Maynooth University
National University of Ireland Maynooth

Novel Developments in Bayesian Modelling Applied to Estimating Abundance in Animal Communities

A dissertation submitted for the degree of
Doctor of Philosophy

By:

Niamh Mimnagh

Under the supervision of:

Dr. Rafael A. Moral

Hamilton Institute
Maynooth University

March 2023

To my family, for their unending love and support.

Declaration

I hereby declare that I have produced this manuscript without the prohibited assistance of any third parties and without making use of aids other than those specified.

The thesis work was conducted from September 2019 to March 2023 under the supervision of Dr. Rafael A. Moral in the Hamilton Institute, Maynooth University.

Niamh Mimnagh.

Maynooth, Ireland,

March 2023.

Sponsor

This work was supported by a Science Foundation Ireland grant number 18/CRT/6049.

Collaborations

Rafael A. Moral: As my supervisor, Dr. Moral (Maynooth University) supervised and collaborated on the work of all chapters.

Andrew C. Parnell: Prof. Parnell (Maynooth University) collaborated on the work of Chapter 3 and Chapter 4 by reading and providing valuable observations and guidance.

Estevão Prado: Dr. Prado (Lancaster University) collaborated on the work of Chapter 3 by providing advice at the model formulation stage and by reading and providing suggestions on the written chapter, and Chapter 4 by collaborating on the writing of Section 4.2, and by reading and providing insights on the remainder of the chapter.

Iuri E.P. Ferreira: Dr. Ferreira (Federal University of São Carlos) collaborated on the work of Chapter 5 by providing suggestions as to the model formulation, and by reading and providing insights on the finished chapter.

Luciano Verdade: Dr. Verdade (University of São Paulo) collaborated on the work of Chapter 5 by reading and providing insights on the finished chapter.

Publications

The chapters contained in this thesis have been either published or submitted to peer-reviewed journals. Chapter 3 has been published in the journal *Environmental and Ecological Statistics* and Chapter 4 has been accepted for publication in *Modelling Insect Populations in Agricultural Landscapes* and is due to appear. Chapter 5 has been submitted and is currently under peer review.

Peer-reviewed journal article:

- Mimnagh, N., Parnell, A., Prado, E., Moral, R.A. Bayesian multi-species n-mixture models for unmarked animal communities. *Environmental and Ecological Statistics*, 29, 755–778 (2022) https://doi.org/10.1007/s10651-022-00542-7

- Mimnagh, N., Parnell, A., and Prado, E. Bayesian N-mixture models applied to estimating insect abundance. *Modelling Insect Populations in Agricultural Landscapes*, Springer (2023).

Submitted articles (under review):

- Mimnagh, N., Ferreira, I.E.P., Verdade, L.M., and Moral, R.A. Counting animals we can't see: a triple count model for scarce vestige data.

Contents

Abstract

In order to understand evolutionary-ecological processes and make decisions concerning wildlife management (i.e., conservation and monitoring, according to Caughley (1994)), the ability to estimate abundances of wild animal species can prove imperative (Verdade et al., 2014). However, methods of data collection that involve direct interaction with wild animals can be invasive and pose risks to both wildlife and humans (Verdade et al., 2013).

In this thesis, we propose methodologies that may be used to estimate animal abundances using different types of data, with an emphasis on data whose collection is relatively low-effort, cost-effective and poses the least risk of danger to the animal and observer. The ability to use these data to estimate abundance may allow for the establishment of large-scale wildlife monitoring programs.

First we present a multivariate extension to the N-mixture model proposed by Royle (2004). This extension allows for the estimation of abundances for multiple species simultaneously, while also estimating the correlation between species abundances. This model is further extended to allow for data collected over long time periods through the addition of a first-order autoregressive term on the abundance. This model is then extended further to allow for the use of zero-inflated data by considering a hurdle-Poisson distribution for the latent abundances.

We then provide an overview of various N-mixture models, aimed at introducing practitioners unfamiliar with statistics to this methodology. We demonstrate a Bayesian implementation of some of these models to estimate foraging bee abundance, with R code provided to allow model implementation by any interested practitioners.

Finally we examine a scenario in which data is not composed of observations of individuals, but rather observations of animal vestiges (i.e., traces that an animal leaves behind as it moves through the environment). Here we present a novel modelling framework, the triple Poisson model, that allows for the estimation of animal abundance using vestige data, even when only very scarce data is available.

Acknowledgements

I would like to give thanks to the following people and organisations that have helped me to undertake this research.

To my supervisor Dr. Rafael de Andrade Moral, for providing his expertise, his enthusiasm for the project, his constant support and his patience.

To Science Foundation Ireland, who financed my research and without whom the undertaking of this project would not have been possible.

To the Centre for Research Training in Foundations of Data Science team, for providing me the opportunity to carry out this research, and to learn and grow in such a diverse and enriching academic environment.

To the faculty at Maynooth University, for providing me with the resources to pursue this graduate research in the Hamilton Institute.

To my fellow cohort members and colleagues in the Hamilton Department, for their kind help and moral support throughout this process.

List of Figures

List of Tables

CHAPTER 1

Introduction

In this chapter, we discuss the motivations behind the work presented in this thesis and provide an overview of the material contained within the following chapters.

1.1 Motivation

Monitoring animal populations is of vital importance for multiple sectors, and is necessary to carry out any plans that involve wildlife conservation, control of wildlife population sizes, or the optimisation of yield for species of economic importance. It has, as a result, been a primary focus for wildlife ecologists for over 20 years (Caughley, 1994).

Abundance in animal communities is of great interest in ecology, particularly in the areas of conservation and wildlife management (Witmer, 2005; Nichols and MacKenzie, 2004). The question of estimating abundance is an important one due in part to the rapid decline in species abundance and diversity occurring globally (Novacek and Cleland, 2001). The ability to estimate abundance facilitates the establishment of large-scale animal monitoring programmes, which impacts not only those working in statistics and ecology, but also conservationists, policy makers, and wider society, as the early warning provided by monitoring programmes will allow for early intervention in species decline. The scenario where animal

populations experience unnoticed decline due to a lack of monitoring programmes is one that would affect society and the environment on a global scale. In addition to conservation, the establishment of wildlife monitoring programmes is of economic importance as these programmes allow for the optimisation of yield of species that are deemed to possess economic value. These can include wild animal populations that provide human populations with products such as meat, clothing, and medicines. Wildlife monitoring programmes are also important due to the ability they provide us with to determine when population sizes of pest species are rapidly increasing, with implications for resulting economic and environmental damage (Witmer, 2007). The ability to estimate animal abundance is central to the development of these long-term wildlife monitoring programmes.

Throughout the remainder of this thesis we make a distinction between direct and indirect data. Direct data refers to data which has been collected or obtained through direct interactions with animal individuals, (e.g., data that is obtained using methods that involve capturing an animal), and indirect data refers to data which is obtained while avoiding direct interactions with an individual (i.e., counts of observed individuals).

Wildlife abundance has typically been estimated using direct methods of observation, which involve close interactions between human and animal, and often involve the capturing of the animal in order to tag or study it. The implementation of these methods can prove to be expensive, time-consuming, and can pose risks of distress or danger to both the animal and observer.

In this thesis we place an emphasis on the use of indirect data to estimate animal abundance, due to certain advantages this type of data possesses over direct data. Data collected indirectly is an attractive option for estimating abundance due to the fact that it does not involve any direct interactions between human and animal, which results in it being relatively affordable to collect when compared with direct methods. There is also a reduced risk of harm to both animals and humans inherent in the collection of these types of data (Verdade et al., 2013). However, this type of data is often imperfect (i.e. the recorded information is usually imperfect in the sense that it does not represent the total abundance). In

order to utilise these types of data to estimate animal population sizes, we must use modelling frameworks that take into account the characteristics of this data. Due to these characteristics, traditional modelling techniques, such as generalised linear models (McCullagh and Nelder, 1989), cannot be applied directly to the data, as they do not allow for the estimation of detection probabilities and so do not accommodate aspects such as imperfect detection.

This thesis presents modelling frameworks that can be used to estimate animal population sizes using different forms of indirect data, namely data composed of counts of animal sightings, and data composed of counts of animal vestiges (e.g. scats, footprints, fur, feathers, among others). In each case this data is collected by carrying out surveys along transect lengths, during which surveyors record every individual animal or vestige observed. Vestige data in particular is currently under-utilised in estimating abundance due to a lack of research into statistical methodologies with the ability to utilise it.

The objective of the research presented in this thesis is twofold. The first objective is the development of novel modelling frameworks that are fit for use by those working in the wildlife monitoring space, including statisticians, population–ecologists and conservation–ecologists. The second objective is to introduce ecologists – and entomologists in particular in Chapter 4 – who may be unfamiliar with statistics, to these types of modelling methodologies. We aim to provide practitioners with the information needed to carry out their own data analysis. We do this by implementing these models in a Bayesian framework, offering comparisons of different methodologies and demonstrating their use using real–world data. It is the overarching aim of this thesis that through the completion of these objectives we might contribute to facilitating the establishment of more large-scale animal monitoring programmes.

1.2 Thesis Outline

The remainder of this thesis is organised as follows. In Chapter 2 we provide an introduction to the case studies examined as part of this thesis. We discuss the characteristics of these datasets, as well as any limitations or difficulties encoun-

tered when working with them. We also discuss the source of each dataset, and provide details for interested practitioners who may wish to fully reproduce the analysis in this thesis using these datasets.

The subsequent chapters are presented in the format of journal articles and were modified where possible to avoid repetition while retaining the consistency and clarity of the thesis.

In Chapter 3 we propose a modelling framework constructed as an extension to the N-mixture model by Royle (2004) that allows the user to estimate abundances for multiple species simultaneously though the inclusion of a species-level random effect. The use of this random effect further provides us with the ability to estimate inter-species correlations, which may allow us to begin to make inferences as to the relationships that these species have with one another. We then propose a further extension to this model that allow us to examine zero-inflated data through the use of a hurdle-Poisson distribution on the abundance. Using an extensive simulation study, we show that this modelling framework provides consistently accurate estimates of abundance in a range of scenarios. We also explore the performance of this modelling framework using a real-world dataset collected as part of the North American Breeding Bird Survey, examining all of the models described above and choosing the model that provides the best fit using BIC values.

In Chapter 4 we provide information on the development of the original N-mixture model for closed populations by Royle (2004). We examine the assumptions made for this framework, along with some limitations associated with it, and issues that can arise when using the N-mixture model, such as those described by Dennis et al. (2015). In Section 4.3, we provide an overview of many important extensions that have been made for this model since its development in 2004. This includes N-mixture models that have been extended to allow for the examination of multiple species simultaneously, models that do not require the population to be closed, and can thus support the occurrence of births, deaths and migration in a population, and models that contain a number of zero counts greater than that expected by the Poisson distributed abundance, which is a common occurrence when working

with count data composed of animal sightings. We assess the advantages and limitations associated with each of these model extensions. In Section 4.4 we provide an implementation of the original N-mixture model by Royle (2004) and the multispecies N-mixture model developed in Chapter 3 to estimate the abundance of foraging bee populations.

In Chapter 5 we propose a novel modelling framework which we refer to as a triple Poisson model. The aim of this model is to allow the use of vestige data to estimate abundance. This is achieved by assuming that the number of vestiges observed, the number of groups of animals in the area and the species abundance may each be estimated using a Poisson distribution. We use simulation studies to show that the predictive performance of the proposed model can rival that of the distance sampling model (Thomas et al., 2006), particularly in scenarios when data is very scarce. We analyse several case studies described in Chapter 2 including data collected on collared peccary which consists of only two counts, data collected on red foxes by Cavallini (1994) and data collected on sika deer by Marques et al. (2001).

All proposed methods in this thesis were implemented using the `R` (R Core Team, 2022) software and are accessible on the author's Github [1] via three public repositories. The repositories `MNM`, `insect_populations_ch11`, and `triple_poisson` are related to Chapters 3, 4, and 5, respectively. Within these repositories we have made available `R` scripts required to produce all analyses and plots presented in this thesis. Additionally, all datasets are publicly available, and their source is provided in Chapter 2.

Finally, in Chapter 6, we present conclusions and final remarks, while indicating topics for future research.

[1] https://github.com/niamhmimnagh

CHAPTER 2

Case Studies

In this chapter, we discuss the various case studies utilised as part of this thesis, provide details as to their notable features, and detail where these datasets may be located, for practitioners interested in reproducing the work detailed within this thesis.

In this thesis we examine data that can be used to estimate animal abundance. The data we examine is collected from several species in diverse locations and under various circumstances. In this Chapter we will present the case studies to be examined in subsequent chapters, examine some of their characteristics and present any obstacles or difficulties that we encountered while using them.

2.1 North American Breeding Bird Survey

In Chapter 3 we consider data collected as part of the North American Breeding Bird Survey (Pardieck et al., 2020). This data is available at a public repository[2]. The North American Breeding Bird Survey is a large scale monitoring programme focused on collecting data on breeding birds in the United States and Canada. It has been conducted annually since 1966, and now provides data on more than 400 bird species at approximately $3,700$ routes (Figure 2.1). Each of these routes

[2] https://www.sciencebase.gov/catalog/item/625f151ed34e85fa62b7f926

is approximately 24.5 miles long and is composed of 50 stops, approximately 0.5 miles apart. At each stop, every bird seen or heard within a 0.25-mile radius is recorded.

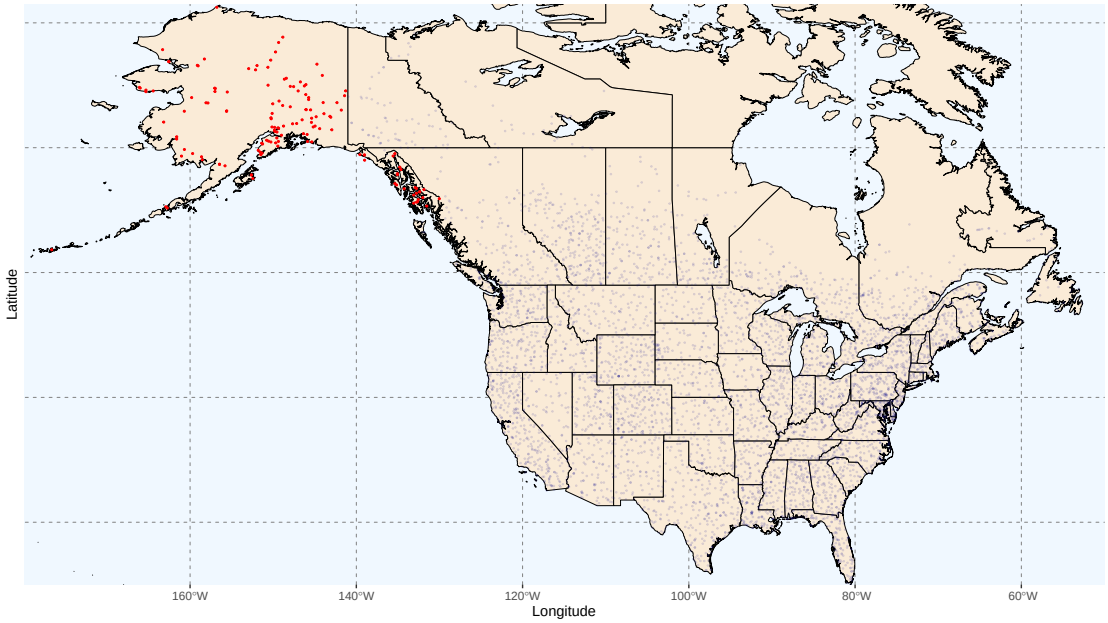


Figure 2.1: The location within the United States and Canada of all North American Breeding Bird Survey sites, with the 94 sites in Alaska that were selected for analysis highlighted in red.

Due to the size of this dataset, we needed to select only certain data for use with our modelling framework. We decided to examine data collected in Alaska in the 10–year period 2010–2019. The result was a dataset that contained 94 routes, each with 50 sampling locations, for a total of $4,700$ observations per bird species. From here the next step was to select bird species to examine. The bald eagle (*Haliaeetus leucocephalus*) was chosen for examination due to the large bald eagle populations present in Alaska (between $8,000$ and $30,000$ birds which accounts for roughly half of the global bald eagle population (Hodges, 2011; Hansen, 1987; King et al., 1972)). Several other species were then selected for examination; these species included waterbirds such as geese, swans and snipes which were chosen for their

relationships with bald eagles, as bald eagles are known to prey on waterbirds such as ducks, geese and grebes when fish are in short supply (Dunstan and Harper, 1975; Todd et al., 1982; McEwan and Hirth, 1980). Additionally, a selection of species with inland habitats, such as thrushes and swallows, were examined. In total, 10 species were selected for analysis, of the 233 total species present in Alaska within the 10-year period. The resulting dataset contained observations collected at 94 sites where each site divided into 50 sections, for 10 species over 10 years. This gave us a total of 470, 000 observations.

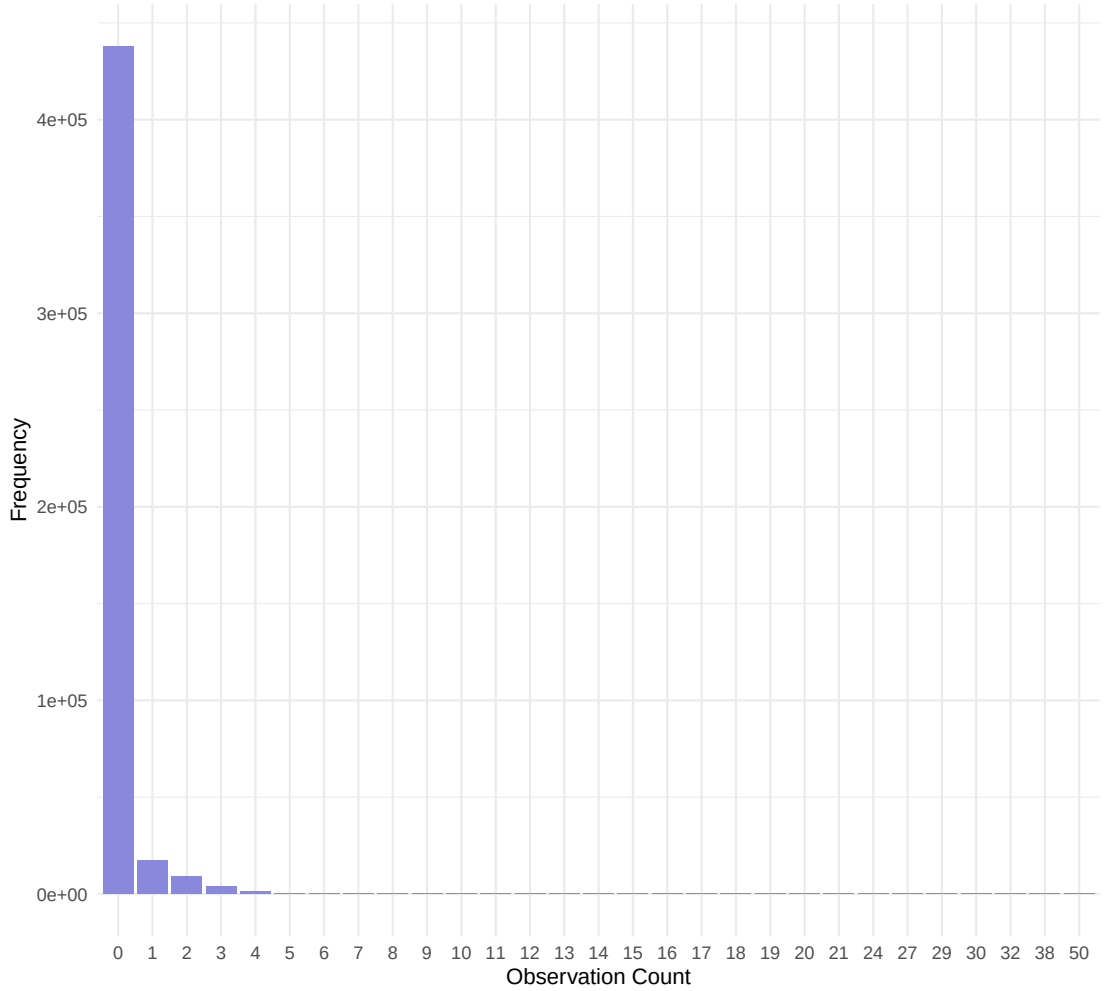


Figure 2.2: The frequency at which counts of observed birds occur at the 94 selected sites between 2010 and 2019 in the North American Breeding Bird Survey.

Initial examination of this dataset revealed that 93.2% of the observations examined for Alaska (438, 040 of a total of 470, 000 observations, Figure 2.2) consisted of zero counts. These counts are decomposed into species-level counts in Figure 2.3. This figure reveals that while zeros dominate the counts for each species examined, no species experienced only zero counts (i.e., no species examined was extinct, or so seldom observed that it did not appear in the dataset at all).

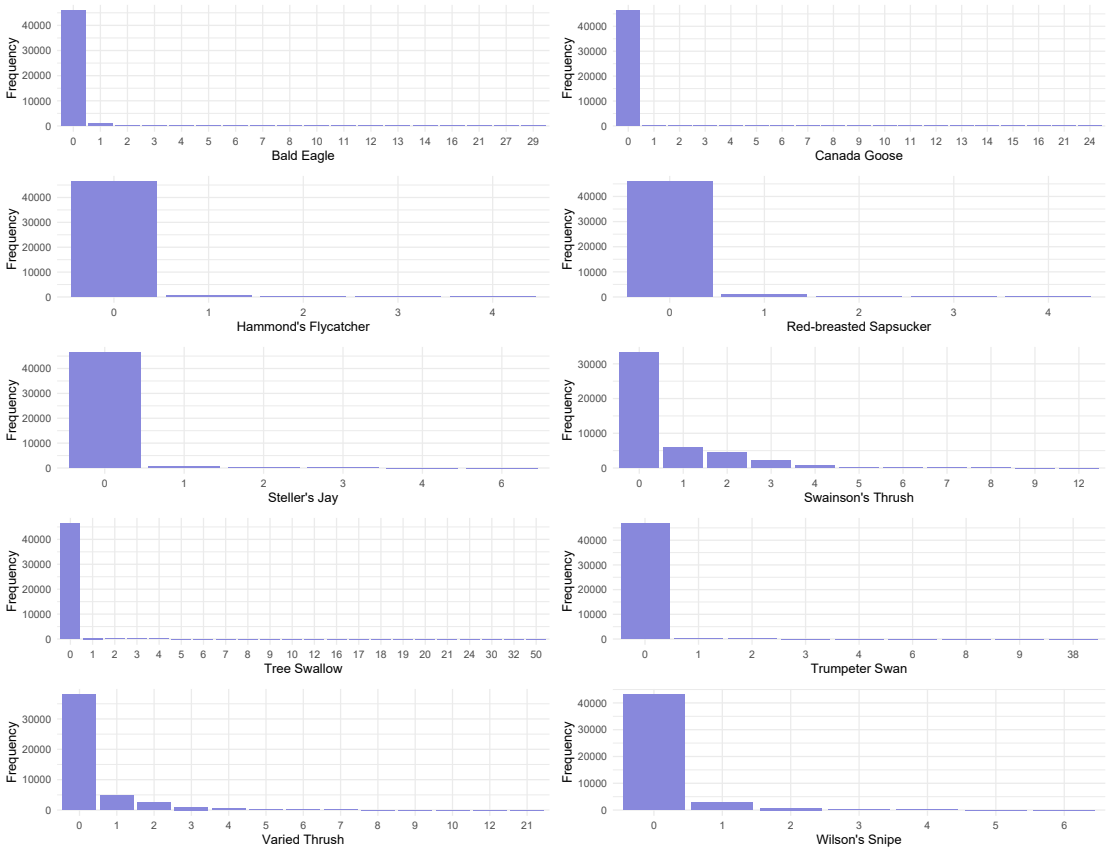


Figure 2.3: The frequency at which counts individual species of observed birds occur at the 94 selected sites between 2010 and 2019 in the North American Breeding Bird Survey.

Zero counts were then temporarily removed to examine frequencies of larger counts (Figure 2.4). This revealed that the majority of bird sightings contained less than four birds, with counts of greater than seven individuals occurring only 170 times, and accounting for only 0.036% of observations.

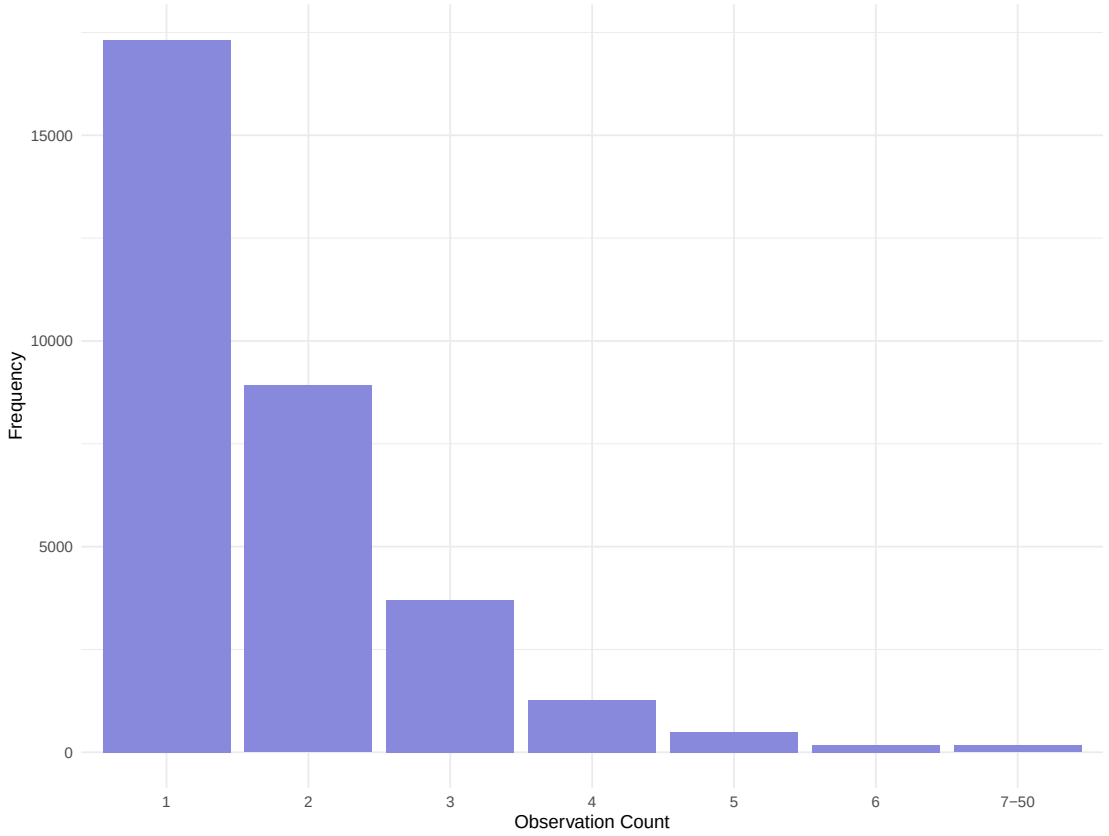


Figure 2.4: The frequency at which counts of observed birds occur at the 94 selected sites between 2010 and 2019 in the North American Breeding Bird Survey, after zero counts have been removed.

