees with seed sets: Ω_2 and B_1^1, Ω_3 and B_1^2, Ω_4 and B_1^3 and so on.

For ease of exposition, in all the case studies shown later, we only consider the most commonly employed constraint in each case. However, similarly to the CJS case described above, different RPT structures can be employed in order to assume different constraints if required.

We note that assuming that the r.v.s are shared across different parts of the tree does not change the conjugate scheme described in Section 4.2.2. In fact, the conjugacy of the PT comes from the conjugacy of the Dirichlet prior with a multinomial likelihood, and replicating different parts of the trees corresponds to having more than one multinomial distributions in the likelihood, which does not affect the conjugacy of the model.

4.5 Bivariate PT models

4.5.1 Bivariate partitions

In this section, we extend the model for CR data conditional on the time of first capture of Section 4.3.1 to two scenarios: CR without conditioning on first capture and RR. To highlight the connection with D18, we briefly describe their model here. D18 define a model for count data by introducing a matrix $\{n_{fl}\}$ of the individuals with first possible detection at time t_f and last possible detection at time t_l. Therefore, inferring matrix n_{fl}, which also includes information on the latent entry intervals, gives rise to an estimate of the entry pattern and the size of the population.

In the CR and RR extensions, we do not condition on first capture as in the CJS but we model the entry intervals of the individuals, and thus instead of working with a set of vectors $\{n_j^k\}$ we will introduce a set of matrices $\{n_{fl}^k\}$, similarly to D18. We note that, as opposed to D18, in the case of uniquely identifiable individuals, as in CR and RR data, the time of first capture t_k is also needed to obtain an analytical expression for the likelihood function. However, the downside of adding another dimension is that it increases the number of parameters of the model. Hence, we share information between individuals with different times of first capture by using the RPT structure, in a similar fashion as this was performed for the CJS.

Although both the RR and the CR models are built in principle by using the same set of matrices $\{n_{fl}^k\}$ to summarise the individuals, the main difference between the two models is that the priors are elicited on different quantities. In particular, in the case of CR data, the prior is elicited on the entry and exit pattern in the case of CR, while in the case of RR data, the prior is elicited on entry and LOS patterns. This is because the CR data we consider here are of the type collected at migratory and breeding sites, where prior information is available on when the individuals are expected to arrive and depart. On the other hand, RR data are collected over multiple years, and hence prior information in this case is available on how long individuals are expected to live.

In order to define the aforementioned models, we need to define partitions for distributions on $\mathbb{R}^2$, as we consider joint distributions on entry and exit or on entry and LOS. Partitions for distributions on $\mathbb{R}^2$ can be defined by using the schemes defined in Section 4.3 separately for each dimension. For example, a bivariate partition can be constructed by applying first FLP in one dimension and

next BLP in the other. If $B^1_{\varepsilon_1,\dots,\varepsilon_k}$ and $B^2_{\varepsilon_1,\dots,\varepsilon_k}$ are the sets generating with the FLP and BLP partition, respectively, the bivariate partition is defined in the following way:

- For the first K levels, we define the partition at the k-th level as $(\tilde{B}^k_0,\dots,\tilde{B}^k_k)$ where

$$\begin{cases} \tilde{B}^k_i = B^1_{\mathbf{1}_{i-1},0} \times (-\infty,\infty) & i = 1,\dots,k-1 \\ \tilde{B}^k_k = B^1_{\mathbf{1}_{i-1},1} \times (-\infty,\infty) & i = k \end{cases}.$$

  This scheme is depicted in the first K levels of Fig. 4.5.

- For the remaining K levels, we define the partition at the k-th level as $(\tilde{B}^k_{0,0},\dots,\tilde{B}^k_{0,k},\dots,\tilde{B}^k_{K,0},\dots,\tilde{B}^k_{K_1,k},)$ where

$$\begin{cases} \tilde{B}^k_{i,j} = B^1_{\mathbf{1}_K} \times B^2_{\mathbf{1}_{j-1},0} & j = 1,\dots,k-1 \\ \tilde{B}^k_{i,j} = B^1_{\mathbf{1}_K} \times B^2_{\mathbf{1}_{j-1},1} & j = k \end{cases}.$$

  This is depicted in the last K levels of Fig. 4.5.

We note that the previous construction is possible as both partitions used for splitting are finite, as is the case for all the applications considered in this paper. To simplify the notation, we will later use B instead of $\tilde{B}$ to indicate the sets of the bivariate partition. We denote this partition as entry and exit length of stay partition (EELP).

This partition is useful when the entry and exit distribution (which is defined on $\{(x,y) : (x,y) \in \mathbb{R}^2, y > x\}$) is of interest, as the marginal distribution of entry is expected to have more variance in the left tail, while for the marginal exit distribution more variance is assumed in the right tail. This is especially the case when modelling stopover data [55] because entries into the population are concentrated at the start of the sampling period while exits from the population are concentrated at the end. This partition is used in D18, who built a model for CD based on a PT. The model is further extended to CR data in Section 4.5.2.

When no prior information is available on the variance of the entry or exit distributions, the length of stay splits can be replaced by the uniform split described in Section 4.3, as shown in Fig. 4.6. A partition built in this way is termed as uniform partition (UP) and is used in Section 4.6.2.

The previous partitions allow us to elicit prior information on the entry and exit distribution. However, as mentioned above, in some applications, prior information is available on the LOS distribution instead and hence it is useful to build a partition that allows us to elicit prior information on LOS instead of exit. For eliciting a prior on the LOS distribution, we define a bivariate partition based on the idea of L94 but split the sample space with respect to the LOS instead of the exit intervals. This is done by splitting the sample space in the set $B_0 = \{(x,y) : t_k < x < y < t_{k+1}\}$, consisting of the individuals exiting before the next sampling occasion, and the set $B_1 = \{(x,y) : t_k < x < t_{k+1} < y\}$, consisting of the individuals exiting after the next sampling occasion. Next, we split B_1 in the set

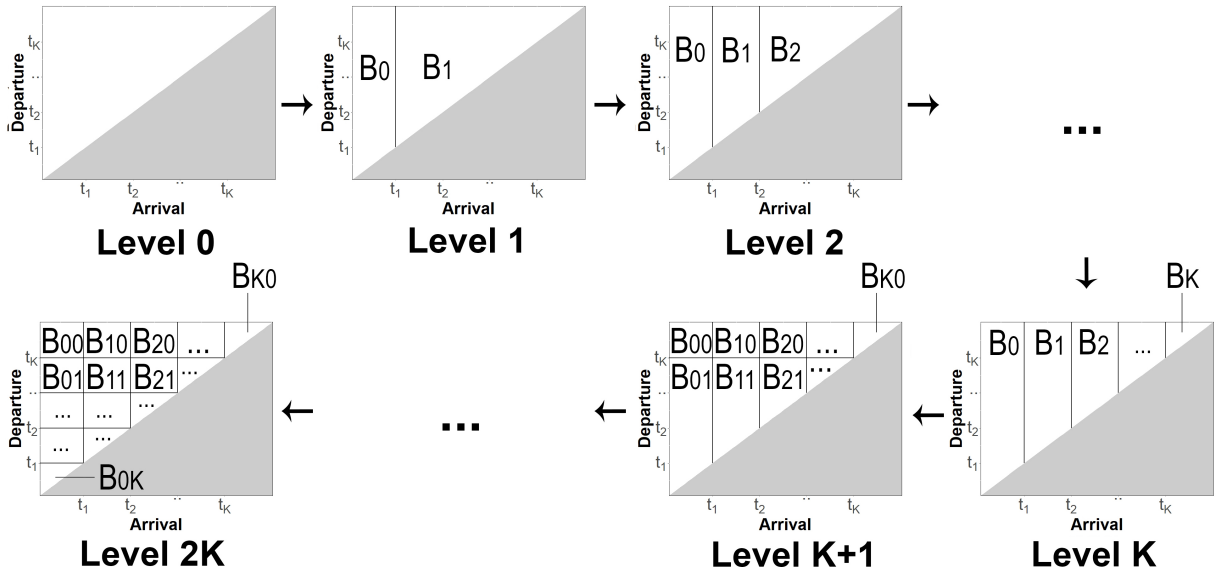


Fig. 4.5 Partition of a PT prior for a distribution defined on $\{(x,y) : (x,y) \in \mathbb{R}^2, y > x\}$, built first using FLP and next using BLP. For simplicity, we have used the notation B for the sets of the bivariate partition instead of $\tilde{B}$ and omitted the superscripts.

$B_{10} = \{(x,y) : t_k < x < t_{k+1} < y < t_{k+2}\}$ of individuals staying for only one sampling occasion and the set $B_1 = \{(x,y) : t_k < x < t_{k+1}, y > t_{k+2}\}$ of individuals remaining for more than one sampling occasions. The process is repeated K times for the remaining sampling occasions. Next, each set is split according to a uniform split using the entry intervals. This partition is termed as bivariate LOS partition (BivLP). A visual representation of this scheme is shown in Fig. 4.7 and an application is presented in Section 4.5.3.

4.5.2 Joint analysis of capture-recapture and count-data

Recent work [2] casted doubt on whether CD alone can be used to provide reliable estimates of population size and suggests that analysis of CD should be augmented by other types of data, such as CR. Therefore, in this section, we demonstrate how to extend D18 to perform a joint analysis of CR and CD using an RPT. As introduced in the previous section, we define the set of matrices $\{\mathbf{n}^k\} = \{n^k_{fl}\}_{f=1,\ldots,K+1, l=1,\ldots,K+1}$, where n^k_{fl} denotes the number of individuals with first possible capture at time t_f, first capture at time t_k and last possible capture at time t_l. For $k = 0$, we define $\mathbf{n}^0$ as the matrix containing the individuals never captured, but possibly detected and hence contributing to the CD.

For each matrix $\mathbf{n}^k$, we define a distribution G_k on the probabilities ω^k_{fl} of an individual belonging to each cell of the matrix $\{n^k_{fl}\}$, and we assume a PT prior for each G_k using the EELP partition of Section 4.5.1. We make the assumption that the distribution of entry and exit intervals is the same for each matrix $\mathbf{n}^k$. If Ω_k is the sample space for the individuals with first capture at time t_k, this is

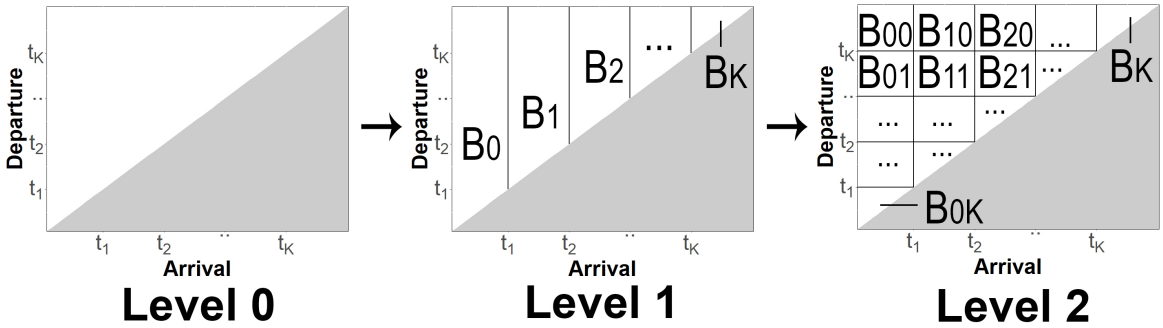


Fig. 4.6 Partition of a PT prior for a distribution defined on $\{(x,y) : (x,y) \in \mathbb{R}^2, y > x\}$ built using the uniform split twice.

achieved by assuming a RPT prior across the sets $\Omega_1, \ldots, \Omega_K$, as in Section 4.3.1. Moreover, we also assume that the exit distribution is independent of the entry interval. This is achieved by assuming, for each slice k, a RPT structure across the sets $B_0 = \{(x,y) : x < t_1\}$, $B_1 = \{(x,y) : t_1 < x < t_2\}$, $\ldots$, $B_{k-1} = \{(x,y) : t_{k-1} < x < t_k\}$. Similarly to the CJS model of Section 4.3.1, the parameters of the Beta r.v.s assigning the mass in each split of the entry and exit distribution are centered on two independent Laplace distributions.

We note that each matrix is defined on a different sample space, as $\mathbf{n}^k$ is defined only for $f \leq k, l > k$, because individuals in the k-th slice have to enter before the k-th sampling occasion and exit after. A representation of one slice is shown in the supplementary material.

We define the sum of the elements of matrix $\mathbf{n}^k$ as D_k. However, while D_k is known for $k = 1, \ldots, K$ (as it is equal to the number of individuals with first capture at time t_k), D_0 is unknown as it includes the number of individuals never captured. Given the number of individuals D_k, the number of individuals in each cell of the matrix is distributed as a multinomial with sample size D_k and vector of probabilities ω^k_{fl}. The total number of individuals N is equal to $D_0 + \sum_{k=1}^K D_k$, where $D = \sum_{k=1}^K D_k$ is the number of caught individuals. Assuming a Poisson(ω) prior distribution for N, the model for the latent matrices n^k_{fl} can be summarised as

$$\begin{cases} n^k_{fl} \sim \text{Multinomial}(D_k, \omega^k_{fl}) \\ D_0 + D \sim \text{Poisson}(\omega) \end{cases}$$

Conditional on the latent matrices n^k_{fl}, we can easily express the distribution of the counts C and the CR matrix H. Defining as N_k the number of individuals available for capture/detection at time t_k the model for the CD, conditional on the latent matrices n^k_{fl}, can be expressed as

$$\begin{cases} C_j \sim \text{Binomial}(N_j, p_D) \\ N_j = \sum_{k=0}^j \sum_{l=j}^K \sum_{f=1}^{l-1} n^k_{fl} \end{cases}$$

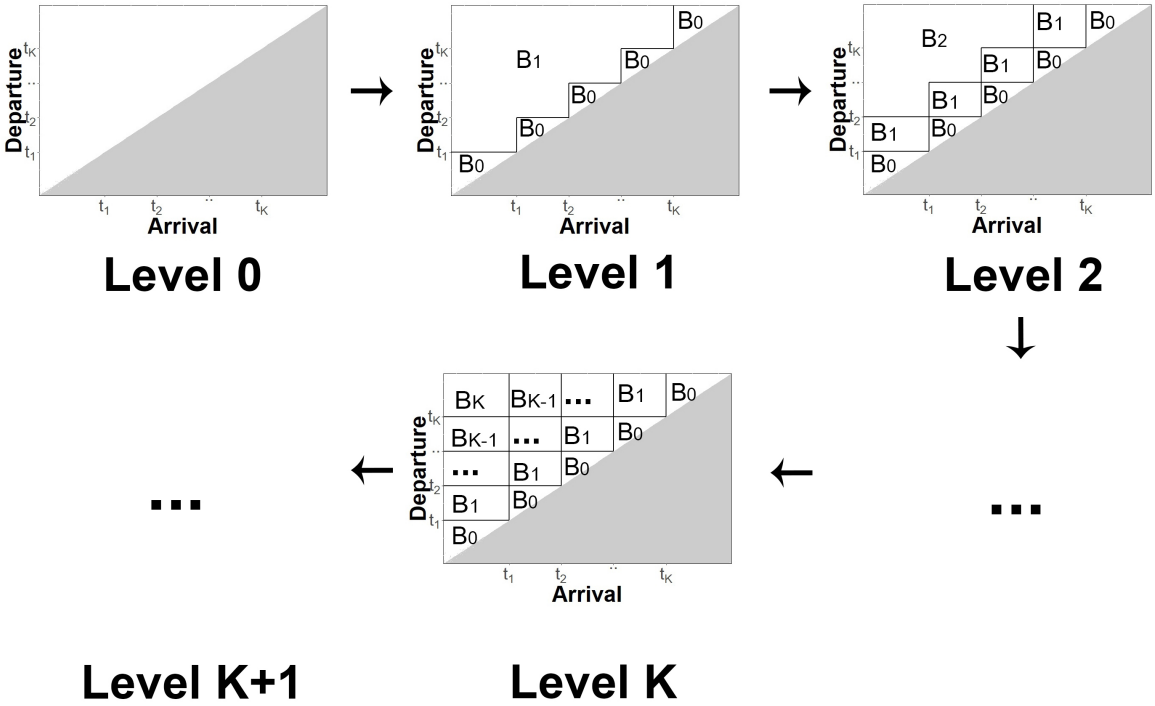


Fig. 4.7 Partition of a PT prior for a distribution defined on $\{(x,y) : (x,y) \in \mathbb{R}^2, y > x\}$ built using BivLP.

where the last equation expresses the relationship between the N_k and n^k_{fl} and p_D and p_C are the probability of detecting and capturing an individual, respectively.

The probability of the capture histories of the individuals with first capture at time t_k, conditional on the matrix $\mathbf{n}^k$, is presented in the supplementary material.

As in the CJS model, we can reformulate this model using individual entry and exit intervals, as performed also in Matechou et al. [55] and Diana et al. (in press). We reformulate the model using the following variables. The D captured individuals are summarised with the variables (t^i_1, t^i_2), describing the latent entry and exit interval, respectively, of the i-th captured individual, while the individuals never captured are summarised in a matrix $\{n_{fl}\}$, where cell (f,l) is the number of individuals with first possible capture at time t_f and last possible capture at time t_l.

The model presented using the set of matrices $\{\mathbf{n}^k\}$ has computational complexity K^3 (as the dimension of the set of matrices $\mathbf{n}^k$ is K^3 up to a proportionality constant), while modelling the individual entry and exit intervals leads to an order of computational complexity equal to the maximum between the number of captured individuals D and the dimension K^2 of the matrix $\{n_{ij}\}$.

Case study

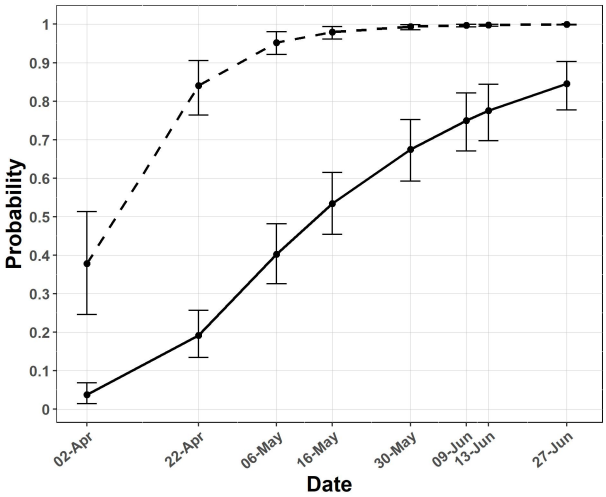
We apply our model to a dataset of counts and resights of a population of Italian spoonbills (*Platalea leucorodia*) collected in the southern Po Delta (NE Italy). Birds captured as chicks in previous years are resighted as adults through photographs obtained using camera traps and by visiting their nests on eight separate sampling occasions, summarised in two capture-recapture matrices H^1 and H^2, respectively. No attempt is made to mark new adult individuals and instead only previously marked individuals can be resighted. In addition, unmarked birds are detected on each sampling occasion. The matrix of individual resight histories in this case corresponds to the matrix of presence histories in the standard CR protocol described in Section 4.2.1.

In this case, there are fewer than 40 resighted individuals, so we choose to perform inference by modelling individual entry and exit intervals for all the marked individuals resighted. However, as opposed to the model described in the previous section, we have two different populations of individuals, the ones already marked and the unmarked, and hence we have to modify the model that has been introduced in the previous section. Similarly to the case with one population, we model the marked individuals resighted using the variables (t_1^i, t_2^i), while the marked individuals never resighted and the unmarked individuals are summarised in two matrices $\{n_{fl}^M\}$ and $\{n_{fl}^U\}$, respectively, where cell (f,l) corresponds to the number of individuals first available for capture at time t_f and last available at time t_l.

The number of marked birds N_M and the number of unmarked birds N_U are assigned two Poisson prior distributions $N_M \sim \text{Poisson}(\omega_M)$ and $N_U \sim \text{Poisson}(\omega_U)$, respectively. We assume that each marked individual can be resighted by visiting its nest with probability p_C and through camera traps with probability p_R. As the unmarked birds are also detected through camera traps, we assume they can be detected with the same probability p_R of resighting an already marked bird. We note that this is a realistic, but also necessary assumption in this case, as otherwise the CD do not include enough information to separately estimate the population size N_U of unmarked birds and the probability of resighting an unmarked bird, p_R, if that is different to marked birds.

The prior distributions and the MCMC scheme are discussed in Section 3.2 of the supplementary material.