This large number of zero-counts would suggest that a Poisson distribution assumption for the abundance would provide a poor fit, and that an alternative model that accounts for large proportions of zeros should be considered. Furthermore, this data was collected over a decade, which suggests that a model that assumes populations are open would provide a better fit than a model that assumes closed populations. A sample of the North American Breeding Bird survey dataset is provided in Table 2.1.

Country	State	Route	Year	Species	Stop 1	Stop 2	...	Stop 50
USA	Alaska	23	2003	Bald Eagle	37	1	...	0
USA	Alaska	24	2005	Bald Eagle	23	3	...	0
USA	Alaska	125	2017	Bald Eagle	3	0	...	2

Table 2.1: A sample of the North American Breeding Bird survey dataset, with the country, state, route, year and species observed at each stop (stops labelled 1–50).

2.2 BeeWalk Survey

In Chapter 4 we examine the N-mixture model proposed by Royle (2004) and revisit the model that we present in Chapter 3, and demonstrate their implementation on a dataset collected as part of the BeeWalk Survey Scheme (Comont et al., 2021).

The BeeWalk Survey Scheme is a survey established in the UK in 2008 by the Bumblebee Conservation Trust, which involves transects being surveyed by volunteers across the UK on a monthly basis. Data up until the end of 2019 is available at a public repository[3]. By the end of the 2019 data-collection period, observations of approximately 70 bee species had been recorded as part of this survey scheme at over 1300 sites in the UK. In Chapter 4 we examine counts of observations for several species of bees collected at 60 sites in 2016 and 2019. To ensure that we are comparing data collected from the same seasonal cycles, we examine data collected in June of both years. The 60 sites examined in Chapter 4 are presented in Figure 2.5.

[3] https://figshare.com/authors/Richard_Comont/97396

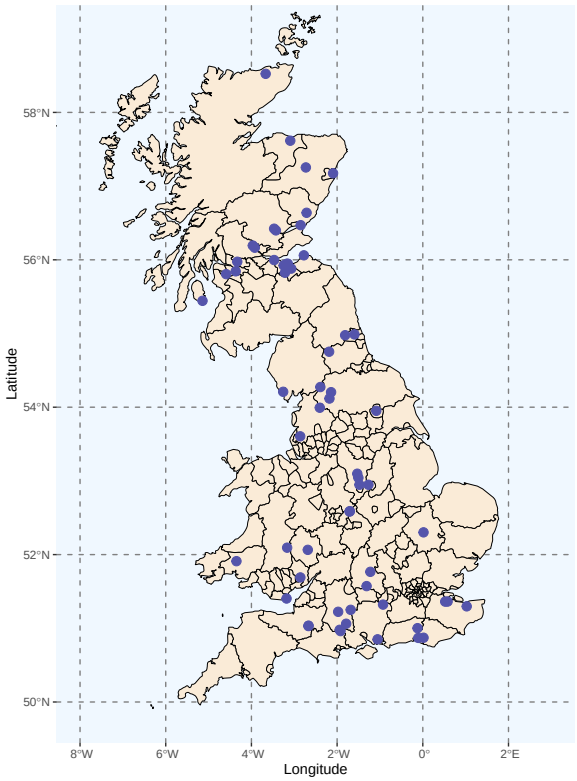


Figure 2.5: A selection of sites in the UK at which bee observations were recorded.

Prior to estimating abundance using an N-mixture model, we confirm that we do not expect to encounter identifiability issues that might lead to infinite estimates of abundance. We do this using the covariance diagnostic proposed by Dennis et al. (2015)(detailed in Chapter 4) . Beewalk survey data is also available for the years 2017 and 2018, but a negative value for the covariance diagnostic revealed that we might expect issues with parameter estimates using the data collected for these two years. For this reason, we decided to examine only data collected in 2016 and 2019.

Figure 2.6 displays the total frequency of counts for observed bees across the 60 selected sites in the BeeWalk survey in 2016 and 2019. From this bar plot it is obvious that zero counts (survey responses which record that no individuals of a particular bee species were observed) compose a large portion of this data. Further examination reveals that 54.6% of the entire dataset is composed of these

zero values, with the majority of the dataset (96.25%) composed of counts of less than 20 bee individuals.

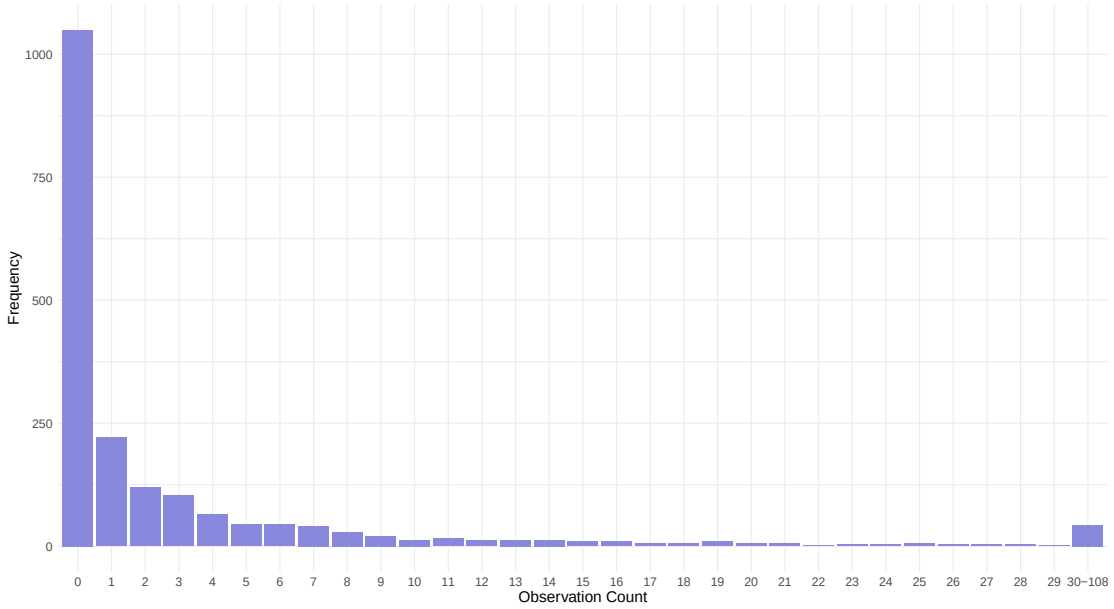


Figure 2.6: The frequency at which counts of observed bees occur at the 60 selected sites in 2016 and 2019 in the BeeWalk survey data.

In Figure 2.7 we decompose the BeeWalk Survey counts into species-level counts. This reveals that while zero counts occur very frequently for all species, no species is composed entirely of zero counts.

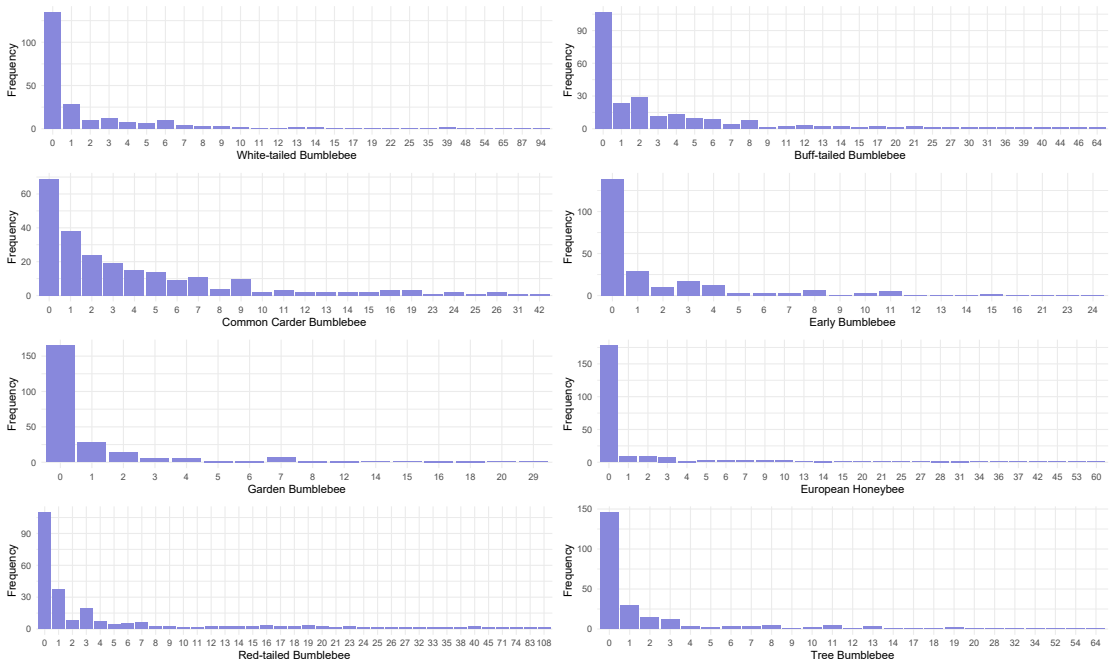


Figure 2.7: The frequency at which counts of individual species of observed bees occur at the 60 selected sites in 2016 and 2019 in the BeeWalk survey data.

An issue that we must take into consideration when using data composed of bee sightings is that at any one time, only a small portion (approximately 30%) of the bee colony's population is currently out foraging. For this reason, any estimate for "abundance" obtained using counts of bees observed is not representative of the entire population. We instead consider these estimates as estimates of the size of the population who are currently out foraging, knowing that there are many more bees in the area than these estimates would suggest. A sample of the BeeWalk survey data is provided in Table 2.2.

Site	Grid Reference	Length (m)	Weather	Wind Speed	Temperature (°C)	Species
Holyrood park: Hunter's Bog S2	NT273733	558	Sunny	Slight smoke drift	14	Common Carder Bumblebee
Holyrood park: Hunter's Bog S1	NT273733	232	Sunny	Slight smoke drift	14	Common Carder Bumblebee
Holyrood park: Hunter's Bog S3	NT273733	383	Sunny/ Cloudy	Wind felt on face, leaves rustle	12	Common Carder Bumblebee
Holyrood park: Hunter's Bog S4	NT273733	283	Sunny/ Cloudy	Smoke rises vertically	17	White-tailed Bumblebee

Table 2.2: A sample of the BeeWalk survey dataset with the site, the site grid reference, the length of the transect in metres, the weather, the wind speed, the temperature in degrees Celsius, and the species observed.

2.3 Collared Peccary Survey

In Chapter 5, several case studies are examined, and are used to demonstrate various aspects of our proposed modelling framework. This modelling framework utilises data composed of counts of animal vestiges, where a vestige may be any trace of an animal in the environment such as fur or tracks. However, for our purposes, all datasets examined here contain counts of observed animal scats.

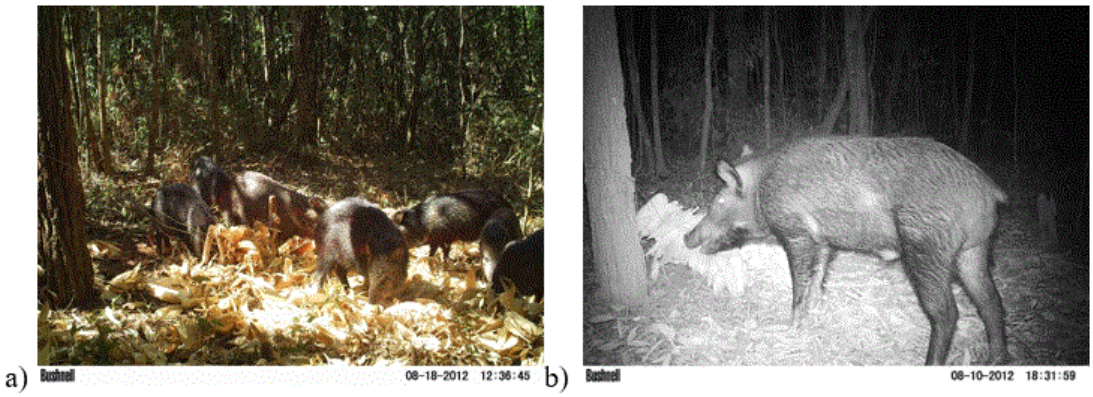


Figure 2.8: Photos of (a) a group of collared peccaries foraging during the day and (b) an individual collared peccary at night, captured at camera traps as part of the collard peccary survey. Reproduced with permission from Prof. Luciano Martins Verdade, Principal Investigator of the BIOTA programme, funded by the Brazilian government.

The first dataset examined is composed of vestige counts collected for collared peccary (*Dicotyles tajacu*) in the state of São Paulo in southeast Brazil (Assis, 2012) between July 2012 and October 2012. A camera trap survey was also carried out at the same sites during this time. While the collared peccary observations collected from camera traps do not form part of the analysis presented in Chapter 5, the data from camera traps was used to inform prior distributions assigned to estimate animal group sizes, and so Figure 2.8 contains photos of collared peccaries captured in August 2012 from some of these camera traps.

Figure 2.9 presents the location at which this survey was carried out, in the municipality of Angatuba in the southeast of São Paulo.

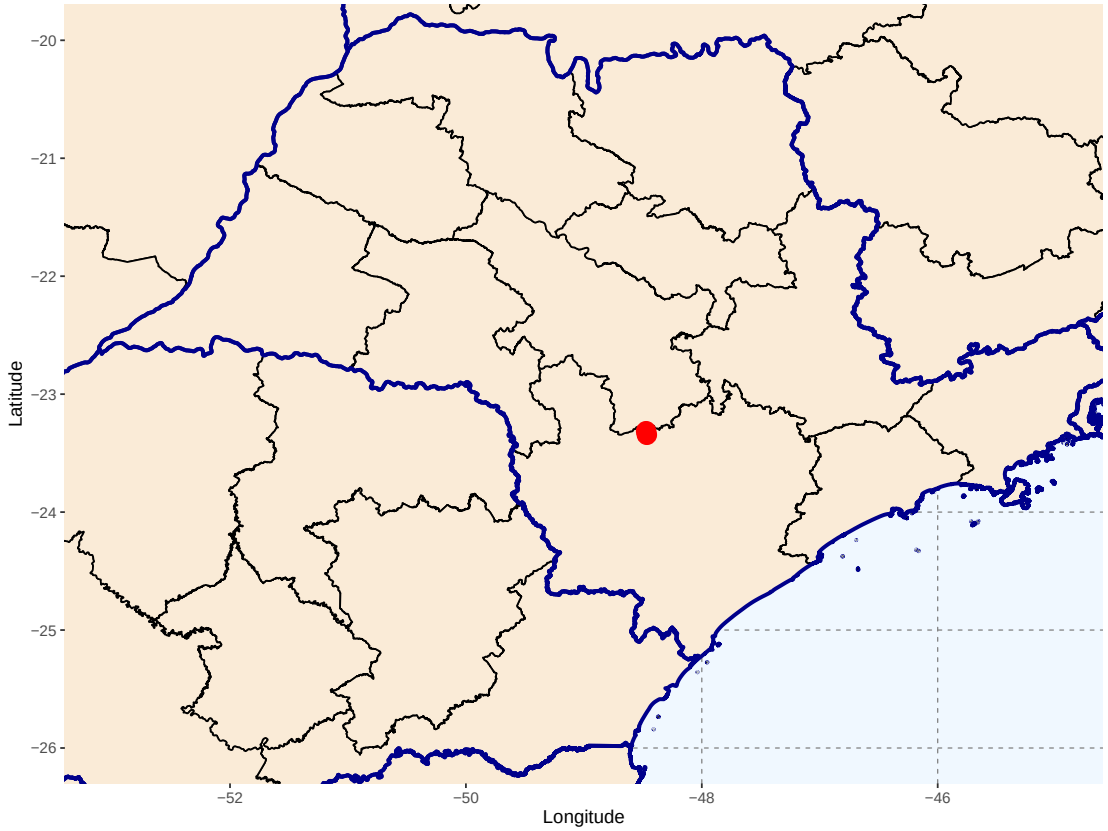


Figure 2.9: The location of the collared peccary survey marked in red in the geographic region of Sorocaba (geographic regions outlined in black) in the state of São Paulo, outlined in dark blue.

This dataset is composed of data collected on a single occasion at only two locations, and so contains only two counts. The first transect (23°20'0"- 23°18'51"S/ 48°27'30"- 48°28'20"W) was 8km in length and seven vestiges were observed along its length, and the second transect (23°22'0"- 23°20'41"S/ 48°28'00"- 48°27'57"W) was 12km long, and only a single vestige was observed along its length. Researchers walked along each transect, with stops approximately every 50m for improved observations. This was the dataset that motivated our interest in determining how the triple Poisson model proposed in Chapter 5 might cope when presented with very scarce datasets. This became one of the central research questions we wished to answer when performing the simulation studies described in Chapter 5.

2.4 Sika Deer Survey

The second dataset examined in Chapter 5 contains vestige counts collected for sika deer (*Cervus nippon*) from eight regions in Scotland from March 1997 to May 1997 (Marques et al., 2001). This data is available as part of the `Distance R` package (Miller et al., 2019), and provides us with the distance between every vestige and the transect, which allows us to compare the predictive performance of the triple Poisson model introduced in Chapter 5 to the distance sampling model (Thomas et al., 2006). This data motivated a simulation study which would compare the performance of the triple Poisson model with a distance sampling model under a range of scenarios. A sample of the dataset is provided in Table 2.3.

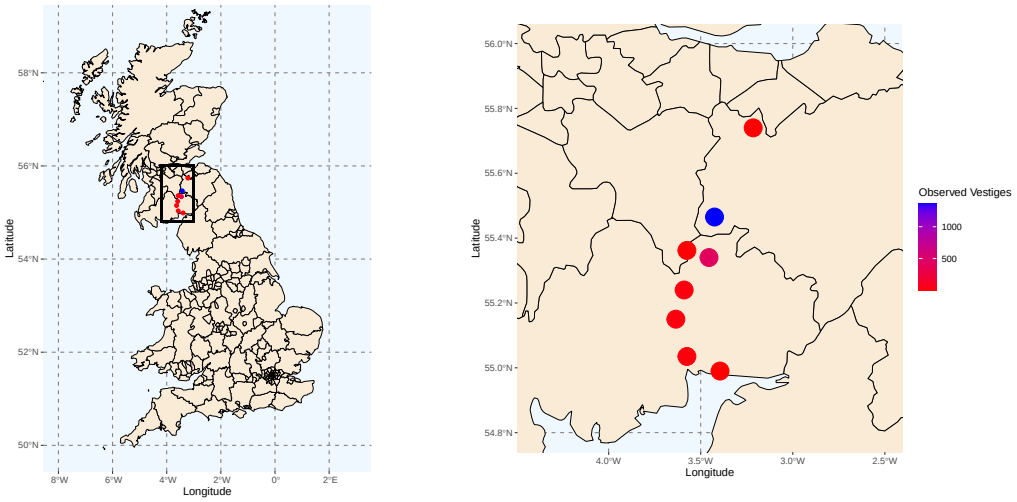
Region	Section	Area (km^2)	Length (km)	Distance (cm)
A	A-15	13.9	0.15	73
B	B-9	10.3	0.20	154
B	B-14	10.3	0.05	108
C	C-1	8.6	0.10	113
J	J-3	9.6	0.10	18

Table 2.3: A sample of the sika deer dataset, with the region label (labelled A–J) for each vestige, the area of the region in square kilometres, the transect section, the length of the transect in kilometres, and the distance in centimetres of the vestige from the transect.

Figure 2.10 displays the location of the study area (2.10a) in the Tweedsmuir region of Scotland, and the number of vestiges observed at each of the eight regions (2.10b). From initial examination, it is apparent that the majority of the vestiges were observed at two sites, with 1366 and 426 vestiges observed at sites 1 and 2 respectively, while the remainder of the sites saw a maximum of 36 vestiges observed. This is due to "greater survey effort being allocated to blocks thought to contain higher deer densities" (Marques et al., 2001). This greater survey effort manifests as greater transect lengths. We anticipate that this will pose no issues for the modelling framework that we will introduce in Chapter 5, as transect length will be incorporated into our abundance estimates.

An issue encountered on initial examination of this data, and one which interested practitioners may keep in mind, is that the sika deer dataset is described

in the `Distance` `R` package as relating to the Peebleshire portion rather than the Tweedsmuir portion of the study by Marques et al. (2001). This caused some confusion initially, as the survey design and data analysis for the Peebleshire portion of the sika deer survey differs from that of the Tweedsmuir region, and did not align with the data presented in the `R` package.

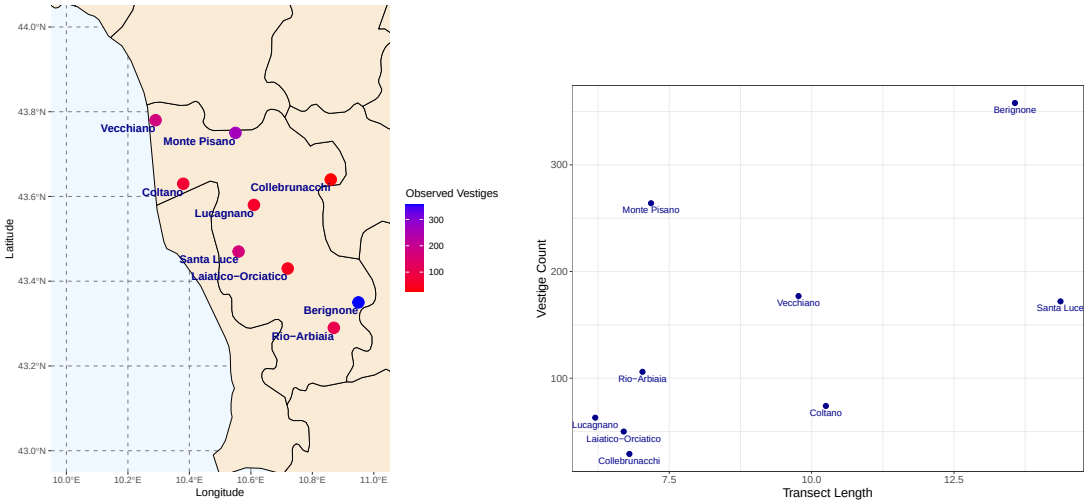


(a) (b)

Figure 2.10: (a) The approximate location of the sika deer study in the Tweedsmuir region of Scotland (b) The number of vestiges observed at each of the eight transects monitored during the study.

2.5 Red Fox Survey

The third and final dataset examined in Chapter 5 is composed of vestige counts collected for red foxes (*Vulpes vulpes*).This data was collected along transects in nine regions in the province of Pisa in central Italy (Cavallini, 1994) between April 1992 and March 1993. The approximate location of each of the nine sites is presented in Figure 2.11, along with the length of each transect in kilometres (Figure 2.11a) and the number of vestiges observed at each site over the course of the 12–month survey (Figure 2.11b).



(a) (b)

Figure 2.11: (a) The number of vestiges observed at transects and (b) a scatter plot of transect length vs. vestige count at transects, at each of nine regions in Pisa, central Italy

In the original paper by Cavallini (1994), the author produced an index of fox abundance, estimated as the number of vestiges observed per kilometre of transect length. This allowed the original paper to estimate this abundance index separately for each of the nine regions examined. Ideally, we would also produce an estimate of red fox abundance per region, for comparison with the original model. However, this is not possible in our case because, as we describe in Chapter 5, the modelling framework that we have developed requires knowledge of the size of each study area. Cavallini (1994) provide the size of the study area as 2448km^2. This is however, the area of the province of Pisa, and so we do not have information as to the area of each region examined within Pisa. For this reason, we are able to obtain only a single value, which is the estimate for red fox abundance in the province of Pisa.

For interested practitioners, this dataset is available in its entirety as a part of the paper by Cavallini (1994). Observations were taken at these transects once each month for a year. This dataset motivated our interest in a modelling framework that can accommodate temporal replicates, and this subsequently formed a large

	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
Berignone	37	22	10	11	25	71	12	23	17	42	22	66
Lucagnano	6	14	6	6	9	8	0	2	4	3	3	2
Coltano	10	11	3	3	4	16	0	3	3	7	6	8

Table 2.4: A sample of the red fox dataset, with the number of vestiges observed each month from April 1992 to March 1993 in three locations of Pisa.

part of the simulation studies described in Chapter 5. A sample of the dataset is provided in Table 2.4.

CHAPTER 3

The Bayesian Multi-Species N-Mixture Model for Unmarked Animal Communities

The N-mixture model proposed by Royle (2004) provides an attractive option for estimating animal abundance using animal count data, as it takes into account the imperfect detection inherent in this form of data. In this chapter, we introduce a multi-species N-mixture (MNM) extension to the original model, which allows us to estimate abundances for multiple species, while simultaneously estimating the correlation between abundances of different species. We also provide further extensions that allow for the use of data that contains a large number of zero-counts and data collected over long time periods. Via simulation studies and real data applications, we demonstrate the utility of the MNM family of models under a range of conditions. `R` *code for the MNM model implementation is available at* `https://github.com/niamhmimnagh/mnm`.

3.1 Introduction

Abundance in animal communities is of great interest in ecology, particularly in the areas of conservation and wildlife management (Witmer, 2005; Nichols and

MacKenzie, 2004). Count data is an attractive option for estimating abundance due to the relative affordability with which it may be collected when compared to methodologies that require the use of technologies such as camera traps and thermal imagery, and the reduced risk of harm to both animals and humans when compared to more direct data collection methods (Verdade et al., 2013), such as a mark-recapture data collection technique. However, count data for animal abundance has a tendency to suffer from imperfect detection (i.e., the recorded information is usually imperfect in the sense that it does not represent the total abundance). Furthermore, when the detection probability is small, there is a tendency towards the underestimation of abundance. Due to the characteristics of these data, traditional modelling techniques, such as generalised linear models (McCullagh and Nelder, 1989), cannot be applied directly to the data, as they do not accommodate imperfect detection through the estimation of detection probabilities. If the detection probability was a known value, it could be incorporated as an offset in a generalised linear model. However, as we do not know the true values for detection probability and wish to estimate it, a generalised linear model was not chosen for this analysis.

N-mixture models (Royle, 2004) constitute a class of models which may be used to estimate abundance from count data. These models assume that the population under analysis is closed, (i.e., it is constant in terms of births, deaths, and migration). The counts at each site and sampling occasion are considered independent and identically distributed (i.i.d) random variables that follow a binomial distribution. The population size at each site is treated as a random effect, with an assumed probability distribution. The distributions that are typically considered for the population size at each site are the Poisson and negative binomial, although any other non-negative discrete distribution could also be considered.

The ability to estimate correlations between species abundances allows us to relax any assumption of independent species abundances. This is the aim of the multi-species N-mixture (MNM) models presented in this chapter – a class of models that estimate abundance for multiple species simultaneously while accounting for imperfect detection, and estimate inter-species correlations, which are intended to allow for inferences about the relationships between different species.

The remainder of this chapter is organised as follows. In Section 3.3, we introduce our novel modelling framework to estimate abundance and inter-species correlations in animal communities based on spatio-temporal count data. We also describe the model formulation, estimation procedure, and the computation of the inter-species correlations. In Section 3.4, we present the data obtained from the North American Breeding Bird Survey (Pardieck et al., 2020), which will be used to illustrate our modelling approach. Later, in Section 3.5, we compare results of model fit on the North American Breeding Bird Survey data to obtain the best fit. Finally, in Section 3.6, we present a general discussion.

3.2 Related Works

Previous examinations of detection probability include those by Fisher (1934), Fisher et al. (1943), and Rao (1965). Fisher (1934) explored the effect of various data collection techniques and ascertainment (the process of choosing individuals for analysis) on the estimation of frequencies. He explored the "detection probability" of albinism occurring in children, the potential for bias in ascertainment and the implication of this bias on resulting population inferences. Rao (1965) also addressed the problem of ascertainment bias, and explored how different ascertainment methods can lead to specific discrete probability distributions. Fisher et al. (1943) explored how different statistical models, including the log-series distribution and the negative binomial distribution, can be used to predict the total number of species in a population based on a population sample.

Several multi-species modelling frameworks have been developed previously which allow for the analysis of occurrence-data (binary data, in which a one represents an occupied site, and a zero represents an unoccupied site) (Dorazio and Royle, 2005; Yamaura et al., 2011) or count-data, which represents the number of individuals of a species of interest that are observed(Yamaura et al., 2012; Golding et al., 2017; Gomez et al., 2018).

Dorazio and Royle (2005) developed a model for estimating the size of a biological community by modelling the probability of detection as a binomial random variable, and the probability of occurrence as a Bernoulli random variable. They allow

rates of detection and occurrence to vary among species, and not every species is assumed to be present at every location. However, the aim of their model is to determine the number of species present, not the number of individuals of each species present, as is the aim of N-mixture models.

Yamaura et al. (2011) developed a multi-species model that estimates animal abundance from occurrence-data. This is an extension of the single-species model developed by Royle and Nichols (2003), in which binary detection/non-detection data is linked to abundance. Yamaura et al. (2012) extended this model to count data. The assumption behind these models is that the abundances or detection probabilities of species in the community might be linked by species-level or functional group-level characteristics. However, inter-species abundance correlations are not explored within these models.

Gomez et al. (2018) developed a multi-species N-mixture model whose aim was to allow for the estimation of abundance of rare species by borrowing strength from other species in the community. This was done by assuming detection probabilities are drawn at random from a Beta distribution. Another multi-species N-mixture model was developed by Golding et al. (2017), which used the dependent double-observer method to create a multi-species dependent double-observer abundance model. This allowed them to address an issue of false-positive errors in detection. The focus of both Gomez et al. (2018) and Golding et al. (2017) was an improvement in detection probability. None of the preceding multi-species models allow us to make inferences as to the relationships within an ecological community, as we propose to do with our multi-species N-mixture model.

Moral et al. (2018) developed an extension to the single-species N-mixture model which allowed for the estimation of abundances of two species, and the correlations between these abundances. However, this model only examines two species, and is therefore not as complete as the model we propose here, which allows us to examine whole communities.

Dorazio and Connor (2014) developed a multi-species N-mixture model which allowed for abundances of species with similar traits to be correlated. However, to guarantee positive definite correlations, they only allow for positive correlations

through the use of a distance metric d coupled with a spatial autocorrelation structure of the type $e^{-\frac{d}{\phi}}$. The framework we present here is more complete in that we guarantee positive definiteness of the correlation matrix via an elegant prior setup. We also explore ways of incorporating zero-inflation and open population dynamics, which is not something attempted by Dorazio and Connor (2014).

Finally, Niku et al. (2019) describe generalised linear latent variable models - a modelling technique which allows for obtaining correlation matrices in an elegant manner. However, these models do not allow for the incorporation of imperfect detection.

Some of the above described modelling frameworks, along with the original N-mixture model by Royle (2004) are detailed in Chapter 4.

3.3 Methods

The models developed in the following Section are a multi-species extension to the original N-mixture model of Royle (2004), which allows for accurate estimation of both the latent abundances and inter-species correlations, while accounting for imperfect detection and relaxing the closure assumption.

3.3.1 Multi-Species N-Mixture Model (MNM Model)

Consider a study which involves the collection of count data Y_{its}, where Y_{its} is the number of individuals observed for S different species ($s = 1, \ldots, S$) from R sites ($i = 1, \ldots, R$). Consider also that these samples are taken from each site on T sampling occasions ($t = 1, \ldots, T$). The parameter of interest is the true abundance, and at site i for species s is, this given by N_{is}. We observe a fraction of N_{is}, with detection probability p_{its}, and it is assumed that species populations are closed with respect to births, deaths and migration (i.e., that the population sizes do not change due to any of these factors, akin to the N-mixture model proposed by Royle (2004)). Our model assumes that N_{is} may be described by a

Poisson distribution, and the model may be written as:

$$\begin{aligned} Y_{its} \mid N_{is}, p_{its} &\sim \text{Binomial}(N_{is}, p_{its}), \\ N_{is} \mid \lambda_{is} &\sim \text{Poisson}(\lambda_{is}), \\ \text{logit}(p_{its}) &= \mathbf{z}_{it}^{\top}\mathbf{b}_s, \\ \log(\lambda_{is}) &= a_{is} + \mathbf{x}_i^{\top}\boldsymbol{\beta}_s, \\ \mathbf{a}_i \mid \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a &\sim \text{MVN}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \end{aligned}$$

where $\mathbf{a}_i = (a_{i1}, \ldots, a_{iS})^{\top}$. The Poisson rate parameter λ_{is} represents the mean abundance at site i, and $\mathbf{a}_i$ is an S-dimensional vector that contains the random effects a_{is} that allow us to estimate an instantaneous inter-species correlation. In the above model, covariates may be incorporated in the detection probability and the abundance, with $\mathbf{z}_{it}^{\top}$ the it-th row of the design matrix $\mathbf{Z}$ of dimension $RT \times q_p$, b_s the $q_p \times 1$ parameter vector for the probability of detection, $\mathbf{x}_i^{\top}$ the i-th row of the design matrix $\mathbf{X}$ of dimension $R \times q_\lambda$, and $\boldsymbol{\beta}$ the $q_\lambda \times 1$ parameter vector for the abundance. Here, q_p and q_λ represent the number of covariates associated with the detection probability and the abundance, respectively. Note that different covariate effects may be estimated per species, and other species-level random effects may also be included.

3.3.2 Hurdle-Poisson Model (MNM-Hurdle Model)

In this Section, we develop a further extension of the multi-species N-mixture model, appropriate for scenarios in which the number of zero-counts exceed those expected under a Poisson distribution. We now allow the counts to follow a Hurdle-Poisson distribution, with λ_{is} defined as in the MNM Model, and θ the probability of obtaining a zero-count.

The Hurdle-Poisson distribution consists of two separate processes. The first is a Bernoulli process, which determines whether a site is occupied (count is non-zero) or unoccupied (count is zero). If the count is non-zero, a second random variable with a zero-truncated Poisson distribution determines the value of the count, i.e.,

$$\begin{aligned} &\text{Occupancy}_{is} \sim \text{Bernoulli}(1-\theta), \\ &\text{Count}_{is} \sim \text{Zero Truncated Poisson}(\lambda_{is}). \end{aligned}$$

We then define the latent abundances N_{is} as

$$N_{is} = \begin{cases} 0 & \text{if Occupancy}_{is} = 0 \\ \text{Count}_{is} & \text{if Occupancy}_{is} = 1 \end{cases},$$

which yields

$$N_{is} \sim \text{Hurdle-Poisson}(\lambda_{is}, \theta).$$

If the Bernoulli process is equal to 0, then the site is unoccupied and N_{is} is equal to 0. However, if the Bernoulli process is equal to 1, then the hurdle is crossed, and the value of N_{is} is determined by the zero-truncated Poisson process. Similar to the MNM model, populations are assumed to be closed.

An alternative to the Hurdle Poisson model would be a zero-inflated Poisson (ZIP) model. This model functions in a way similar to the Hurdle Poisson model described above, with the substitution of a Poisson distribution for the count process, rather than a zero-truncated Poisson distribution as is described above. The result being that in a ZIP model, the count process is also capable of producing zero counts, and so the zero values can be decomposed into "true" zeros (produced because the site is unoccupied) and "false" zeros (produced because, though the site is occupied, the animal is not observed). This categorisation of zero counts was not of interest to us in this case, and so we choose to implement the Hurdle Poisson model instead.

We assume a single probability of obtaining a zero count θ. However, θ may also be allowed to vary by site and/or species, and may depend on covariates through a logit link. All other parameters are as described in the MNM model in Section 3.3.1.

3.3.3 Autoregressive Model (MNM-AR Model)

In order to model populations over multiple years, a further extension to the multi-species N-mixture model is proposed, which allows us to relax the assumption that species populations are closed with respect to births, deaths and migration (i.e., that the latent abundance N does not change during the data-collection period due to any of these factors). We do this through the inclusion of a first-order autoregressive term in the abundance parameter.

The study design now consists of data collected over K years ($k = 1, \ldots, K$) for S species at R locations, each site sampled on T occasions. The observed abundance (Y) and actual abundance (N) are now allowed to vary by year:

$$Y_{itks} \sim \text{Binomial}(N_{iks}, p_s),$$
$$N_{iks} \sim \text{Poisson}(\lambda_{iks}).$$

If $k = 1$, then λ_{i1s} is defined as before:

$$\log(\lambda_{i1s}) = a_{is} + \mathbf{x}_i^\top \boldsymbol{\beta}_s.$$

However, for $k > 1$, we allow λ_{iks} to depend on the latent abundance at year $k-1$:

$$\log(\lambda_{iks}) = a_{is} + \mathbf{x}_i^\top \boldsymbol{\beta}_s + \phi_s \log(N_{i(k-1)s} + 1).$$

We regress the mean abundance at time k, λ_{iks}, on the actual abundance at time $k-1$, $N_{i(k-1)s}$, which is, in essence, a Poisson ARCH (autoregressive conditional heteroscedasticity) model, as examined by Zeger and Qaqish (1988) and Fahrmeir et al. (1994). An alternative, proposed by Fokianos and Tjøstheim (2011) and Ferland et al. (2006) would be the Poisson INGARCH (integer-valued generalised autoregressive conditional heteroscedasticity) process, which regresses λ on both past values of N and λ. By regressing λ on past values of N, we capture the effect of short-term dependence on the previous time-point, while regressing on past values of λ might allow us to capture long-term patterns in the average count rate, and may be useful for count data with varying trends or seasonality. For parsimony in model design, ease of interpretation and lower computational complexity, we choose to examine the framework that regresses on only past values of N in this chapter.

The term $\phi_s \log(N_{i(k-1)s} + 1)$ is used rather than the simpler $\phi_s N_{i(k-1)s}$ to avoid a known issue with this type of modelling framework, (Fokianos and Tjøstheim, 2011), that is, the tendency for sampled λ values to increase rapidly when λ_{iks} is regressed on $N_{i(k-1)s}$.

3.3.4 Hurdle-Autoregressive Model (MNM-Hurdle-AR Model)

A straightforward combination of the MNM-Hurdle model and the MNM-AR model produces the MNM-Hurdle-AR model. This model accommodates excess zeros while also accounting for an autoregressive structure in the data. The zero-inflation is introduced as in Section 3.3.2, i.e.,

$$\begin{aligned} Y_{itks} &\sim \text{Binomial}(N_{iks}, p_s), \\ N_{iks} &\sim \text{Hurdle-Poisson}(\lambda_{iks}, \theta), \end{aligned}$$

and the autoregressive structure is incorporated as in Section 3.3.3,

$$\log(\lambda_{iks}) = \begin{cases} a_{is} + \mathbf{x}_i^\top \boldsymbol{\beta}_s, & \text{for } k = 1 \\ a_{is} + \mathbf{x}_i^\top \boldsymbol{\beta}_s + \phi_s \log(N_{i(k-1)s} + 1), & \text{for } k > 1 \end{cases}.$$

3.3.5 Model Estimation

The models described in this chapter are implemented using a Bayesian framework. Each of the above models were implemented in `R` (R Core Team, 2022) through the probabilistic programming software JAGS (Plummer, 2003, 2017) using four chains with 50,000 iterations each, of which the first 10,000 were discarded as burn-in, with a thinning of five to reduce autocorrelation in the MCMC samples. Parameter convergence was determined using the potential scale reduction factor ($\hat{R}$), a diagnostic criteria proposed by Gelman and Rubin (1992). An $\hat{R}$ value that is very close to one is an indication that the four chains have mixed well. If $\hat{R}$ value was less than 1.05, the chains were considered to have mixed properly, and the posterior estimates of the parameters were considered reliable.

Prior distributions were assigned as follows: $\boldsymbol{\mu}_a$, the vector of means of the random effect $\mathbf{a}$, was assigned a multivariate Normal prior with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}_0$ and mean vector $\boldsymbol{\mu}_0$. $\boldsymbol{\Sigma}_a$, the variance-covariance matrix of $\mathbf{a}$ was assigned an inverse-Wishart prior with a diagonal scale matrix $\boldsymbol{\Omega}$, and $S+1$ degrees of freedom v which results in a Uniform$(-1, 1)$ prior on the correlations

(Plummer, 2017):

$$\boldsymbol{\mu}_a \sim \text{MVN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),$$
$$\boldsymbol{\Sigma}_a \sim \text{Inverse-Wishart}(\boldsymbol{\Omega}, v).$$

An inverse-Wishart distribution is specified as the prior distribution for the covariance matrix of the random effect $\mathbf{a}$. Criticisms of the inverse-Wishart prior include the dependency imposed between correlations and variances, and the fact that there is a single degree of freedom parameter which determines the uncertainty for all variance parameters. It is demonstrated by Alvarez et al. (2014) that when the variance is small relative to the mean, the correlation is biased towards zero, and the variance is biased towards larger values. However, it is not anticipated that this issue will arise with count data, as when working with count data, typically variances are large relative to the mean. Despite these issues, the inverse-Wishart distribution is a prior distribution commonly assigned to a covariance matrix in Bayesian analysis due to its conjugacy with the Normal distribution, and for these models the inverse-Wishart distribution provides a good solution due to its guarantee of providing a positive definite covariance matrix.

In the Hurdle and Hurdle-AR models, θ is the non-negative probability of obtaining a zero count, and so is assigned a beta prior with the value of both shape parameters equal to one, which is equivalent to a non-informative uniform prior distribution:

$$\theta \sim \text{Beta}(1, 1).$$

In the AR and Hurdle-AR models, ϕ is assigned a multivariate normal prior distribution, with hyperpriors $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ and diagonal matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$:

$$\boldsymbol{\phi_s} \sim \text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\phi}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}}),$$

Extensive simulation studies were carried out to examine the accuracy of parameter estimates; see Appendix 3.A for more details.

3.3.6 Inter-Species Correlations

The presence of the multivariate normal random effect $\mathbf{a}$ in the abundance provides a link between species' abundances. The correlation matrix for the random effect, $\mathbf{\Sigma}_a$, may be estimated directly from the Bayesian model. Using these estimated values, the inter-species correlations for the latent abundances N_s and $N_{s'}$, for all $s \neq s'$, are calculated for each model as:

$$\rho(N_s, N_{s'}) = \frac{\text{Cov}(N_s, N_{s'})}{\sqrt{(\text{Var}(N_s))(\text{Var}(N_{s'}))}}.$$

The derivation of $\rho(N_s, N_{s'})$ for each of the models described in this Section can be found in Appendix 3.B.

The inter-species correlations for the MNM model and Hurdle model are assumed not to vary by year, so these models have a single analytic correlation matrix ρ. However, in the AR and Hurdle-AR model, we assume latent abundances change by year, which requires the computation of K analytic correlation matrices. Note that the MNM and AR models required the use of properties of conditional variance and covariance to determine analytic correlations. In the Hurdle and Hurdle-AR models, the properties of conditional variance and covariance were merged with second-order Taylor approximations to make their computation feasible.

3.4 Case Study: North American Breeding Bird Survey

In this section, we briefly revisit the North American Breeding Bird Survey described in Chapter 2, and describe the application of the multi-species N-mixture models to this case study, to examine bird populations in Alaska, in the northwest of the United States of America.

The North American Breeding Bird Survey (Pardieck et al., 2020) was first conducted in 1966, and now provides data annually on more than 400 bird species across 3700 routes in the United States and Canada. Each of these routes is approximately 24.5 miles long and is composed of 50 stops, approximately 0.5 miles apart. At each stop, every bird seen or heard within a 0.25-mile radius is recorded.

For the sake of our models, each of these routes is considered a site, and each of the 50 stops along a route is a sampling occasion.

We examine data collected in Alaska in the 10-year period 2010-2019. There are 94 routes in Alaska (Fig. 3.1) at which data was collected during this time, and each of these routes are composed of 50 sampling locations, totalling 4,700 observations per bird species.

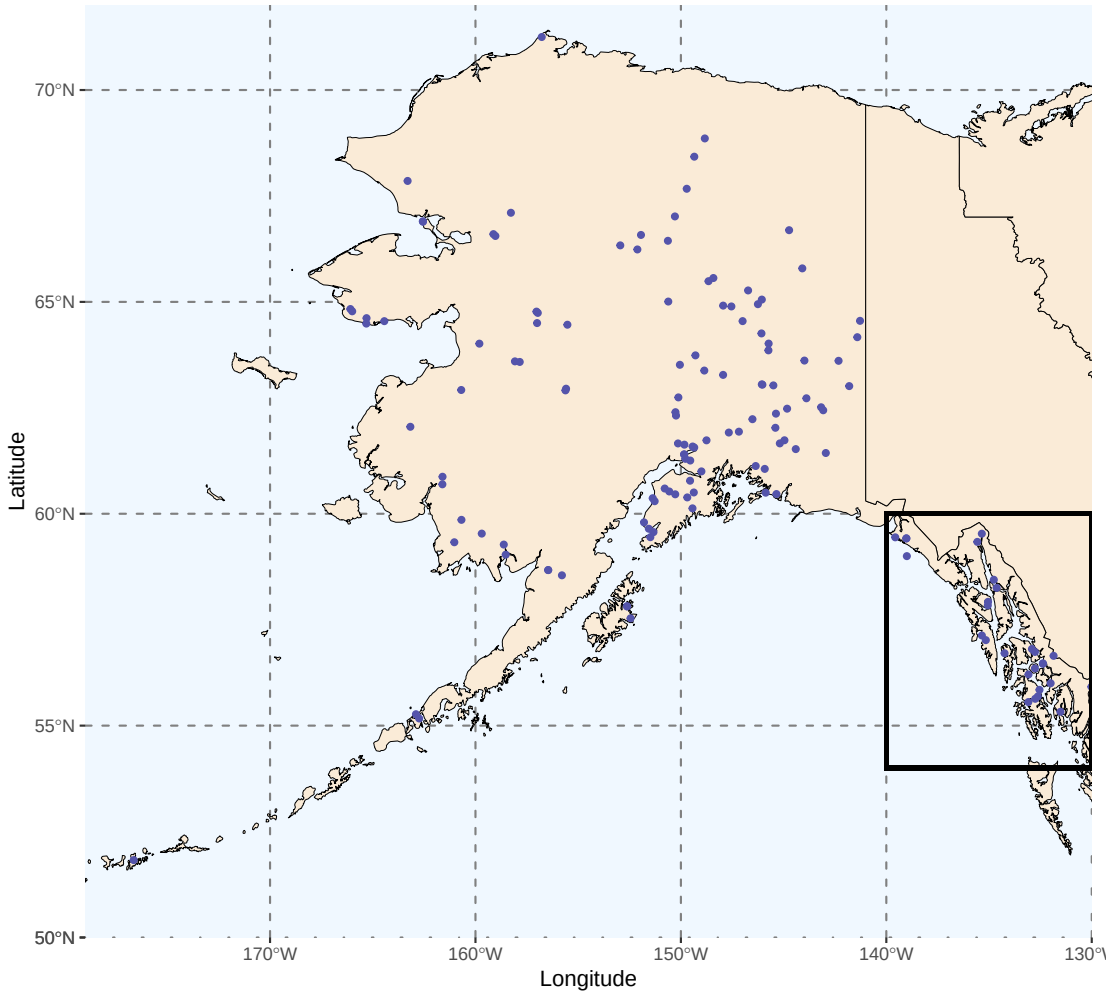


Figure 3.1: Location of sites in Alaska marked in blue, with the Alexander Archipelago - the location of the Alaskan Bald Eagle population - outlined in black.

Bald eagle populations in Alaska are estimated at between 8,000 and 30,000 birds,

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Bald Eagle	21	20	25	20	22	19	23	30	30	23
Canada Goose	14	18	17	19	8	11	12	15	16	14
Hammond's Flycatcher	16	16	15	12	12	11	12	15	17	14
Red-breasted Sapsucker	12	11	10	9	10	9	10	12	12	12
Steller's Jay	16	16	14	11	12	13	12	14	15	13
Swainson's Thrush	57	52	56	50	50	49	48	63	62	55
Tree Swallow	27	26	27	25	22	27	25	31	27	30
Trumpeter Swan	13	9	14	14	11	12	12	14	12	8
Varied Thrush	62	58	62	52	55	52	50	68	62	58
Wilson's Snipe	52	47	51	47	44	48	42	53	50	47

Table 3.1: Number of routes each species appears at by year, out of a total 94 routes.

which accounts for roughly half of the global population (Hodges, 2011; Hansen, 1987; King et al., 1972). For this reason, bald eagles were chosen as a species of interest for this analysis. Several other species were chosen; these included waterbirds such as geese, swans and snipes which were chosen for their relationships with bald eagles, as bald eagles are known to prey on waterbirds such as ducks, geese and grebes when fish are in short supply (Dunstan and Harper, 1975; Todd et al., 1982; McEwan and Hirth, 1980). Additionally, a selection of species with inland habitats, such as thrushes and swallows, were examined. In total, 10 species were selected for analysis, of the 233 total species present in Alaska within the 10-year period. The full list of species selected and the frequency with which they were observed is given in Table 3.1.

The models described in Section 3.3 were fitted to the North American Breeding Bird Survey data. Each was fitted three times, varying the dimension of the detection probability. Initially, the detection probability was allowed to vary by site, species and year. Subsequently, models were fitted in which detection probability varies only by site and species, and then by species alone.

Initially, the models were fitted without covariates, and results were compared using their Bayesian Information Criterion (BIC) (Delattre et al., 2014) and Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) values. Subsequently, lat-

itude, longitude and their interaction term latitude × longitude were included in the linear predictors for the abundance parameters, and models were again compared using BIC and DIC values. All covariates were scaled to have zero-mean and unit variance.

Initial examination of this data revealed that 93.2% of observations (438,040 of a total of 470,000 observations) consisted of zero counts. This suggested that a model with a hurdle component might provide an appropriate framework for this data. Furthermore, this data was collected over the course of a decade. It would be unrealistic to expect that populations remain closed for this length of time. For this reason, we might expect that an autoregressive term may be useful to incorporate time dependence in abundance estimates.

Each model was fitted using four chains with 50,000 iterations each, of which the first 10,000 were discarded as burn-in, using a thinning value of five. All prior distributions were assigned as described in Section 3.3.5.