The posterior distribution of the entry and exit cumulative distribution functions (cdf) are shown in Fig. 4.8a. The cdf of the departure does not reach 1 within the study period since some birds remain at the site using the colony as a roost. Similarly, the cdf of the arrival intervals is at around 0.4 at the start of the study, suggesting that a large proportion of individuals are present at the site when sampling starts, with over 80% of individuals estimated to have arrived by the second sampling occasion towards the end of April. The posterior distributions of the two population sizes and resighting probabilities are shown in Table 4.8b. The camera trap resight probability, p_C, is slightly lower than the resight probability through nest visits, p_R, since cameras are pointed toward the nest, where it is unlikely to see floaters and prospectors. The difference between the population sizes of marked and unmarked birds is due to the fact that only a small proportion of chicks are marked each year, and out of those



(a)

	2.5%	Mean	97.5%
N_M	39	48	59
N_U	45	65	91
p_R	.30	.38	.48
p_C	.21	.29	.38

(b)

Fig. 4.8 Posterior mean of the cumulative distribution function of entry intervals (dashed line) and exit intervals (solid line), with confidence bars show the 95% PCIs. The ticks on the x-axis represent the sampling occasions (a). Posterior summaries for the spoonbills data-set (b).

marked, only a small proportion return to breed as adults. In fact, local recruitment rate is thought to be around 0.12, while the proportion of immigrants on total number of recruits is known to range from 0.49 to 0.83 [85].

4.5.3 Ring-Recovery

In this section, we show how to build a model for RR data using a RPT prior. Similarly to what was mentioned in Section 4.5.1, in order to jointly model entry and exit patterns we work with a set of matrices $\mathbf{n}^k = \{n^k_{u_f u_l}\}$, where in this case $n^k_{u_f u_l}$ represent the number of individuals marked on sampling occasion k that remained in the population for u_f sampling occasions before being marked and u_l sampling occasions after being marked. We assume that individuals marked at time k could have entered the population for up to U sampling occasions before they were marked and could have stayed for up to U sampling occasions after being marked, and thus we assume the dimension of each matrix $\mathbf{n}^k$ is $(U+1)\times(U+1)$.

For each matrix $\mathbf{n}^k$, we assume a PT prior for the probabilities $\omega^k_{u_f u_l}$ for an individual to belong to each cell $n^k_{u_f u_l}$, with partition taken to be the BivLP partition built in Section 4.5.1. A visual representation of the partition for a matrix is presented in the supplementary material.

Similarly to Section 4.3.1, we make the assumption that the survival probability is age-dependent, that is, the probability of an individual surviving until the next sampling occasion depends only on the age of the individual and not on calendar time. If we define Ω_k to be the sample space for the individuals marked at time t_k, this assumption can be achieved by assuming an RPT across $\Omega_1,\ldots,\Omega_K$. This forces the probabilities $\omega^k_{u_f u_l}$ to be the same for each k, that is $\omega^k_{u_f u_l} \equiv \omega_{u_f u_l}$. Similarly to the CJS

model, if we define $\phi_{j,a}$ as the probability of an individual of age a that is present at time t_j to remain until the following sampling occasion, assuming an RPT prior over the different slices is equivalent to assuming that $\phi_{j,a} \equiv \phi_a$.

For the tree corresponding to the first U survival splits the masses are assigned according to Beta distributions, with parameters that are determined by centering the tree on a truncated Laplace distribution. For the remaining U splits, each set B_j, corresponding to the individuals surviving for j sampling occasions, is split into the sets of individuals arriving in the intervals $(t_{k-j}, t_{k-j+1}), \ldots, (t_{k-1}, t_k)$ and the masses are sampled from a Dirichlet$(\alpha_0, \ldots, \alpha_0)$ distribution.

The number of individuals marked at time t_k which can be recovered that at time t_{k+j} can be obtained by the matrix $n^k_{u_f u_l}$ as $\sum_{u_f=0}^{U} n^k_{u_f j}$. Hence, the number of individuals marked in year t_k recovered dead in year t_{k+j}, $R_{k,k+j}$, is distributed as a Binomial$(\sum_{u_f=0}^{U} n^k_{u_f j}, \lambda)$. Assuming also a Beta prior for the recovery probability λ, the model can be summarised as

$$\begin{cases} R_{k,k+j} \sim \text{Binomial}(\sum_{u_f=0}^{U} n^k_{u_f j}, \lambda) \\ n^k_{u_f u_l} \sim \text{Multinomial}(m_k, \omega_{u_f u_l}) \\ \{\omega_{u_f u_l}\} \sim \text{PT}(\Pi, \{\alpha\}) \\ \lambda \sim \text{Beta}(a_0, b_0) \end{cases} \tag{4.5.1}$$

.

Case study

We apply our model to a dataset collected in Minnesota, USA. A total of 100127 female mallards have been banded throughout the course of the study, which lasted 51 years. Marking occurs from July to September each year, while recoveries occur during the hunting season immediately following marking, from September to January. Newly caught individuals can be recognized as juveniles if their age is less than 1 year at the time of capture and as adults otherwise. As we can recognize the juveniles at the time of capture, the age at time of death of the recovered juveniles is known. The entry and exit of the individuals corresponds in this case to births and deaths, while the sampling occasions corresponds to the markings and recoveries.

We summarise the data in two $K \times K$ matrices, R^A and R^J, for adults and juveniles, respectively. The total number of juveniles and adults marked in year k, are denoted by m^J_k and m^A_k, respectively.

The first row of the matrix $\mathbf{n}^k$, $\{n^k_{0u_l}\}_{u_l=0,\ldots,U}$, consists of the juveniles, and its sum is thus equal to m^J_k, while the rows from 2 to $U+1$ consist of the individuals marked as adults, and their sum is equal to m^A_k. Assuming a Beta prior for the recovery probability, λ, the model can be summarised as

$$\begin{cases} R^y_{k,k+l} \sim \text{Binomial}(n^k_{0u_l}, \lambda) & R^a_{k,k+l} \sim \text{Binomial}(\sum_{u_f=1}^{U} n^k_{u_f u_l}, \lambda) \\ n^k_{0u_l} \sim \text{Multinomial}(m^J_k, \omega_{0u_l}) & \{n^k_{u_f u_l}\}_{f>0} \sim \text{Multinomial}(m^A_k, \omega_{u_f u_l}) \\ \{\omega_{u_f u_l}\} \sim \text{PT}(\Pi, \{\alpha\}) & \lambda \sim \text{Beta}(a_0, b_0) \end{cases} \tag{4.5.2}$$

As opposed to Eq. 4.5.1, the first row of the matrix $n^k_{u_f u_l}$ contributes to the recoveries R^y for the juveniles and the rest of the matrix contributes to the recoveries R^a of the adults.

We perform inference by updating the random matrices $n^k_{u_f u_l}$, the probabilities $\omega_{u_f u_l}$ and the recovery probability λ. The prior specifications and details of the MCMC sampler are presented in the supplementary material.

The posterior mean of the survival probabilities is presented in Fig 4.9. The result agrees with the general pattern observed for bird populations, with survival being lower in very young and older ages. A similar pattern was observed by [57] when analysing RR data for mallards (*Anas platyrhynchos*) and by [42] for blackbirds (*Turdus merula*). As expected, uncertainty increases for older age because of the sparseness of the data. The 95% PCI of the recovery probability is $(0.1209, 0.1250)$.

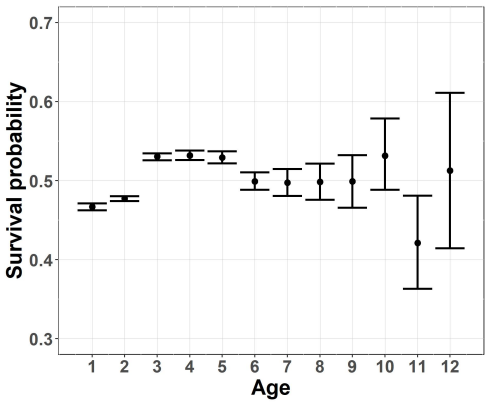


Fig. 4.9 Posterior means of the probabilities ϕ_j of surviving to the age $j+1$ conditional on being alive at age j.

4.6 The hierarchical Logistic PT

The RPT defined in Section 4.4 allows us to impose constraints across different distributions or conditional distributions by assuming the same r.v.s across different trees or across different branches of the same tree, which in the case of density estimation forces the different distributions to be the same. This assumption is however too restrictive in many real applications, as for example when jointly modelling data collected across different years or sites. In these cases, it may be more reasonable to employ a hierarchical approach, which enables us to borrow strength across the different data-sets but without assuming that the distributions describing the different data-sets are exactly the same.

In the context of data collected at different units, Christensen and Ma [13] propose a Hierarchical PT by assuming a PT prior in each unit, where each PT is centered on a common baseline PT prior. However, because of the lack of conjugacy of a Beta prior with a Beta likelihood, the sampling scheme relies on quadrature techniques. To employ a hierarchical approach while still performing inference with a Gibbs sampler, we suggest an alternative based on the logistic PT (LPT) of Jara and Hanson

[40]. The LPT is defined by replacing the Dirichlet distributions for the $V_{\varepsilon_1\ldots\varepsilon_i}$ in (4.2.1) with logistic multinomial distributions.

We define a hierarchical model for distributions in the following way. If we have distributions $G_1,\ldots,G_S$, we assume an LPT prior for each distribution, where the r.v.s of each distribution G_j assigning the mass in each split, $\beta^j_{\varepsilon_1\ldots\varepsilon_m}$, are centered on a common prior $N(\mu^0_{\varepsilon_1\ldots\varepsilon_m},\sigma^2_{\varepsilon_1\ldots\varepsilon_m})$, where μ^0 are the parameters of a PT centered on a distribution $G_0(\cdot|\eta)$ and an additional hyperprior P_0 is assumed on η. The model can be summarised hierarchically as

$$\begin{cases} \log\left(\frac{Y_{\varepsilon_1\ldots\varepsilon_m,i_1}}{Y_{\varepsilon_1\ldots\varepsilon_m,i_2}}\right) = \begin{cases} \beta_{\varepsilon_1\ldots\varepsilon_m,i_1} - \beta_{\varepsilon_1\ldots\varepsilon_m,i_2} & i_1,i_2 = 0,\ldots,K-2 \\ \beta_{\varepsilon_1\ldots\varepsilon_m,i_1} & i_2 = K-1 \end{cases} \\ \beta^j_{\varepsilon_1\ldots\varepsilon_m} \sim \mathrm{N}(\mu^0_{\varepsilon_1\ldots\varepsilon_m},\sigma^2_{\varepsilon_1\ldots\varepsilon_m}) \qquad j=1,\ldots,S \\ \mu^0_{\varepsilon_1\ldots\varepsilon_m} = \mathrm{logit}\left(\frac{G_0(B_{\varepsilon_1\ldots\varepsilon_m}|\eta)}{G_0(B_{\varepsilon_1\ldots\varepsilon_{m-1}}|\eta)}\right) \qquad \eta \sim P_0 \end{cases}$$

The first equation expresses how the masses are assigned according to the LPT structure. A set of PTs following this structure will be termed as Hierarchical Logistic PT (HLPT). If $\sigma^2_\varepsilon \to 0\ \forall\varepsilon$, we have $\beta^j_\varepsilon \equiv \mu^0_\varepsilon$ and hence the distributions for each group are the same.

As we mentioned above, the LPT is defined in terms of logistic transform of normal distributions and hence a Gibbs sampler, based on the Polya-Gamma scheme of Polson et al. [69], is available to sample from the posterior distribution, details of which can be found in Section 3.4 of the supplementary material.

4.6.1 A hierarchical model across different data-sets

In this section we extend D18 by defining a model for data collected at different sites using the HLPT defined in the previous section. The data consist of S different datasets which, for ease of exposition, we take to be CD, but the same rationale applies for any other sampling schemes described before. We define C_{js} to be the number of individuals detected at time t_j in the s-th data-set.

We arrange all the individuals in S matrices $\mathbf{n}^s = \{n^s_{fl}\}_{f=0,\ldots,K,l=0,\ldots,K}$, where n^s_{fl} is the number of individuals in data-set s having first and last possible detection at times t_f and t_l, respectively. For the s-th data-set, individuals entry and exit times are drawn from a Poisson process with intensity $\nu_s = \omega_s \times G_s$, where ω_s is the overall intensity and G_s is the normalized density. In order to borrow information between different data-sets, the normalized densities $(G_1,\ldots,G_S)$ are assumed to follow a HLPT prior, where the partition is taken to be the EELP of Section 4.5.1.

Assuming also that each individual in dataset s can be detected with probability p_s and assuming a Gamma prior distribution for the overall intensities ω_s, the model can be summarised in the following hierarchical structure

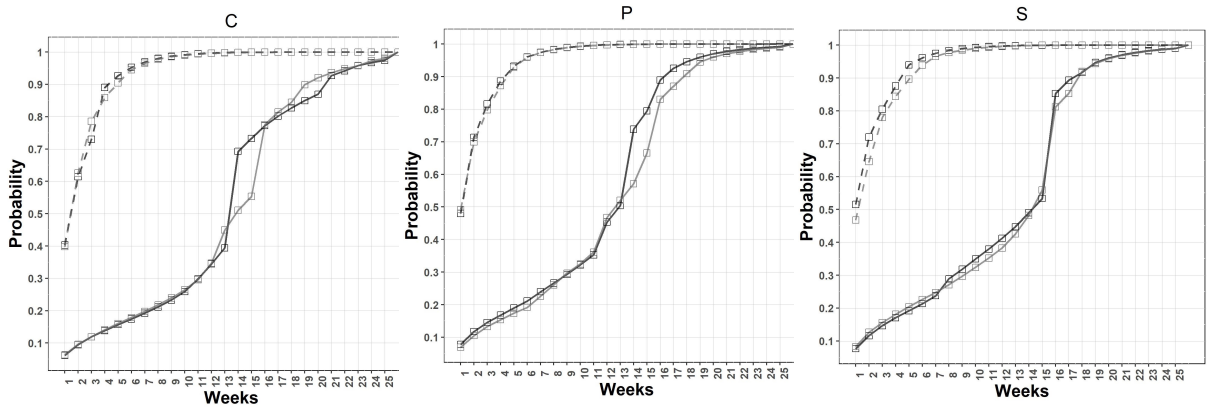


Fig. 4.10 Posterior c.d.f.s of the entry (dotted line) and exit (solid line) distribution for great crested newts (C), palmate (P) and smooth (S). The male are represented in black and the female in grey.

$$\begin{cases} C_{js} \sim \text{Binomial}(N_{js}, p_s) & N_{js} = \sum_{l=j}^{K} \sum_{f=0}^{j-1} n_{fl}^s \\ n_{fl}^s \sim \text{Poisson}(\omega_s \times \omega_{fl}^s) & \{\omega_{fl}^s\} \sim \text{HLPT}(\eta, \sigma) \\ \eta \sim \text{P}_0 & \omega_s \sim \text{Gamma}(a_\omega, b_\omega) \end{cases}$$

where N_{ks} is the number of individuals available for sampling on occasion k and data-set s.

Case study

We consider a data-set of weekly detections of 3 species of newts separated for gender, for a total of 6 different sets of counts, performed at a local site at the University of Kent in 2014. Newts are usually sampled weekly during their breeding period using aquatic traps. However, some visits might be missed on occasion because of weather conditions. All three species migrate to ponds in late winter-early spring for breeding, and leave the breeding site in late spring or summer to spend the rest of the year on land. The data are shown in Section 1 of the supplementary material.

The posterior estimates of the entry and exit distributions for the 6 data-sets are shown in Fig. 4.10. From the plots, a slight pattern showing males arriving earlier and females departing later can be detected, which could be due to female staying longer to complete egg-laying. Moreover, it seems that among the three species, great crested newts tend to arrive and depart later than smooth or palmate newts.

4.6.2 Modelling long time-series of ecological data

The last scenario we consider is a long time series of counts collected daily. Difficulties arise in this case as individuals are not present during the whole sampling period but only for a small part of the sampling period. Therefore, building a partition by considering all K intervals would increase the computational complexity without necessarily providing a better description of the data. A naive

modelling approach is to merge sampling occasions in order to obtain a shorter time series. However, the decision of how to merge is arbitrary and the same merging may not be meaningful for the duration of the study period. Instead, we let the model infer which sampling occasions to merge using the Optional Polya Tree of Wong et al. [89].

As in the model of Section 4.6.1, we build a prior distribution on the number of individuals present in each sampling occasions by defining a prior distribution on the number of individuals arriving and departing in each sampling occasion, which is done by assuming a PT prior on the individual entry and exit intervals. The partition of the PT is built using the UP of Section 4.5.1, where the splits are performed according to some nested periods of time, which for example can be taken as the natural scales at which data are collected, such as months, weeks and days. Clearly, the periods of time of the partition are defined up to the resolution at which the data are collected, in order to define a prior on the number of individuals arriving and departing in each sampling occasion. As the OPT allows us to stop further partitioning, the choice of the UP will lead the partition to end in one of these periods of time.

We divide the overall study period in K_0 periods of time of equal length, $\Omega_1,\dots,\Omega_{K_0}$, called 0-periods, and we assume that individuals enter and exit only in the same 0-period. Next, for each 0-period, Ω_{i_0}, we define a distribution G_{i_0} on the entry and exit intervals of the individuals. The 0-periods can be thought of as the periods of time in which the migration patterns seasonally repeat, such as the years. Next, as mentioned above, we define a set of splitting rules according to which the sample space of G_{i_0}, Ω_{i_0} is first split into the sets $\mathcal{I}^1=\{I_1^1,\dots,I_{K_1}^1\}$, and next, each set at level i, I_j^i, is split into $\mathcal{I}_j^{i+1}=(I_1^{i+1},\dots,I_{K_{i+1}}^{i+1})$. The sets in $\mathcal{I}^i$ are called i-periods. Given $\varepsilon=(\varepsilon_1,\dots,\varepsilon_m)$, we define as T_ε the set obtained by taking first the ε_1-th 1-period, then the ε_2-th 2-period and so on. For example, the years can be taken as the 0-periods and the splitting rules can be built using first the months as 1-periods, weeks as 2-periods and so on.

Next, given two sequence of indexes ε and η, we define as $n^{i_0}(\varepsilon,\eta)$ the number of individuals in the 0-period Ω_{i_0} entering in T_ε and exiting in T_η. We also define as $N_\varepsilon^{i_0}$ and $C_\varepsilon^{i_0}$ the number of individuals available for sampling and detected, respectively, in T_ε and in the i_0-th 0-period. Even if an explicit formula is cumbersome to present, we note that it is straightforward to obtain the latent individuals $N_\varepsilon^{i_0}$ from the $n^{i_0}(\varepsilon,\eta)$. Thanks to this construction, we have defined a prior on the probabilities $\omega^{i_0}(\varepsilon,\eta)$ of an individual to belong to the cell $n^{i_0}(\varepsilon,\eta)$.

Having defined this set of splitting rules, we define the partition of sample space Ω_{i_0} by first splitting Ω_{i_0} in $(B_1^{i_0},\dots,B_{K_1}^{i_0})$ where $B_i^{i_0}=\{(x,y)\in\Omega_{i_0},x\in I_i^1\}$, that is, the sample space is split using the dimension of entry according to the 1-periods. Next, each $B_i^{i_0}$ is split into $(B_{i,1}^{i_0},\dots,B_{i,K_1}^{i_0})$, where $B_{i,j}^{i_0}=\{(x,y)\in B_i^{i_0},y\in I_j^1\}$, that is, each set is now split using the dimension of exit according to again the 1-periods. Analogously, in the next steps, given two sequences of indexes $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)$ and $\eta=(\eta_1,\dots,\eta_m)$, each set $B_{\varepsilon,\eta}^{i_0}$ is split into

$$\begin{cases}(B_{\varepsilon 1,\eta}^{i_0},\dots,B_{\varepsilon K_n,\eta}^{i_0}) \text{ where } B_{\varepsilon i,\eta}^{i_0}=\{(x,y)\in B_{\varepsilon,\eta}^{i_0},x\in I_i^n\} & \text{if } n=m\\ (B_{\varepsilon,\eta 1}^{i_0},\dots,B_{\varepsilon,\eta K_m}^{i_0}) \text{ where } B_{\varepsilon,\eta i}^{i_0}=\{(x,y)\in B_{\varepsilon,\eta}^{i_0},y\in I_i^m\} & \text{if } n=m+1\end{cases}$$

A representation of this scheme is shown in the supplementary material. We also denote by $\beta^{i_0}_{\varepsilon,\eta}$ the r.v.s assigning the masses in each level of the partition, that is, $G_{i_0}(\beta^{i_0}_{\varepsilon i,\eta}) = \frac{e^{\beta^{i_0}_{\varepsilon i,\eta}}}{1+\sum_{k=0}^{K-2} e^{\beta^{i_0}_{\varepsilon k,\eta}}}$ and $G_{i_0}(\beta^{i_0}_{\varepsilon,\eta j}) = \frac{e^{\beta^{i_0}_{\varepsilon,\eta j}}}{1+\sum_{k=0}^{K-2} e^{\beta^{i_0}_{\varepsilon,\eta k}}}$. We also make the following assumptions on the r.v.s G_{i_0}:

- We assume an HLPT structure across the distributions G_{i_0}. This implies that the r.v.s $\beta^{i_0}_{\varepsilon,\eta}$ have a common prior $\beta^{i_0}_{\varepsilon,\eta} \sim \mathrm{N}(\mu_{\varepsilon,\eta}, \sigma^2)$.
- We assume that $\beta^{i_0}_{\varepsilon i,\eta j}$ is the same across the different entry periods, *i.e.* $\beta^{i_0}_{\varepsilon i,\eta j} = \gamma^{j_0}_{\varepsilon,\eta j} \forall i$.
- We assume that the masses in periods that are in close proximity are more strongly correlated than the mass of periods further away from each other, hence we assume that the means $(\mu_{\varepsilon 1,\eta}, \ldots, \mu_{\varepsilon K,\eta})$ of the common prior are drawn jointly from a Gaussian process $\mathrm{GP}((\mu^0_{\varepsilon 1,\eta}, \ldots, \mu^0_{\varepsilon K,\eta}), \sigma_0^2 k_l)$, where k_l is the Gaussian covariance function $k(x_i, x_j) = \exp(-\frac{(x_i - x_j)^2}{l^2})$ with parameter l.

As a result, σ^2 tunes the variability across 0-periods, σ_0^2 tunes the variability across i-periods and l the correlation across i-periods.

Assuming a Poisson(ω) prior distribution on the number of individuals N_{i_0} in each 0-period, the model can be summarised in the following hierarchical structure

$$\begin{cases} C^{i_0}_{\varepsilon} \sim \mathrm{Binomial}(N^{i_0}_{\varepsilon}, p) \\ n^{i_0}(\varepsilon, \eta) \sim \mathrm{Poisson}(\omega \cdot \omega^{i_0}(\varepsilon, \eta)) \\ \omega^{i_0}(\varepsilon, \eta) \sim \mathrm{HLPT}(\mu^0, \sigma, \sigma_0) \end{cases}$$

where p is the probability of detecting an individual.

The downside of this model is that the partition is built for 1-period, 2-period, etc., even if no observations are available in some periods. Ideally, we would like to use a modelling approach where, instead of building the partition for each period, we keep partitioning the PT only in the periods of time where the entry and exit density has significant variation. To do so, we change slightly the definition of the PT and allow for different branches of the tree to not partition further, by borrowing ideas from the OPT of Wong et al. [89].

Specifically, we assume that in each step, partitioning of the PT into the further level happens with probability ρ which, as in the OPT, is a-priori the same for each level. We then introduce a variable S_{B_ε} which is equal to 1 if B_ε is not partitioned further and 0 otherwise. The distribution in the set B_ε can thus be expressed recursively from its descendant sets as

$$G(x|x \in B_\varepsilon) = S_{B_\varepsilon} U(x|x \in B_\varepsilon) + (1 - S_{B_\varepsilon}) \sum_{i=1}^{K_\varepsilon} G(x|x \in B_{\varepsilon,\varepsilon_i}) 1(x \in B_{\varepsilon,\varepsilon_i})$$

where U is the uniform distribution and $(B_{\varepsilon,\varepsilon_1},\ldots,B_{\varepsilon,\varepsilon_{K_\varepsilon}})$ are the K_ε descendant set of B_ε. Hence, when no splitting is required we replace the PT distribution with a uniform distribution.

To perform inference with this new structure, we first sample the variables S_{B_ε} for all the sets in the next level, as well as the parameters β, next we sample S_{B_ε} and β for all the sets in the upper level and so on.

Case study

We consider a long series of daily CD of two species of moths, Canadian Melanolophia (*Melanolophia canadaria*) and Signate Melanolophia (*Melanolophia signataria*), collected across 5 years [67]. The data are shown in Section 1 of the supplementary materials. Detection of the individuals is possible only during the flight period, which is known to happen only in a limited amount of time inside the calendar year. Therefore, we consider the years to be the 0-periods. The year is then split into 13 periods of time of equal duration (1-periods), then each 1-period is split into 4 periods of time of equal duration (2-periods) and finally each of these periods in 7 (8 for a leap year) equal periods of time (3-periods), where each of these 3-periods corresponds to a day. In order to simplify the computation, we summarise the data across the 2-periods by using the sum of the individuals captured in a 2-period as the count.

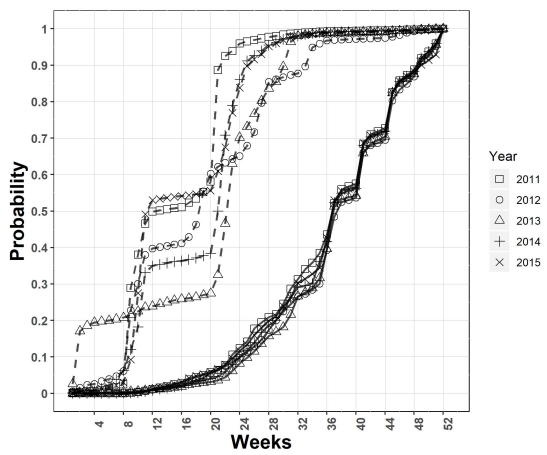


Fig. 4.11 Posterior mean of the entry (dashed line) and exit (straight line) cumulative distribution functions for the moths data. Each year is represented by a different symbol.

The posterior means of the entry and exit cumulative distribution functions are shown in Fig. 4.11. The exit pattern is practically identical between the different years, with the c.d.f suggesting a smooth exit pattern with a few bigger exit waves around weeks 36, 40 and 44 (start of September). On the other hand, entry patterns present a bit more variation across years. The species is known to have two broods [15, 87], which is evident by the two big entry waves each year around weeks 20 and 24 (mid May) for all years apart from 2013, when the first brood emerged much earlier. This is likely due to the unusual warm temperature at the start of 2013, as the table in the supplementary material suggests.

4.7 Conclusion

We introduced a framework for defining models for ecological data based on the PT prior. The advantage of this framework is that a wide variety of assumptions can be made on the model structure by changing the tree structure, as a result of the flexibility of the PT. We have applied our framework to different types of commonly collected ecological data, such as CR, CD and RR, either in isolation or in combination by building joint models. We have also introduced the RPT, which allows us to place constraints on the model parameters, and the HLPT, which allows us to impose a hierarchical structure and share information when modelling different data-sets. We have also presented a way to infer the partition of a PT using the OPT of Wong et al. [89], which allows us to build trees with depth informed by the data.

It is easy to extend the model to other protocols, as for example removal data [65]. Removal data are collected by repeatedly visiting the site and removing all caught individuals. In this case, an approach similar to CR and RR can be employed, by assuming that the time of exit is known for the removed individuals and estimating entry and exit of the individuals not removed.

We have used the HLPT to perform sharing because of its computational advantage over the HPT of [13], which is due to the use of the logistic normal distribution in the place of the Beta distribution. However, there are additional advantages in using a logistic normal distribution in place of a Beta distribution. In fact, as the model relies on the normal structure, it is easy to extend the model while keeping the same posterior inference strategy. For example, it is easy to introduce covariates across the different units, extending the model from a standard exchangeable scenario to one of conditional exchangeability, or extending the model to a scenario where the effect of the different units are modelled using a time series. The model can also be easily extended to more levels, which can be applied, for example, to scenarios when the data are collected on multiple species and at multiple sites and a joint modelling is required across both levels.

In the OPT, although it is possible to sample from the full conditional thanks to a conjugacy property, sampling from the posterior involves recursive calculations across the whole tree. This is computationally inefficient if the tree is large and there are many layers and this has to be repeated for each step of the MCMC. In that case, a more efficient scheme is to prune the tree and in each step of the MCMC only propose to update/add/remove branches of the tree.

In our modelling framework, different ecological assumptions correspond to different trees or changes in the dependence structure of the r.v.s on the tree. The question of evaluating evidence in favour of different assumptions can be addressed using standard Bayesian model selection methods. First, model selection can be performed by simply evaluating the evidence of each model by using Bayes factors. Alternatively, as a change in the ecological assumption entails a change in the structure of the trees, model selection can be performed by employing an additional prior on these different tree structures and computing the posterior of the structure of trees as part of the parameter space.

Chapter 5

A vector of softcore processes for modelling interactions between species using capture-recapture data.

Abstract

Capture-recapture (CR) data and corresponding models have been used extensively to perform inference on several quantities of interest of a wildlife population. When more than one traps is used, CR models can be extended to spatial CR models to also provide information on the spatial patterns of individual activity centres. Although the spatial patterns of the individual activity centres have usually been assumed to be independent, recent work relaxes this assumption by considering interactions between individuals of a single species. In this paper, we develop a model for CR data that can take into account interactions between and within multiple species. This is achieved by considering a vector of point processes from the general class of interaction point processes. These processes present a challenge from an inferential perspective because of the untractability of the normalising constant of the likelihood function, and hence standard MCMC procedures cannot be applied. Instead, we adopt an inference procedure based on the Monte-Carlo Metropolis Hastings algorithm of [50], which scales well also when the number of species is large. We apply the model two CR data-sets on leopards and tigers.

5.1 Introduction

Conservation scientists, researchers and charities, as well as government organisations worldwide, are interested in monitoring sizes of populations, distributions of species in the landscape and spatial patterns of individuals in the wild. Capture-recapture (CR) is one of the most commonly employed surveying tools to achieve this goal. According to the CR protocol, an attempt is made to capture individuals on repeated sampling occasions at one or more traps, and all newly caught individuals are uniquely marked. Several models have been developed for estimating population sizes [77], arrival and departure patterns [68] or survival patterns [57] using CR data. When more than one trap is used, CR data can also be used to infer the spatial pattern of the locations of centres of the activity area (CAA), defined as the locations around which the individuals move when performing their daily activity [79], using spatial capture-recapture (SCR) models [24, 5, 78]. Even in cases when the spatial pattern is not of main interest, SCR models are still able to provide less biased estimates as they relax some assumptions compared to standard CR models. Specifically, standard CR models assume that all individuals are equally likely to be caught at each trap, which is often violated in practice [25]. On the other hand, SCR models account for the fact that individuals are more likely to be caught in traps

that overlap with their activity area. For a nice review of the literature for SCR models, see Royle et al. [79].

When using SCR models, the locations of the individual activity centres are latent, which gives rise to challenges from an inferential perspective. SCR models can be developed in a classical [5] or Bayesian [79] framework. The classical approach relies on maximizing the marginal likelihood computed by integrating out the activity centres, while the Bayesian approach relies on sampling from the posterior distribution of the activity centres as well as of the other parameters of interest. Estimation through marginal likelihood has been usually performed in the case where the activity centres are assumed to come from a tractable distribution, which makes the computation of the marginal likelihood analytically possible. As the processes considered in this paper are quite complex, this computation cannot be performed analytically, and hence we consider the Bayesian approach.

A common assumption in SCR models is that the activity centres of the individuals are independent of each other. An extension is the work by [74], who used a Strauss process [83] to model the activity centres of a single species while accounting for interaction within the species. By interaction, we mean that the probability density of a given point (i.e. the activity centre of one member of the species) depends on the location of the other points (the activity centres of the other members). This dependence can manifest either as *repulsion*, where the presence of a point at a particular location decreases the probability of having another point at a nearby location, and as *attraction*, where the opposite is true and points tend to be clustered. In our case, we are more interested in modelling repulsion, as repulsion is more likely to be present in the species considered in our application. In this paper, we extend the model of [74] by introducing a vector of spatial point processes (PPs) for modelling the locations of the activity centres and build a multi-species SCR model that can account for interaction within and between species.

Spatial PPs are a major area of research in spatial statistics and a useful modelling tool when the objects of interest are locations of objects in space [75, 16]. The simplest example of a PP is the Poisson process, where the locations of the individual points are drawn independently of each other and they do not exhibit interaction. One extension of the Poisson process is the class of Cox processes, which are processes for clustered or aggregated point patterns [63]. Another extension of point processes is the class of interaction PPs (IPP) [80, 70], which are especially suited to modelling interactions between points. We choose to work with interaction PPs, because they are better suited to modelling repulsion between points.

The computational challenge that arises when working with interaction PPs is that the normalising constant of the likelihood function, which is a function of the model parameters, is usually unavailable in closed form apart from trivial cases [59]. Distributions of this kind are known in the literature as *doubly-intractable* distributions, as both the marginal likelihood of the data, with parameters integrated out, and the likelihood function are intractable, and a variety of methods have been proposed to perform inference, as we describe next. In a frequentist setting, this is usually dealt with by replacing the complete likelihood with a pseudo-likelihood [4], for which asymptotic results have been provided [41]. In a Bayesian setting, the problem usually involves computing a ratio of normalising constants,

as the ratio arises naturally when using a Metropolis-Hastings (MH) algorithm for sampling from the posterior distribution of the parameters of the PP. Popular inference procedures include the auxiliary variable method of [61], path sampling [64], exchange sampling [62] and the Monte Carlo MH (MCMH) algorithm [50]. In this paper, we choose to use the MCMH algorithm, which essentially approximates the ratio of normalising constants with an importance sampling estimate, because this approach is easily scalable as the number of parameters grows.

As mentioned above, as we work with several species, we do not use a single PP but a vector of PPs (VPP). Although an extensive literature exists for univariate interaction PPs, VPPs have only recently gained attention in the literature [86, 72]. For example, Waagepetersen et al. [86] developed a model that takes into account correlation between more than one species using a Log-Gaussian Cox process, while Rajala et al. [72] proposed a model to take into account interactions between many species using a Gibbs model, which is a particular case of IPP.

We apply our new model to two CR data-sets on tigers and leopards. However, for confidentiality reasons, we cannot reveal the location where the data were collected. The main interest is in estimating the population density of the individuals while accounting for interaction between and within species. In fact, tigers and leopards are territorial animals and this leads to interaction both between and within the species.

The paper is organised as follows. In Section 5.2 we briefly summarise the theory of PPs, as well as the extension to VPPs and the main approaches available for performing inference. In Section 5.3 we define the spatial model for CR data. In Section 5.4 we present results of a simulation study and in Section 5.5 we apply the model to the tigers and leopards CR data-sets. Section 5.6 concludes the paper and introduces some potential future directions.

5.2 Interaction point processes

A spatial PP X on a space $S \subset R^d$ is a random countable subset of S. It is usually assumed that the realisations of X are locally finite and points are not duplicated, that is, for any subset $B \subset S$, the number of points of X in B is finite or, equivalently, that X takes values in the space $N_{lf} = \{x \subset S \mid \#(x \cap B) < \infty \; \forall B \subset S\}$. A nice review of PPs is presented in [59].

The most basic example of a PP is the Poisson process. A Poisson process is a PP with intensity function $\rho : S \to [0, \infty)$ if, for any $B \subset S$ such that $|B| < \infty$, the number of points in B, $N(B)$, is distributed as a Poisson$(\mu(B))$ with $\mu(B) = \int_B \rho(x)dx$. A Poisson process is said to be homogeneous if $\rho(x) \equiv \rho$, and inhomogeneous otherwise.

According to the Poisson process, points are independent to each other, in the sense that the location of a point does not affect the location of another point. In order to introduce interaction between points, a common strategy is to define a process with density f with respect to (w.r.t.) another PP, for example the Poisson process, and let f determine the interaction between points. To this end, we say that a PP X_1 has density f w.r.t. another PP X_2 if $\mathbb{P}(X_1 \in F) = \mathbb{E}[1(X_2 \in F) f(X_2)]$ for each

$F \in N_{fl}$. It can be shown that if X has density f w.r.t. a homogeneous Poisson process with density 1 it holds that

$$\mathbb{P}(X \in F) = \sum_{n=0}^{\infty} \frac{\exp(-|S|)}{n!} \int_{S^n} 1[\{x_1, \ldots, x_n\} \in F] f(x_1, \ldots, x_n) dx_1, \ldots, dx_n.$$

The normalising constant of f,

$$\mathbb{P}(X \in N_{fl}) = \sum_{n=0}^{\infty} \frac{\exp(-|S|)}{n!} \int_{S^n} f(x_1, \ldots, x_n) dx_1, \ldots, dx_n,$$

is not available in analytical form (except in the case of a Poisson process) and hence f is usually expressed up to proportionality $f \propto h$. Next, we consider the case of pairwise interaction PP, where f is defined w.r.t. a unit rate Poisson process and is of the form

$$f(\{x_1, \ldots, x_n\}) \propto \prod_{i=1}^{n} \phi_1(x_i) \prod_{i<j} \phi_2(x_i, x_j | \theta) \tag{5.2.1}$$

where ϕ_1 is called intensity function and ϕ_2 is called interaction function. Whenever $\phi_2 < (>)\ 1$, there is repulsion (attraction) between points. In the following, we set $\phi_1 \equiv \beta$.

Pairwise interaction PPs can thus be defined by simply specifying the function ϕ_2 that determines the interaction between two points. For example, the Strauss process [83] is obtained when $\phi_2(x_1, x_2 | \lambda, R) \propto \lambda^{1[|x_1 - x_2| < R]}$. The interpretation of this process is that points have an interaction only if they are at a distance less than R. The Strauss process is well-defined only for $\lambda \leq 1$ (for $\lambda = 1$ it corresponds to the Poisson process), as for $\lambda > 1$ the number of points explodes. Another example of interaction PP is the soft-core process [64], defined by the interaction function

$$\phi_2(x_1, x_2 | \gamma) \propto 1 - \exp\left(-\frac{(x_1 - x_2)^2}{\gamma}\right) \tag{5.2.2}$$

which, as opposed to the Strauss process, has an infinite range of interaction, in the sense that points interact with each other regardless of their distance. Examples of interaction functions for the two processes are shown in Fig. 5.1.

5.2.1 Vectors of point processes

Similarly to a single PP, in the case of a vector of D PPs, a VPP can be defined as a random countable set of S^D. Analogously to a single PP, we can define a VPP $(X_1, \ldots, X_D)$ with density f w.r.t. to another VPP $(Y_1, \ldots, Y_D)$ if $\mathbb{P}((X_1, \ldots, X_D) \in (F_1, \ldots, F_D)) = \mathbb{E}[1((X_1, \ldots, X_D) \in (F_1, \ldots, F_D)) f(X_1, \ldots, X_D)]$ for each $(F_1, \ldots, F_D) \in N_{fl}^D$.

If f can be factorised as $f(X_1, \ldots, X_D) = f_1(X_1) \ldots, f_D(X_D)$, the VPP reduces to a product of D independent single PPs. We consider, as before, pairwise interaction functions w.r.t. a vector of

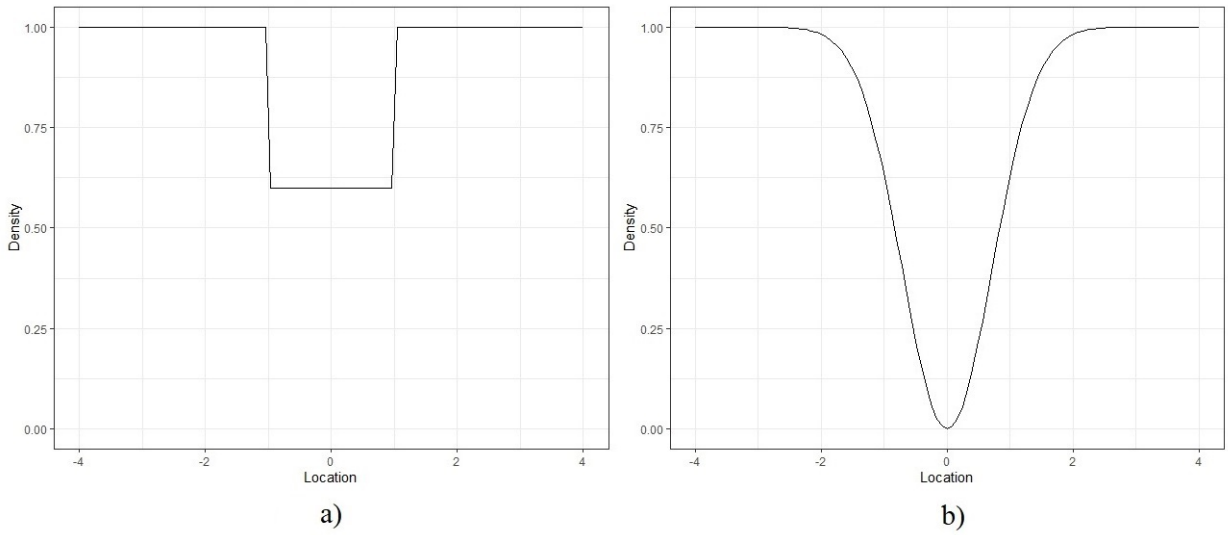


Fig. 5.1 Interaction function $\phi_2(x,0)$ for a Strauss process with parameters $\lambda = .6$, $R = 1$ (a) and for a soft-core process with parameter $\gamma = 1$ (b).

independent unit rate Poisson processes, that is, we choose f as

$$f(\mathbf{x}_1,\ldots,\mathbf{x}_D) \propto \underbrace{\prod_{d=1}^{D}\prod_{i=1}^{N_d}\phi_1(x_i^d)}_{\text{single point intensities}} \underbrace{\prod_{d=1}^{D}\prod_{i=1}^{N_{d_1}}\prod_{j=i+1}^{N_{d_2}}\phi_2(x_i^d,x_j^d|\theta_d^1)}_{\text{within PP interactions}} \underbrace{\prod_{d_1=1}^{D}\prod_{d_2=2}^{d_1-1}\prod_{i=1}^{N_{d_1}}\prod_{j=1}^{N_{d_2}}\phi_2(x_i^{d_1},x_j^{d_2}|\theta_{d_1,d_2}^2)}_{\text{between PP interactions}}.$$

where $(\mathbf{x}_1,\ldots,\mathbf{x}_D) = ((x_1^1,\ldots,x_{N_1}^1),\ldots,(x_1^D,\ldots,x_{N_D}^D))$, θ_d^1 is the parameter of the within-points interactions of $\mathbf{x}_d$ and θ_{d_1,d_2}^2 is the parameter of the between-points interaction of $\mathbf{x}_{d_1}$ and $\mathbf{x}_{d_2}$. For example, ϕ_2 can be taken as the interaction function of the Strauss or soft-core process. Similarly to the Strauss process, a sufficient condition to guarantee that the PP does not explode is $\phi_2(x_1,x_2) < 1\ \forall x_1,x_2$.