3.5 Results

Initially, the models were fitted without covariates and were compared using BIC and DIC values. The result of this comparison was that BIC values suggested that the Hurdle-AR model, in which detection probability varies by species (Hurdle-AR(C) model in Table 3.2), provided the best fit for the North American Breeding Bird Survey data. DIC values, on the other hand, chose the Hurdle model, in which detection probability varies by site and species (Hurdle(B) model in Table 3.2) as the model of best fit. This is due to the difference in penalty terms used in the calculation of BIC and DIC values. The BIC value penalises model complexity more heavily than the DIC value, and as a result chooses a model with far fewer parameters. The subsequent addition of a response surface for latitude and longitude in the linear predictors for the abundance parameter results in the Hurdle model in which detection probability varies by species (Hurdle(C) model in Table 3.2) producing the lowest BIC value overall, and the DIC value choosing the Hurdle-AR model in which detection probability varies by site and species (Hurdle-AR(B) in Table 3.2). As a result, we could choose to examine either the

Hurdle(C) model or the Hurdle-AR(B) model as the model of best fit. Prioritising parsimony in our model choice, we will examine the Hurdle(C) model for the remainder of this analysis. The variance that was initially explained by the addition of the autoregressive term (in model Hurdle-AR(C)) is now explained by the latitude and longitude covariates, which render the autoregressive component unnecessary. The full details of this model comparison are given in Table 3.2, along with the number of parameters in each model. Because the models differ in terms of dimension of detection probability p, models in which p varies by 94 sites, 10 species, and 10 years (models labelled (A))contain far more parameters than models in which p varies only by site and species (models labelled (B)), or by species alone (models labelled (C)).

	No Covariates				Covariates			
Model	P	BIC	EP	DIC	P	BIC	EP	DIC
MNM(A)	9,465	291,860	3,968	191,563	9,471	291,899	3,974	191,558
MNM(B)	1,005	194,592	4,181	191,597	1,011	194,743	4,044	191,419
MNM(C)	75	183,587	5,555	193,989	81	183,940	5,260	193,696
AR(A)	9,467	291,648	4,204	191,803	9,473	291,808	4,096	191,692
AR(B)	1,007	194,767	4,029	191,450	1,013	194,805	4,040	191,455
AR(C)	77	183,922	5,236	193,668	83	183,580	5,674	194,108
Hurdle(A)	9,466	291,699	3,820	191,094	9,472	291,596	3,988	191,272
Hurdle(B)	1,006	194,964	3,092	**189,779**	1,012	194,915	3,203	189,897
Hurdle(C)	76	183,890	4,091	191,352	82	**183,349**	4,771	192,116
Hurdle-AR(A)	9,468	291,597	3,950	191,233	9,474	291,652	3,952	191,237
Hurdle-AR(B)	1,008	194,859	3,217	189,906	1,014	195,044	3,094	**189,789**
Hurdle-AR(C)	78	**183,572**	4,515	191,810	84	183,644	4,493	191,837

Table 3.2: BIC values and associated number of parameters (P), DIC values and associated effective number of parameters (EP), comparing model fits on North American Breeding Bird Survey data. Models labelled (A) contain detection probability which varies by site, species and year, (B) contain detection probability which varies by site and species, and (C) contain detection probability which varies only by species. Smallest BIC and DIC values for each case are indicated in bold.

The latent inter-species correlations are given in Fig. 3.2, while the derivation of analytic correlations, which vary by site and year, are given in Appendix 3.B. These latent correlations are obtained after probability of detection and other covariates

are taken into account, and they may be interpreted as an interaction strength metric, which allows for the study of the influence of one species' abundance on the abundance of others (Berlow et al., 2004; Moral et al., 2018). The strength of the correlation between the abundances of two species may be due to environmental factors that affect both species, or may be due to direct interactions between them. This might inform ecologists as to whether conservation efforts made for one species could have an effect on another species. For example, the bald eagle has a positive correlation with the other water birds examined (trumpeter swan, Wilson's snipe, Canada goose). This suggests that there may be environmental factors at play that would have similar affects on the abundances of these species, and conservation efforts to increase bald eagle abundances may also increase populations of these other water birds.

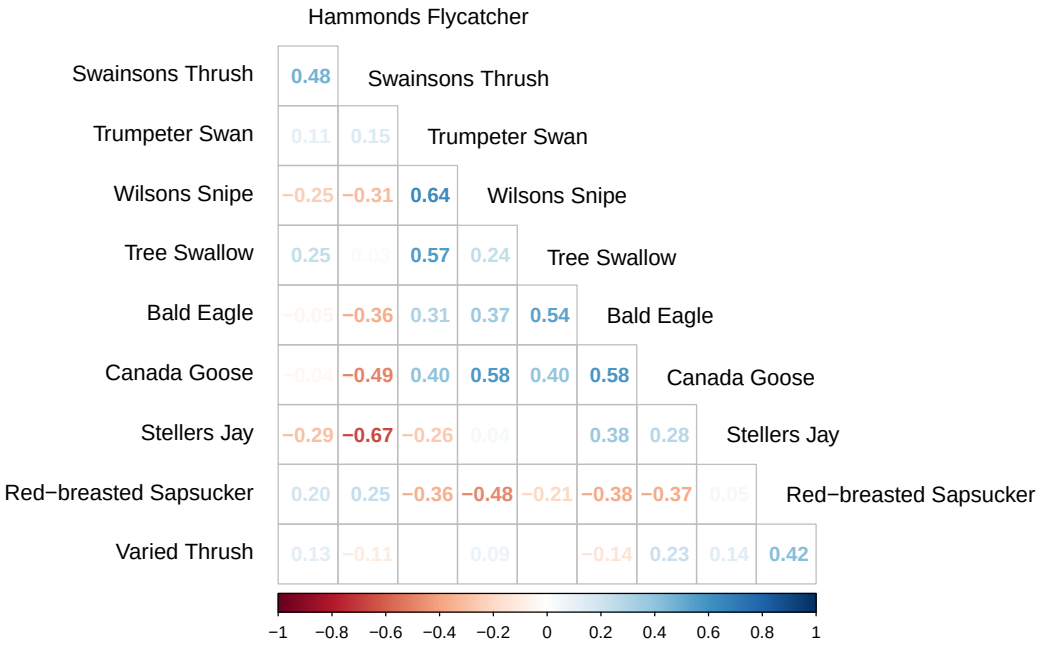


Figure 3.2: Estimated latent inter-species correlation matrix, produced by the Hurdle(C) model fitted to the North American Breeding Bird Survey data including covariates in the linear predictor for the abundance parameter.

3.6 Discussion

We have proposed a multi-species extension to the N-mixture model which allows for the estimation of inter-species abundance correlations through the addition of a random variable in the abundance. Results of simulation studies (see Appendix 3.A) reveal that this model performs well under a range of scenarios, with abundances and detection probabilities that range from low to high. For this reason, we believe that this approach represents an attractive framework for examining multi-species abundances.

Issues with parameter convergence were encountered when fitting the Hurdle and Hurdle-AR models. When zero-inflation and abundance are large, and detection probability is small, issues with convergence occurred in up to 20% of parameters. While this convergence issue does not appear to negatively affect the relative biases of parameter estimates (as can be seen in Appendix 3.A, Table 3.A.1 and Table 3.A.2), coverage probability for detection probability p and random effect mean μ_a is negatively impacted (Appendix 3.D). In the same models, we see larger coverage for N. This is to be expected, and is due to zero counts being perfectly predicted.

Previous works have demonstrated that N-mixture models can sometimes suffer from issues with identifiability (Dennis et al., 2015) wherein probability of detection estimates are very close to zero and abundance estimates are infinite. To address this issue, we have performed extensive simulation studies, detailed in Appendix 3.A, in which we assess the estimates of abundance and detection probability for a large range of sample sizes, detection probabilities, abundance sizes, and in the case of the Hurdle and Hurdle-AR models, zero-count probabilities. We have thus far not witnessed the occurrence of these identifiability issues in any of the simulation studies performed.

The models presented here all use the Poisson distribution to model the latent abundances. However, any other count distribution might instead be used. For example, if overdispersion is to be expected in latent abundances, the negative binomial distribution may provide a better fit to estimate N. Our calculations for the analytic correlations, however, reflect only the use of the Poisson distribution.

We have also examined an implementation of this modelling framework using real-world data collected as part of the North American Breeding Bird Survey. Results reveal that the model with the lowest associated BIC value is the Hurdle model, in which detection probability varies only by species, and a latitude $\times$ longitude is included in the abundance linear predictor. The difference in BIC value between this model, and the model with the second-lowest associated BIC value is 223. This suggests that the Hurdle(C) model with latitude and longitude covariates provides the best fit to our data, of the models examined.

Case study detection probability values range from 0.047 (Tree Swallow) to 0.564 (Swainson's Thrush). Estimates of the maximum latent abundance N per species are provided in Appendix 3.C. We do not see any excessively large estimates for N or exceedingly small estimates for p, which suggests that while N-mixture models occasionally suffer from identifiability issues as described by Dennis et al. (2015), this does not appear to be an issue for this case study.

The estimates for bald eagle abundance produced by this model are plotted by site and year in Fig. 3.3. Of the 94 possible sites in Alaska, the bald eagle population is concentrated at 18 sites at the southeastern coast, along a 300-mile stretch of islands called the Alexander Archipelago. Examination of this figure suggested a possible increase in bald eagle abundance in this area between 2010 and 2019. The mean abundance was calculated per year (Fig. 3.4), and a one-sided Mann-Kendall test (Mann, 1945; Kendall, 1948) for an increasing trend in time series data was performed. The result of this was a Kendall's τ value of 0.6 and a p-value of 0.0082, indicating that it was appropriate to reject the null hypothesis that no increasing trend exists. We can therefore conclude that bald eagle abundances increased in the area of the Alexander Archipelago in the decade between 2010 and 2019.

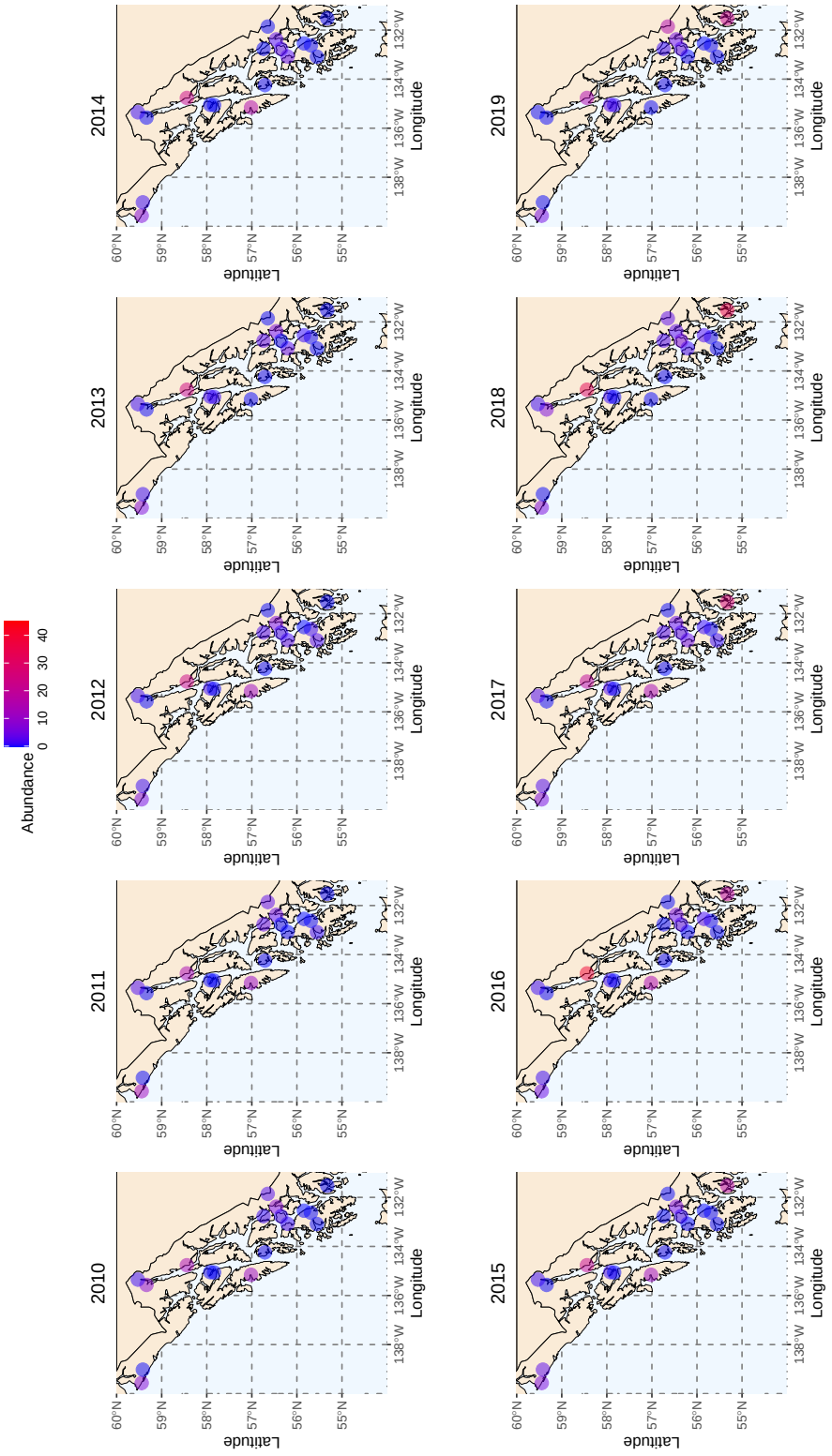


Figure 3.3: Estimated bald eagle abundances at sites in the Alexander Archipelago, produced by the Hurdle(C) model fitted to the North American Breeding Bird Survey data.

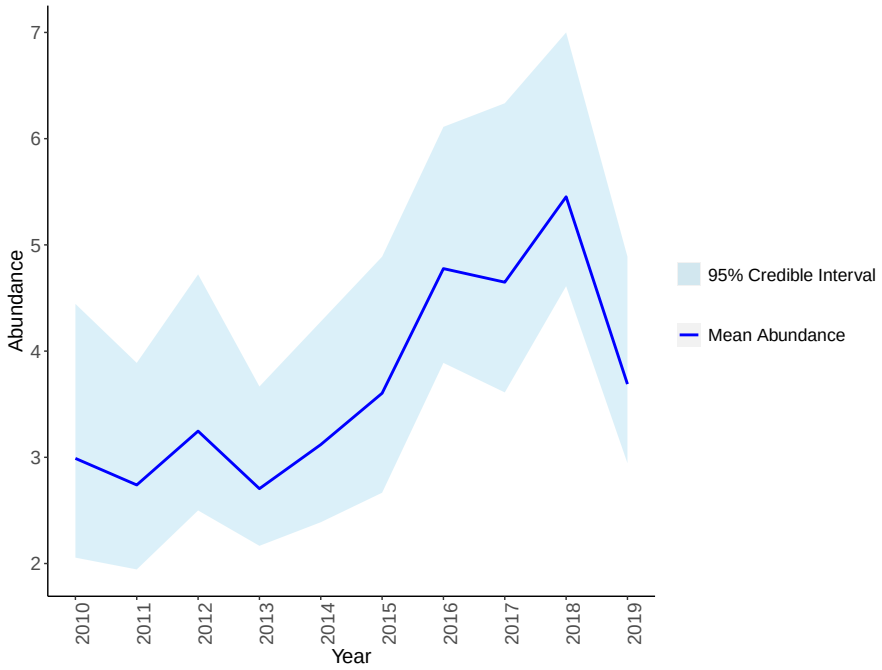


Figure 3.4: Annual posterior mean abundance for bald eagles in the Alexander Archipelago, estimated by the Hurdle(C) model fitted to the North American Breeding Bird Survey data. The light-blue ribbon represents the 95% credible interval for the mean.

In the models that contain an autoregressive component, we obtain separate correlations $\rho(N_s, N_{s'})$ per year. As a feature of model formulation (the autocorrelation coefficient $\mathbf{\Sigma}_{\phi}$ is currently a diagonal matrix) the correlation between two species do not change sign from year to year. We can accommodate a change in sign by allowing for an unstructured covariance matrix of the autocorrelation coefficient $\mathbf{\Sigma}_{\phi}$, and this particular extension is subject of ongoing work. Furthermore, the models presented in this paper assume that sites are independent of one another. A further extension we are currently working on is the incorporation of spatial dependence in abundance estimates, which aims to relax this assumption of site independence.

Appendix

3.A Simulation Study

In this section, we describe the simulation studies that were used to determine the accuracy of the estimates produced by the multi-species N-mixture models.

To determine if our modelling framework produces accurate estimates at contrasting sample sizes, a series of simulations were run in which we varied the number of sites, $R \in \{10, 100\}$, the number of sampling occasions, $T \in \{5, 10\}$, and the number of species observed, $S \in \{5, 10\}$. Within these simulations, we varied the detection probability p, and the mean number of individuals per site λ. Small values for p lay between 0.1 and 0.4, while large values for p lay between 0.5 and 0.9. Small values for λ had a median value of 7 and standard deviation of 10, while large values for λ had a median value of 55 and standard deviation of 74.

In the case of the Hurdle and Hurdle-AR models, we also varied the probability of a zero-count occurring, $\theta \in \{0.2, 0.7\}$. For each combination of parameters, we simulated 100 datasets and estimated N, $\boldsymbol{\Sigma}_a$, and p. We also estimated values for θ and ϕ, in the case of the Hurdle and AR models, respectively. Relative mean bias was calculated for the estimated probability of obtaining a zero count $\hat{\theta}$, autocorrelation coefficient $\hat{\phi}$, probability of detection p, and mean of the abundance random effects $\boldsymbol{\mu}_a$. The smaller the value for relative bias, the closer to the true value our estimated parameters were. We compared abundance estimates $\hat{N}$ to true values N using the concordance correlation coefficient (Lin, 1989), which is

given by the formula:

$$\rho_c = \frac{2\rho\hat{\sigma}\sigma}{\hat{\sigma^2} + \sigma^2 + (\hat{\mu} - \mu)^2},$$

where ρ is Pearson's correlation coefficient, σ and μ are the standard deviation and mean of the true values of N, and $\hat{\sigma}$ and $\hat{\mu}$ are the standard deviation and mean of the estimated values of N. The Pearson correlation coefficient is a measure of the strength of a linear association between two variables. However, the Pearson correlation is invariant under changes in location and scale. If two variables exhibit a linear relationship, but are very different in terms of their location or scale, the Pearson correlation coefficient will not reveal this. The concordance correlation coefficient, however, does take into account differences in location and scale. For this reason, the concordance correlation coefficient was chosen as a measure of the linear relationship between the true abundance and estimated abundance, rather than the Pearson correlation coefficient. The higher the value of the concordance correlation coefficient, the closer our estimates for N were to the true values.

We compared our estimated correlation matrix to the true value using the correlation matrix distance (Herdin et al., 2005), which is given by the following formula:

$$\text{CMD}(\mathbf{X_1}, \mathbf{X_2}) = 1 - \frac{\text{tr}(\mathbf{X_1}\mathbf{X_2})}{\|\mathbf{X_1}\|_f \|\mathbf{X_2}\|_f},$$

where $\mathbf{X_1}$ and $\mathbf{X_2}$ are two correlation matrices, $\text{tr}(\mathbf{X_1}\mathbf{X_2})$ is the trace of the product of these two matrices, and $\|.\|_f$ denotes the Frobenius norm.

Additionally, the coverage probabilities for each parameter were determined as the proportion of simulations in which the 50% credible interval contained the true parameter value. We expect that approximately 50% of the time, the estimated 50% credible interval for the parameter will contain the true value of that parameter (Appendix 3.D). Each of these scenarios were simulated 100 times. All data was simulated using the `R` statistical software version 4.0.2 (R Core Team, 2022), and all Bayesian models were implemented using the `R2jags` package (Su and Yajima, 2020).

3.A.1 Simulation Study Results

The results of the small-scale simulation study, which was composed of data simulated for five species at 10 sites, over five years, is shown in Table 3.A.1. The results of the large-scale simulation study, which contained 10 species, 100 sites and 10 years, is shown in Table 3.A.2.

3.A.1.1 MNM Model

The large-scale simulation study (Table 3.A.2) produced reliable estimates for latent abundance N at every combination of p and λ, with CCC values between 0.97 and 0.99. Estimates of N from the small-scale simulation study (Table 3.A.1) appear more dependent on the detection probability p, with greater CCC values associated with larger detection probabilities.

From Table 3.A.2, the relative bias for the estimate of p shows that when R, T and S are large, the model produces estimates for p which are accurate to two decimal places. When, R, T and S are small (Table 3.A.1), the relative bias for the estimate of p is larger for small median p. When R, T and S are small, larger values of p produce more reliable estimates of p.

Estimates for the correlation matrix and $\boldsymbol{\mu}_a$ improve with larger values of λ. In both Table 3.A.1 and Table 3.A.2, the relative bias for $\boldsymbol{\mu}_a$ and the CMD decrease when λ is larger. Larger values of R, T and S produce more accurate estimates of the inter-species correlations and $\boldsymbol{\mu}_a$, as can be seen by the decrease in the sizes of the CMD and RB($\boldsymbol{\mu}_a$) between Table 3.A.1 and Table 3.A.2.

Coverage probabilities (Appendix 3.D) for this model reveal that both small- and large-scale simulations produce parameters whose true value lie within the 50% credible interval approximately 50% of the time, as expected.

3.A.1.2 Autoregressive Model

At both small-scale simulations (Table 3.A.1) and large-scale simulations (Table 3.A.2), the autoregressive model produced reliable estimates for N, with CCC values above 0.9 for all simulations. Both the Table 3.A.1 and Table 3.A.2 con-

tain CMD values accurate to two decimal places. Relative bias for p decreases as median p increases. This can be observed for both small (Table 3.A.1) and large (Table 3.A.2) values of R, T, and S. In Table 3.A.1, relative bias for the autocorrelation coefficient ϕ is much larger when abundance is small. In this situation, the estimates for the autocorrelation coefficient ϕ cannot be relied upon. This is an issue that persists, though not as severely, as R, T and S increase in size in Table 3.A.2.

All parameters in this model have coverage probabilities of approximately 50%, as is expected for the 50% credible intervals (Appendix 3.D).

Median p	Median λ	θ	CCC	CMD	RB(p)	RB($\boldsymbol{\mu}_a$)	RB(θ)	RB(ϕ)
MNM								
0.3	7	-	0.7871	0.1037	0.2512	0.1269	-	-
0.3	55	-	0.8689	0.0641	0.2256	0.0554	-	-
0.8	7	-	0.9847	0.0768	0.0612	0.0379	-	-
0.8	55	-	0.9878	0.0522	0.0579	0.013	-	-
Hurdle								
0.3	7	0.2	0.772	0.146	0.437	0.2499	0.241	-
0.3	7	0.7	0.808	0.196	0.418	0.3485	0.116	-
0.3	55	0.2	0.799	0.081	0.418	0.1065	0.191	-
0.3	55	0.7	0.816	0.152	0.469	0.1194	0.134	-
0.8	7	0.2	0.960	0.114	0.071	0.1385	0.242	-
0.8	7	0.7	0.928	0.167	0.094	0.1923	0.130	-
0.8	55	0.2	0.927	0.074	0.085	0.0712	0.201	-
0.8	55	0.7	0.946	0.125	0.103	0.0811	0.114	-
AR								
0.3	7	-	0.9475	0.0638	0.1529	0.140	-	1.9807
0.3	55	-	0.9818	0.0515	0.1479	0.0663	-	0.0515
0.8	7	-	0.999	0.0652	0.0153	0.1149	-	1.8982
0.8	55	-	0.9999	0.0598	0.0168	0.0538	-	0.0598
Hurdle-AR								
0.3	7	0.2	0.9227	0.0602	0.1919	0.1475	0.1145	2.758
0.3	7	0.7	0.8899	0.1262	0.3228	0.2084	0.0329	12.365
0.3	55	0.2	0.9716	0.0709	0.1866	0.0752	0.0959	0.649
0.3	55	0.7	0.9223	0.0918	0.3575	0.0941	0.0356	4.206
0.8	7	0.2	0.994	0.0718	0.0386	0.1183	0.1024	2.088
0.8	7	0.7	0.974	0.0972	0.0787	0.1457	0.0316	6.459
0.8	55	0.2	0.9977	0.0636	0.0355	0.0497	0.1112	0.419
0.8	55	0.7	0.9322	0.1101	0.0986	0.0775	0.0308	5.505

Table 3.A.1: Results of small-scale simulation ($(R, T, S, K) = (10, 5, 5, 5)$): Concordance Correlation Coefficient (CCC) for the estimates of latent abundance N, Correlation Matrix Distance (CMD) for the estimate of the inter-species correlations, and relative biases for probability of detection (p), probability of obtaining a zero count (θ), and autocorrelation coefficient (ϕ).

Median p	Median λ	θ	CCC	CMD	RB(p)	RB($\boldsymbol{\mu}_a$)	RB(θ)	RB(ϕ)
MNM								
0.3	7	-	0.973	0.0315	0.0661	0.0163	-	-
0.3	55	-	0.990	0.0197	0.0613	0.0110	-	-
0.8	7	-	0.997	0.0261	0.0162	0.0104	-	-
0.8	55	-	0.999	0.0176	0.0171	0.0041	-	-
Hurdle								
0.3	7	0.2	0.953	0.049	0.183	0.0881	0.058	-
0.3	7	0.7	0.948	0.193	0.278	0.1310	0.016	-
0.3	55	0.2	0.964	0.022	0.192	0.0420	0.061	-
0.3	55	0.7	0.928	0.152	0.462	0.0879	0.015	-
0.8	7	0.2	0.997	0.041	0.019	0.0389	0.058	-
0.8	7	0.7	0.997	0.179	0.033	0.0569	0.015	-
0.8	55	0.2	0.999	0.023	0.024	0.018	0.046	-
0.8	55	0.7	0.999	0.206	0.034	0.0546	0.013	-
AR								
0.3	7	-	0.9923	0.0174	0.0520	0.0443	-	0.2799
0.3	55	-	0.9971	0.0182	0.0469	0.0192	-	0.0802
0.8	7	-	0.9997	0.0182	0.0044	0.0320	-	0.2433
0.8	55	-	0.9999	0.0204	0.0054	0.0160	-	0.0704
Hurdle-AR								
0.3	7	0.2	0.9937	0.02	0.0567	0.0417	0.0248	0.269
0.3	7	0.7	0.9892	0.0257	0.0679	0.0535	0.0109	0.434
0.3	55	0.2	0.9779	0.0143	0.0662	0.0223	0.0276	0.069
0.3	55	0.7	0.9678	0.0168	0.0943	0.0258	0.0128	0.109
0.8	7	0.2	0.9994	0.0178	0.0078	0.0335	0.0227	0.315
0.8	7	0.7	0.9992	0.0241	0.0096	0.0371	0.0123	0.396
0.8	55	0.2	0.9967	0.0187	0.0136	0.0158	0.0245	0.049
0.8	55	0.7	0.9976	0.0166	0.0179	0.0180	0.0179	0.112

Table 3.A.2: Results of large-scale simulation ($(R, T, S, K) = (100, 10, 10, 10)$): Concordance Correlation Coefficient (CCC) for the estimates of latent abundance N, Correlation Matrix Distance (CMD) for the estimate of the inter-species correlations, and relative biases for probability of detection (p), probability of obtaining a zero count (θ), and autocorrelation coefficient (ϕ).