In this paper, we choose to work with the soft-core process interaction function (5.2.2) as it is summarised with a single parameter and thus it is easier to estimate than the Strauss process. In fact, for $D = 2$ using a Strauss process would lead to a total of 8 parameters (each pair (λ,R) is repeated for the two between and the within species interaction plus the two intensities β). Therefore, since in the case of CR data the locations of the points correspond to the activity centres and hence are latent, we need to limit the number of parameters to infer in the PP.

If $D = 2$ and the interaction function is taken to be the one of the soft-core process, we term the resulting PP as a bivariate soft-core process and we use the notation $\text{BivSC}(\gamma_1,\gamma_2,\gamma_{12},\beta_1,\beta_2)$ for the distribution of this process. That is, we take f as

$$f(\{x_1^1,\ldots,x_n^1\},\{x_1^2,\ldots,x_m^2\}|\beta_1,\beta_2,\gamma_1,\gamma_2,\gamma_{12}) \propto$$
$$\beta_1^n\beta_2^m\prod_{i<j}\left\{1-\exp\left(-\frac{(x_i^1-x_j^1)^2}{\gamma_1}\right)\right\}\prod_{i<j}\left\{1-\exp\left(-\frac{(x_i^2-x_j^2)^2}{\gamma_2}\right)\right\}\prod_{i,j}\left\{1-\exp\left(-\frac{(x_i^1-x_j^2)^2}{\gamma_{12}}\right)\right\} \tag{5.2.3}$$

We note that γ_1 and γ_2 model the interaction within each of the two species and γ_{12} models the interaction between the species. In Fig. 5.2 we show several sets of simulations from the process. Increasing the interaction parameters leads to a reduction in the number of points that can be located within a given area. Hence, to show how the pattern of points changes when the interaction parameters vary, we show simulations from a *conditional* point process, which is constructed by fixing the number of points and then sampling points with density proportional to (5.2.3). The realizations of the process presented in the two plots on the diagonal of Fig. 5.2 are equivalent to realizations from a single soft-core PP, as the interaction is the same for each pair of points, regardless of colour, with the only difference that in the second row there is less chance of having points that are close to each other as the interaction is stronger. As the parameter θ_{12} increases, the two colours of points become completely separated (bottom-left plot), while if the within PPs interaction is stronger than the between PPs interaction the two colours mix more, and it is less likely to observed clusters of points of the same colour (top-right plot).

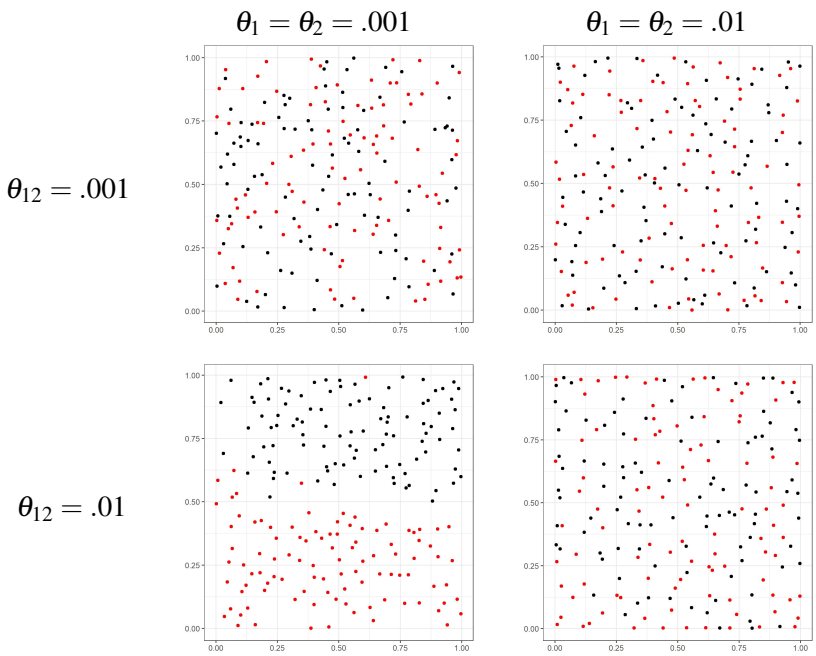


Fig. 5.2 Simulations from a conditional BivSC process with $N_1 = N_2 = 100$.

5.2.2 Bayesian inference for interaction point processes

As mentioned in the introduction, posterior inference for the parameters $\theta = (\{\theta_1^d\}_{d=1,\dots,D}, \{\theta_2^{d_1,d_2}\}_{d_1,d_2=1,\dots,D})$ of a PP X is usually difficult in the case of an interaction PP as the normalising constant is not available in closed form. Given a sample from a PP $y \sim X|\theta$ and a prior π for θ, to make posterior inference we can run a MH sampler to sample from the posterior distribution $p(\theta|y) = \pi(\theta)\frac{h(y|\theta)}{Z_\theta}$, where we have written the likelihood of the PP in an unnormalized form $f(y|\theta) = \frac{h(y|\theta)}{Z_\theta}$. The MH ratio for the parameter θ takes the form

$$H(\theta'|\theta) = \frac{\pi(\theta')h(y|\theta')q(\theta|\theta')}{\pi(\theta)h(y|\theta)q(\theta'|\theta)} \frac{Z_\theta}{Z_{\theta'}} \tag{5.2.4}$$

where q is the proposal distribution. The unavailability of the ratio $\frac{Z_\theta}{Z_{\theta'}}$ makes the use of standard Markov chain Monte Carlo (MCMC) techniques infeasible and thus a different sampling procedure has to be used. For ease of notation, we sometimes express $f(y|\theta)$ and $h(y|\theta)$ as $f_\theta(y)$ and $h_\theta(y)$, respectively.

An ingenious method to avoid calculating the ratio is the auxiliary variable method of Møller et al. [61], which relies on the introduction of an auxiliary variable to simplify the ratio of normalising constants (RNC) and essentially replaces the problem of computing the ratio with the problem of simulating from the process. The downside of this approach is that the MCMC offers good acceptance rates only if the marginal distribution density of the auxiliary variable has a known normalising constant while being sufficiently close to the distribution of interest, which is a difficult task.

Another useful method is the exchange sampling (ES) algorithm of Murray et al. [62], which we recall:

Algorithm 1: Exchange sampling algorithm

1. Sample $\theta' \sim q(\cdot|\theta)$.
2. Sample $x \sim f(\cdot|\theta')$.
3. Accept θ' with probability $\min(1, H(\theta'|\theta))$ where

$$H(\theta'|\theta) = \frac{\pi(\theta')h(y|\theta')q(\theta|\theta')}{\pi(\theta)h(y|\theta)q(\theta'|\theta)} \frac{h(x|\theta)}{h(x|\theta')}$$

The simplicity of the ES algorithm is both its strength and its weakness, as although being easy to implement does not offer good mixing properties. Moreover, it obviously requires to be able to simulate from the process of interest.

Another interesting inference procedure is path sampling [64], which is also used in the CR model of [74]. Path sampling is based on the identity $\frac{\partial \log(Z_\theta)}{\partial \theta_i} = \mathbb{E}_{x \sim X_{\theta_i}}\left[\frac{\partial h_\theta(x)}{\partial \theta_i}\right] =: G_i(\theta)$ where θ_i is the i-th

component of θ. Using this identity, the ratio can be computed as

$$\log\left(\frac{Z_{\theta_i}}{Z'_{\theta_i}}\right) = \int_0^1 G_i(\theta(s))s(t)dt$$

where $s(t)$ is a path connecting θ and θ', that is, s is a differentiable function such that $s(0) = \theta'$ and $s(1) = \theta$. By computing the function $G_i(\cdot)$ on a grid of values of θ in advance, the integral can be approximated numerically and used in (5.2.4). However, this quickly become infeasible if the dimension of the parameter θ is greater than 2, as for example in the case of an VPP, as the grid becomes prohibitively large.

A more scalable approach is based on the MCMH algorithm of Liang and Jin [50] (LJ13), which is based on replacing the RNC with an importance sampling estimate. LJ13 propose three versions of the MCMH algorithm and we adopt here the following algorithm, termed as MCMH-I:

Algorithm 2: Monte Carlo MH algorithm-I

1. Sample $\theta' \sim q(\cdot|\theta)$.
2. Accept θ' with probability $\min(1, H(\theta'|\theta))$ where

$$H(\theta'|\theta) = \frac{\pi(\theta')h(y|\theta')q(\theta|\theta')}{\pi(\theta)h(y|\theta)q(\theta'|\theta)} R_M(\theta,\theta',X)$$

   where $R_M(\theta,\theta',y) = \frac{1}{M}\sum_{m=1}^{M} \frac{h(X_m|\theta)}{h(X_m|\theta')}$ and M is the number of samples used to compute the Monte Carlo approximation.
3. If θ' is accepted, sample $(X_1,\ldots,X_m)$ from $f(x|\theta')$.

The MCMH algorithm can be seen as a generalization of the ES algorithm, as the latter is obtained when $M = 1$ in the MCMH-III algorithm of LJ13. However, the advantage of the MCMH over the ES algorithm is that by drawing the samples $X_1,\ldots,X_M$ in parallel, the MCMH can run at comparable speed to the ES, while providing a substantially better estimate of the RNC. A further advantage of the MCMH is that it does not depend on the dimension of θ, which makes it useful for application to VPPs.

The samples $(X_1,\ldots,X_m)$ in step 3 can be obtained through either MH [33] or through perfect simulation [45]. A review of the existing algorithms for simulating interaction PPs is presented in Møller and Waagepetersen [59]. We choose to use an MH algorithm as, when a new θ' is accepted, the current samples $(X_1,\ldots,X_m)$ from $f(x|\theta)$ can be used as a warm start for the MH algorithm to sample from $f(x|\theta')$, significantly reducing the burn-in period. In fact, if θ and θ' are sufficiently close, we also expect samples from PPs with densities $f_{\theta'}$ and f_θ to be close. On the other hand, perfect simulation is computationally very intensive and hence not feasible to perform at each iteration.

5.3 Spatial capture-recapture model

We now define the spatial model for CR data collected on two or more species. To define the prior for the activity centres of individuals, we are going to use the soft-core VPP model defined in the previous section. For simplicity, we define here a model for two species but the definition for more than two species is straightforward.

We assume that there are K traps, placed in the study area at locations $(x_1,\ldots,x_K)$, and that sampling is performed for T sampling occasions in each trap. For two species, the data can be summarised in two matrices, H^1 and H^2, where H^l_{ik} is the number of times individual i of species l is captured at trap k.

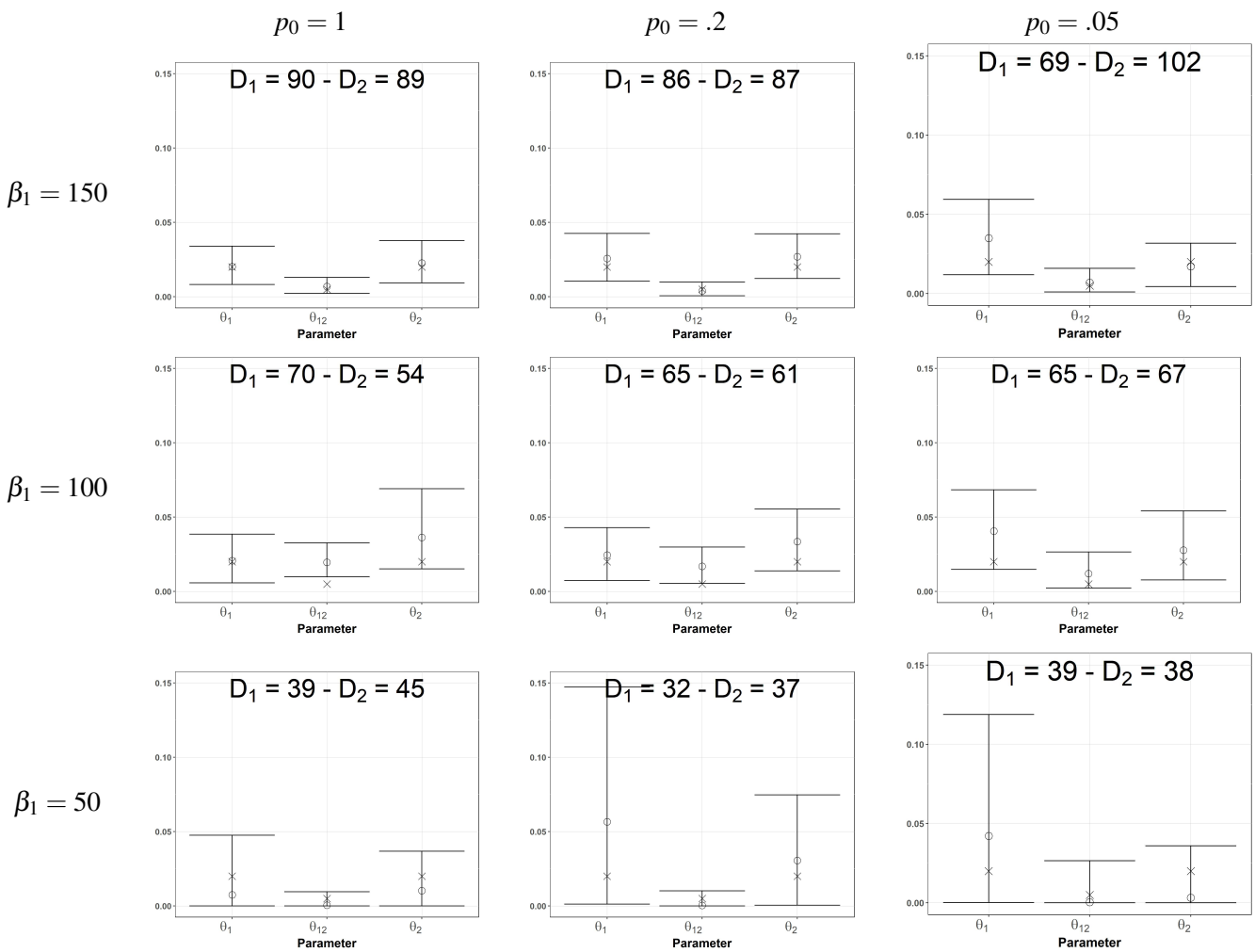


Fig. 5.3 Simulation study: 95% PCI of the interaction parameters, with crosses indicating the true values. The title indicates the number of uniquely captured individuals for each of the two species.

Each individual is assumed to have an activity center s^l_i for the duration of the study. Given the activity centre s^i_l, the probability of capturing individual i of species l at trap k is a decreasing

function of the distance between the location of the activity centre of the individual, s_i^l, and the trap location x_k. We choose here an exponential function, which leads to the probability of capture $p(s,x) = p_0 \exp(-\alpha||s-x||^2)$ between activity center s and trap location x, where p_0 is the baseline probability of detection if the distance is 0 and α is a scale parameter. We assume the two populations are closed during the study, that is, individuals do not enter or leave the area, and we define by N_1 and N_2 the number of individuals in the first and second population, respectively. The two vectors of activity centres $(\mathbf{s}^1, \mathbf{s}^2) = (s_1^1, \ldots, s_{N_1}^1), (s_1^2, \ldots, s_{N_2}^2)$ are assumed to have a BivSC$(\beta_1, \beta_2, \gamma_1, \gamma_2, \gamma_{12})$ prior. We assume Gamma prior distributions for γ_1, γ_2 and γ_{12}, and a Beta prior distribution and truncated normal prior distribution for p_0 and α, respectively. The model can be summarised hierarchically as

$$\begin{cases} H_{ik}^l = \text{Binomial}(T, p_{ik}^l) & i = 1, \ldots, N_l, \quad k = 1, \ldots, K \quad l = 1,2 \\ p_{ik}^l = p_0 \exp(-\alpha||s_i^l - x_k||^2) & i = 1, \ldots, N_l, \quad k = 1, \ldots, K \quad l = 1,2 \\ (\mathbf{s}^1, \mathbf{s}^2) \sim \text{BivSC}(\gamma_1, \gamma_2, \gamma_{12}, \beta_1, \beta_2) & \\ \beta_j \sim \text{Gamma}(a_\beta, b_\beta) & j = 1,2 \\ \gamma_j \sim \text{Gamma}(a_\gamma, b_\gamma) & j = 1,2,12 \\ p_0 \sim \text{Beta}(a_p, b_p) & \\ \alpha \sim \text{N}(\mu_\alpha, \sigma_\alpha^2) & \end{cases}$$

The MCMC sampler is presented in the Appendix.

5.4 Simulation study

In order to assess the performance of the model, we performed several simulations on the CR model, varying the detection probabilities and the intensity of the population sizes. We simulated data from the model of Section 5.3 with $\gamma_1 = \gamma_2 = .02$, $\gamma_{12} = .005$ and $\alpha = .5$. The locations of the traps and the size of the study area is the same as the case study of Section 5.5. We performed 9 sets of simulations, where p_0 takes values in $\{1, .2, .05\}$ and $\beta_1 = \beta_2$ take values in $\{50, 100, 150\}$. We report in Fig. 5.3 the 95% posterior credible intervals (PCI) of interaction parameters $(\theta_1, \theta_{12}, \theta_2)$. As expected, the accuracy of the estimates decreases as the population sizes and/or the capture probability decrease. The PCIs for the interaction parameter are quite wide in some cases, which is also due to the fact that these parameters are strongly correlated with the intensities β_1 and β_2, which causes uncertainty in the estimates as for a combination of values of $(\beta_1, \beta_2, \theta_1, \theta_2, \theta_{12})$ draws from the process are almost identical. However, although the PCIs overlap, the model is almost always able to estimate that the between interactions, θ_1 and θ_2, are larger than the within interaction θ_{12}.

5.5 Application: CR data on leopards and tigers

We apply the model to two CR data of leopards (*Panthera pardus fusca*) and tigers (*Panthera tigris tigris*). Because of confidentiality, we cannot provide further information on the survey area. Camera traps were placed in 253 locations covering an area of 925 km^2. A total of 38 leopards and 24 tigers have been detected during the study.

The posterior distributions of the interaction parameters are shown in Fig. 5.4. The distributions suggest that leopards and tigers tend to repel within the same species more than they repel with the other species, although the effect is not significant as the PCI overlap each other. The differences in these interactions might be due to the fact that the two species hunt different prey. In fact, tigers prefer larger prey such as sambar deer (*Rusa unicolor)* and gaur (*Bos gaurus*), while leopards prefer prey such as chital (*Axis axis*) and wild pig (*Sus scrofa*). However, it is important to mention that given these small sample sizes the posterior distributions are quite sensitive to prior specification, as shown in the simulation study. The posterior distributions of the population densities are shown in Fig. 5.5. The posterior mode is at the sample size for each species, that is, the number of individual detected at least once, and then it sharply decreases as the number of individuals per kilometer squared increases.

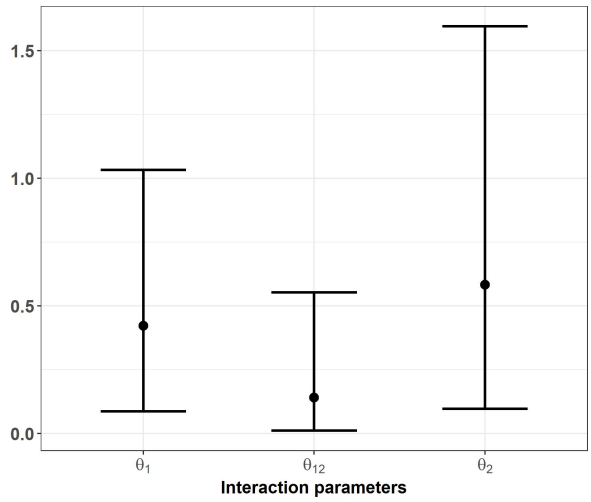
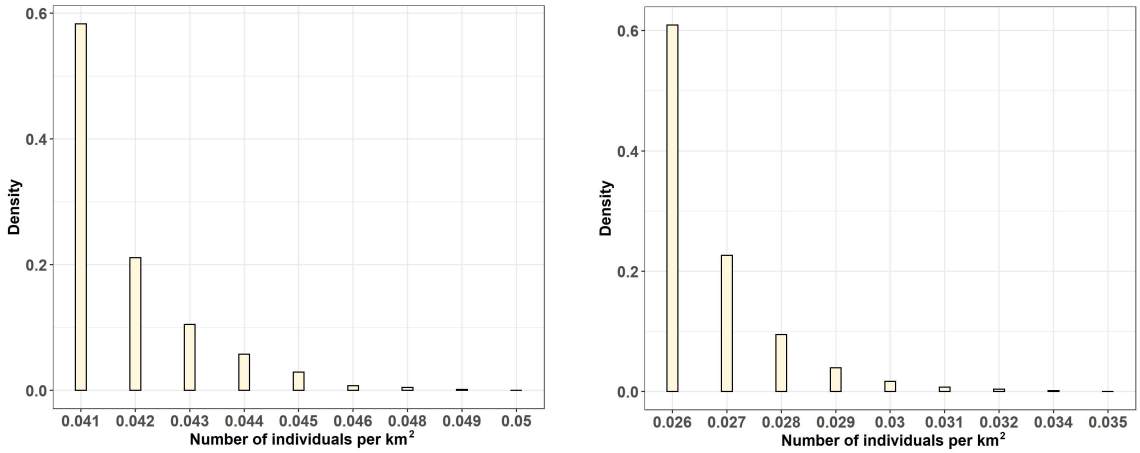


Fig. 5.4 Case study: 90% posterior credible interval of the interaction parameters of the BivSC process, with filled circles showing the posterior median in each case.

5.6 Discussion

We have developed a SCR model for capture-recapture data collected on two or more species accounting for interactions between and within the species using a VPPs. Inference for the parameters of the point process is performed using the MCMH algorithm of Liang and Jin [50], which in the context of VPP we think provides the best balance in terms of speed, accuracy and scalability.



a) b)

Fig. 5.5 Case study: posterior distributions of the population densities of leopards (a) and tigers (b).

Although the model is developed for CR data, it is in theory possible to extend it to other sampling protocols where individual information is not present. The most popular sampling method for unmarked individuals is count data. [11] defined a Bayesian spatial model for count data, which works by inferring the latent capture histories of individuals. However, this is highly inefficient in practise as reconstructing all these latent variables is very difficult.

We used the soft-core process for modelling the interaction between individuals, as this process is summarised using a single parameter and is able to model repulsion, although is not able to model attraction between points. It is generally difficult to model attraction using interaction PPs as if the interaction function ϕ_2 in (5.2.1) is greater than 1 there is the potential for the number of points to explode. Being able to model attraction and repulsion at the same time is useful in the case of particular wildlife populations, as some species present attraction in small groups, such as for example within a family of individuals, and repulsion in bigger groups, such as between different families. This suggests for example a modelling approach where a pairwise and a triple-wise interaction function are considered together in the interaction function, and hence the process is bounded from exploding as the two interaction functions balance each other. Alternatively, a modelling approach is to work with a *conditional* PP, where the number of points is sampled first from another distribution and then the locations of the points are sampled conditional on the total number. However, assuming that the number of points is independent of the spatial structure of the process is often unrealistic.

We have also chosen to assume that the locations of the activity centres were fixed over time, as there is not enough information to estimate the evolution of the activity centres across time because of the low number of recaptures of the individuals. However, in the case where more recaptures are present, a spatio-temporal approach were the locations of the points are allowed to change over time is worth considering. Although spatio-temporal models for log-Gaussian Cox process are available [8, 9], analogous models for Markov point process are not available.

Chapter 6

Discussion

In this thesis, we have presented four manuscripts, described in the second to fifth chapter.

The second chapter presents a BNP model for CR data collected at different sites and for several years. The idea of the paper is to build an extension of the DP, which we termed as HDDP. The new structure developed allows to easily perform density estimation jointly across different units and in the presence of covariates. The advantage of the construction is that the Chinese Restaurant Franchise (CRF) process sampler of [84] can be used to perform inference easily. The drawback of using the CRF and more generally Chinese Restaurant based samplers is that the Gibbs sampler has to update the cluster allocations for all the observations, which can be computationally intensive when the number of observations is large. In addition, as the clusters allocations are updated iteratively, the algorithm is not parallelisable. Moreover, as the HDDP is based on the DDP, it allows to introduce covariates only in the cluster locations but not on the cluster weights. To introduce covariates in the cluster weights, a more flexible structure such as the compound random measures of [35] can be used.

An additional limitation of the MCMC scheme is the slow mixing, which is due to the fact that the latent arrival and departure times of each individual have to be sampled at each iteration. A much more efficient approach, although challenging, is to integrate out the latent times from the MCMC.

The simulation study shows that the model can accurately estimate the entry and exit density for each year, even when capture probabilities are low. A limitation of the case study considered is that for many sites sampling starts after individuals have already entered the site, which causes uncertainty for most of the mass of the arrival density.

The third chapter presents a model for estimating entry and exit patterns, as well as the population size, using a PT prior in the presence of CD. However, as CD do not provide information at the individual level, we have used informative prior distributions to obtain estimates of the capture probability, population size and entry/exit patterns. The advantage of the model is that the computational complexity does not depend on the number of individuals detected but only on the number of sampling occasions, which is advantageous when large counts are considered.

The fourth chapter addresses the potential extensions of the third chapter. More specifically, we extend the model to CR, RR and develop a joint model for CR and CD, and we build a joint model for

different data-sets. The framework developed allows to easily build models for entry and exit patterns using ecological data, as well as showing the connection between the existing models for the different protocols.

To perform inference jointly across different data-sets, we introduce the HLPT, which is based on the logistic normal distribution. The use of the logistic normal distribution has a twofold advantage with respect to the use of Beta distributions. First, inference can be performed with a Gibbs sampler using the Polya-Gamma scheme. Secondly, it is easy to extend the model, adding covariates or considering a time-series model for the site-specific random effects, while keeping the same inference procedure.

As was the case in the third chapter, inference can be performed with a computational complexity that does not depend on the number of individuals but only on the number of sampling occasions. However, for the CR case, implementing the inference procedure requires the computation of a complex likelihood function. An alternative approach is to adopt a likelihood-free approach, such as approximate Bayesian computation [3], which only requires us to simulate data from the model. If the number of individuals is low, we also propose an inference procedure that considers the arrival and departure interval of each individual, similarly to what is performed in the second chapter.

An additional limitation in the inference procedure arises when considering the OPT for the long time series of counts. In fact, inference using the OPT requires to sample first from the full tree and then pruning the tree, which can be inefficient if many prunings have to be performed. Instead, an alternative idea is to work with an already pruned tree and in each iteration only propose to add, delete or change a branch.

The fifth chapter presents a spatial model to estimate interactions between multiple species using CR data. The model can be seen as an extension of the model of [74] through the use of a vector of interaction point process (IPP), which allows us to estimate interactions between and within species. As using an IPP leads to an intractable ratio of normalising constants (RNC), we use the Monte Carlo MH algorithm to approximate the RNC with an importance sampling estimate. Although the model can provide reliable estimates of the interaction parameters when the samples sizes are moderately high, we found that with the sample sizes of our case studies the estimates of the effect are sensitive to the choice of prior distributions. This suggests to potentially increase the number of uniquely captured individuals by using a higher number of traps.

The sampling procedure scales linearly with the number of species as, according to the MCMH, increasing the number of species only requires the computation of the unnormalised interaction function. However, as the number of species increase, it is more difficult for the model to separately estimate each interaction parameter.

Another downside of our approach is that the activity centres of each individual have to be sampled from their posterior distribution, which causes the MCMC algorithm to mix poorly. This happens especially if there are not many recaptures on each individual, as there is not enough information in the data to estimate the activity centre. An alternative approach is to integrate out the latent activity centres. However, this is not possible in our case as the integration with respect to the vector

of interaction point processes is intractable. Instead, a possible approach which could be further investigated is to use a likelihood-free approach for the inference, such as approximate Bayesian computation.

References

[1] Mauricio A Álvarez and Neil D Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12(May):1459–1500, 2011.

[2] Richard J Barker, Matthew R Schofield, William A Link, and John R Sauer. On the reliability of N-mixture models for count data. *Biometrics*, 74(1):369–377, 2018.

[3] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[4] Julian Besag. Some methods of statiscal analysis for spatial data. *Bull. Int. Statist. Inst.*, 47(2): 77–92, 1978.

[5] David L Borchers and MG Efford. Spatially explicit maximum likelihood methods for capture–recapture studies. *Biometrics*, 64(2):377–385, 2008.

[6] Christiaan Both, Sandra Bouwhuis, CM Lessells, and Marcel E Visser. Climate change and population declines in a long-distance migratory bird. *Nature*, 441(7089):81, 2006.

[7] Anders Brix. Generalized gamma measures and shot-noise cox processes. *Advances in Applied Probability*, pages 929–953, 1999.

[8] Anders Brix and Peter J Diggle. Spatiotemporal prediction for log-gaussian cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):823–841, 2001.

[9] Anders Brix and Jesper Moller. Space-time multi type log gaussian cox processes with a view to modelling weeds. *Scandinavian Journal of Statistics*, 28(3):471–488, 2001.

[10] Federico Camerlenghi, Antonio Lijoi, Peter Orbanz, and Igor Prünster. Distribution theory for hierarchical processes. *Annals of Statistics*, 2018.

[11] Richard B Chandler and J Andrew Royle. Spatially explicit models for inference about density in unmarked or partially marked populations. *The Annals of Applied Statistics*, 7(2):936–954, 2013.

[12] Zexun Chen, Bo Wang, and Alexander N Gorban. Multivariate Gaussian and Student-t process regression for multi-output prediction. *arXiv preprint arXiv:1703.04455*, 2017.

[13] Jonathan Christensen and Li Ma. A Bayesian hierarchical model for related densities by using Polya trees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 2020.

[14] RM Cormack. Estimates of survival from the sighting of marked animals. *Biometrika*, 51(3/4): 429–438, 1964.

[15] Charles V Covell Jr et al. *A field guide to the moths of eastern North America.* Houghton Mifflin Co., 1984.

[16] Noel AC Cressie. Statistics for spatial data. *John Wiley and Sons, Inc., New York*, 1993.

[17] Maria De Iorio, Peter Müller, Gary L Rosner, and Steven N MacEachern. An anova model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215, 2004.

[18] Emily B Dennis, Byron JT Morgan, and Martin S Ridout. Computational aspects of n-mixture models. *Biometrics*, 71(1):237–246, 2015.

[19] Emily B Dennis, Byron JT Morgan, Tom M Brereton, David B Roy, and Richard Fox. Using citizen science butterfly counts to predict species population trends. *Conservation biology*, 31 (6):1350–1361, 2017.

[20] Alex Diana, Jim Griffin, and Eleni Matechou. A Polya tree based model for unmarked individuals in an open wildlife population. In *International Conference on Bayesian Statistics in Action*, pages 3–11. Springer, 2018.

[21] Alex Diana, Jim Griffin, and Eleni Matechou. Supplementary material: A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK. 2018.

[22] Robert M Dorazio, Bhramar Mukherjee, Li Zhang, Malay Ghosh, Howard L Jelks, and Frank Jordan. Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics*, 64(2):635–644, 2008.

[23] Dirk Eddelbuettel, Romain François, J Allaire, Kevin Ushey, Qiang Kou, N Russel, John Chambers, and D Bates. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[24] Murray Efford. Density estimation in live-trapping studies. *Oikos*, 106(3):598–610, 2004.

[25] Murray G Efford, Deanna K Dawson, and David L Borchers. Population density estimated from locations of individuals on a passive detector array. *Ecology*, 90(10):2676–2682, 2009.

[26] Sarah M Eglington, Romain Julliard, Gabriel Gargallo, Henk P van der Jeugd, James W Pearce-Higgins, Stephen R Baillie, and Robert A Robinson. Latitudinal gradients in the productivity of e uropean migrant warblers have not shifted northwards during a period of climate change. *Global Ecology and Biogeography*, 24(4):427–436, 2015.

[27] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[28] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[29] Thomas S Ferguson et al. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629, 1974.

[30] Jessica H Ford, Toby A Patterson, and Mark V Bravington. Modelling latent individual heterogeneity in mark-recapture data with Dirichlet process priors. *arXiv preprint arXiv:1511.07103*, 2015.

[31] Stephen N Freeman and Byron J T Morgan. A modelling strategy for recovery data from birds ringed as nestlings. *Biometrics*, 48:217–236, 1992.

[32] Hong Ge, Yutian Chen, Moquan Wan, and Zoubin Ghahramani. Distributed inference for dirichlet process mixture models. In *International Conference on Machine Learning*, pages 2276–2284, 2015.

[33] Charles J Geyer and Jesper Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian journal of statistics*, pages 359–373, 1994.

[34] Phillip Gienapp, Roosa Leimu, and Juha Merilä. Responses to climate change in avian migration time—microevolution versus phenotypic plasticity. *Climate Research*, 35(1/2):25–35, 2007.

[35] Jim E Griffin and Fabrizio Leisen. Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (2):525–545, 2017.

[36] Timothy E Hanson. Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101(476):1548–1565, 2006.

[37] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

[38] Ary A Hoffmann and Carla M Sgro. Climate change and evolutionary adaptation. *Nature*, 470 (7335):479, 2011.

[39] Sonia Jain and Radford M Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of computational and Graphical Statistics*, 13(1): 158–182, 2004.

[40] Alesjandro Jara and Timothy Hanson. A class of mixtures of dependent tail-free processes. *Biometrika*, 98(3):553–566, 2011.

[41] Jens Ledet Jensen and Jesper Møller. Pseudolikelihood for exponential family models of spatial point processes. *The Annals of Applied Probability*, 1(3):445–461, 1991.

[42] Marina Jiménez-Muñoz, Diana J Cole, Stephen N Freeman, Robert A Robinson, Stephen R Baillie, and Eleni Matechou. Estimating age-dependent survival from age-aggregated ringing data extending the use of historical records. *Ecology and Evolution*, 9(2):769–779, 2019.

[43] Alison Johnston, Robert A Robinson, Gabriel Gargallo, Romain Julliard, Henk Jeugd, and Stephen R Baillie. Survival of afro-palaearctic passerine migrants in western Europe and the impacts of seasonal weather variables. *Ibis*, 158(3):465–480, 2016.

[44] George M Jolly. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52(1/2):225–247, 1965.

[45] Wilfrid S Kendall and Jesper Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, 32 (3):844–865, 2000.

[46] John Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

[47] John Frank Charles Kingman. *Poisson processes*, volume 3. Clarendon Press, 1993.

[48] Michael Lavine et al. Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235, 1992.

[49] Michael Lavine et al. More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 22(3):1161–1176, 1994.

[50] Faming Liang and Ick-Hoon Jin. A monte carlo metropolis-hastings algorithm for sampling from distributions with intractable normalizing constants. *Neural computation*, 25(8):2199–2234, 2013.

[51] Steven N MacEachern. Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, pages 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999, 1999.

[52] Darryl I MacKenzie, James D Nichols, Gideon B Lachman, Sam Droege, J Andrew Royle, and Catherine A Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

[53] Darryl I MacKenzie, James D Nichols, Gideon B Lachman, Sam Droege, J Andrew Royle, and Catherine A Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

[54] Daniel Manrique-Vallier. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254, 2016.

[55] Eleni Matechou, François Caron, et al. Modelling individual migration patterns using a Bayesian nonparametric approach for capture–recapture data. *The Annals of Applied Statistics*, 11(1): 21–40, 2017.

[56] Rachel S McCrea and Byron JT Morgan. *Analysis of capture-recapture data*. Chapman and Hall/CRC, 2014.

[57] Rachel S McCrea, Byron JT Morgan, and Diana J Cole. Age-dependent mixture models for recovery data on animals marked at unknown age. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):101–113, 2013.

[58] Anders Pape Møller, Diego Rubolini, and Esa Lehikoinen. Populations of migratory bird species that did not show a phenological response to climate change are declining. *Proceedings of the National Academy of Sciences*, 2008.

[59] Jesper Møller and Rasmus Plenge Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, 2003.

[60] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

[61] Jesper Møller, Anthony N Pettitt, Robert Reeves, and Kasper K Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

[62] Iain Murray, Zoubin Ghahramani, and David MacKay. MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*, 2012.

[63] Jerzy Neyman and Elizabeth L Scott. Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(1):1–29, 1958.

[64] Yosihiko Ogata and Masaharu Tanemura. Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):496–518, 1984.

[65] David L Otis, Kenneth P Burnham, Gary C White, and David R Anderson. Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, (62):3–135, 1978.

[66] WJ Peach, ST Buckland, and SR Baillie. The use of constant effort mist-netting to measure between-year changes in the abundance and productivity of common passerines. *Bird Study*, 43 (2):142–156, 1996.

[67] John Pickering. Find your dark side: Invitation to join discover life's mothing project. *S. Lep. News*, 37(4):205–208, 2015.

[68] Shirley Pledger, Murray Efford, Kenneth Pollock, Jaime Collazo, and James Lyons. Stopover duration analysis with departure probability dependent on unknown time since arrival. In *Modeling demographic processes in marked populations*, pages 349–363. Springer, 2009.

[69] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Polya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–1349, 2013.

[70] Chris Preston. *Random Fields*, volume 534. Springer-Verlag New York, 1976.

[71] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[72] Tuomas Rajala, DJ Murrell, and SC Olhede. Detecting multivariate interactions in spatial point patterns with gibbs models and variable selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1237–1273, 2018.

[73] Carl Edward Rasmussen. Gaussian processes in machine learning. MIT Press, 2006.

[74] Brian J Reich and Beth Gardner. A spatial capture-recapture model for territorial species. *Environmetrics*, 25(8):630–637, 2014.

[75] Brian D Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192, 1977.

[76] David Robson and Carles Barriocanal. Ecological conditions in wintering and passage areas as determinants of timing of spring migration in trans-saharan migratory birds. *Journal of Animal Ecology*, 80(2):320–331, 2011.

[77] J Andrew Royle. N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115, 2004.

[78] J Andrew Royle and Kevin V Young. A hierarchical model for spatial capture–recapture data. *Ecology*, 89(8):2281–2289, 2008.

[79] J Andrew Royle, Richard B Chandler, Rahel Sollmann, and Beth Gardner. *Spatial capture-recapture*. Academic Press, 2013.

[80] David Ruelle. *Statistical mechanics: Rigorous results*. World Scientific, 1969.

[81] George AF Seber. A note on the multiple-recapture census. *Biometrika*, 52(1/2):249–259, 1965.

[82] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[83] David J Strauss. A model for clustering. *Biometrika*, 62(2):467–475, 1975.

[84] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2012.

[85] Simone Tenan, Mauro Fasola, Stefano Volponi, and Giacomo Tavecchia. Conspecific and not performance-based attraction on immigrants drives colony growth in a waterbird. *Journal of Animal Ecology*, 86(5):1074–1081, 2017.

[86] Rasmus Waagepetersen, Yongtao Guan, Abdollah Jalilian, and Jorge Mateu. Analysis of multispecies point patterns by using multivariate log-gaussian cox processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):77–96, 2016.

[87] David L Wagner. *Caterpillars of eastern North America: a guide to identification and natural history*. Princeton University Press, 2005.

[88] Robert L Wolpert and Katja Ickstadt. Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267, 1998.

[89] Wing H Wong, Li Ma, et al. Optional Pólya tree and Bayesian inference. *The Annals of Statistics*, 38(3):1433–1459, 2010.