3.A.1.3 Hurdle Model

Similar to the MNM model, when R, T and S are large (Table 3.A.2), consistently accurate estimates of latent abundance N are produced, with CCC values between

0.948 and 0.999. In Table 3.A.1 we see that CCC values depend more heavily on the detection probability, with more accurate estimates of N produced when detection probability is high. Both Table 3.A.1 and Table 3.A.2 show higher accuracy in estimates of the inter-species correlations when zero-inflation is small, and abundance is large. CMD values are greater when $\theta = 0.7$ or median $\lambda = 7$ than for $\theta = 0.2$ or median $\lambda = 55$. From both Table 3.A.1 and Table 3.A.2, the Hurdle model sees much smaller relative bias for p when median p is large compared to when median p is small. Relative bias for θ decreases when θ increases, indicating that θ is estimated with more accuracy when zero-inflation is large. Table 3.A.2 sees smaller relative bias for θ than Table 3.A.1, revealing that the strength of zero-inflation θ is estimated more accurately when R, T and S are large.

Issues with parameter convergence were encountered when fitting the Hurdle model. When zero-inflation and abundance are large, and detection probability is small, issues with convergence occurred in up to 20% of parameters. While this convergence issue does not appear to negatively affect the relative biases of parameter estimates, as can be seen in Table 3.A.1 and Table 3.A.2, coverage probabilities for detection probability p and random effect mean μ_a are negatively impacted (Appendix 3.D). We also see coverage for N which is larger than 50%. This is to be expected, and is due to zero counts being perfectly predicted.

3.A.1.4 Hurdle-Autoregressive Model

In Table 3.A.1, CCC values demonstrate that the Hurdle-AR model produces estimates for N with greater accuracy when the probability of obtaining a zero count is small. However, increasing R, T, and S (Table 3.A.2) reduces this dependence on θ, and all CCC values produced are greater than 0.95.

The small-scale simulation (Table 3.A.1) has CMD values and relative biases for p and $\boldsymbol{\mu}_a$ which increase when the probability of obtaining a zero count increases, indicating that the inter-species correlations, p and $\boldsymbol{\mu}_a$ are estimated more accurately when the degree of zero-inflation is low. The same is true for the large-scale simulation (Table 3.A.2), though the differences in CMD and relative biases be-

tween small θ and large θ are not as large, revealing that the increase in R, T and S renders the increase in zero-inflation less important in the estimation of these parameters.

The Hurdle-AR model suffers with the same issue estimating ϕ when the probability of obtaining a zero count is high. This issue is more severe in Table 3.A.1, and estimates of ϕ cannot be trusted when R, T and S are small but θ is large. Like the AR model, this issue is not as acute in Table 3.A.2, as an increase in R, T and S appears to compensate for the problems caused by large zero-inflation.

3.B Analytic Correlations

While the random effect a allows for the incorporation of the species-level correlations, in order to examine the correlations between species abundances $\rho(N_s, N_{s'})$ (rather than the correlations of the species-level random effects themselves), the analytic correlations must be derived. An `R` implementation of these analytic correlations is available in the public GitHub repository https://github.com/niamhmimnagh/mnm.

MNM Model

In this section, we present the analytical expressions for the correlation between the latent abundances (N_s and $N_{s'}$) for all $s \neq s'$ for the MNM model described in Section 3.3.1. For convenience of notation, we drop the dependence on i and t, $Y_s = (\{Y_{its}\})$, $N_s = (\{N_{is}\})$, $p_s = (\{p_{its}\})$, $\lambda_s = (\{\lambda_{is}\})$.

The MNM model in this case may be described as follows:

$$
\begin{aligned}
Y_s \mid N_s &\sim \text{Binomial}(N_s, p_s) \\
N_s \mid \lambda_s &\sim \text{Poisson}(\lambda_s) \\
a_s &\sim \text{Normal}_s(\mu, \Sigma) \\
\lambda_s &\sim \text{log-Normal}_s(e^{\mu_s + \frac{1}{2}\Sigma_{ss}}, e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})} e^{\Sigma_{ss'}} - 1)
\end{aligned}
$$

To begin our analysis we require the expectation, variance and covariance of the log-Normal parameter λ_s. These are as follows:

$$\mathrm{E}[\lambda_s] = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$$

$$\mathrm{Var}(\lambda_s) = e^{\mu_s + \mu_s + \frac{1}{2}(\Sigma_{ss} + \Sigma_{ss})}(e^{\Sigma_{ss}} - 1) = e^{2\mu_s + \Sigma_{ss}}(e^{\Sigma_{ss}} - 1)$$

$$\mathrm{Cov}(\lambda_s, \lambda_{s'}) = e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})} e^{\Sigma_{ss'}} - 1)$$

As N is described with a Poisson distribution, we may write the conditional expectation and variance directly:

$$\mathrm{E}[N_s \mid \lambda_s] = \mathrm{Var}(N_s \mid \lambda_s) = \lambda_s$$

The conditional expectation and variance of the binomial Y_s are given by:

$$\mathrm{E}[Y_s \mid N_s] = N_s p_s$$

$$\mathrm{Var}(Y_s \mid N_s) = N_s p_s (1 - p_s)$$

The unconditional expectation of N_s can be derived using the law of total expectation as follows:

$$\begin{aligned} \mathrm{E}[N_s] &= \mathrm{E}_{\lambda_s}(\mathrm{E}_{N_s}(N_s \mid \lambda_s)) \\ &= \mathrm{E}_{\lambda_s}(\lambda_s) \\ &= e^{\mu_s + \frac{1}{2}\Sigma_{ss}} \end{aligned}$$

The unconditional expectation of Y_s may be similarly derived:

$$\begin{aligned} \mathrm{E}[Y_s] &= \mathrm{E}_{N_s}(\mathrm{E}_{Y_s}(Y_s \mid N_s)) \\ &= \mathrm{E}_{N_s}(N_s p_s) \\ &= \lambda_s p_s \end{aligned}$$

We can then derive the unconditional variance of N_s using the law of total variance:

$$\begin{aligned}\mathrm{Var}(N_s) &= \mathrm{E}_{\lambda_s}(\mathrm{Var}_{N_s}(N_s \mid \lambda_s)) + \mathrm{Var}_{\lambda_s}(\mathrm{E}_{N_s}(\mathrm{N}_s \mid \lambda_s)) \\ &= \mathrm{E}_{\lambda_s}(\lambda_s)\mathrm{Var}_{\lambda_s}(\lambda_s) \\ &= e^{\mu_s+\frac{1}{2}\Sigma_{ss}} + e^{2\mu_s+\Sigma_{ss}}(e^{\Sigma_{ss}} - 1)\end{aligned}$$

Similarly, the unconditional variance of Y_s can be derived using the law of total variance:

$$\begin{aligned}\mathrm{Var}(Y_s) &= \mathrm{E}_{N_s}(\mathrm{Var}_{Y_s}(Y_s \mid N_s)) + \mathrm{Var}_{N_s}(\mathrm{E}_{Y_s}(Y_s \mid N_s)) \\ &= \mathrm{E}_{N_s}(N_s p_s(1-p_s)) + \mathrm{Var}_{N_s}(N_s p_s) \\ &= \lambda_s p_s(1-p_s) + \lambda_s p_s^2 \\ &= \lambda_s p_s - \lambda_s p_s^2 + \lambda_s p_s^s \\ &= \lambda_s p_s\end{aligned}$$

Finally, the unconditional covariance between N_s and $N_{s'}$ can be derived using the law of total covariance:

$$\mathrm{Cov}(N_s, N_{s'}) = \mathrm{E}_{\lambda_s\lambda_{s'}}(\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'}) + \mathrm{Cov}_{\lambda_s\lambda_{s'}}(\mathrm{E}_{Ns}(N_s \mid \lambda_s), \mathrm{E}_{N_{s'}}(N_{s'} \mid \lambda_{s'}))$$

We assume that, given the correlated effects $\mathbf{a}$, the abundances are independent, which means:

$$\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'}) = 0$$

The result of this is as follows:

$$\begin{aligned}
\mathrm{Cov}(N_s, N_{s'}) &= \mathrm{Cov}_{\lambda_s \lambda_{s'}}(\mathrm{E}_{Ns}(N_s \mid \lambda_s), \mathrm{E}_{N_{s'}}(N_{s'} \mid \lambda_{s'})) \\
&= \mathrm{Cov}(\lambda_s, \lambda_{s'}) \\
&= e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})}(e^{\Sigma_{ss'}} - 1) \\
&= (e^{\mu_s + \frac{1}{2}\Sigma_{ss}})(e^{\mu_{s'} + \frac{1}{2}\Sigma_{s's'}})(e^{\Sigma_{ss'}} - 1) \\
&= \mathrm{E}[N_s]\mathrm{E}[N_{s'}](e^{\Sigma_{ss'}} - 1)
\end{aligned}$$

The correlation between latent abundances N_s and $N_{s'}$ can then be easily computed using the following formula:

$$\rho(N_s, N_{s'}) = \frac{\mathrm{Cov}(N_s, N_{s'})}{\sqrt{\mathrm{Var}(N_s), \mathrm{Var}(N_{s'})}}$$

The unconditional covariance between observed abundances Y_s and $Y_{s'}$ can be derived using the law of total covariance:

$$\mathrm{Cov}(Y_s, Y_{s'}) = \mathrm{E}_{N_s N_{s'}}(\mathrm{Cov}(Y_s, Y_{s'} \mid N_s, N_{s'}) + \mathrm{Cov}_{N_s N_{s'}}(\mathrm{E}_{Y_s}(Y_s \mid N_s), \mathrm{E}_{Y_{s'}}(Y_{s'} \mid N_{s'}))$$

We assume that the observed abundances are independent given the actual abundances, i.e.

$$\mathrm{Cov}(Y_s, Y_{s'} \mid N_s, N_{s'}) = 0$$

This results in the following simplification of the unconditional covariance between observed abundances Y_s and $Y_{s'}$:

$$\begin{aligned}
\mathrm{Cov}(Y_s, Y_{s'}) &= \mathrm{Cov}_{N_s N_{s'}}(\mathrm{E}_{Y_s}(Y_s \mid N_s), \mathrm{E}_{Y_{s'}}(Y_{s'} \mid N_{s'})) \\
&= \mathrm{Cov}_{N_s, N_{s'}}(N_s p_s, N_{s'} p_{s'}) \\
&= \mathrm{Cov}_{N_s, N_{s'}}(N_s, N_{s'}) p_s p_{s'}
\end{aligned}$$

We have shown previously that $\text{Cov}(N_s, N_{s'}) = \text{E}[N_s]\text{E}[N_{s'}](e^{\Sigma_{ss'}} - 1)$

Therefore

$$\text{Cov}(Y_s, Y_{s'}) = \text{E}[N_s]\text{E}[N_{s'}](e^{\Sigma_{ss'}} - 1)p_s p_{s'}$$

$$\rho(Y_s, Y_{s'}) = \rho(N_s, N_{s'})\sqrt{p_s p_{s'}}$$

$$\mid \rho(Y_s, Y_{s'}) \mid \leq \mid \rho(N_s, N_{s'}) \mid$$

We may therefore conclude that the correlation between latent abundances N_s and $N_{s'}$ can be written as follows:

$$\rho(N_s, N_{s'}) = \frac{(e^{\mu_s + \frac{1}{2}\Sigma_{ss}})(e^{\mu_{s'} + \frac{1}{2}\Sigma_{s's'}})(e^{\Sigma_{ss'}} - 1)}{(\sqrt{e^{\mu_s + \frac{1}{2}\Sigma_{ss}} + e^{2\mu_s + \Sigma_{ss}}(e^{\Sigma_{ss}} - 1)})(e^{\mu_{s'} + \frac{1}{2}\Sigma_{s's'}} + e^{2\mu_{s'} + \Sigma_{s's'}}(e^{\Sigma_{s's'}} - 1))}$$

Hurdle Model

In this section, we present the analytical expressions for the correlation between the latent abundances (N_s and $N_{s'}$) for all $s \neq s'$ for the MNM hurdle-Poisson model described in Section 3.3.2. For convenience of notation, we drop the dependence on i and t, $Y_s = (\{Y_{its}\})$, $N_s = (\{N_{is}\})$, $p_s = (\{p_{its}\})$, $\lambda_s = (\{\lambda_{is}\})$.

The MNM hurdle-Poisson model in this case may be described as follows:

$$\begin{aligned}
Y_s \mid N_s &\sim \text{Binomial}(N_s, p_s) \\
N_s &\sim \text{Hurdle Poisson}(\lambda_s, \theta) \\
\lambda_s &= e^{a_s} \\
a_s &\sim \text{Normal}_s(\mu, \Sigma) \\
\lambda_s &\sim \text{log-Normal}_s(e^{\mu_s + \frac{1}{2}\Sigma_{ss}}, e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})}(e^{\Sigma_{ss'}} - 1))
\end{aligned}$$

The analytical expressions for the correlation between latent abundances for this model are derived by following similar steps to those followed in the previous section. The major difference encountered here is the need to approximate a number of parameters using Taylor expansions. For completeness, all steps necessary to derive these analytical correlations will be provided here.

To begin our analysis we require the expectation, variance and covariance of the log-Normal parameter λ_s. These are as follows:

$$\mathrm{E}[\lambda_s] = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$$

$$\mathrm{Var}(\lambda_s) = e^{\mu_s + \mu_s + \frac{1}{2}(\Sigma_{ss} + \Sigma_{ss})} e^{\Sigma_{ss}} - 1) = e^{2\mu_s + \Sigma_{ss}}(e^{\Sigma_{ss}} - 1)$$

$$\mathrm{Cov}(\lambda_s, \lambda_{s'}) = e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})} e^{\Sigma_{ss'}} - 1)$$

As the latent abundance N_s follows a Hurdle Poisson distribution, the conditional expectation and variance are as follows:

$$\mathrm{E}[N_s \mid \lambda_s] = \frac{(1-\theta)\lambda_s}{1 - e^{-\lambda_s}}$$

$$\mathrm{Var}(N_s \mid \lambda_s) = (1-\theta)\left[\frac{\lambda_s}{1 - e^{-\lambda_s}} - e^{\lambda_s}(\frac{\lambda_s}{1 - e^{-\lambda_s}})^2\right] + \theta(1-\theta)(\frac{\lambda_s}{1 - e^{-\lambda_s}})^2$$

The conditional expectation and variance of the binomial observed abundance Y_s is given by:

$$\mathrm{E}[Y_s \mid N_s] = N_s p_s$$
$$\mathrm{Var}(Y_s \mid N_s) = N_s p_s (1 - p_s)$$

The unconditional expectation of N_s can be derived using the law of total expectation as follows:

$$\begin{aligned}\mathrm{E}[N_s] &= \mathrm{E}_{\lambda_s}[\mathrm{E}_{N_s}[N_s \mid \lambda_s]] \\ &= \mathrm{E}_{\lambda_s}\left[\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right]\end{aligned}$$

This is a function f of the random variable λ_s. We can approximate the expected value of this function using second-order Taylor expansions around the point $\lambda_s = \mu_{\lambda_s}$ (where μ_{λ_s} is the expected value of λ_s) as follows:

$$\begin{aligned}f(\lambda_s) &\approx f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2 \\ \mathrm{E}[f(\lambda_s)] &\approx \mathrm{E}[f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2] \\ &= \mathrm{E}[f(\mu_{\lambda_s})] + f_{\lambda_s}(\mu_{\lambda_s})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})] + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})^2] \\ &= f(\mu_{\lambda_s}) + \frac{1}{2} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})\sigma^2_{\lambda_s}\end{aligned}$$

Where $\sigma^s_{\lambda_s}$ is the variance of the expected abundance λ_s

In this case we have:

$$\begin{aligned}f(\lambda_s) &= \frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}} \\ f_{\lambda_s,\lambda_s}(\lambda_s) &= \frac{e^{-\lambda_s}(e^{-\lambda_s}\lambda_s + \lambda_s + 2e^{-\lambda_s} - 2)(1-\theta)}{(1-e^{-\lambda_s})^3} \\ \mathrm{E}[\lambda_s] &= e^{\mu_s + \frac{1}{2}\Sigma_{ss}} \\ \mathrm{Var}(\lambda_s) &= (e^{2\mu_s + \Sigma_{ss}})(e^{\Sigma_{ss}} - 1)\end{aligned}$$

So:

$$\mathrm{E}[N_s] \approx \frac{(1-\theta)(\mu_{\lambda_s})}{1-e^{-\mu_{\lambda_s}}} + \frac{\sigma^2_{\lambda_s}}{2}\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}$$

where $\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$ and $\sigma^2_{\lambda_s} = (e^{2\mu_s + \Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

The unconditional variance of N_s can be similarly derived. We begin by using the law of total variance:

$$\begin{aligned}
\mathrm{Var}(N_s) &= \mathrm{E}_{\lambda_s}[\mathrm{Var}_{N_s}(N_s \mid \lambda_s)] + \mathrm{Var}_{\lambda_s}(\mathrm{E}_{N_s}[N_s \mid \lambda_s]) \\
&= \mathrm{E}_{\lambda_s}\left[(1-\theta)\left[\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{\lambda_s}(\frac{\lambda_s}{1-e^{-\lambda_s}})^2\right] + \theta(1-\theta)(\frac{\lambda_s}{1-e^{-\lambda_s}})^2)\right] + \mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right)
\end{aligned}$$

The variance of a function f of variable λ_s can also be approximated at the point μ_{λ_s} using a second-order Taylor expansion as follows:

$$\begin{aligned}
f(\lambda_s) &\approx f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2 \\
\mathrm{Var}[f(\lambda_s)] &\approx \mathrm{Var}[f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2] \\
&= \mathrm{Var}[f(\mu_{\lambda_s})] + f_{\lambda_s}(\mu_{\lambda_s})^2 \mathrm{Var}[(\lambda_s - \mu_{\lambda_s})] + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2 \mathrm{Var}[(\lambda_s - \mu_{\lambda_s})^2] \\
&= f_{\lambda_s}(\mu_{\lambda_s})^2 \mathrm{Var}(\lambda_s) + \frac{1}{2} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2 \mathrm{Var}(\lambda_s^2 - 2\mu_{\lambda_s}\lambda_s + \mu_{\lambda_s}) \\
&= f_{\lambda_s}(\mu_{\lambda_s})^2 \mathrm{Var}(\lambda_s) + \frac{1}{2} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2 \mathrm{Var}(\lambda_s^2 - 2\mu_{\lambda_s}\lambda_s) \\
&= f_{\lambda_s}(\mu_{\lambda_s})^2 \mathrm{Var}(\lambda_s) + \frac{1}{2} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2 (\mathrm{Var}(\lambda_s^2) + 4\mu^2_{\lambda_s}\mathrm{Var}(\lambda_s) - 4\mu_{\lambda_s}\mathrm{Cov}(\lambda_s^2, \lambda_s))
\end{aligned}$$

The variance of N_s can be approximated using two separate Taylor series expansions:

For the first Taylor series approximation let:

$$
\begin{aligned}
f(\lambda_s) &= (1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{-\lambda_s}\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2\right) + \theta(1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2 \\
f_{\lambda_s\lambda_s}(\lambda_s) &= \frac{e^{-2\lambda_s}(-e^{-\lambda_s}\lambda_s^2 - 4\lambda_s^2 - e^{\lambda_s}\lambda_s^2 + 5e^{\lambda_s}\lambda_s - 5e^{-\lambda_s}\lambda_s - 4e^{\lambda_s} - 4e^{-\lambda_s} + 8)(1-\theta)}{(1-e^{-\lambda_s})^4} \\
&\quad + \frac{2\theta(2e^{-2\lambda_s}\lambda_s^2 + e^{-\lambda_s}\lambda_s^2 + 4e^{-2\lambda_s}\lambda_s - 4e^{-\lambda_s}\lambda_s + e^{-2\lambda_s} - 2e^{-\lambda_s} + 1)(1-\theta)}{(1-e^{-\lambda_s})^4} \\
\mathrm{E}[\lambda_s] &= e^{\mu_s + \frac{1}{2}\Sigma_{ss}} \\
\mathrm{Var}(\lambda_s) &= (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}} - 1)
\end{aligned}
$$

We then have:

$$
\begin{aligned}
&\mathrm{E}_{\lambda_s}\left((1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{-\lambda_s}\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2\right) + \theta(1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2\right) \\
&\approx (1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}} - e^{-\mu_{\lambda_s}}\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2\right) + \theta(1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2 \\
&+ \frac{\sigma_{\lambda_s}^2}{2}\left(\frac{e^{-2\mu_{\lambda_s}}(-e^{-\mu_{\lambda_s}}\mu_{\lambda_s}^2 - 4\mu_{\lambda_s}^2 - e^{-\mu_{\lambda_s}}\mu_{\lambda_s}^2 + 5e^{\mu_{\lambda_s}}\mu_{\lambda_s} - 5e^{-\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{\mu_{\lambda_s}} - 4e^{-\mu_{\lambda_s}} + 8)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right. \\
&\left. + \frac{2\theta(2e^{-2\mu_{\lambda_s}}\mu_{\lambda_s}^2 + e^{-\mu_{\lambda_s}}\mu_{\lambda_s}^2 + 4e^{-2\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + e^{-2\mu_{\lambda_s}} - 2e^{-\mu_{\lambda_s}} + 1)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right)
\end{aligned}
$$

Where $\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$ and $\sigma_{\lambda_s}^2 = (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

For the second Taylor series approximation let:

$$
\begin{aligned}
f(\lambda_s) &= \frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}} \\
f_{\lambda_s}(\lambda_s) &= \frac{(1-e^{-\lambda_s} - e^{-\lambda_s}\lambda_s)(1-\theta)}{(1-e^{-\lambda_s})^2} \\
f_{\lambda_s,\lambda_s}(\lambda_s) &= \frac{e^{-\lambda_s}(e^{-\lambda_s}\lambda_s + \lambda_s + 2e^{-\lambda_s} - 2)(1-\theta)}{(1-e^{-\lambda_s})^3} \\
\mathrm{E}[\lambda_s] &= e^{\mu_s + \frac{1}{2}\Sigma_{ss}} \\
\mathrm{Var}(\lambda_s) &= (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}} - 1)
\end{aligned}
$$

$$\mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right) \approx \left(\frac{(1-e^{-\mu_{\lambda_s}}-e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^2}\right)^2 \sigma^2_{\lambda_s}$$
$$+\frac{1}{2}\left(\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s}+\mu_{\lambda_s}+2e^{-\mu_{\lambda_s}}-2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}\right)^2 (\mathrm{Var}(\lambda_s^2)+4\mu^2_{\lambda_s}\sigma^2_{\lambda_s}-4\mu_{\lambda_s}\mathrm{Cov}(\lambda_s^2,\lambda_s))$$

Where $\mu_{\lambda_s} = e^{\mu_s+\frac{1}{2}\Sigma_{ss}}$ and $\sigma^2_{\lambda_s} = (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}}-1)$

We know

$$\mathrm{Var}(\lambda_s^2) = \mathrm{E}[(\lambda_s^2 - \mathrm{E}[\lambda_s^2])^2] = \mathrm{E}[\lambda_s^4] - \mathrm{E}[\lambda_s^2]^2$$

$$\mathrm{Cov}(\lambda_s^2, \lambda_s) = \mathrm{E}[\lambda_s^3] - \mathrm{E}[\lambda_s^2]\mathrm{E}[\lambda_s]$$

Now we need the following moments about the origin: $\mathrm{E}[\lambda_s^4]$ and $\mathrm{E}[\lambda_s^3]$

We can find both of these moments about the origin using their respective central moments:

$$\mathrm{E}[(\lambda_s-\mu_{\lambda_s})^3] = (\sqrt{e^{\Sigma_{ss}}})(2+e^{\Sigma_{ss}})$$
$$\mathrm{E}[(\lambda_s-\mu_{\lambda_s})^4] = -3+3e^{2\Sigma_{ss}}+2e^{3\Sigma_{ss}}+e^{4\Sigma_{ss}}$$

Therefore:

$$\mathrm{E}[\lambda_s^3] = \mathrm{E}[(\lambda_s-\mu_{\lambda_s})^3] + 3\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^2] - 2(\mathrm{E}[\lambda_s])^3$$
$$\mathrm{E}[\lambda_s^4] = \mathrm{E}[(\lambda_s-\mu_{\lambda_s})^4] + 4\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^3] - 6(\mathrm{E}[\lambda_s])^2\mathrm{E}[\lambda_s^2] + 3(\mathrm{E}[\lambda_s])^4$$

$$
\begin{aligned}
\mathrm{Var}(\lambda_s^2) &= \mathrm{E}[\lambda_s^4] - \mathrm{E}[\lambda_s^2]^2 \\
&= \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^4] + 4\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^3] - 6(\mathrm{E}[\lambda_s])^2\mathrm{E}[\lambda_s^2] + 3(\mathrm{E}[\lambda_s])^4 - \mathrm{E}[\lambda_s^2]^2 \\
&= \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^4] + 4\mathrm{E}[\lambda_s](\mathrm{E}[(\lambda_s - \mu_{\lambda_s})^3 + 3\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^2 - 2\mathrm{E}[\lambda_s]^2) \\
&- 6(\mathrm{E}[\lambda_s])^2\mathrm{E}[\lambda_s^2] + 3(\mathrm{E}[\lambda_s])^4 - \mathrm{E}[\lambda_s^2]^2 \\
&= -3 + 3e^{2\Sigma_{ss}} + 2e^{3\Sigma_{ss}} + e^{4\Sigma_{ss}} + 4\mu_{\lambda_s}((\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_{\lambda_s}^3) \\
&- 6\mu_{\lambda_s}^2(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) + 3\mu_{\lambda_s}^4 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(\lambda_s^2, \lambda_s) &= \mathrm{E}[\lambda_s^3] - \mathrm{E}[\lambda_s^2]\mathrm{E}[\lambda_s] \\
&= \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^3] + 3\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^2] - 2(\mathrm{E}[\lambda_s])^3 - \mathrm{E}[\lambda_s^2]\mathrm{E}[\lambda_s] \\
&= (\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\mathrm{Var}(\lambda_s) + \mathrm{E}[\lambda_s]^2) - 2\mu_{\lambda}^3 - (\mathrm{Var}(\lambda_s) + \mathrm{E}[\lambda_s]^2)\mu_{\lambda_s} \\
&= (\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_{\lambda}^3 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)\mu_{\lambda_s}
\end{aligned}
$$

Thus we have:

$$
\begin{aligned}
\mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1 - e^{-\lambda_s}}\right) \approx &\left(\frac{(1 - e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)}{(1 - e^{-\mu_{\lambda_s}})^2}\right)^2 \sigma_{\lambda_s}^2 \\
&+ \frac{1}{2}\left(\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1 - e^{-\mu_{\lambda_s}})^3}\right)^2 \\
&\times (-3 + 3e^{2\Sigma_{ss}} + 2e^{3\Sigma_{ss}} + e^{4\Sigma_{ss}} + 4\mu_{\lambda_s}((\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_{\lambda}^3 \\
&- 6\mu_{\lambda_s}^2(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) + 3\mu_{\lambda_s}^4 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)^2 + 4\mu_{\lambda_s}^2\sigma_{\lambda_s}^2 \\
&- 4\mu_{\lambda_s}(\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_{\lambda}^3 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)\mu_{\lambda_s})
\end{aligned}
$$

Putting these together gives:

$$
\begin{aligned}
&\mathrm{Var}(N_s) = \mathrm{E}_{\lambda_s}(\mathrm{Var}_{N_s}(N_s \mid \lambda_s)) + \mathrm{Var}_{\lambda_s}(\mathrm{E}_{N_s}(N_s \mid \lambda_s)) \\
&= \mathrm{E}_{\lambda_s}\left((1-\theta)\left[\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{\lambda_s}(\frac{\lambda_s}{1-e^{-\lambda_s}})^2\right] + \theta(1-\theta)(\frac{\lambda_s}{1-e^{-\lambda_s}})^2)\right) + \mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right) \\
&\approx (1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}} - e^{-\mu_{\lambda_s}}\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2\right) + \theta(1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2 \\
&+ \frac{\sigma^2_{\lambda_s}}{2}\left(\frac{e^{-2\mu_{\lambda_s}}(-e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} - 4\mu^2_{\lambda_s} - e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} + 5e^{\mu_{\lambda_s}}\mu_{\lambda_s} - 5e^{-\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{\mu_{\lambda_s}} - 4e^{-\mu_{\lambda_s}} + 8)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right. \\
&\left. + \frac{2\theta(2e^{-2\mu_{\lambda_s}}\mu^2_{\lambda_s} + e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} + 4e^{-2\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + e^{-2\mu_{\lambda_s}} - 2e^{-\mu_{\lambda_s}} + 1)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right) \\
&+ \left(\frac{(1 - e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^2}\right)^2 \sigma^2_{\lambda_s} \\
&+ \frac{1}{2}\left(\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}\right)^2 \\
&\times (-3 + 3e^{2\Sigma_{ss}} + 2e^{3\Sigma_{ss}} + e^{4\Sigma_{ss}} + 4\mu_{\lambda_s}((\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma^2_{\lambda_s} + \mu^2_{\lambda_s}) - 2\mu^3_{\lambda_s}) \\
&- 6\mu^2_{\lambda_s}(\sigma^2_{\lambda_s} + \mu^2_{\lambda_s}) + 3\mu^4_{\lambda_s} - (\sigma^2_{\lambda_s} + \mu^2_{\lambda_s})^2 + 4\mu^2_{\lambda_s}\sigma^2_{\lambda_s} \\
&- 4\mu_{\lambda_s}(\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma^2_{\lambda_s} + \mu^2_{\lambda_s}) - 2\mu^3_{\lambda} - (\sigma^2_{\lambda_s} + \mu^2_{\lambda_s})\mu_{\lambda_s})
\end{aligned}
$$

Where $\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$ and $\sigma^2_{\lambda_s} = (e^{2\mu_s + \Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

Finally, the unconditional covariance between N_s and $N_{s'}$ can be derived using the law of total covariance:

$$\mathrm{Cov}(N_s, N_{s'}) = \mathrm{E}_{\lambda_s\lambda_{s'}}(\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'}) + \mathrm{Cov}_{\lambda_s\lambda_{s'}}(\mathrm{E}_{Ns}(N_s \mid \lambda_s), \mathrm{E}_{N_{s'}}(N_{s'} \mid \lambda_{s'}))$$

We assume that, given λ, the abundances are independent: $\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'}) = 0$

Therefore:

$$
\begin{aligned}
\mathrm{Cov}(N_s, N_{s'}) &= \mathrm{Cov}_{\lambda_s\lambda_{s'}}(\mathrm{E}_{Ns}(N_s \mid \lambda_s), \mathrm{E}_{N_{s'}}(N_{s'} \mid \lambda_{s'})) \\
&= \mathrm{Cov}_{\lambda_s,\lambda_{s'}}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}, \frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right)
\end{aligned}
$$

To find this covariance we can use the fact that

$$\mathrm{Cov}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}, \frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right) = \mathrm{E}\left[\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}} \times \frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right] - \mathrm{E}\left[\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right]\mathrm{E}\left[\frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right]$$

A second-order Taylor series expansion in two variables near the point $(\mu_{\lambda_s}, \mu_{\lambda_s})$ has the following form:

$$\begin{aligned} f(\lambda_s, \lambda_{s'}) \approx\ & f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) + f_{\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s}) + f_{\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}}) \\ & + \frac{1}{2}(f_{\lambda_s\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})^2 + f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}})^2) \\ & + f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}}) \end{aligned}$$

$$\begin{aligned} \mathrm{E}[f(\lambda_s, \lambda_{s'})] \approx\ & \mathrm{E}[f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) + f_{\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s}) + f_{\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}}) \\ & + \frac{1}{2}(f_{\lambda_s\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})^2 + f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}})^2) \\ & + f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}})] \\ =\ & \mathrm{E}[f(\mu_{\lambda_s}, \mu_{\lambda_{s'}})] + f_{\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})] + f_{\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_{s'} - \mu_{\lambda_{s'}})] \\ & + \frac{1}{2}f_{\lambda_s\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})]^2 + \frac{1}{2}f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_{s'} - \mu_{\lambda_{s'}})]^2 \\ & + f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}})] \\ =\ & f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) + \frac{1}{2}f_{\lambda_s,\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{Var}(\lambda_s) + \frac{1}{2}f_{\lambda_{s'},\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{Var}(\lambda_{s'}) \\ & + f_{\lambda_s,\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{Cov}(\lambda_s, \lambda_{s'}) \end{aligned}$$

As $\mathrm{E}[\lambda_s - \mu_{\lambda_s}] = 0$, $\mathrm{E}[(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}})] = \mathrm{Cov}(\lambda_s, \lambda_{s'})$, $\mathrm{E}[(\lambda_s - \mu_{\lambda_s})^2] = \mathrm{Var}(\lambda)$

We know $\mathrm{Var}(\lambda_s)$ and $\mathrm{Cov}(\lambda_s, \lambda_{s'})$ from the lognormal distribution, so now we need the second derivatives of $f(\lambda_s, \lambda_{s'})$

$$
\begin{aligned}
f(\lambda_s, \lambda_{s'}) &= \left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right)\left(\frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right) \\
f_{\lambda_s\lambda_s}(\lambda_s\lambda_{s'}) &= \frac{e^{-\lambda_s}\lambda_{s'}(1-\theta)^2(e^{-\lambda_s}\lambda_s + \lambda_s + 2e^{-\lambda_s} - 2)}{(1-e^{-\lambda_{s'}})(1-e^{-\lambda_s})^3} \\
f_{\lambda_{s'}\lambda_{s'}}(\lambda_s, \lambda_{s'}) &= \frac{e^{-\lambda_{s'}}\lambda_s(1-\theta)^2(e^{-\lambda_{s'}}\lambda_{s'} + \lambda_{s'} + 2e^{-\lambda_{s'}} - 2)}{(1-e^{-\lambda_s})(1-e^{-\lambda_{s'}})^3} \\
f_{\lambda_s\lambda_{s'}}(\lambda_s, \lambda_{s'}) &= \frac{(1-e^{-\lambda_s} - e^{-\lambda_s}\lambda_s)(1-\theta)^2(1-e^{-\lambda_{s'}} - e^{-\lambda_{s'}}\lambda_{s'})}{(1-e^{-\lambda_s})^2(1-e^{-\lambda_{s'}})^2}
\end{aligned}
$$

Evaluating these at the point $(\mu_{\lambda_s}, \mu_{\lambda_{s'}})$ gives:

$$
\begin{aligned}
f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) &= \left(\frac{(1-\theta)\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)\left(\frac{(1-\theta)\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}}\right) \\
f_{\lambda_s\lambda_s}(\mu_{\lambda_s}\mu_{\lambda_{s'}}) &= \frac{e^{-\mu_{\lambda_s}}\mu_{\lambda_{s'}}(1-\theta)^2(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)}{(1-e^{-\mu_{\lambda_{s'}}})(1-e^{-\mu_{\lambda_s}})^3} \\
f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) &= \frac{e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_s}(1-\theta)^2(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)}{(1-e^{-\mu_{\lambda_s}})(1-e^{-\mu_{\lambda_{s'}}})^3} \\
f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) &= \frac{(1-e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)^2(1-e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})}{(1-e^{-\mu_{\lambda_s}})^2(1-e^{-\mu_{\lambda_{s'}}})^2}
\end{aligned}
$$

where $\mu_{\lambda_s} = \mathrm{E}[\lambda_s] = e^{\mu_s + \Sigma_{ss}}$, from the log-normal distribution.

We then have:

$$
\begin{aligned}
\mathrm{E}[f(\lambda_s, \lambda_{s'})] \approx{}& \left(\frac{(1-\theta)\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)\left(\frac{(1-\theta)\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}}\right) \\
&+ \frac{\sigma_{\lambda_s}^2}{2}\frac{e^{-\mu_{\lambda_s}}\mu_{\lambda_{s'}}(1-\theta)^2(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)}{(1-e^{-\mu_{\lambda_{s'}}})(1-e^{-\mu_{\lambda_s}})^3} \\
&+ \frac{\sigma_{\lambda_{s'}}^2}{2}\frac{e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_s}(1-\theta)^2(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)}{(1-e^{-\mu_{\lambda_s}})(1-e^{-\mu_{\lambda_{s'}}})^3} \\
&+ \frac{\Sigma_{\lambda_s,\lambda_{s'}}(1-e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)^2(1-e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})}{(1-e^{-\mu_{\lambda_s}})^2(1-e^{-\mu_{\lambda_{s'}}})^2}
\end{aligned}
$$

Where:

$\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$,

$\sigma^2_{\lambda_s} = (e^{2\mu_s + \Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

$\Sigma_{\lambda_s,\lambda_{s'}} = e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})}(e^{\Sigma_{ss'}} - 1)$ are the mean, variance and covariance of the log-normal distribution.

$$
\begin{aligned}
\mathrm{Cov}(N_s, N_{s'}) = &\left(\frac{(1-\theta)\mu_{\lambda_s}}{1 - e^{-\mu_{\lambda_s}}}\right)\left(\frac{(1-\theta)\mu_{\lambda_{s'}}}{1 - e^{-\mu_{\lambda_{s'}}}}\right) \\
&+ \frac{\sigma_{\lambda_s^2}}{2}\frac{e^{-\mu_{\lambda_s}}\mu_{\lambda_{s'}}(1-\theta)^2(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)}{(1 - e^{-\mu_{\lambda_{s'}}})(1 - e^{-\mu_{\lambda_s}})^3} \\
&+ \frac{\sigma_{\lambda_{s'}^2}}{2}\frac{e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_s}(1-\theta)^2(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)}{(1 - e^{-\mu_{\lambda_s}})(1 - e^{-\mu_{\lambda_{s'}}})^3} \\
&+ \frac{\Sigma_{\lambda_s,\lambda_{s'}}(1 - e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)^2(1 - e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})}{(1 - e^{-\mu_{\lambda_s}})^2(1 - e^{-\mu_{\lambda_{s'}}})^2} \\
&- \left(\frac{(1-\theta)(\mu_{\lambda_s})}{1 - e^{-\mu_{\lambda_s}}} + \frac{\sigma^2_{\lambda_s}}{2}\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1 - e^{-\mu_{\lambda_s}})^3}\right) \\
&\times \left(\frac{(1-\theta)(\mu_{\lambda_{s'}})}{1 - e^{-\mu_{\lambda_{s'}}}} + \frac{\sigma^2_{\lambda_{s'}}}{2}\frac{e^{-\mu_{\lambda_{s'}}}(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)(1-\theta)}{(1 - e^{-\mu_{\lambda_{s'}}})^3}\right)
\end{aligned}
$$

The correlation between species' abundances can then be calculated using:

$$\rho(N_s, N_{s'}) = \frac{\mathrm{Cov}(N_s, N_{s'})}{\sqrt{\mathrm{Var}(N_s), \mathrm{Var}(N_{s'})}}$$

$$\rho(N_s, N_{s'}) = \left(\frac{(1-\theta)\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)\left(\frac{(1-\theta)\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}}\right) + \frac{\sigma_{\lambda_s^2}}{2}\frac{e^{-\mu_{\lambda_s}}\mu_{\lambda_{s'}}(1-\theta)^2(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)}{(1-e^{-\mu_{\lambda_{s'}}})(1-e^{-\mu_{\lambda_s}})^3}$$
$$+ \frac{\sigma_{\lambda_{s'}^2}}{2}\frac{e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_s}(1-\theta)^2(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)}{(1-e^{-\mu_{\lambda_s}})(1-e^{-\mu_{\lambda_{s'}}})^3}$$
$$+ \frac{\Sigma_{\lambda_s,\lambda_{s'}}(1-e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)^2(1-e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})}{(1-e^{-\mu_{\lambda_s}})^2(1-e^{-\mu_{\lambda_{s'}}})^2}$$
$$- \left(\frac{(1-\theta)(\mu_{\lambda_s})}{1-e^{-\mu_{\lambda_s}}} + \frac{\sigma_{\lambda_s}^2}{2}\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}\right)$$
$$\times \left(\frac{(1-\theta)(\mu_{\lambda_{s'}})}{1-e^{-\mu_{\lambda_{s'}}}} + \frac{\sigma_{\lambda_{s'}}^2}{2}\frac{e^{-\mu_{\lambda_{s'}}}(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_{s'}}})^3}\right)$$
$$\times \left[(1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}} - e^{-\mu_{\lambda_s}}\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2\right) + \theta(1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2\right.$$
$$+ \frac{\sigma_{\lambda_s}^2}{2}\left(\frac{e^{-2\mu_{\lambda_s}}(-e^{-\mu_{\lambda_s}}\mu_{\lambda_s}^2 - 4\mu_{\lambda_s}^2 - e^{-\mu_{\lambda_s}}\mu_{\lambda_s}^2 + 5e^{\mu_{\lambda_s}}\mu_{\lambda_s} - 5e^{-\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{\mu_{\lambda_s}} - 4e^{-\mu_{\lambda_s}} + 8)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right.$$
$$\left.\left.+ \frac{2\theta(2e^{-2\mu_{\lambda_s}}\mu_{\lambda_s}^2 + e^{-\mu_{\lambda_s}}\mu_{\lambda_s}^2 + 4e^{-2\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + e^{-2\mu_{\lambda_s}} - 2e^{-\mu_{\lambda_s}} + 1)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right)\right.$$
$$+ \left(\frac{(1-e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^2}\right)^2\sigma_{\lambda_s}^2 + \frac{1}{2}\left(\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}\right)^2$$
$$\times (-3 + 3e^{2\Sigma_{ss}} + 2e^{3\Sigma_{ss}} + e^{4\Sigma_{ss}} + 4\mu_{\lambda_s}((\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_{\lambda_s}^3)$$
$$- 6\mu_{\lambda_s}^2(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) + 3\mu_{\lambda_s}^4 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)^2 + 4\mu_{\lambda_s}^2\sigma_{\lambda_s}^2$$
$$\left.- 4\mu_{\lambda_s}(\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_{\lambda}^3 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)\mu_{\lambda_s})\right]$$
$$\times \left[(1-\theta)\left(\frac{\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}} - e^{-\mu_{\lambda_{s'}}}\left(\frac{\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}}\right)^2\right) + \theta(1-\theta)\left(\frac{\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}}\right)^2\right.$$
$$+ \frac{\sigma_{\lambda_{s'}}^2}{2}\left(\frac{e^{-2\mu_{\lambda_{s'}}}(-e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}}^2 - 4\mu_{\lambda_{s'}}^2 - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}}^2 + 5e^{\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} - 5e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} - 4e^{\mu_{\lambda_{s'}}} - 4e^{-\mu_{\lambda_{s'}}} + 8)(1-\theta)}{(1-e^{-\mu_{\lambda_{s'}}})^4}\right.$$
$$\left.\left.+ \frac{2\theta(2e^{-2\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}}^2 + e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}}^2 + 4e^{-2\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} - 4e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + e^{-2\mu_{\lambda_{s'}}} - 2e^{-\mu_{\lambda_{s'}}} + 1)(1-\theta)}{(1-e^{-\mu_{\lambda_{s'}}})^4}\right)\right.$$
$$+ \left(\frac{(1-e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})(1-\theta)}{(1-e^{-\mu_{\lambda_{s'}}})^2}\right)^2\sigma_{\lambda_{s'}}^2 + \frac{1}{2}\left(\frac{e^{-\mu_{\lambda_{s'}}}(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_{s'}}})^3}\right)^2$$
$$\times (-3 + 3e^{2\sigma_{s's'}} + 2e^{3\sigma_{s's'}} + e^{4\sigma_{s's'}} + 4\mu_{\lambda_{s'}}((\sqrt{e^{\sigma_{s's'}}})(2 + e^{\sigma_{s's'}}) + 3\mu_{\lambda_{s'}}(\sigma_{\lambda_{s'}}^2 + \mu_{\lambda_{s'}}^2) - 2\mu_{\lambda_{s'}}^3)$$
$$- 6\mu_{\lambda_{s'}}^2(\sigma_{\lambda_{s'}}^2 + \mu_{\lambda_{s'}}^2) + 3\mu_{\lambda_{s'}}^4 - (\sigma_{\lambda_{s'}}^2 + \mu_{\lambda_{s'}}^2)^2 + 4\mu_{\lambda_{s'}}^2\sigma_{\lambda_{s'}}^2$$
$$\left.\left.- 4\mu_{\lambda_{s'}}(\sqrt{e^{\sigma_{s's'}}})(2 + e^{\sigma_{s's'}}) + 3\mu_{\lambda_{s'}}(\sigma_{\lambda_{s'}}^2 + \mu_{\lambda_{s'}}^2) - 2\mu_{\lambda}^3 - (\sigma_{\lambda_{s'}}^2 + \mu_{\lambda_{s'}}^2)\mu_{\lambda_{s'}})\right)^{\frac{1}{2}}\right]^{-1}$$

AR Model

In this section, we present the analytical expressions for the correlation between the latent abundances (N_s and $N_{s'}$) for all $s \neq s'$ for the MNM AR model described in Section 3.3.3. In this model, indices are separated using commas to make it easier for the autoregressive component to be read and understood.

The MNM-AR model may be described as follows:

$$\begin{aligned} Y_{i,t,k,s} \mid N_{i,k,s} &\sim \text{Binomial}(N_{i,k,s}, p_{i,s}) \\ N_{i,k,s} \mid \lambda_{i,k,s} &\sim \text{Poisson}(\lambda_{i,k,s}) \\ \log(\lambda_{i,k,s}) &= a_{i,s} + \phi_s \times \log(N_{i,k-1,s} + 1) \\ \phi &\sim \text{MVN}_s(\mu_\phi, \Sigma_\phi) \end{aligned}$$

with $\Sigma_\phi = 100(I_S)$ and $\mu_\phi = [0, 0, ..., 0]^T$ of length S, where I_S is the identity matrix of dimension S.

We first need to determine the mean and covariance associated with the log of the mean abundance λ. This may be achieved as follows.

$$\begin{aligned} \mu_{\log(\lambda_{i,k-1,s})} &= \text{E}[a + \log(N_{i,k-1,s} + 1)\phi] \\ &= \text{E}[a] + \log(N_{i,k-1,s} + 1)\text{E}[\phi] \\ &= \mu_a + \log(N_{i,k-1,s} + 1)\mu_\phi \end{aligned}$$

$$\begin{aligned} \Sigma_{\log(\lambda_{i,k-1,s})} &= \text{Cov}(a + \phi\log(N_{i,k-1,s} + 1), a + \phi \times \log(N_{i,k-1,s} + 1)) \\ &= \text{Cov}(a) + 2\log(N_{i,k-1,s} + 1)\text{Cov}(a, \phi) + \log(N_{i,k-1,s} + 1)^2\text{Cov}(\phi) \\ &= \Sigma_a + \log(N_{i,k-1,s} + 1)^2\Sigma_\phi \end{aligned}$$

In the models described in Appendices above, we have $\log(\lambda_{i,s}) \sim \text{MVN}(\mu_a, \Sigma_a)$, where μ_a is an $S \times 1$ vector and Σ_a is an SxS matrix. This means that though $\log(\lambda)$ varies by site, the analytic covariance matrix does not.

However, in the AR Model we have:

$$\log(\lambda) \sim \text{MVN}(\mu_a + \log(N_{i,k-1,s}+1)\mu_\phi, \Sigma_a + \log(N_{i,k-1,s}+1)^2\Sigma_\phi)$$

At year k=1, the mean and covariance are the same at every site:

$$\mu_{\log(\lambda_{i,1,s})} = \mu_a$$

$$\Sigma_{\log(\lambda_{i,1,s})} = \Sigma_a$$

At k > 1 the mean and covariance are affected by the $\log(N_{i,k-1,s}+1)$ term, which means we have separate SxS matrices for each combination of site and year.

Thus:

$$\lambda \sim \text{MVLN}(e^{\mu_s+\frac{1}{2}\Sigma_{ss}}, (e^{\mu_s+\mu_{s'}+\frac{1}{2}(\Sigma_{ss}+\Sigma_{s's'})})(e^{\Sigma_{ss'}}-1))$$

where $\mu = \mu_a + \log(N_{i,k-1,s}+1)\mu_\phi$

and $\Sigma = \Sigma_a + \log(N_{i,k-1,s}+1)\Sigma_\phi$

The rest of the analytic correlations for the AR model are carried out similarly to those of the MNM model, with the above substitutions for μ and Σ. For completeness, we will provide all steps necessary to obtain the analytic correlations here.