Appendix A

Supplementary material for A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK

A.1 Posterior distributions

A.1.1 Posterior distributions for μ and Σ

The joint distribution $p(\mu,\beta,\Sigma|y,b,B,K)$ can be computed as:

$$p(\mu,\Sigma|y,b,B,K,\nu_0,\Sigma_0) \propto p(y|\mu,\Sigma)p(\mu|\beta,K,\Sigma)p(\beta|b,B,\Sigma)p(\Sigma|\nu,\Sigma_0) \propto$$

$$\exp\left\{-\frac{1}{2}tr\left(\Sigma^{-1}\underbrace{\left(\sum_{i=1}^{n}(y_i-\mu_{c_i})^T(y_i-\mu_{c_i})+(\mu-X\beta)^TK^{-1}(\mu-X\beta)+(\beta-b)^TB^{-1}(\beta-b)\right)}_{\star}\right)\right\}p(\Sigma|\nu_0,\Sigma_0) \propto$$

where:

$$\star = \sum_{i=1}^{n}(y_i-\mu_{c_i})^T(y_i-\mu_{c_i})+\mu K^{-1}\mu-2\mu^TK^{-1}X\beta+\beta^TX^TK^{-1}X\beta+\beta^TB^{-1}\beta-2\beta^TB^{-1}b+b^TB^{-1}b \propto$$

$$\sum_{i=1}^{n}(y_i-\mu_{c_i})^T(y_i-\mu_{c_i})+\mu K^{-1}\mu+b^TB^{-1}b+\beta^T\underbrace{(X^TK^{-1}X+B^{-1})}_{\Lambda_\beta^{-1}}\beta-2\beta^T(B^{-1}b+X^TK^{-1}\mu)$$

Completing the square and integrating over β, the marginal of μ and Σ is proportional to:

$$\sum_{i=1}^{n}(y_i-\mu_{c_i})^T(y_i-\mu_{c_i})+\mu K^{-1}\mu+b^TB^{-1}b-(B^{-1}b+X^TK^{-1}\mu)^T\Lambda_\beta(B^{-1}b+X^TK^{-1}\mu)\propto$$

$$\sum_{i=1}^{n}(y_i-\mu_{c_i})^T(y_i-\mu_{c_i})+\mu K^{-1}\mu+b^TB^{-1}b-(B^{-1}b)^T\Lambda_\beta(B^{-1}b)-2\mu^T(X^TK^{-1})^T\Lambda_\beta B^{-1}b-\mu^T(X^TK^{-1})^T\Lambda_\beta(X^TK^{-1})\mu\propto$$

$$\sum_{i=1}^{n}y_i^Ty_i-2\mu^T\sum_{i=1}^{n}V_i+\mu^T\sum_{i=1}^{n}D_i\mu+\mu K^{-1}\mu+b^TB^{-1}b-(B^{-1}b)^T\Lambda_\beta(B^{-1}b)-$$
$$2\mu^T(X^TK^{-1})^T\Lambda_\beta B^{-1}b-\mu^T(X^TK^{-1})^T\Lambda_\beta(X^TK^{-1})\mu\propto$$

$$\underbrace{\sum_{i=1}^{n}y_i^Ty_i+b^TB^{-1}b-(B^{-1}b)^T\Lambda_\beta(B^{-1}b)}_{S}-2\mu^T\underbrace{(\sum_{i=1}^{n}V_i+(X^TK^{-1})^T\Lambda_\beta B^{-1}b)}_{m_\mu}+\mu^T\underbrace{(\sum_{i=1}^{n}D_i+K^{-1}-(X^TK^{-1})^T\Lambda_\beta(X^TK^{-1}))}_{\Lambda_\mu}\mu$$

Summarising, we have:

$$p(y|\mu,\Sigma)p(\mu|\beta,K,\Sigma)p(\beta|b,B,\Sigma)\propto$$

$$\exp\left\{-\frac{1}{2}tr\left(\Sigma^{-1}\left((\mu-(\Lambda_\mu)^{-1}m_\mu)^T(\Lambda_\mu)^{-1}(\mu-(\Lambda_\mu)^{-1}m_\mu)\right)\right)\right\}\exp\left\{-\frac{1}{2}tr\left(\Sigma^{-1}\left(S-(m_\mu)^T(\Lambda_\mu)^{-1}(m_\mu)\right)\right)\right)$$

$$\Rightarrow\begin{cases}\Sigma|y\sim IW(\nu_0+n,\Sigma_0+S-(m_\mu)^T(\Lambda_\mu)^{-1}(m_\mu)\\ \mu|y,\Sigma\sim MN(\Lambda_\mu^{-1}(m_\mu),\Lambda_\mu^{-1},\Sigma)\end{cases}$$

A.1.2 Posterior for b

The conditional distribution $p\left(b|\{\mu_j\},\{\delta_j\},\{\Sigma_j\},b_0,B,B_0\right)$ can be computed as:

$$p\left(b|\{\mu_j\},\{\delta_j\},\{\Sigma_j\},b_0,B,B_0\right)\propto p(b|b_0,B_0,\Sigma_0)\left(\prod_{i=1}^{n}p(\delta_j|\mu_j,b,\Sigma_j,K)\right)=$$

$$\mathrm{MN}(b|b_0,B_0,\Sigma_0)\prod_{k=1}^{K}MN(\mu_k|X(b+\delta_k),K,\Sigma_k)\propto\exp\left\{-\frac{1}{2}tr\left(\star\right)\right\}$$

where

$$\star=(vec(b)-vec(b_0))^T(\Sigma_0\otimes B_0)^{-1}(vec(b)-vec(b_0))+\sum_k(\underbrace{vec(\mu_j)-vec(X\delta_j)}_{vec(\tilde{\delta}_k)}-vec(Xb)))^T(\Sigma_j\otimes K)^{-1}(vec(\mu_j)-vec(X(b+\delta_j)))=$$

$$vec(b)^T(\Sigma_0\otimes B_0)^{-1}vec(b)-2vec(b)^T\left((\Sigma_0\otimes B_0)^{-1}\vec{(}b_0)\right)+\sum_k(vec(\tilde{\delta}_k)-(I_p\otimes X)vec(b))^T(\Sigma_j\otimes K)^{-1}(vec(\tilde{\delta}_k)-(I_p\otimes X)vec(b))=$$

$$vec(b)^T((\Sigma_0\otimes B_0)^{-1}+\sum_k(I_p\otimes X)^T(\Sigma_k\otimes K)^{-1}(I_p\otimes X))vec(b)-2vec(b)^T\left((\Sigma_0\otimes B_0)^{-1}\vec{(}b_0)+\sum_k(I_p\otimes X)^T(\Sigma_k\otimes K)^{-1}vec(\tilde{\delta}_k)\right)$$

from which it follows that the posterior distribution is a normal.

The posterior distribution $p(\delta_k|\mu_k, B, \Sigma_k, b)$ can be written as

$$p(\delta_k|B,\Sigma_k)p(\mu_k|b+\delta_k,K,\Sigma_k) \propto \mathrm{MN}(\delta_k|0,B,\Sigma_k)\mathrm{MN}(\mu_k|b+\delta_k,K,\Sigma_k) \propto \exp\left\{-\frac{1}{2}tr\left(\Sigma_k^{-1}(\star)\right)\right\}$$

where

$$\star = \delta_k^T B^{-1}\delta_k + (\underbrace{\mu_k - Xb}_{\tilde{\delta}_k} - X\delta_k)^T K^{-1}(\mu_k - Xb - X\delta_k) = \delta_k^T B^{-1}\delta_k + \delta_k^T X^T K^{-1} X\delta_k - 2\delta_k^T X^T K^{-1}\tilde{\delta}_k + C$$

$$\Rightarrow \exp\left\{-\frac{1}{2}tr\left(\Sigma_k^{-1}(\star)\right)\right\} \propto \exp\left\{-\frac{1}{2}tr\left(\Sigma_k^{-1}\left(\delta_k^T \underbrace{(B^{-1}+X^T K^{-1}X)}_{\Lambda_\beta}\delta_k - 2\delta_k^T \underbrace{(X^T K^{-1}\tilde{\delta}_k)}_{\mu_\delta}\right)\right)\right\}$$

$$\Rightarrow \beta \sim MN(\Lambda_\beta^{-1}\mu_\delta, \Lambda_\delta^{-1}, \Sigma_k)$$

A.2 Posterior update of α and γ

In a single DP, the probability of having k different clusters given the concentration parameter α and the sample size n is

$$p(k|\alpha,n) \propto \alpha^{k-1}(\alpha+n)\beta(\alpha+1,n)$$

In order to update the parameter γ of an HDP, we use the update presented in Escobar and West [27], using the number of different dishes as the number of clusters, and the number of tables as the sample size.

In the case of the parameter α of the HDP, if k_i is the number of different tables in restaurant i, the posterior is

$$p(\alpha|\{k_i\}) \propto p(\alpha)\prod_{i=1}^{J} p(k_i|\alpha,n_i) \propto p(\alpha)\prod_{i=1}^{J}\alpha^{k_i-1}(\alpha+n_i)\beta(\alpha+1,n_i)$$

$$\propto p(\alpha)\prod_{i=1}^{J}\alpha^{k_i-1}(\alpha+n_i)\int_0^1 x^\alpha(1-x)^{n_i}$$

$$\Rightarrow p(\alpha|\{k_i\},\{\eta_i\}) \propto p(\alpha)\prod_{i=1}^{J}\alpha^{k_i-1}(\alpha+n_i)\eta_i^\alpha$$

where $\eta_i \sim \beta(\alpha+1,n_i)$. Thus:

$$p(\alpha)\prod_{i=1}^{J}\alpha^{k_i-1}(\alpha+n_i)\eta_i^{\alpha} \propto \alpha^{a-1}e^{-b\alpha}\prod_{i=1}^{J}\alpha^{k_i-1}(\alpha+n_i)e^{\alpha\log(\eta_i)} =$$

$$\alpha^{a-1}e^{-b\alpha}\alpha^{\sum k_i-J}e^{\alpha\sum\log(\eta_i)}\prod_{i=1}^{J}(\alpha+n_i) = \alpha^{a-1+\sum k_i-J}e^{-\alpha(b-\sum\log(\eta_i))}\prod_{i=1}^{J}(\alpha+n_i) =$$

$$\alpha^{a-1+\sum k_i-J}e^{-\alpha(b-\sum\log(\eta_i))}\left(\alpha^J+\alpha^{J-1}\left(\sum_{i=1}^{J}n_i\right)+\alpha^{J-2}\left(\sum_{i\neq j}^{J}n_in_j\right)+\cdots+\prod_{i=1}^{J}n_i\right) =$$

$$\alpha^{a-1+\sum k_i}e^{-\alpha(b-\sum\log(\eta_i))}\left(1+\alpha^{-1}\left(\sum_{i=1}^{J}n_i\right)+\alpha^{-2}\left(\sum_{i\neq j}^{J}n_in_j\right)+\cdots+\alpha^{-J}\prod_{i=1}^{J}n_i\right) \tag{A.2.1}$$

which reduces to the following mixture of gamma distributions:

$$\pi_1\mathrm{G}(a+\sum k_i, b-\sum\log(\eta_i)) + \pi_2\mathrm{G}(a+\sum k_i-1, b-\sum\log(\eta_i)) + \ldots + \pi_{J+1}\mathrm{G}(a+\sum k_i-J, b-\sum\log(\eta_i))$$

The weights can be found equating the coefficient in A.2.1 to the coefficient of the mixture of Gamma distributions. Defining as $C_a = a+\sum k$ and $C_b = b-\sum\log(\eta_i)$:

$$\begin{cases} \pi_1\left(\frac{C_b^{C_a}}{\Gamma(C_a)}\right) = 1 \\ \pi_2\left(\frac{C_b^{C_a-1}}{\Gamma(C_a-1)}\right) = \sum_{i=1}^{J}n_i \\ \pi_3\left(\frac{C_b^{C_a-2}}{\Gamma(C_a-2)}\right) = \sum_{i\neq j}n_in_j \\ \vdots \\ \pi_{J+1}\left(\frac{C_b^{C_a-J}}{\Gamma(C_a-J)}\right) = \prod_{i=1}^{J}n_i \end{cases}$$

$$\Rightarrow \begin{cases} \pi_1\left(\frac{C_b^{C_a}}{(C_a-1)\ldots(C_a-J)\Gamma(C_a-J)}\right) = 1 \\ \pi_2\left(\frac{C_b^{C_a-1}}{(C_a-2)\ldots(C_a-J)\Gamma(C_a-J)}\right) = \sum_{i=1}^{J}n_i \\ \pi_3\left(\frac{C_b^{C_a-2}}{(C_a-3)\ldots(C_a-J)\Gamma(C_a-J)}\right) = \sum_{i\neq j}n_in_j \\ \vdots \\ \pi_{J+1}\left(\frac{C_b^{C_a-J}}{\Gamma(C_a-J)}\right) = \prod_{i=1}^{J}n_i \end{cases}$$

Observing that the term $\frac{C_b^{C_a}}{\Gamma(C_a-J)}$ is common to all the terms, the weights are given by:

$$\begin{cases} \pi_1 \propto (a+\sum k_i - 1)\dots(a+\sum k_i - J) \\ \pi_2 \propto (a+\sum k_i - 2)\dots(a+\sum k_i - J)(b-\sum \log(\eta_i))\sum_{i=1}^J n_i \\ \pi_3 \propto (a+\sum k_i - 3)\dots(a+\sum k_i - J)(b-\sum \log(\eta_i))^2 \sum_{i\neq j} n_i n_j \\ \vdots \\ \pi_{J+1} \propto (b-\sum \log(\eta_i))^J \prod_{i=1}^J n_i \end{cases}$$

A.3 Results on Multivariate Gaussian process

As for the univariate Gaussian process, if we assume a matrix normal prior distribution $\beta \sim MN(b,B,\Sigma)$ this can be integrated out obtaining another multivariate Gaussian process:

$$(y_1,\dots,y_n) \sim \text{MGP}(Xb, K+I+XBX^T, \Sigma)$$

If we consider the joint distribution of μ and β:

$$p(\mu,\beta) \propto exp(-\frac{1}{2}tr(\Sigma^{-1}(\mu - X\beta)^T K^{-1}(\mu - X\beta) + (\beta - b)^T B^{-1}(\beta - b)))$$

$$= exp(-\frac{1}{2}tr(\Sigma^{-1}(\mu^T K^{-1}\mu) + \beta^T (X^T K^{-1} X)\beta - 2\beta^T (X^T K^{-1})\mu +$$

$$\beta^T B^{-1}\beta + b^T B^{-1} b - 2\beta^T B^{-1} b))$$

$$\propto exp(-\frac{1}{2}tr(\Sigma^{-1}((\mu^T K^{-1}\mu) + \beta^T (\underbrace{X^T K^{-1} X + B^{-1}}_{\Lambda_\beta})\beta - 2\beta^T (\underbrace{X^T K^{-1}\mu + B^{-1} b}_{m_\beta})))$$

We now complete the square with respect to β

$$\propto exp(-\frac{1}{2}tr(\Sigma^{-1}((\mu^T K^{-1}\mu) - m_\beta^T \Lambda_\beta^{-1} m_\beta + (\beta - \Lambda_\beta^{-1} m_\beta)^T \Lambda_\beta (\beta - \Lambda_\beta^{-1} m_\beta)) =$$

Integrating over β, the last term does not give any contribution as it is the integral of a matrix normal distribution.

$$= \text{(after integrating over } \beta) \quad exp(-\frac{1}{2}tr(\Sigma^{-1}((\mu^T K^{-1}\mu) -$$

$$(X^T K^{-1}\mu + B^{-1} b)^T (X^T K^{-1} X + B^{-1})^{-1} (X^T K^{-1}\mu + B^{-1} b)) =$$

$$= exp(-\frac{1}{2}tr(\Sigma^{-1}(\mu^T (K^{-1} - (X^T K^{-1})^T (X^T K^{-1} X + B^{-1} b)^{-1} (X^T K^{-1}))\mu -$$

$$2\mu^T (K^{-1} X \underbrace{(X^T K^{-1} X + B^{-1})^{-1}}_{B^\star} B^{-1} b)) =$$

Using the Woodbury matrix identity on the matrix $K^{-1} - (X^T K^{-1})^T B^\star (X^T K^{-1})$, this will be equal to $K + XBX^T =: (K^\star)^{-1}$, which thus correspond to the precision matrix.

It thus follows that the marginal distribution of μ is $\mathrm{MGP}(K^{\star}(K^{-1}XB^{\star})B^{-1}b, K^{\star}, \Sigma)$

A.4 Prior distributions

Below we present the prior distributions for the capture probability and on the intensity of the population size.

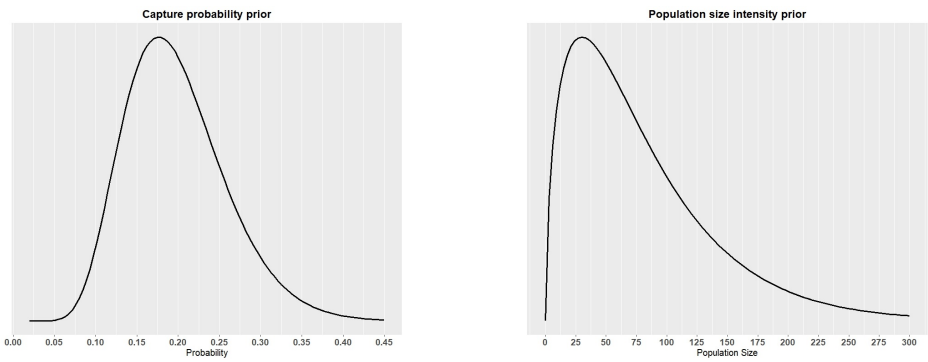


Fig. A.1 Prior distributions on the capture probability and on the intensity of the population size.

A.5 Results

In this section we present results for the remaining sites and year, not shown in the paper.

Fig. A.2 Arrival distribution - Posterior distributions, with site names given on top of each column. The black line shows the posterior mean density and the grey area shows the 95% PCI. The sampling occasions are shown in bold on the x-axis.

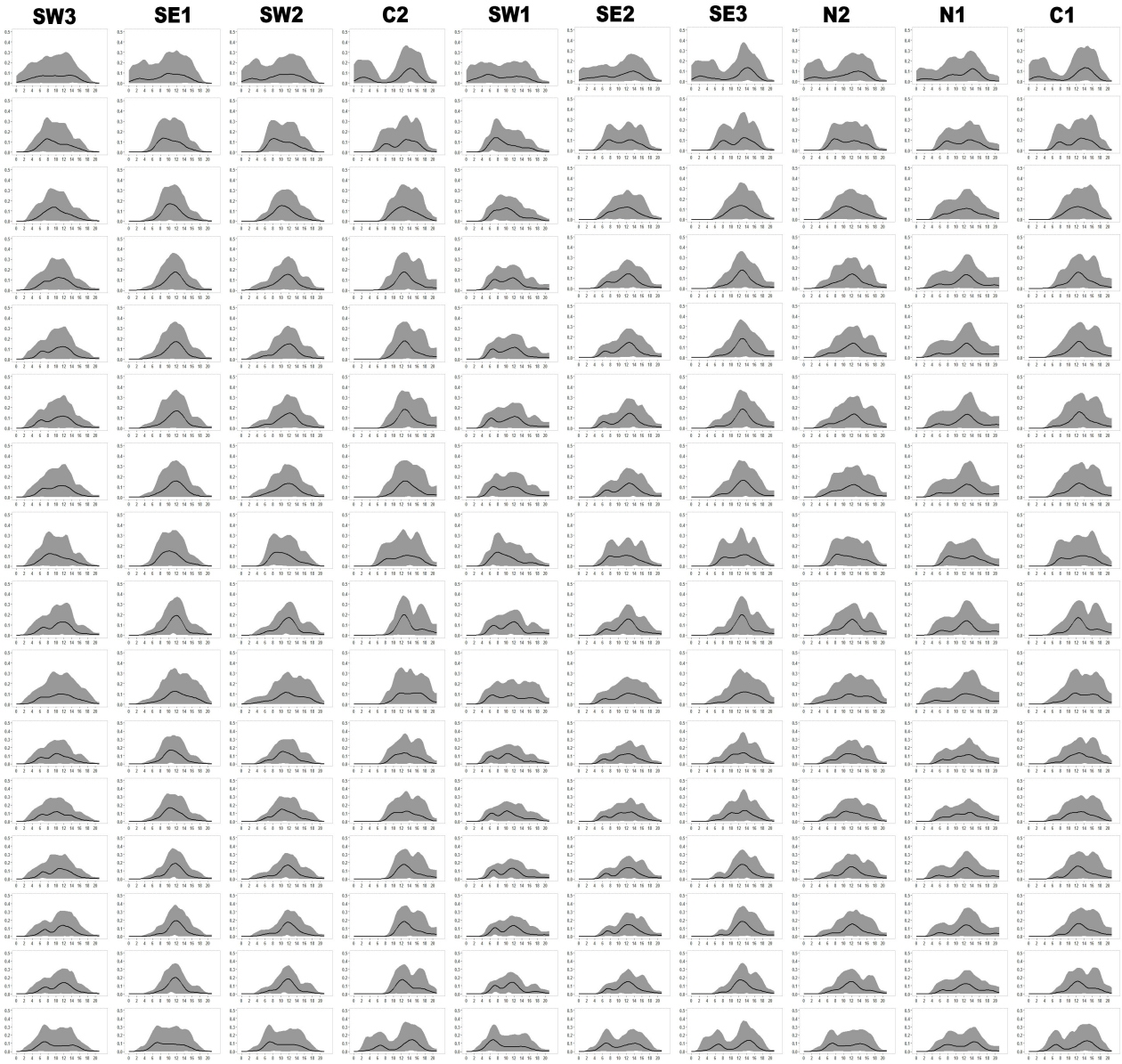


Fig. A.3 Length of stay - Posterior distributions, with site names given on top of each column. The black line shows the posterior mean density and the grey area shows the 95% PCI.