As N is Poisson distributed, we can write the conditional expectation and variance:

$$\text{E}[N_s \mid \lambda_s] = \text{Var}(N_s \mid \lambda_s) = \lambda_s$$

The conditional expectation and variance of the binomial Y_s is given by:

$\text{E}[Y_s \mid N_s] = N_s p_s$

$\text{Var}(Y_s \mid N_s) = N_s p_s(1-p_s)$

The unconditional expectation of N_s can be derived using the law of total expectation as follows:

$$\begin{aligned}
\mathrm{E}[N_s] &= \mathrm{E}_{\lambda_s}[\mathrm{E}_{N_s}[N_s \mid \lambda_s]] \\
&= \mathrm{E}_{\lambda_s}[\lambda_s] \\
&= e^{\mu_s + \frac{1}{2}\Sigma_{ss}}
\end{aligned}$$

The unconditional expectation of Y_s can be similarly derived:

$$\begin{aligned}
\mathrm{E}[Y_s] &= \mathrm{E}_{N_s}[\mathrm{E}_{Y_s}[Y_s \mid N_s]] \\
&= \mathrm{E}_{N_s}[N_s p_s] \\
&= \lambda_s p_s
\end{aligned}$$

The unconditional variance of N_s can be derived using the law of total variance:

$$\begin{aligned}
\mathrm{Var}(N_s) &= \mathrm{E}_{\lambda_s}[\mathrm{Var}_{N_s}(N_s \mid \lambda_s)] + \mathrm{Var}_{\lambda_s}(\mathrm{E}_{N_s}[N_s \mid \lambda_s]) \\
&= \mathrm{E}_{\lambda_s}[\lambda_s]\mathrm{Var}_{\lambda_s}(\lambda_s) \\
&= e^{\mu_s + \frac{1}{2}\Sigma_{ss}} + e^{2\mu_s + \Sigma_{ss}}(e^{\Sigma_{ss}} - 1)
\end{aligned}$$

Similarly, the unconditional variance of Y_s can be derived using the law of total variance:

$$\begin{aligned}
\mathrm{Var}(Y_s) &= \mathrm{E}_{N_s}[\mathrm{Var}_{Y_s}(Y_s \mid N_s)] + \mathrm{Var}_{N_s}(\mathrm{E}_{Y_s}[Y_s \mid N_s]) \\
&= \mathrm{E}_{N_s}[N_s p_s(1 - p_s)] + \mathrm{Var}_{N_s}(N_s p_s) \\
&= \lambda_s p_s(1 - p_s) + \lambda_s p_s^2 \\
&= \lambda_s p_s - \lambda_s p_s^2 + \lambda_s p_s^s \\
&= \lambda_s p_s
\end{aligned}$$

The unconditional covariance between N_s and $N_{s'}$ can be derived using the law of total covariance:

$$\mathrm{Cov}(N_s, N_{s'}) = \mathrm{E}_{\lambda_s\lambda_{s'}}[\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'})] + \mathrm{Cov}_{\lambda_s\lambda_{s'}}(\mathrm{E}_{Ns}[N_s \mid \lambda_s], \mathrm{E}_{N_{s'}}[N_{s'} \mid \lambda_{s'}])$$

We assume, given $\mathbf{a}$, abundances are independent: $\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'}) = 0$, so:

$$\begin{aligned}
\mathrm{Cov}(N_s, N_{s'}) &= \mathrm{Cov}_{\lambda_s\lambda_{s'}}(\mathrm{E}_{Ns}[N_s \mid \lambda_s], \mathrm{E}_{N_{s'}}[N_{s'} \mid \lambda_{s'}]) \\
&= \mathrm{Cov}(\lambda_s, \lambda_{s'}) \\
&= e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})}(e^{\Sigma_{ss'}} - 1) \\
&= (e^{\mu_s + \frac{1}{2}\Sigma_{ss}})(e^{\mu_{s'} + \frac{1}{2}\Sigma_{s's'}})(e^{\Sigma_{ss'}} - 1) \\
&= \mathrm{E}[N_s]\mathrm{E}[N_{s'}](e^{\Sigma_{ss'}} - 1)
\end{aligned}$$

Then the correlation between N_s and $N_{s'}$ can be easily computed using:

$$\rho(N_s, N_{s'}) = \frac{\mathrm{Cov}(N_s, N_{s'})}{\sqrt{\mathrm{Var}(N_s), \mathrm{Var}(N_{s'})}}$$

$$\rho(N_s, N_{s'}) = \frac{(e^{\mu_s + \frac{1}{2}\Sigma_{ss}})(e^{\mu_{s'} + \frac{1}{2}\Sigma_{s's'}})(e^{\Sigma_{ss'}} - 1)}{(\sqrt{e^{\mu_s + \frac{1}{2}\Sigma_{ss}} + e^{2\mu_s + \Sigma_{ss}}(e^{\Sigma_{ss}} - 1)})(e^{\mu_{s'} + \frac{1}{2}\Sigma_{s's'}} + e^{2\mu_{s'} + \Sigma_{s's'}}(e^{\Sigma_{s's'}} - 1))}$$

where $\mu = \mu_a + \log(N_{i,k-1,s} + 1)\mu_\phi$, and $\Sigma = \Sigma_a + \log(N_{i,k-1,s} + 1)\Sigma_\phi$

Hurdle-AR Model

$$
\begin{aligned}
Y_{i,t,k,s} \mid N_{i,k,s} &\sim \text{Binomial}(N_{i,k,s}, p_s) \\
N_{i,k,s} &\sim \text{Poisson-Hurdle}(\lambda_{i,k,s}, \theta) \\
\log(\lambda_{i,k,s}) &= a_{i,s} + \log(N_{i,k-1,s}+1)\phi_s \\
a_{i,s} &\sim \text{Normal}_s(\mu_a, \Sigma_a) \\
\phi_{i,s} &\sim \text{Normal}_s(\mu_\phi, \Sigma_\phi) \\
\lambda_s &\sim \text{logNormal}_s(e^{\mu_i+\frac{1}{2}\Sigma_{ii}}, e^{\mu_i+\mu_j+\frac{1}{2}(\Sigma_{ii}+\Sigma_{jj})}(e^{\Sigma_{ij}}-1))
\end{aligned}
$$

Where $\mu = \mu_a + \log(N_{i,k-1,s}+1)\mu_\phi$ and $\Sigma = \Sigma_a + \log(N_{i,k-1,s}+1)\Sigma_\phi$, from the AR analytic correlations. The hurdle model and AR model can be combined through the substitution of these μ and Σ from the AR model into the Hurdle model approximations below.

First we need the expectation, variance and covariance of the log-normal λ_s

$$E[\lambda_s] = e^{\mu_s+\frac{1}{2}\Sigma_{ss}}$$

$$\text{Var}(\lambda_s) = e^{\mu_s+\mu_s+\frac{1}{2}(\Sigma_{ss}+\Sigma_{ss})}e^{\Sigma_{ss}}-1) = e^{2\mu_s+\Sigma_{ss}}(e^{\Sigma_{ss}}-1)$$

$$\text{Cov}(\lambda_s, \lambda_{s'}) = e^{\mu_s+\mu_{s'}+\frac{1}{2}(\Sigma_{ss}+\Sigma_{s's'})}(e^{\Sigma_{ss'}}-1)$$

As N follows a Poisson-Hurdle distribution, the conditional expectation and variance are as follows:

$$E[N_s \mid \lambda_s] = \frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}$$

$$\mathrm{Var}(N_s \mid \lambda_s) = (1-\theta)\left[\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{\lambda_s}(\frac{\lambda_s}{1-e^{-\lambda_s}})^2\right] + \theta(1-\theta)(\frac{\lambda_s}{1-e^{-\lambda_s}})^2$$

The conditional expectation and variance of the binomial Y_s is given by:

$$\begin{aligned} E[Y_s \mid N_s] &= N_s p_s \\ \mathrm{Var}(Y_s \mid N_s) &= N_s p_s (1-p_s) \end{aligned}$$

The unconditional expectation of N_s can be derived using the law of total expectation as follows:

$$\begin{aligned} E[N_s] &= \mathrm{E}_{\lambda_s}(\mathrm{E}_{N_s}(N_s \mid \lambda_s)) \\ &= \mathrm{E}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right) \end{aligned}$$

This is a function f of the random variable λ_s. We can approximate the expected value of this function using second-order Taylor expansions around the point $\lambda_s = \mu_{\lambda_s}$ (where μ_{λ_s} is the expected value of λ_s) as follows:

$$\begin{aligned} f(\lambda_s) &\approx f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2 \\ \mathrm{E}[f(\lambda_s)] &\approx \mathrm{E}[f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2] \\ &= \mathrm{E}[f(\mu_{\lambda_s})] + f_{\lambda_s}(\mu_{\lambda_s})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})] + \frac{1}{2!} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})^2] \\ &= f(\mu_{\lambda_s}) + \frac{1}{2} f_{\lambda_s\lambda_s}(\mu_{\lambda_s})\sigma^2_{\lambda_s} \end{aligned}$$

Where $\sigma^s_{\lambda_s}$ is the variance of λ_s

In this case we have:

$$
\begin{aligned}
f(\lambda_s) &= \frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}} \\
f_{\lambda_s,\lambda_s}(\lambda_s) &= \frac{e^{-\lambda_s}(e^{-\lambda_s}\lambda_s + \lambda_s + 2e^{-\lambda_s} - 2)(1-\theta)}{(1-e^{-\lambda_s})^3} \\
\mathrm{E}[\lambda_s] &= e^{\mu_s + \frac{1}{2}\Sigma_{ss}} \\
\mathrm{Var}(\lambda_s) &= (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}} - 1)
\end{aligned}
$$

So:

$$
E[N_s] \approx \frac{(1-\theta)(\mu_{\lambda_s})}{1-e^{-\mu_{\lambda_s}}} + \frac{\sigma^2_{\lambda_s}}{2}\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}
$$

where $\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$ and $\sigma^2_{\lambda_s} = (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

The unconditional variance of N_s can be similarly derived. We begin by using the law of total variance:

$$
\begin{aligned}
\mathrm{Var}(N_s) &= \mathrm{E}_{\lambda_s}(\mathrm{Var}_{N_s}(N_s \mid \lambda_s)) + \mathrm{Var}_{\lambda_s}(\mathrm{E}_{N_s}(N_s \mid \lambda_s)) \\
&= \mathrm{E}_{\lambda_s}\left((1-\theta)\left[\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{\lambda_s}(\frac{\lambda_s}{1-e^{-\lambda_s}})^2\right] + \theta(1-\theta)(\frac{\lambda_s}{1-e^{-\lambda_s}})^2)\right) + \mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right)
\end{aligned}
$$

The variance of a function f of variable λ_s can also be approximated at the point μ_{λ_s} using a second-order Taylor expansion as follows:

$$
\begin{aligned}
f(\lambda_s) &\approx f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!}f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2 \\
Var[f(\lambda_s)] &\approx Var[f(\mu_{\lambda_s}) + f_{\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s}) + \frac{1}{2!}f_{\lambda_s\lambda_s}(\mu_{\lambda_s})(\lambda_s - \mu_{\lambda_s})^2] \\
&= Var[f(\mu_{\lambda_s})] + f_{\lambda_s}(\mu_{\lambda_s})^2 Var[(\lambda_s - \mu_{\lambda_s})] + \frac{1}{2!}f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2 Var[(\lambda_s - \mu_{\lambda_s})^2] \\
&= f_{\lambda_s}(\mu_{\lambda_s})^2\mathrm{Var}(\lambda_s) + \frac{1}{2}f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2\mathrm{Var}(\lambda_s^2 - 2\mu_{\lambda_s}\lambda_s + \mu_{\lambda_s}) \\
&= f_{\lambda_s}(\mu_{\lambda_s})^2\mathrm{Var}(\lambda_s) + \frac{1}{2}f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2\mathrm{Var}(\lambda_s^2 - 2\mu_{\lambda_s}\lambda_s) \\
&= f_{\lambda_s}(\mu_{\lambda_s})^2\mathrm{Var}(\lambda_s) + \frac{1}{2}f_{\lambda_s\lambda_s}(\mu_{\lambda_s})^2(\mathrm{Var}(\lambda_s^2) + 4\mu^2_{\lambda_s}\mathrm{Var}(\lambda_s) - 4\mu_{\lambda_s}\mathrm{Cov}(\lambda_s^2, \lambda_s))
\end{aligned}
$$

The variance of N_s can be approximated using two separate Taylor series expansions. For the first Taylor series approximation let:

$$
\begin{aligned}
f(\lambda_s) &= (1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{-\lambda_s}\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2\right) + \theta(1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2 \\
f_{\lambda_s\lambda_s}(\lambda_s) &= \frac{e^{-2\lambda_s}(-e^{-\lambda_s}\lambda_s^2 - 4\lambda_s^2 - e^{\lambda_s}\lambda_s^2 + 5e^{\lambda_s}\lambda_s - 5e^{-\lambda_s}\lambda_s - 4e^{\lambda_s} - 4e^{-\lambda_s} + 8)(1-\theta)}{(1-e^{-\lambda_s})^4} \\
&\quad + \frac{2\theta(2e^{-2\lambda_s}\lambda_s^2 + e^{-\lambda_s}\lambda_s^2 + 4e^{-2\lambda_s}\lambda_s - 4e^{-\lambda_s}\lambda_s + e^{-2\lambda_s} - 2e^{-\lambda_s} + 1)(1-\theta)}{(1-e^{-\lambda_s})^4} \\
\mathrm{E}[\lambda_s] &= e^{\mu_s + \frac{1}{2}\Sigma_{ss}} \\
\mathrm{Var}(\lambda_s) &= (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}} - 1)
\end{aligned}
$$

We then have:

$$
\begin{aligned}
&\mathrm{E}_{\lambda_s}\left((1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{-\lambda_s}\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2\right) + \theta(1-\theta)\left(\frac{\lambda_s}{1-e^{-\lambda_s}}\right)^2\right) \\
&\approx (1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}} - e^{-\mu_{\lambda_s}}\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2\right) + \theta(1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2 \\
&+ \frac{\sigma^2_{\lambda_s}}{2}\left(\frac{e^{-2\mu_{\lambda_s}}(-e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} - 4\mu^2_{\lambda_s} - e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} + 5e^{\mu_{\lambda_s}}\mu_{\lambda_s} - 5e^{-\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{\mu_{\lambda_s}} - 4e^{-\mu_{\lambda_s}} + 8)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right. \\
&\left. + \frac{2\theta(2e^{-2\mu_{\lambda_s}}\mu^2_{\lambda_s} + e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} + 4e^{-2\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + e^{-2\mu_{\lambda_s}} - 2e^{-\mu_{\lambda_s}} + 1)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right)
\end{aligned}
$$

Where $\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$ and $\sigma^2_{\lambda_s} = (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

For the second Taylor series approximation let:

$$
\begin{aligned}
f(\lambda_s) &= \frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}} \\
f_{\lambda_s}(\lambda_s) &= \frac{(1-e^{-\lambda_s}-e^{-\lambda_s}\lambda_s)(1-\theta)}{(1-e^{-\lambda_s})^2} \\
f_{\lambda_s,\lambda_s}(\lambda_s) &= \frac{e^{-\lambda_s}(e^{-\lambda_s}\lambda_s+\lambda_s+2e^{-\lambda_s}-2)(1-\theta)}{(1-e^{-\lambda_s})^3} \\
\mathrm{E}[\lambda_s] &= e^{\mu_s+\frac{1}{2}\Sigma_{ss}} \\
\mathrm{Var}(\lambda_s) &= (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}}-1)
\end{aligned}
$$

$$
\mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right) \approx \left(\frac{(1-e^{-\mu_{\lambda_s}}-e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^2}\right)^2 \sigma^2_{\lambda_s}
$$
$$
+\frac{1}{2}\left(\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s}+\mu_{\lambda_s}+2e^{-\mu_{\lambda_s}}-2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}\right)^2 (\mathrm{Var}(\lambda_s^2)+4\mu^2_{\lambda_s}\sigma^2_{\lambda_s}-4\mu_{\lambda_s}\mathrm{Cov}(\lambda_s^2,\lambda_s))
$$

Where $\mu_{\lambda_s} = e^{\mu_s+\frac{1}{2}\Sigma_{ss}}$ and $\sigma^2_{\lambda_s} = (e^{2\mu_s+\Sigma_{ss}})(e^{\Sigma_{ss}}-1)$

$$
\mathrm{Var}(\lambda_s^2) = \mathrm{E}[(\lambda_s^2-\mathrm{E}[\lambda_s^2])^2] = \mathrm{E}[\lambda_s^4]-\mathrm{E}[\lambda_s^2]^2
$$

and

$$
\mathrm{Cov}(\lambda_s^2,\lambda_s) = \mathrm{E}[\lambda_s^3]-\mathrm{E}[\lambda_s^2]\mathrm{E}[\lambda_s]
$$

Now we need $E[\lambda_s^4]$ and $\mathrm{E}[\lambda_s^3]$

We can find both of these moments about the origin using their respective central moments.

$\mathrm{E}[(\lambda_s-\mu_{\lambda_s})^3] = (\sqrt{e^{\Sigma_{ss}}})(2+e^{\Sigma_{ss}})$

$\mathrm{E}[(\lambda_s-\mu_{\lambda_s})^4] = -3+3e^{2\Sigma_{ss}}+2e^{3\Sigma_{ss}}+e^{4\Sigma_{ss}}$

$$\mathrm{E}[\lambda_s^3] = \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^3] + 3\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^2] - 2(\mathrm{E}[\lambda_s])^3$$
$$\mathrm{E}[\lambda_s^4] = \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^4] + 4\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^3] - 6(\mathrm{E}[\lambda_s])^2\mathrm{E}[\lambda_s^2] + 3(\mathrm{E}[\lambda_s])^4$$

$$\begin{aligned}
\mathrm{Var}(\lambda_s^2) &= \mathrm{E}[\lambda_s^4] - \mathrm{E}[\lambda_s^2]^2 \\
&= \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^4] + 4\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^3] - 6(\mathrm{E}[\lambda_s])^2\mathrm{E}[\lambda_s^2] + 3(\mathrm{E}[\lambda_s])^4 - \mathrm{E}[\lambda_s^2]^2 \\
&= \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^4] + 4\mathrm{E}[\lambda_s](\mathrm{E}[(\lambda_s - \mu_{\lambda_s})^3 + 3\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^2 - 2\mathrm{E}[\lambda_s]^2) \\
&- 6(\mathrm{E}[\lambda_s])^2\mathrm{E}[\lambda_s^2] + 3(\mathrm{E}[\lambda_s])^4 - \mathrm{E}[\lambda_s^2]^2 \\
&= -3 + 3e^{2\Sigma_{ss}} + 2e^{3\Sigma_{ss}} + e^{4\Sigma_{ss}} + 4\mu_{\lambda_s}((\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_{\lambda_s}^3) \\
&- 6\mu_{\lambda_s}^2(\sigma_{\lambda_s}^2 + mu_{\lambda_s}^2) + 3\mu_{\lambda_s}^4 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)^2
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\lambda_s^2, \lambda_s) &= \mathrm{E}[\lambda_s^3] - \mathrm{E}[\lambda_s^2]\mathrm{E}[\lambda_s] \\
&= \mathrm{E}[(\lambda_s - \mu_{\lambda_s})^3] + 3\mathrm{E}[\lambda_s]\mathrm{E}[\lambda_s^2] - 2(\mathrm{E}[\lambda_s])^3 - \mathrm{E}[\lambda_s^2]\mathrm{E}[\lambda_s] \\
&= (\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\mathrm{Var}(\lambda_s) + E[\lambda_s]^2) - 2\mu_\lambda^3 - (\mathrm{Var}(\lambda_s) + E[\lambda_s]^2)\mu_{\lambda_s} \\
&= (\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_\lambda^3 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)\mu_{\lambda_s}
\end{aligned}$$

Thus we have:

$$\begin{aligned}
\mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1 - e^{-\lambda_s}}\right) \approx &\left(\frac{(1 - e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)}{(1 - e^{-\mu_{\lambda_s}})^2}\right)^2 \sigma_{\lambda_s}^2 \\
&+ \frac{1}{2}\left(\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1 - e^{-\mu_{\lambda_s}})^3}\right)^2 \\
&\times (-3 + 3e^{2\Sigma_{ss}} + 2e^{3\Sigma_{ss}} + e^{4\Sigma_{ss}} + 4\mu_{\lambda_s}((\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_\lambda^3 \\
&- 6\mu_{\lambda_s}^2(\sigma_{\lambda_s}^2 + mu_{\lambda_s}^2) + 3\mu_{\lambda_s}^4 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)^2 + 4\mu_{\lambda_s}^2\sigma_{\lambda_s}^2 \\
&- 4\mu_{\lambda_s}(\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2) - 2\mu_\lambda^3 - (\sigma_{\lambda_s}^2 + \mu_{\lambda_s}^2)\mu_{\lambda_s})
\end{aligned}$$

Putting these together gives:

$$
\begin{aligned}
&\mathrm{Var}(N_s) = \mathrm{E}_{\lambda_s}(\mathrm{Var}_{N_s}(N_s \mid \lambda_s)) + \mathrm{Var}_{\lambda_s}(\mathrm{E}_{N_s}(N_s \mid \lambda_s)) \\
&= \mathrm{E}_{\lambda_s}\left((1-\theta)\left[\frac{\lambda_s}{1-e^{-\lambda_s}} - e^{\lambda_s}(\frac{\lambda_s}{1-e^{-\lambda_s}})^2\right] + \theta(1-\theta)(\frac{\lambda_s}{1-e^{-\lambda_s}})^2)\right) + \mathrm{Var}_{\lambda_s}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right) \\
&\approx (1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}} - e^{-\mu_{\lambda_s}}\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2\right) + \theta(1-\theta)\left(\frac{\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)^2 \\
&+ \frac{\sigma^2_{\lambda_s}}{2}\left(\frac{e^{-2\mu_{\lambda_s}}(-e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} - 4\mu^2_{\lambda_s} - e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} + 5e^{\mu_{\lambda_s}}\mu_{\lambda_s} - 5e^{-\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{\mu_{\lambda_s}} - 4e^{-\mu_{\lambda_s}} + 8)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right. \\
&\left. + \frac{2\theta(2e^{-2\mu_{\lambda_s}}\mu^2_{\lambda_s} + e^{-\mu_{\lambda_s}}\mu^2_{\lambda_s} + 4e^{-2\mu_{\lambda_s}}\mu_{\lambda_s} - 4e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + e^{-2\mu_{\lambda_s}} - 2e^{-\mu_{\lambda_s}} + 1)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^4}\right) \\
&+ \left(\frac{(1 - e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^2}\right)^2 \sigma^2_{\lambda_s} \\
&+ \frac{1}{2}\left(\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1-e^{-\mu_{\lambda_s}})^3}\right)^2 \\
&\times (-3 + 3e^{2\Sigma_{ss}} + 2e^{3\Sigma_{ss}} + e^{4\Sigma_{ss}} + 4\mu_{\lambda_s}((\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma^2_{\lambda_s} + \mu^2_{\lambda_s}) - 2\mu^3_{\lambda_s}) \\
&- 6\mu^2_{\lambda_s}(\sigma^2_{\lambda_s} + mu^2_{\lambda_s}) + 3\mu^4_{\lambda_s} - (\sigma^2_{\lambda_s} + \mu^2_{\lambda_s})^2 + 4\mu^2_{\lambda_s}\sigma^2_{\lambda_s} \\
&- 4\mu_{\lambda_s}(\sqrt{e^{\Sigma_{ss}}})(2 + e^{\Sigma_{ss}}) + 3\mu_{\lambda_s}(\sigma^2_{\lambda_s} + \mu^2_{\lambda_s}) - 2\mu^3_{\lambda} - (\sigma^2_{\lambda_s} + \mu^2_{\lambda_s})\mu_{\lambda_s})
\end{aligned}
$$

Where $\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$ and $\sigma^2_{\lambda_s} = (e^{2\mu_s + \Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

Finally, the unconditional covariance between N_s and $N_{s'}$ can be derived using the law of total covariance:

$$\mathrm{Cov}(N_s, N_{s'}) = \mathrm{E}_{\lambda_s\lambda_{s'}}(\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'}) + \mathrm{Cov}_{\lambda_s\lambda_{s'}}(\mathrm{E}_{Ns}(N_s \mid \lambda_s), \mathrm{E}_{N_{s'}}(N_{s'} \mid \lambda_{s'}))$$

We assume that, given λ, abundances are independent: $\mathrm{Cov}(N_s, N_{s'} \mid \lambda_s, \lambda_{s'}) = 0$

Therefore:

$$
\begin{aligned}
\mathrm{Cov}(N_s, N_{s'}) &= \mathrm{Cov}_{\lambda_s\lambda_{s'}}(\mathrm{E}_{Ns}(N_s \mid \lambda_s), \mathrm{E}_{N_{s'}}(N_{s'} \mid \lambda_{s'})) \\
&= \mathrm{Cov}_{\lambda_s,\lambda_{s'}}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}, \frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right)
\end{aligned}
$$

To find this covariance we can use the fact that

$$\mathrm{Cov}\left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}, \frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right) = E\left[\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}} \times \frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right] - E\left[\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right] E\left[\frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right]$$

A second-order Taylor series expansion in two variables near the point $(\mu_{\lambda_s}, \mu_{\lambda_s})$ has the following form:

$$\begin{aligned} f(\lambda_s, \lambda_{s'}) \approx\ & f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) + f_{\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s}) + f_{\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}}) \\ & + \frac{1}{2}(f_{\lambda_s\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})^2 + f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}})^2) \\ & + f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}}) \end{aligned}$$

$$\begin{aligned} \mathrm{E}[f(\lambda_s, \lambda_{s'})] \approx\ & \mathrm{E}[f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) + f_{\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s}) + f_{\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}}) \\ & + \frac{1}{2}(f_{\lambda_s\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})^2 + f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_{s'} - \mu_{\lambda_{s'}})^2) \\ & + f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}})] \\ =\ & \mathrm{E}[f(\mu_{\lambda_s}, \mu_{\lambda_{s'}})] + f_{\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})] + f_{\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_{s'} - \mu_{\lambda_{s'}})] \\ & + \frac{1}{2} f_{\lambda_s\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})]^2 + \frac{1}{2} f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_{s'} - \mu_{\lambda_{s'}})]^2 \\ & + f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{E}[(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}})] \\ =\ & f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) + \frac{1}{2} f_{\lambda_s,\lambda_s}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{Var}(\lambda_s) + \frac{1}{2} f_{\lambda_{s'},\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{Var}(\lambda_{s'}) \\ & + f_{\lambda_s,\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}})\mathrm{Cov}(\lambda_s, \lambda_{s'}) \end{aligned}$$

As $\mathrm{E}[\lambda_s - \mu_{\lambda_s}] = 0$, $\mathrm{E}[(\lambda_s - \mu_{\lambda_s})(\lambda_{s'} - \mu_{\lambda_{s'}})] = \mathrm{Cov}(\lambda_s, \lambda_{s'})$

and $\mathrm{E}[(\lambda_s - \mu_{\lambda_s})^2] = \mathrm{Var}(\lambda)$

We know $\mathrm{Var}(\lambda_s)$ and $\mathrm{Cov}(\lambda_s, \lambda_{s'})$ from the lognormal distribution, so now we need the second derivatives of $f(\lambda_s, \lambda_{s'})$

$$
\begin{aligned}
f(\lambda_s, \lambda_{s'}) &= \left(\frac{(1-\theta)\lambda_s}{1-e^{-\lambda_s}}\right)\left(\frac{(1-\theta)\lambda_{s'}}{1-e^{-\lambda_{s'}}}\right) \\
f_{\lambda_s\lambda_s}(\lambda_s\lambda_{s'}) &= \frac{e^{-\lambda_s}\lambda_{s'}(1-\theta)^2(e^{-\lambda_s}\lambda_s + \lambda_s + 2e^{-\lambda_s} - 2)}{(1-e^{-\lambda_{s'}})(1-e^{-\lambda_s})^3} \\
f_{\lambda_{s'}\lambda_{s'}}(\lambda_s, \lambda_{s'}) &= \frac{e^{-\lambda_{s'}}\lambda_s(1-\theta)^2(e^{-\lambda_{s'}}\lambda_{s'} + \lambda_{s'} + 2e^{-\lambda_{s'}} - 2)}{(1-e^{-\lambda_s})(1-e^{-\lambda_{s'}})^3} \\
f_{\lambda_s\lambda_{s'}}(\lambda_s, \lambda_{s'}) &= \frac{(1-e^{-\lambda_s} - e^{-\lambda_s}\lambda_s)(1-\theta)^2(1-e^{-\lambda_{s'}} - e^{-\lambda_{s'}}\lambda_{s'})}{(1-e^{-\lambda_s})^2(1-e^{-\lambda_{s'}})^2}
\end{aligned}
$$

Evaluating these at the point $(\mu_{\lambda_s}, \mu_{\lambda_{s'}})$ gives:

$$
\begin{aligned}
f(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) &= \left(\frac{(1-\theta)\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)\left(\frac{(1-\theta)\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}}\right) \\
f_{\lambda_s\lambda_s}(\mu_{\lambda_s}\mu_{\lambda_{s'}}) &= \frac{e^{-\mu_{\lambda_s}}\mu_{\lambda_{s'}}(1-\theta)^2(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)}{(1-e^{-\mu_{\lambda_{s'}}})(1-e^{-\mu_{\lambda_s}})^3} \\
f_{\lambda_{s'}\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) &= \frac{e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_s}(1-\theta)^2(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)}{(1-e^{-\mu_{\lambda_s}})(1-e^{-\mu_{\lambda_{s'}}})^3} \\
f_{\lambda_s\lambda_{s'}}(\mu_{\lambda_s}, \mu_{\lambda_{s'}}) &= \frac{(1-e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)^2(1-e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})}{(1-e^{-\mu_{\lambda_s}})^2(1-e^{-\mu_{\lambda_{s'}}})^2}
\end{aligned}
$$

where $\mu_{\lambda_s} = \mathrm{E}[\lambda_s] = e^{\mu_s + \Sigma_{ss}}$, from the lognormal distribution.