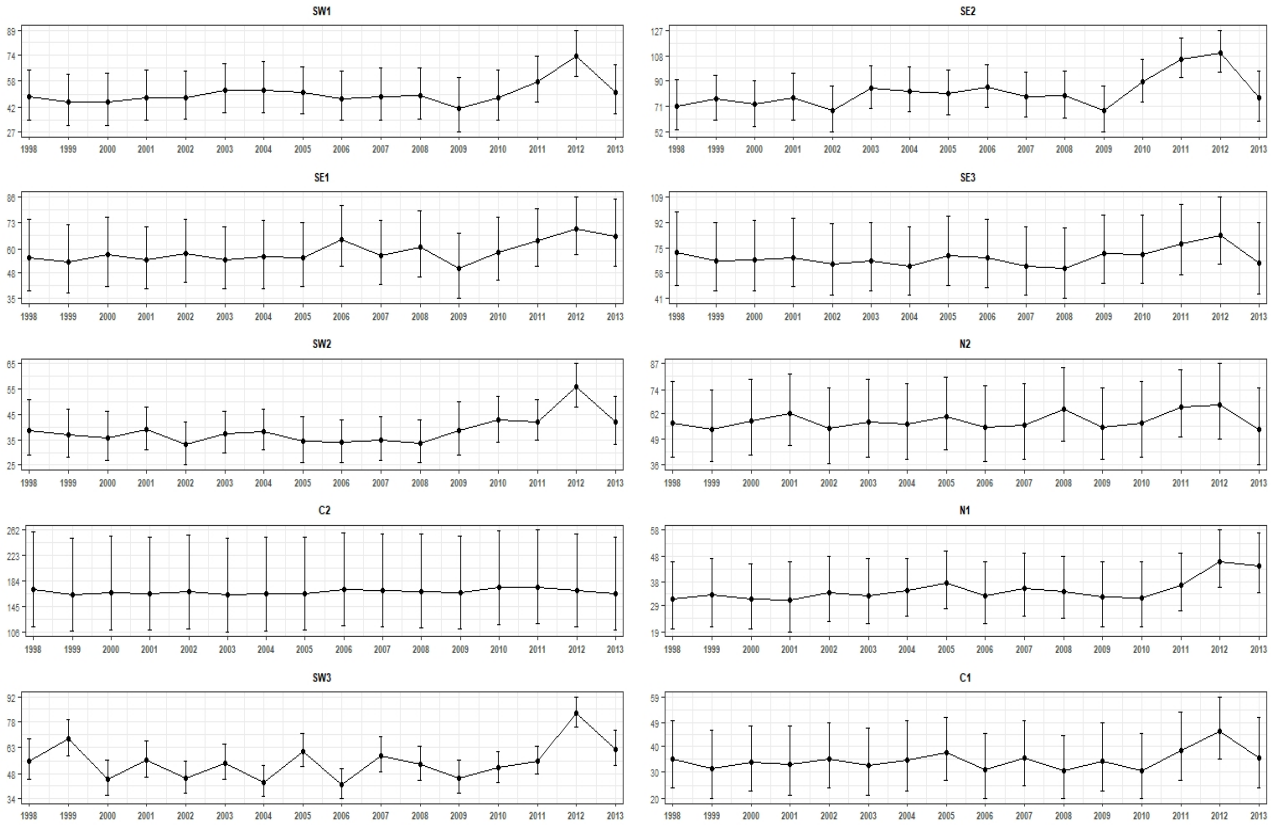


Fig. A.4 Posterior densities of the population size. The bars show the 95% credible interval and the red dots show the posterior means.

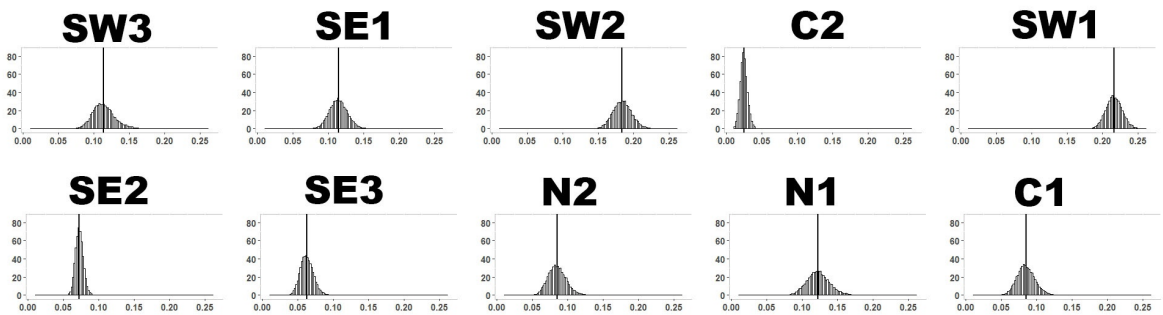


Fig. A.5 Posterior densities of the capture probabilities. The black lines show the posterior means.

Appendix B

Supplementary material for A general modelling framework for ecological data based on the Polya Tree prior

B.1 Partitions

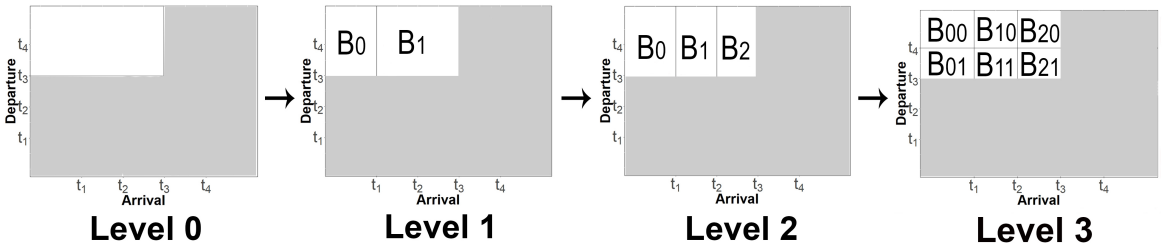


Fig. B.1 Entry and exit partition for the slice $k = 3$ consisting of the individuals first captured at time t_k.

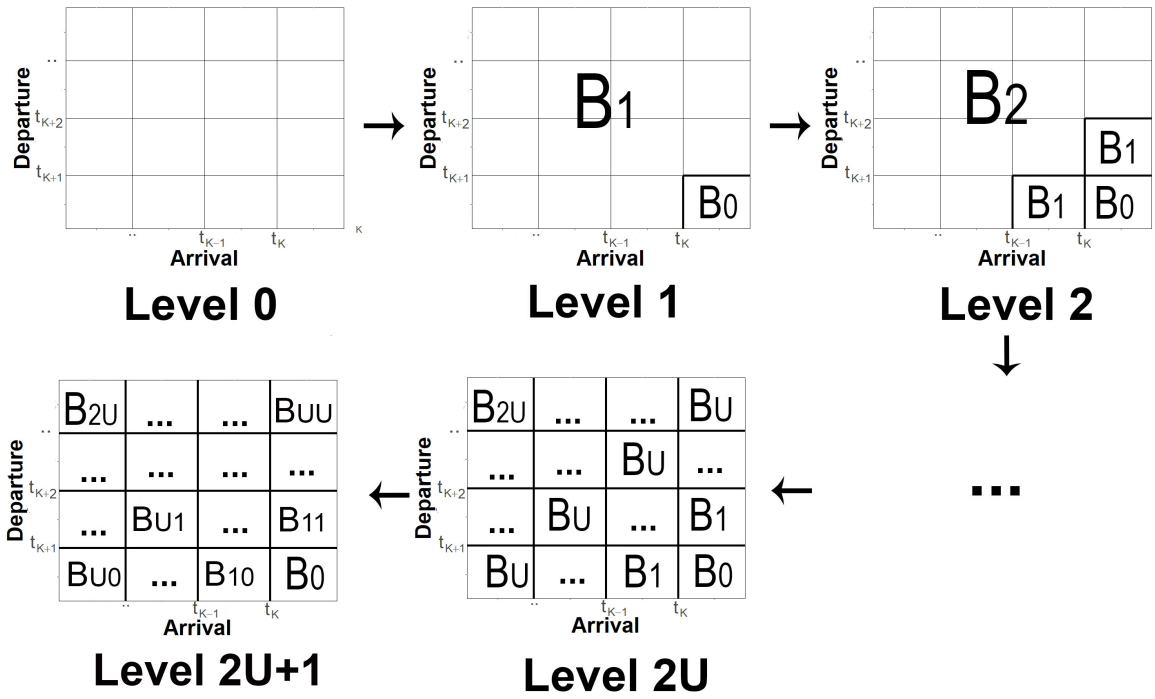


Fig. B.2 Partition of the PT in the case of RR data for the matrix $\mathbf{n}^k$ corresponding to the individuals marked at time t_k. The sample space is split in the set of individuals staying for less than one year, B_0, and the set of individuals staying for more than one year, B_1. After this process is repeated for $2U$ years, each set is split according to the exit intervals.

B.2 Proof of Lemma 1

$$\mathrm{Var}(X_1X_2) = \mathrm{Var}(X_1)\mathrm{Var}(X_2) + \mathrm{Var}(X_1)\mathbb{E}[X_2]^2 + \mathrm{Var}(X_2)\mathbb{E}[X_1]^2 =$$

$$= \frac{\alpha^2}{(2\alpha)^2(2\alpha+1)}\frac{\beta^2}{(2\beta)^2(2\beta+1)} + \frac{1}{4}\left(\frac{\alpha^2}{(2\alpha)^2(2\alpha+1)} + \frac{\beta^2}{(2\beta)^2(2\beta+1)}\right) =$$

$$= \frac{\alpha^2\beta^2}{(2\alpha)^2(2\alpha+1)(2\beta)^2(2\beta+1)} + \frac{1}{4}\left(\frac{\alpha^2(2\beta)^2(2\beta+1)+\beta^2(2\alpha)^2(2\alpha+1)}{(2\alpha)^2(2\alpha+1)(2\beta)^2(2\beta+1)}\right) =$$

$$= \frac{\alpha^2\beta^2+\alpha^2\beta^2(2\beta+1)+\alpha^2\beta^2(2\alpha+1)}{(2\alpha)^2(2\alpha+1)(2\beta)^2(2\beta+1)} = \frac{\alpha^2\beta^2(2\alpha+2\beta+3)}{(2\alpha)^2(2\alpha+1)(2\beta)^2(2\beta+1)} =$$

$$\mathrm{Var}(X_1)\frac{2\alpha+2\beta+3}{4(2\beta+1)}$$

and $\frac{2\alpha+2\beta+3}{4(2\beta+1)} \leq 1$ as

$$\frac{2\alpha+2\beta+3}{4(2\beta+1)} \leq 1 \iff \beta \geq \frac{\alpha}{3} - \frac{1}{6}$$

which holds true as $\beta \geq \alpha$.

B.3 Likelihoods for capture-recapture data

B.3.1 Likelihood of the CJS

Before deriving the likelihood of the general model, we first present a simple case with only 2 sampling occasions. By definition, individuals with capture history 01 can only be assigned to n_2^2, from which it follows that the number of possible combinations for assigning the Z_{22} individuals are $\binom{n_2^2}{Z_{22}}$. The individuals with capture history 11 can only be assigned to n_2^1, which gives $\binom{n_2^1}{Z_{12}}$ combinations, while the individuals with capture history 10 can be assigned to either n_2^1 or n_1^1, and as Z_{12} are already assigned to n_2^1, the number of combinations is equal to $\binom{(n_2^1-Z_{12})+n_1^1}{Z_{11}}$. If we define as N_k and C_k the individuals with first capture on sampling occasion k that are available for capture and captured, respectively, the probability of the observed capture histories is

$$\underbrace{\binom{n_{12}}{Z_{12}}\binom{(n_{12}-Z_{12})+n_{11}}{Z_{11}}p^{C_1}(1-p)^{N_1-C_1}}_{\text{individuals captured on the } 1^{st} \text{ sampling occasion}} \times \underbrace{\binom{n_{22}}{Z_{22}}p^{C_2}(1-p)^{N_2-C_2}}_{\text{individuals captured on the } 2^{nd} \text{ sampling occasion}}$$

Using the same rationale for the case with K sampling occasions, the likelihood for the individuals with first capture on sampling occasion k, can be expressed as

$$f_k(H_k|\mathbf{n}^k) = \binom{n_{K-k}^k}{Z_{kK}}\binom{(n_{K-k}^k - Z_{k,K}) + n_{K-k-1}^k}{Z_{k,K-1}}\cdots\binom{n_1^k + \sum_{l=2}^{K-k}(n_l^k - Z_{l+k,K})}{Z_{k,k}}p^{C_k}(1-p)^{N_k-C_k}$$

B.3.2 Likelihood of open-population CR

We write down the likelihood for the individuals with first capture on sampling occasion k.

- As the individuals in the cell Z_{kK} have to arrive before the k-th sampling occasion and depart after the K-th, they can only be assigned to the cells $(n_{1K}^k, \ldots, n_{kK}^k)$. Hence, their capture histories can be assigned to the presence histories in $\binom{\sum_{l=1}^k n_{lK}^k}{Z_{kK}}$ possible ways.
- Similarly, the individuals in cell $Z_{k,K-1}$ have to arrive before the k-th sampling occasion and depart after the $K-1$-th and can only be assigned to the cells $(n_{1,K-1}^k, \ldots, n_{k,K-1}^k)$ and to the cells $(n_{1K}^k, \ldots, n_{kK}^k)$ not already assigned to the birds in the cells Z_{kK}. Hence, the number of combinations is $\binom{\left(\left[\sum_{l=1}^k n_{lK}^k\right] - Z_{kK}\right) + \sum_{l=1}^k n_{l,K-1}^k}{Z_{k,K-1}}$

Applying the same rationale for each cell, the likelihood can be written as

$$f_k(H|\mathbf{n}^k) = \binom{\sum_{l=1}^k n_{lK}^k}{Z_{kK}} \binom{\left(\left[\sum_{l=1}^k n_{lK}^k\right] - Z_{kK}\right) + \sum_{l=1}^k n_{l,K-1}^k}{Z_{k,K-1}} \dots \binom{\sum_{i=k+1}^K \left(\left[\sum_{l=1}^k n_{l,i}^k\right] - Z_{ki}\right) + \left(\sum_{l=1}^k n_{l,k}^k\right)}{Z_{k,k}} p^{C_k}(1-p)^{N_k - C_k} \quad \text{(B.3.1)}$$

B.4 Case studies

B.4.1 Spoonbills

Prior specification

We choose a uniform prior distribution for the resighting probabilities and a Gamma prior distribution with mean 50 and variance 4000 for the intensities ω_M and ω_U of the two population sizes. Inference is performed using an MCMC algorithm by sampling from the posterior distribution of the individual entry and exit intervals (t_1^i, t_2^i), the two matrices $\{n_{fl}^M\}$ and $\{n_{fl}^U\}$, as well as the probabilities ω_{ij} of the PT and the resighting probabilities.

B.4.2 Ring-Recovery

Prior specification

Instead of assuming a centering distribution on the tree, we assume uniform prior distributions directly on the parameters of the Beta distribution assigning the masses in each split of the PT. Moreover, we assume a Beta$(1,1)$ prior distribution for the recovery probability λ and we set the upper bound U on the age of the individuals to 18.

B.4.3 Newts

Prior specification

We center the HLPT on a bivariate distribution with double exponential marginal distributions, with parameters $\mu = (\mu_1, \mu_2)$ and $\lambda = (\lambda_1, \lambda_2)$. Prior information on the entry and exit distribution is elicited through the prior on $\eta = (\mu, \lambda)$. Prior knowledge suggests that a considerable number of individuals tend to enter the site between the beginning of March and the end of April and exit between the end of May and the end of July, and hence we choose hyperpriors for $\eta = (\mu, \lambda)$ such that 95% of the prior mass of the entry and exit density is in the aforementioned ranges.

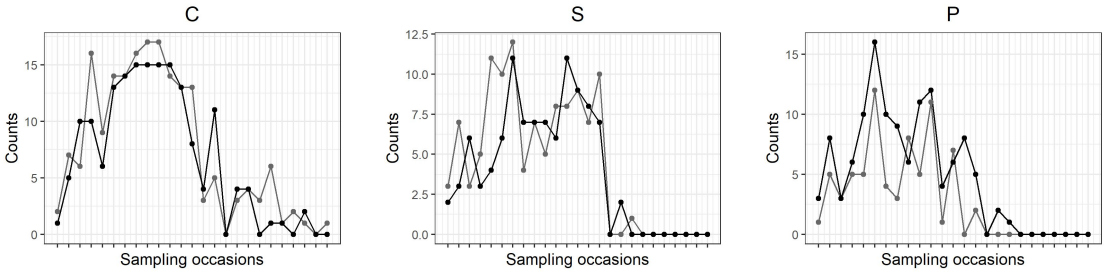


Fig. B.3 Number of newts detected on each sampling occasion shown for great crested newts, smooth and palmate newts, respectively. In each plot, the male are represented in grey and the female in black.

Figures

B.4.4 Moths

Prior specifications

We center the PT for the entry and exit distribution on two normal distributions with parameter $\mu = 2$ and 11, respectively, and variance $\lambda = 4$, where the unit is taken to be a month. We take a uniform prior distribution for the detection probability p and for the mean ω of the population size we take a Gamma$(40, .2)$, where the parameters have been chosen so that the mean of the distribution is 200 and the variance is 1000. Moreover, we take $\sigma_0 = \sigma = l = 1$ and the prior ρ on the probability of stopping equal to 0.1.

Figures

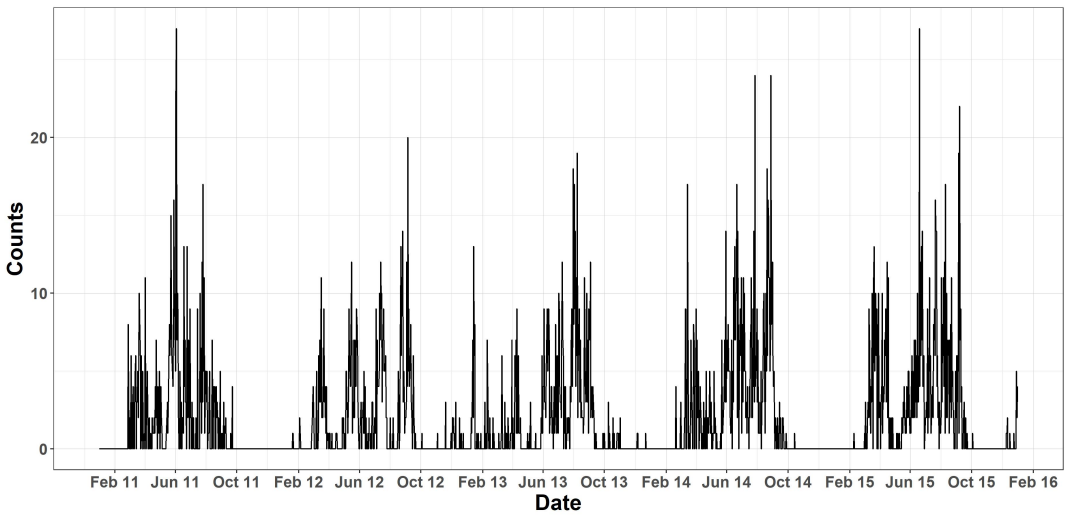


Fig. B.4 Daily counts of moths detected at the site.

Year	2011	2012	2013	2014	2015
Temperature	27.14	35.28	42.98	27.42	30.21

Table B.1 Average daily minimum temperature from the 1st to the 14th of January for each year, measured in Farenheit.

B.5 Posterior inference

B.5.1 Cormack-Jolly-Seber model

The parameters to be updated are the latent number of individuals $\{n_j^k\}$, the random variables V_j^k and the capture probability p.

To ease the computation, we introduce the probabilities ω_j^k of an individual with first capture at time t_k to have the time of possible last capture at time t_j. These probabilities can be computed straightforwardly from the V_j^k.

We remind that the capture histories of each individual are summarised in a matrix H, where H_{ik} is 1 if the i-th individual was captured on sampling occasion k and 0 otherwise.

- Update $\{n_j^k\}$

  We update each element of the vector with a MH. Reminding that the number of individual with first capture at time t_k, F_k, is fixed, the posterior distribution given the data H can be written as

$$p(n_j^k|H,\omega_j^k,p) \propto p(n_j^k|F_k,\omega_j^k)p(H|n_j^k,p) \propto$$
$$\left(\prod_j \text{Multinomial}(n_j^k|F_k,\omega_j^k))\right) f_k(H_k|\{n_j^k\})$$

  where $f_k(H_k|\{n_j^k\})$ is shown in Section 4.3.1.

- Update V_j^k

  As mentioned in Section 4.3.1, these are the r.v.s assigning the masses in a PT, and therefore they can be sampled from the posterior distribution of a PT.

- Update p

  This parameter is sampled straightforwardly from its full conditional.

B.5.2 Joint model for capture-recapture and count data

The captures are arranged in the matrices H^1 and H^2, respectively. The parameters to be updated are the entry and exit intervals of the already marked and captured individuals (t_1^i, t_2^i), the latent number of not resighted already marked individuals $\{n_{ij}^M\}$, the latent number of unmarked individuals $\{n_{ij}^U\}$, the detection and resightings probabilities p_R and p_C and the probabilities ω_{ij} sampled from the PT.

- Update (t_1^i, t_2^i)

  We update these variable with a random walk MH algorithm, where we for each individual we propose a new elemement $((t_1^i)^\star, (t_2^i))^\star)$ from $\{t_1^i - 1, t_1^i, t_1^i + 1\} \times \{t_2^i - 1, t_2^i, t_2^i + 1\}$.

  The posterior distribution can be written as

$$p(t_1^i, t_2^i | H_i^1, H_i^2, \{D_k\}, \{\omega_{ij}), p^R, p^C\}) \propto$$
$$p(H_i^1 | t_1^i, t_2^i | p_R)\ p(H_i^2 | t_1^i, t_2^i | p_C) \left(\prod_{k=1}^{K} p(C_k | N_k, p_R) \right)\ p(t_1^i, t_2^i | \{\omega_{ij})\})$$

  where $p(H_i^1 | t_1^i, t_2^i | p) = p^{\sum_{k=1}^K H_{ik}^1} (1-p)^{t_2^i - t_1^i + 1 - \sum_{k=1}^K H_{ik}^1}$.

- Update $\{n_{ij}^M\}$ and $\{n_{ij}^U\}$

  These variables are updated with a MH algorithm, where for each cell the number of elements we propose to add/remove is sampled from a geometric distribution.

  The posterior distribution can be written similarly to the (t_1^i, t_2^i).

- Update ω_{ij}.

  If we summarise all the individuals in a matrix n_{ij}, where cell (ij) has the number of individuals with first possible capture at time t_i and last possible capture at time t_j, the probabilities ω_{ij} can be sampled from the posterior of a PT given the variables n_{ij}.

- Update p_R and p_C

  These probabilities can be sampled directly from their full conditional.

B.5.3 Ring Recovery model

The parameters to be updated are the elements $\{n_{ij}^k\}$ in the entry and exit matrix, the probabilities $\{\omega_{ij}\}$ and the recovery probabilities p_R.

- Update $\{n_{ij}^k\}$

  These variables can be updated with a MH algorithm. The posterior can be written as

$$p(n_{ij}^k | H^A, H^J, \omega_{ij}, m_k^A, m_k^J, p_R) \propto p(n_{ij}^k | m_k, \omega_{ij})\ p(H_{kj}^J | n_{ij}^k, \lambda)$$

  where the first term is a multinomial distribution and the second term is a product of binomial distributions.

- Update $\{\omega_{ij}\}$

These parameter are sampled using the standard update of the PT. As we use a RPT across the different matrices $\mathbf{n}^k$, in order to use the standard update of the PT we summarise all the individuals in a single matrix n_{ij} by collapsing the set of matrices n_{ij}^k across the dimension k.

- Update μ and σ

  It is useful, in order to write down the posterior distribution of these parameters, to introduce the variables n_ε, representing the number of elements belonging to the set B_ε of the partition. The marginal likelihood of n_ε can be expressed as

$$p(\{n_\varepsilon\}|\{\alpha_\varepsilon\}) = \prod_\varepsilon \int \mathrm{Bin}(n_{\varepsilon 0}|n_\varepsilon, q_{\varepsilon 0})\mathrm{Beta}(q_{\varepsilon 0}|\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})dq_{\varepsilon 0}.$$

  Integrating out the probabilities $q_{\varepsilon 0}$ gives as a result a beta-binomial distribution, from which it follows that the marginal likelihood of the latent variable n_ε given the hyperparameter η is

$$p(\{n_\varepsilon\}|\eta) \propto \prod_\varepsilon \frac{B(\alpha_{\varepsilon 0} + n_{\varepsilon 0}, \alpha_{\varepsilon 1} + n_{\varepsilon 1})}{B(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})}.$$

  Given the marginal likelihood, the parameters can be updated using a random walk.

- Update λ

  The parameter can be sampled easily from the full conditional as the prior is a Beta distribution and the likelihood is a product of Binomial likelihoods.

B.5.4 Hierarchical model

The parameters to be updated are the latent number of individuals n_{ij}^s, the parameters β_ε^s assigning the probabilities of entering and exiting for each site, the hyperparameters η and the capture probabilities p_s.

To write the posterior of the βs and the hyperparameter η (or equivalently of the parameters μ as there is a deterministic relationship between the parameters μ and η) given the latent variable n_{ij}^s, we work with the variables $\tilde{n}_{\varepsilon_1,\ldots,\varepsilon_m}^s$, representing the number of elements in the s-th dataset belonging to the set $B_{\varepsilon_1\ldots\varepsilon_m}$ of the partition. The posterior can be written as:

$$p(\beta, \eta|\{\tilde{n}\}) \propto p(\eta)\prod_{s=1}^{S}\mathrm{Pois}(n_\Omega^s, \omega_s)\left(\prod_\varepsilon \frac{(e^{\beta_\varepsilon^s})^{n_{\varepsilon 0}^s}}{(1+e^{\beta_\varepsilon^s})^{n_\varepsilon^s}} p(\beta_\varepsilon^s|\eta)\right)$$

where n_Ω^s is the number of elements in the s-th dataset. Now as in Polson et al. [69] we add a variable $\omega_\varepsilon \sim \mathrm{PG}(\tilde{n}_\varepsilon, 0)$ and write the joint posterior distribution of all the variables as:

$$p(\beta, \eta|\{\tilde{n}\}, \omega_s) \propto p(\eta)\left(\prod_{s=1}^{S}\mathrm{Pois}(n_\Omega^s, \omega_s)\prod_\varepsilon e^{k_\varepsilon^s \beta_\varepsilon^s} e^{-\frac{\omega_\varepsilon^s(\beta_\varepsilon^s)^2}{2}} p(\omega_\varepsilon^s|\tilde{n}_\varepsilon^s, 0)p(\beta_\varepsilon^s|\mu_\varepsilon)\right)$$

where $k_\varepsilon^s = \tilde{n}_{\varepsilon 0}^s - \frac{\tilde{n}_\varepsilon^s}{2}$

The MCMC sampling scheme is performed as follows.

- Update β

  The posterior for (each) β given ω is a normal as:

$$p(\beta_\varepsilon|\omega_\varepsilon,\{\tilde{n}_\varepsilon\}) \propto e^{k_\varepsilon\beta_\varepsilon} e^{-\beta_\varepsilon^2\frac{\omega_\varepsilon}{2}} e^{\beta_\varepsilon\frac{\mu_\varepsilon}{\sigma_\varepsilon^2}} e^{-\frac{\beta_\varepsilon^2}{2\sigma_\varepsilon^2}} = e^{-\frac{\beta_\varepsilon^2}{2}(\omega_\varepsilon+\frac{1}{\sigma_\varepsilon^2})+\beta_\varepsilon(k_\varepsilon+\frac{\mu_\varepsilon}{\sigma_\varepsilon^2})}$$

  Thus:

$$\beta_\varepsilon|\omega_\varepsilon,\{\tilde{n}_\varepsilon\} \sim N(\left(k_\varepsilon+\frac{\mu_\varepsilon}{\sigma_\varepsilon^2}\right)\frac{1}{\omega_\varepsilon+\frac{1}{\sigma_\varepsilon^2}},\frac{1}{\omega_\varepsilon+\frac{1}{\sigma_\varepsilon^2}})$$

- Update ω

  As in Polson et al. [69], the posterior for (each) ω is:

$$p(\omega_\varepsilon|\beta_\varepsilon,\{\tilde{n}_\varepsilon\}) \propto e^{-\frac{\omega_{ij}\beta_\varepsilon^2}{2}} p(\omega_\varepsilon|\tilde{n}_\varepsilon,0)$$

  Thus:

$$\omega_\varepsilon|(\beta_\varepsilon,\{\tilde{n}_\varepsilon\}) \sim \mathrm{PG}(n_\varepsilon,\beta_\varepsilon)$$

- Update η

  In order to sample η more efficiently we integrate out β from the posterior (we drop the subscripts for simplicity):

$$p(\eta|\omega,\{\tilde{n}\}) = \int \prod_{s=1,\dots,S}\prod_{\varepsilon_1\dots\varepsilon_m} p(\eta|\omega,\beta,\{\tilde{n}\})p(\beta)d\beta = p(\eta)\prod_{s=1,\dots,S}\prod_{\varepsilon_1\dots\varepsilon_m}\int e^{k\beta}e^{-\frac{\omega\beta^2}{2}}\frac{e^{-\frac{(\beta-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}d\beta$$

  where:

$$\int e^{k\beta}e^{-\frac{\omega\beta^2}{2}}\frac{e^{-\frac{(\beta-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}d\beta = \frac{e^{-\frac{\mu}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\int e^{-\frac{\beta^2}{2}(\omega+\frac{1}{\sigma^2})+\beta(k+\frac{\mu}{\sigma^2})}d\beta \propto e^{-\frac{\mu}{2\sigma^2}}e^{\frac{(k+\frac{\mu}{\sigma^2})^2}{2(\omega+\frac{1}{\sigma^2})}}$$

  However we can still write the posterior on η given the variables β as:

$$p(\eta,|\beta) \propto p(\eta) \prod_{\varepsilon_1\ldots\varepsilon_m} \prod_{s=1}^{S} p(\beta^s_{\varepsilon_1\ldots\varepsilon_m}|\mu_{\varepsilon_1\ldots\varepsilon_m}) \propto$$

$$\underbrace{N(\mu|\mu_0,\Sigma_0)\text{Gamma}(\lambda_1|a_\lambda,b_\lambda)\text{Gamma}(\lambda_2|a_\lambda,b_\lambda)}_{p(\eta)} \prod_{\varepsilon_1\ldots\varepsilon_m} \prod_{s=1}^{S} N(\beta^s_{\varepsilon_1\ldots\varepsilon_m}|\mu_{\varepsilon_1\ldots\varepsilon_m},(\sigma_{\varepsilon_1\ldots\varepsilon_m})^2)$$

- Update $\{n^s_{ij}\}$

  These variables can be updated with a MH algorithm. The posterior can be written as

$$p(n^s_{ij}|C_{ks},\omega_{sij},p_s) \propto p(n^s_{ij}|\omega_{ij},\omega_s)\ p(C_{ks}|n^s_{ij},p_s)$$

  where the first term is a Poisson distribution and the second term is a product of binomial distributions.

- Update p_s

  The parameter can be sampled easily from the full conditional as the prior is a Beta distribution and the likelihood is a product of Binomial likelihoods.

B.5.5 Long count data model

The parameters to be updated are the coefficient $\beta^{i,a}$ and $\beta^{i,d}$ assigning the masses to the entry and exit distribution for 1-periods and 2-periods, the means μ^a and μ^d of the prior, the stopping variables S_{B_ε}, the number of latent individuals $n^{i_0}(i_1,j_1,i_2,j_2)$ and the detection probability p.

- Update β^i_j (entry coefficients)

  Each β^i_j can be updated condition on the other parameters using the Polya-Gamma scheme for multinomial outcomes. We define as n_i the sample size for β^i_j (which is equal to $\#(\Omega_i)$) and as a_{ij} the number of successes (equal to $\#(B^i_j)$).

  Following Polson et al. [69], the posterior of β (we drop the index i in the following) can be found as

$$p(\beta_j|a_j,n) \propto p(a_j|\beta_j,n)p(\beta_j) \propto \frac{(e^{\beta_j})^{a_j}}{(\sum_{k=1}^{K} e^{\beta_k})^n} e^{-\frac{(\beta_j-\beta_j^0)^2}{\sigma_0^2}} =$$

$$\frac{(e^{\beta_j})^{a_j}}{(\underbrace{\sum_{k=1,k\neq j}^{K} e^{\beta_k}}_{e^{c_j}} + e^{\beta_j})^n} e^{-\frac{(\beta_j-\beta_j^0)^2}{\sigma_0^2}} \propto \frac{(e^{\beta_j-c_j})^{a_j}}{(1+e^{\beta_j-c})^n} e^{-\frac{(\beta_j-\beta_j^0)^2}{\sigma_0^2}} =$$

$$2^{-n}e^{k(\beta_j-c_j)}e^{-\frac{(\beta_j-\mu)^2}{\sigma^2}}\int e^{-\omega_j\frac{(\beta_j-c_j)^2}{2}}p(\omega_j)d\omega_j$$

where $\omega_j \sim \mathrm{PG}(n,0)$, $k = a_j - \frac{n}{2}$ and $c_j = \log(1+\sum_{k\neq j}e^{\beta_k})$, from which it follows that a Gibbs sampler can be constructed as

$\omega_j|\beta_j \sim \mathrm{PG}(n,\beta_j - c_j)$
$\beta_j|\omega_j \sim N((\frac{\sigma^2}{1+\omega_j\sigma^2})^{-1}(k+\omega_j c_j + \frac{\beta_j^0}{\sigma_0^2}), \frac{\sigma_0^2}{1+\omega_j\sigma_0^2})$

- Update $\beta_{j,k}^i$ (exit coefficient)

  If n_{kj} is the number of elements in $B_{k,j}^i$, the likelihood of $\beta_{k,j}^i$ comes from assigning n_{kj} elements to the set $B_{k,j}^i$ out of the a_k elements in set B_k^i.

  The posterior is (dropping the subscript i and k)

$$e^{-\frac{(\beta_j-\beta_j^0)^2}{\sigma_0^2}}\prod_{k=1}^{j}p(\beta_j|n_{ij},a_i) \propto e^{-\frac{(\beta_j-\beta_j^0)^2}{\sigma_0^2}}\prod_{k=1}^{j}\frac{(e^{\beta_j})^{n_{kj}}}{(\sum_{l=k}^{13}e^{\beta_l})^{a_k}} \propto$$
$$e^{-\frac{(\beta_j-\beta_j^0)^2}{\sigma_0^2}}\prod_{k=1}^{j}2^{-a_k}e^{k_{kj}(\beta_j-c_{jk})}\int e^{-\omega\frac{(\beta_j-c_{jk})^2}{2}}p(\omega_k)d\omega_k$$

  where $\omega_k \sim \mathrm{PG}(a_k,0)$, $k_{kj} = n_{kj} - \frac{a_k}{2}$ and $c_{ji} = \log(\sum_{k\geq i,k\neq j}^{K}e^{\beta_k})$. The posterior of β_j can thus be expressed as

$$e^{-\frac{(\beta_j-\beta_j^0)^2}{\sigma_0^2}}\prod_{k=1}^{j}e^{k_{kj}(\beta_j-c_{jk})}e^{-\omega_k\frac{(\beta_j-c_{jk})^2}{2}}$$

  from which it follows that a sample from the posterior can be performed by sampling

$$\begin{cases}\omega_k \sim PG(a_k,\beta_j-c_{jk}) \qquad k=1,\dots,j\\ \beta_j|\omega_k \sim N((\frac{\sigma_0^2}{1+\sum_k\omega_k\sigma_0^2})^{-1}(\sum_k k_{kj}+\sum_k(\omega_k c_{jk})+\frac{\beta_j^0}{\sigma_0^2}), \frac{\sigma_0^2}{1+\sum_k\omega_k\sigma_0^2})\end{cases}$$

- Update μ

  These parameters can be sampled straightforwardly from the full conditional, as the prior is $(\mu_1,\dots,\mu_{K_1}) \sim \mathrm{GP}((\mu_1^0,\dots,\mu_{K_1}^0),\sigma_0^2 k_l)$ and the likelihood is $\beta_i^j \sim N(\mu_i,\sigma^2)$.

- Update ρ

  If G is the joint entry and exit distribution and S_{B_ε} is the indicator variable which is 1 if B_ε is not further partitioned and 0 otherwise, we have

$$G(x|x\in B_\varepsilon) = S_{B_\varepsilon}U_{B_\varepsilon}(x) + (1-S_{B_\varepsilon})\sum_{i=1}^{K_\varepsilon}G(x|x\in B_{\varepsilon,\varepsilon_i})1(x\in B_{\varepsilon,\varepsilon_i})$$

To sample each S_{B_ε}, we first sample all the S and β for the "children" sets of B and then sample S_{B_ε} from its posterior, which can be written as

$$p(S_{B_\varepsilon}=1|\beta,x)=p(S_{B_\varepsilon}\frac{p(x|\beta,S_{B_\varepsilon}=1)}{p(x|\beta)}=1)=\rho\frac{\mathrm{U}_{B_i^1}(x)}{\rho U_{B_\varepsilon}(x)+(1-\rho)\sum_{i=1}^{K_\varepsilon}G(x|x\in B_{\varepsilon,\varepsilon_i})1(x\in B_{\varepsilon,\varepsilon_i})}$$

- Update $n_{d,w,m,y}$

  We update this parameter with a MH step, where the posterior distribution is

$$p(n^{i_0}(i_1,j_1,i_2,j_2)|\omega^{i_0}(i_1,j_1,i_2,j_2),\omega,\{C^{i_0}_{i_1,i_2}\},p)\propto$$
$$p(n^{i_0}(i_1,j_1,i_2,j_2)|\omega,\omega^{i_0}(i_1,j_1,i_2,j_2))\ p(\{C^{i_0}_{i_1,i_2}\}|\{N^{i_0}_{i_1,i_2}\},p)$$

- Update p

  This parameter is sampled straightforwardly from its full conditional

Appendix C

Supplementary material for A vector of softcore processes for modelling interactions between species using capture-recapture data.

C.1 Simulation algorithm for interaction point processes

Theorem 2. *Given $X_m = \{x_1, \ldots, x_n\}$, generate $X_{m+1}|X_m$ by applying with equal probability one of the following steps:*

1. *Move: Draw $i \sim$ Uniform$(\{1, \ldots, n\})$, propose $\xi \sim q_i(X_m, \cdot)$, where q_i is the proposal distribution for the i-th point and set*

$$X_{m+1} = \begin{cases} \{x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n\} & \text{if } U([0,1]) \leq r_i(X_m, \xi) \\ X_m & \text{otherwise} \end{cases}$$

   where $r_i(X, \xi)$ is the Hastings ratio

$$r_i(X, \xi) = \frac{f((X \setminus x_i) \cup \xi) q_i(\{x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n\}, x_i)}{f(X) q_i(X, \xi)}$$

2. *Add: Propose a new point $\xi \sim q_b(X_m, \cdot)$ and set*

$$X_{m+1} = \begin{cases} X_m \cup \xi & \text{if } U([0,1]) \leq r_b(X_m, \xi) \\ X_m & \text{otherwise} \end{cases}$$

   where $r_i(X, \xi)$ is the Hastings ratio

$$r_b(X,\xi) = \frac{f(X \cup \xi) q_d(X_m \cup \xi, \xi)}{f(X) q_b(X_m, \xi)}$$

3. *Delete: Sample a point x_i from $q_d(X_m, \cdot)$ and set*

$$X_{m+1} = \begin{cases} X_m \setminus x_i & \text{if } U([0,1]) \leq r_d(X_m, x_i) \\ X_m & \text{otherwise} \end{cases}$$

where $r_d(X,\xi)$ is the Hastings ratio

$$r_d(X,\xi) = \frac{f(X \setminus \xi) q_b(X_m \setminus \xi, \xi)}{f(X) q_d(X_m, \xi)}$$

C.2 MCMC Algorithm

Update s_{il}

Similarly to the previous Section, we update the pair of point processes $(\mathbf{s}^1, \mathbf{s}^2) = (s_1^1, \ldots, s_{N_1}^1), (s_1^2, \ldots, s_{N_2}^2)$ by using with equal probability a move/add/delete step for each of the two processes.

The posterior distribution of $(\mathbf{s}^1, \mathbf{s}^2)$ can be written as

$$p(\mathbf{s}^1, \mathbf{s}^2 | \beta_1, \beta_2, \gamma_1, \gamma_2, \gamma_{12}, H^1, H^2, p_0, \alpha) \propto p(\mathbf{s}^1, \mathbf{s}^2 | \beta_1, \beta_2, \gamma_1, \gamma_2, \gamma_{12})\ p(H^1 | \mathbf{s}^1, p_0, \alpha) / p(H^2 | \mathbf{s}^2, p_0, \alpha)$$

where

$$p(\mathbf{s}^1, \mathbf{s}^2 | \beta_1, \beta_2, \gamma_1, \gamma_2, \gamma_{12}) \propto$$
$$\beta_1^{N_1} \beta_2^{N_2} \prod_{i<j} 1 - \exp\left(-\frac{(s_i^1 - s_j^1)^2}{\gamma_1}\right) \prod_{i<j} 1 - \exp\left(-\frac{(s_i^2 - s_j^2)^2}{\gamma_2}\right) \prod_{i,j} 1 - \exp\left(-\frac{(s_i^1 - s_j^2)^2}{\gamma_{12}}\right) \quad \text{(C.2.1)}$$

and $p(H^l | s_i^l)$ is the probability of the captures locations.

Update γ_j and β_j

These variables are updated using the MCMH algorithm described in Section 5.2.2.

Update p_0 and α

We update these variables using a simple Metropolis-Hastings update.