We then have:

$$
\begin{aligned}
\mathrm{E}[f(\lambda_s, \lambda_{s'})] \approx{}& \left(\frac{(1-\theta)\mu_{\lambda_s}}{1-e^{-\mu_{\lambda_s}}}\right)\left(\frac{(1-\theta)\mu_{\lambda_{s'}}}{1-e^{-\mu_{\lambda_{s'}}}}\right) \\
&+ \frac{\sigma_{\lambda_s^2}}{2}\frac{e^{-\mu_{\lambda_s}}\mu_{\lambda_{s'}}(1-\theta)^2(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)}{(1-e^{-\mu_{\lambda_{s'}}})(1-e^{-\mu_{\lambda_s}})^3} \\
&+ \frac{\sigma_{\lambda_{s'}^2}}{2}\frac{e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_s}(1-\theta)^2(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)}{(1-e^{-\mu_{\lambda_s}})(1-e^{-\mu_{\lambda_{s'}}})^3} \\
&+ \frac{\Sigma_{\lambda_s,\lambda_{s'}}(1-e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)^2(1-e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})}{(1-e^{-\mu_{\lambda_s}})^2(1-e^{-\mu_{\lambda_{s'}}})^2}
\end{aligned}
$$

Where:

$\mu_{\lambda_s} = e^{\mu_s + \frac{1}{2}\Sigma_{ss}}$,

$\sigma^2_{\lambda_s} = (e^{2\mu_s + \Sigma_{ss}})(e^{\Sigma_{ss}} - 1)$

$\Sigma_{\lambda_s,\lambda_{s'}} = e^{\mu_s + \mu_{s'} + \frac{1}{2}(\Sigma_{ss} + \Sigma_{s's'})}(e^{\Sigma_{ss'}} - 1)$ are the mean, variance and covariance of the log-normal distribution.

$$
\begin{aligned}
\mathrm{Cov}(N_s, N_{s'}) = &\left(\frac{(1-\theta)\mu_{\lambda_s}}{1 - e^{-\mu_{\lambda_s}}}\right)\left(\frac{(1-\theta)\mu_{\lambda_{s'}}}{1 - e^{-\mu_{\lambda_{s'}}}}\right) \\
&+ \frac{\sigma_{\lambda_s^2}}{2}\frac{e^{-\mu_{\lambda_s}}\mu_{\lambda_{s'}}(1-\theta)^2(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)}{(1 - e^{-\mu_{\lambda_{s'}}})(1 - e^{-\mu_{\lambda_s}})^3} \\
&+ \frac{\sigma_{\lambda_{s'}^2}}{2}\frac{e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_s}(1-\theta)^2(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)}{(1 - e^{-\mu_{\lambda_s}})(1 - e^{-\mu_{\lambda_{s'}}})^3} \\
&+ \frac{\Sigma_{\lambda_s,\lambda_{s'}}(1 - e^{-\mu_{\lambda_s}} - e^{-\mu_{\lambda_s}}\mu_{\lambda_s})(1-\theta)^2(1 - e^{-\mu_{\lambda_{s'}}} - e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}})}{(1 - e^{-\mu_{\lambda_s}})^2(1 - e^{-\mu_{\lambda_{s'}}})^2} \\
&- \left(\frac{(1-\theta)(\mu_{\lambda_s})}{1 - e^{-\mu_{\lambda_s}}} + \frac{\sigma^2_{\lambda_s}}{2}\frac{e^{-\mu_{\lambda_s}}(e^{-\mu_{\lambda_s}}\mu_{\lambda_s} + \mu_{\lambda_s} + 2e^{-\mu_{\lambda_s}} - 2)(1-\theta)}{(1 - e^{-\mu_{\lambda_s}})^3}\right) \\
&\times \left(\frac{(1-\theta)(\mu_{\lambda_{s'}})}{1 - e^{-\mu_{\lambda_{s'}}}} + \frac{\sigma^2_{\lambda_{s'}}}{2}\frac{e^{-\mu_{\lambda_{s'}}}(e^{-\mu_{\lambda_{s'}}}\mu_{\lambda_{s'}} + \mu_{\lambda_{s'}} + 2e^{-\mu_{\lambda_{s'}}} - 2)(1-\theta)}{(1 - e^{-\mu_{\lambda_{s'}}})^3}\right)
\end{aligned}
$$

The correlation between species' abundances can then be calculated using:

$$
\rho(N_s, N_{s'}) = \frac{\mathrm{Cov}(N_s, N_{s'})}{\sqrt{\mathrm{Var}(N_s), \mathrm{Var}(N_{s'})}}
$$

3.C Estimated Abundances

In this section, we provide a comparison of the maximum observed abundance with the maximum abundance estimated from the Hurdle(C) Model, fitted to the North American Breeding Bird Survey data, as discussed in Section 3.4. This table demonstrates that this dataset is not producing excessively large estimates for abundance N, as described by Dennis et al. (2015).

Species	Maximum Y	Maximum N
Bald Eagle	29	39
Canada Goose	24	46
Hammond's Flycatcher	4	10
Red-breasted Sapsucker	4	22
Steller's Jay	6	14
Swainson's Thrush	12	71
Tree Swallow	50	53
Trumpeter Swan	38	39
Varied Thrush	21	70
Wilson's Snipe	6	43

Table 3.C.1: Maximum observed and estimated abundances per species, produced by the Hurdle(C) model.

3.D Coverage Probabilities

True Median Value			Coverage					
p	λ	θ	N	Σ	p	μ_a	θ	ϕ
MNM								
0.3	7	-	0.55	0.53	0.49	0.49	-	-
0.3	55	-	0.49	0.51	0.48	0.45	-	-
0.8	7	-	0.56	0.52	0.47	0.54	-	-
0.8	55	-	0.52	0.52	0.51	0.58	-	-
Hurdle								
0.3	7	0.2	0.61	0.57	0.45	0.52	0.54	-
0.3	7	0.7	0.72	0.60	0.47	0.50	0.6	-
0.3	55	0.2	0.59	0.51	0.49	0.51	0.45	-
0.3	55	0.7	0.70	0.52	0.51	0.52	0.50	-
0.8	7	0.2	0.64	0.51	0.53	0.51	0.42	-
0.8	7	0.7	0.73	0.55	0.53	0.55	0.54	-
0.8	55	0.2	0.60	0.52	0.52	0.52	0.53	-
0.8	55	0.7	0.69	0.56	0.46	0.56	0.54	-
AR								
0.3	7	-	0.59	0.53	0.54	0.52	-	0.52
0.3	55	-	0.53	0.54	0.53	0.49	-	0.45
0.8	7	-	0.54	0.51	0.52	0.48	-	0.50
0.8	55	-	0.52	0.52	0.51	0.49	-	0.47
Hurdle-AR								
0.3	7	0.2	0.66	0.54	0.54	0.56	0.42	0.47
0.3	7	0.7	0.85	0.53	0.51	0.56	0.46	0.46
0.3	55	0.2	0.62	0.49	0.50	0.51	0.52	0.49
0.3	55	0.7	0.84	0.51	0.52	0.50	0.48	0.50
0.8	7	0.2	0.66	0.52	0.55	0.50	0.54	0.51
0.8	7	0.7	0.87	0.51	0.46	0.47	0.54	0.50
0.8	55	0.2	0.63	0.52	0.58	0.49	0.48	0.48
0.8	55	0.7	0.84	0.52	0.47	0.50	0.54	0.41

Table 3.D.1: Coverage probabilities for the small-scale MNM simulations – i.e. the proportion of small-scale simulations ($(R, T, S, K) = (10, 5, 5, 5)$) in which the true parameter value lies within the estimated 50% credible interval.

True Median Value			Coverage					
p	λ	θ	N	Σ	p	μ_a	θ	ϕ
MNM								
0.3	7	-	0.55	0.53	0.46	0.50	-	-
0.3	55	-	0.49	0.52	0.45	0.47	-	-
0.8	7	-	0.52	0.52	0.49	0.43	-	-
0.8	55	-	0.51	0.53	0.50	0.42	-	-
Hurdle								
0.3	7	0.2	0.61	0.53	0.38	0.39	0.43	-
0.3	7	0.7	0.84	0.54	0.37	0.38	0.51	-
0.3	55	0.2	0.51	0.54	0.36	0.36	0.46	-
0.3	55	0.7	0.78	0.55	0.30	0.25	0.54	-
0.8	7	0.2	0.62	0.51	0.53	0.52	0.44	-
0.8	7	0.7	0.85	0.53	0.51	0.51	0.57	-
0.8	55	0.2	0.61	0.55	0.49	0.48	0.58	-
0.8	55	0.7	0.85	0.56	0.50	0.41	0.6	
AR								
0.3	7	-	0.58	0.56	0.48	0.45	-	0.48
0.3	55	-	0.52	0.50	0.51	0.52	-	0.51
0.8	7	-	0.52	0.53	0.48	0.51	-	0.52
0.8	55	-	0.49	0.47	0.53	0.47	-	0.46
Hurdle-AR								
0.3	7	0.2	0.66	0.53	0.46	0.53	0.45	0.50
0.3	7	0.7	0.78	0.51	0.50	0.45	0.56	0.52
0.3	55	0.2	0.60	0.56	0.49	0.49	0.40	0.48
0.3	55	0.7	0.74	0.54	0.47	0.52	0.40	0.53
0.8	7	0.2	0.63	0.56	0.47	0.48	0.48	0.45
0.8	7	0.7	0.76	0.53	0.48	0.47	0.44	0.50
0.8	55	0.2	0.59	0.54	0.49	0.49	0.42	0.49
0.8	55	0.7	0.74	0.55	0.41	0.44	0.40	0.50

Table 3.D.2: Coverage probabilities for the large-scale MNM simulations – i.e. the proportion of large-scale simulations ($(R, T, S, K) = (100, 10, 10, 10)$) in which the true parameter value lies within the estimated 50% credible interval.

CHAPTER 4

A Review and Comparison of N-Mixture Models and Extensions

In this chapter, we discuss the original N-mixture model, detail its advantages and assumptions, and present notation and terminologies used. We then discuss a range of extensions that have been made to this original model that allow for the estimation of animal abundance in scenarios not supported by the original model, such as when observations are available for multiple species, or when a large number of zero counts are present in the data. Finally we demonstrate how this original model may be used to estimate abundances for bee species using data collected at sites in the UK. For those interested, `R` *code that implements this approach is available at* https://github.com/niamhmimnagh/insect_populations_ch11.

4.1 Introduction

The sizes of a certain animal populations are often of interest to many professionals that work in conservation, such as biologists, ecologists and environmentalists (Krebs, 1972; Norris, 2004). This information is important to evaluate whether

the population size is growing, stable, diminishing or at risk of extinction. In the context of conservation policies and monitoring programmes, it can give an idea whether the actions that have been carried out are proving successful (Jenkins et al., 2003). Animal abundance data normally consists of counts at different locations (referred to generally as sites or transects) and periods (sampling occasions), which are obtained using methods which may or may not involve the capture of the animal (Williams et al., 2002). However, the recorded information is usually imperfect in the sense that it does not represent the total abundance, which prevents traditional statistical techniques based on linear models from being used in the analysis of these data.

This chapter presents the N-mixture model proposed by Royle (2004) from the frequentist and Bayesian perspectives. This model assumes that the animal counts follow a binomial distribution and that the population under analysis is closed (i.e., it does not change over time due to births, deaths or migration). This chapter will also explore extensions to this original model, which allow for the relaxation of this assumption of a closed population, the analysis of multiple species simultaneously, and the use of data which contains large numbers of zero-counts. This chapter will finish by presenting how some of these models may be implemented to estimate abundances of foraging bees, on data collected as part of the BeeWalk Survey.

4.2 N-Mixture Model for a Closed Population

Let Y_{it} denote the counts associated with distinct animals from the same species at site i and sampling occasion t, where $i = 1, 2, \ldots, R$ and $t = 1, 2, \ldots, T$. In the N-mixture model proposed by Royle (2004), the Y_{it} are assumed to be independent and identically distributed (i.i.d) realisations from a binomial distribution with denominator N_i, with the likelihood function given by

$$L(\{N_i\}_1^R, p|\mathbf{Y}) = \prod_{i=1}^{R} \left[\prod_{t=1}^{T} \binom{N_i}{Y_{i,t}} p^{Y_{it}} (1-p)^{N_i - Y_{it}} \right], \tag{4.1}$$

where N_i is the latent population size at site i, p corresponds to the probability of detecting the animals of interest and $\mathbf{Y}$ denotes an $R \times T$ matrix which contains

the counts Y_{it}. One advantage of this model is that it accounts for imperfect detection as $Y_{i,t}$ represents the number of animals observed at site i and time t, while N_i is actually the quantity of interest since it is the true (and latent) number of animals at site i. In the form presented in (4.1), the N-mixture model assumes that the true population, N_i, does not change over time, which may be reasonable if studies are carried out over a short time period, and depending on the species being observed.

To estimate the latent abundance, N_i, and the detection probability, p, N_i is considered as a random effect (or a nuisance parameter) and then standard optimisation methods can be used to estimate both parameters. In this case, a distribution $\pi(N_i|\boldsymbol{\theta})$ is assumed for N_i, which can be any positive discrete distribution, and the estimation of the parameters $\boldsymbol{\theta}$ and p is obtained by maximising the integrated likelihood function

$$L(p, \boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^{R} \left[\sum_{N_i=\max_t\{Y_{it}\}}^{\infty} \left\{ \left(\prod_{t=1}^{T} \binom{N_i}{Y_{it}} p^{Y_{it}} (1-p)^{N_i - Y_{it}} \right) \pi(N_i|\boldsymbol{\theta}) \right\} \right]. \quad (4.2)$$

The Poisson and negative binomial distributions are reasonable choices for $\pi(N_i|\boldsymbol{\theta})$. For the Poisson distribution, $\boldsymbol{\theta} = \lambda$ is a scalar. For the negative binomial distribution, $\boldsymbol{\theta}$ is a vector. It is also possible to use covariates to model $\boldsymbol{\theta}$ (e.g., $\boldsymbol{\theta} = \lambda$ can be modelled via log-linear predictor when $\pi(N_i|\boldsymbol{\theta})$ is a Poisson) in order to evaluate whether external information might help to predict the total abundance. In practice, N-mixture models are available through the R (R Core Team, 2022) package `unmarked` (Fiske and Chandler, 2011).

From the Bayesian viewpoint, the likelihood function of the N-mixture model is the same as presented in (4.1). However, to estimate the parameters of interest (N_i and p), the likelihood does not require any marginalisation as in (4.2). Instead, prior distributions are placed on N_i and p and the inference is carried out through the posterior distribution. The prior distributions describe the knowledge that one possesses about these parameters beforehand (e.g., due to experience or previous studies) and their specification should reflect it. In the Bayesian N-mixture model, the joint posterior distribution is proportional to the likelihood function times the

priors as

$$\pi(\{N_i\}_{i=1}^{R}, p|\mathbf{Y}) \propto \left[\prod_{i=1}^{R}\prod_{t=1}^{T}\binom{N_i}{Y_{it}}p^{Y_{it}}(1-p)^{N_i-Y_{it}}\right] \times \left[\prod_{i=1}^{R}\pi(N_i|\boldsymbol{\theta})\right]\pi(\boldsymbol{\theta})\ \pi(p), \quad (4.3)$$

where $\pi(\cdot)$ represent the prior distributions. Similar to its frequentist counterpart, the Bayesian N-mixture model can be easily implemented through `R` packages, such as `R2jags` (Su and Yajima, 2020), via probabilistic programming languages like `JAGS` (Plummer, 2003) and `Stan` (Carpenter et al., 2017).

Though the N-mixture models are interesting modelling alternatives to estimate latent abundances, they present some limitations. First, these models assume that the population under study does not evolve over time, which is an assumption that can be appropriate for short-term studies but not for medium- and long-term programmes where the animals are followed for years or even decades. Second, these models allow only one species to be analysed at a time. That is, if there is an interest in analysing how two or more species interact jointly in a given environment (e.g., prey and predator), the original N-mixture model cannot be used because it does not account for multiple species nor for any sort of correlation measure that might exist between different species. Third, the counts Y_{it} are assumed to be independent which implies that the sites where the counts are collected are also independent. In practice, the latter assumption is satisfied when the sites under study are reasonably far apart so that the animals in one site can not easily access other sites.

Some of these limitations have been recognised, and N-mixture models have been extended to deal with different applications. Its theoretical and computational aspects have also been explored. The closure assumption has been relaxed in models by Dail and Madsen (2011); Hostetler and Chandler (2015); Mimnagh et al. (2022), and in Section 4.3.2 we provide further details about these N-mixture models for open populations, which allow the species under study to change over time. In addition, we highlight the work of Martin et al. (2011) who consider that both p and N_i are random effects in a Bayesian approach to account for non-independent detection. Thus, a beta distribution with fixed hyperparameters is adopted for p. In some cases, however, it may be of interest to estimate the detection probability through covariates specified in a linear predictor, but under this framework this

is not possible. For instance, characteristics inherent to the environment, such as temperature, vegetation and observer experience, may influence in the detection of the animals (Kéry et al., 2009). Also, the value of the detection probability p impacts directly the estimates of the total abundance, and there is not a default mechanism to choose ideal values for the hyperparameters of the distribution on p.

Haines (2016) shows that the infinite sum in (4.2) can be re-expressed in a closed form by using a hypergeometric function, thus avoiding the need to set an upper limit. This is an important contribution to the estimation of the parameters in the frequentist N-mixture model as, in practice, an upper bound needs to be established in such way that the sum of the remainder terms is negligible. In addition, the specification of the upper bound in the sum over N_i and the equivalence between the N-mixture model proposed by Royle (2004) and the multivariate Poisson model were explored by Dennis et al. (2015). Via simulation studies, they show that the N-mixture model can produce 'infinite' estimates for the latent abundance (N_i) when both the detection probability (p) and the number of periods of observations (T) are small. For these situations, they proposed some diagnostics based on the method of moments to identify beforehand if the estimate of the abundance will potentially be unrealistic. In short, the diagnostics suppose that N_i follows a mixed-Poisson distribution in the form of

$$\begin{aligned}\pi(N_i) &= \int_0^\infty \pi(N_i|\lambda)g(\lambda)d\lambda \\ &= \int_0^\infty \frac{\lambda_i^N e^{-\lambda}}{N_i!}g(\lambda)d\lambda,\end{aligned}$$

where $g(\lambda)$ is the mixing distribution, $\mathrm{E}[N_i] = \lambda$ and $\mathrm{V}(N_i) = \sigma^2 \geq \lambda$. Considering that $Y_{it}|N_i, p \sim \text{Binomial}(N_i, p)$ and that $\mathrm{E}[Y_{it}|N_i, p] = N_i p$, $\mathrm{E}[Y_{it}^2|N_i, p] = N_i p(1-p) + N^2p^2$ and $\mathrm{E}[Y_{it}, Y_{it'}|N_i, p] = N_i^2 p^2$, for all $t \neq t'$, the following unconditional expectations are given by

$$\begin{aligned}\mathrm{E}[Y_{it}] &= \lambda p, \\ \mathrm{E}[Y_{it}^2] &= \lambda p(1-p) + (\lambda^2 + \sigma^2)p^2, \\ \mathrm{E}[Y_{it}, Y_{it'}] &= (\lambda^2 + \sigma^2)p^2.\end{aligned}$$

Hence, the method of moments estimators for p, λ and σ^2 are

$$\begin{aligned}\hat{p} &= (m_1 - m_2 + m_{12})/m_1,\\ \hat{\lambda} &= m_1/\hat{p},\\ \hat{\sigma}^2 &= (m_{12} - m_1^2)/\hat{p}^2,\end{aligned}$$

where

$$\begin{aligned}m_1 &= \frac{1}{RT}\sum_{t=1}^{T}\sum_{i}^{R} Y_{it},\\ m_2 &= \frac{1}{RT}\sum_{t=1}^{T}\sum_{i}^{R} Y_{it}^2,\\ m_{12} &= \frac{2}{RT(T-1)}\sum_{t=1}^{T}\sum_{t'=t+1}^{T}\sum_{i}^{R} Y_{it}Y_{t't}.\end{aligned}$$

Since $0 \leq p \leq 1$ represents a probability, λ a positive rate and σ^2 a measure of variance, the diagnostics are based on the following inequalities:

$$\begin{aligned}&m_1 - m_2 + m_{12} > 0,\\ &m_1 - m_2 + m_{12} \leq m_1 \text{ and}\\ &m_{12} - m_1^2 \geq 0.\end{aligned}$$

Dennis et al. (2015) highlight that when one of the inequalities above is violated, in most of the cases the estimate for λ is very large, whereas the detection probability, p, is very small. In their simulations, the greater the detection probability and the number of periods of observations, the less the inequalities are violated. However, there are cases where the inequalities are satisfied, but the estimate for λ is still large. In contrast, situations where λ is finite and the inequalities are violated are also found. Nonetheless, they recommend using the following covariance diagnostic to determine whether unrealistic estimates may arise.

For more than two sampling occasions ($T > 2$),

$$\mathrm{cov}(\mathrm{y}_1, ..., \mathrm{y}_T) = \frac{2(\overline{\mathrm{y}_1\mathrm{y}_2}, \ldots, \overline{\mathrm{y}_{T-1}\mathrm{y}_T})}{T(T-1)} - \left(\frac{\overline{\mathrm{y}_1} + \ldots + \overline{\mathrm{y}_T}}{T}\right)^2$$

where $\overline{\mathrm{y}_1\mathrm{y}_2}$ denotes the mean of the product $\mathrm{y}_1\mathrm{y}_2$ over R sites. A negative value for this covariance diagnostic suggests that the issue of infinite estimates of λ

may arise. An implementation of this covariance diagnostic in R is available in Appendix 4.C

4.3 Model Extensions

Following the introduction of the N-mixture model by Royle (2004), many extensions have been proposed to estimate abundances using a broad range of data. This includes, but is not limited to, data in which species abundances are permitted to vary with time, data containing abundances for multiple species, data pertaining to species that are observed very rarely, and data with a large proportion of zero-counts. In this section, we will examine the methodologies proposed for some of these extensions.

4.3.1 N-Mixture Models for Multiple Species

The N-mixture model introduced by Royle (2004) provides estimates of abundance for single species. There have since been several extensions made to this model that allow us to examine abundances of multiple species simultaneously.

Golding et al. (2017) proposed a multi-species extension to the N-mixture where the imperfect detection is addressed in the form of false-positive errors, by combining the N-mixture model with a dependent double-observer data collection framework. This framework involves a primary observer recording the number of individuals that they observe, and a secondary observer verifying the observations of the first observer. This observation method has three possible outcomes, with three associated probabilities:

1. The primary observer observes an individual with probability p_1.
2. The secondary observer observes an individual that the primary observer missed with probability $(1 - p_1)p_2$.
3. Both observers miss an individual with probability $(1 - p_1)(1 - p_2)$.

Because this process has multiple possible outcomes, the observation process in this case cannot be modelled using a binomial distribution as in the Royle (2004)

paper, but is instead a multinomial process. The abundance of species s at site i and sampling occasion t is then modelled as:

$$Y_{its} \sim \text{Multinomial}(N_{i,}\pi_{its}),$$

where π_{its} represents the survey outcomes. This model may prove useful in scenarios where the possibility of misidentification or double-counting of individuals is a concern, though it requires the presence of a second observer, which may not be practical for every data-collection programme.

Gomez et al. (2018) proposed a multi-species extension to the N-mixture model which involves the use of a beta distribution for the detection probability of low-abundance species. The abundance of species s at site i and sampling occasion t is in this modeling framework represented by:

$$\begin{aligned} Y_{its} &\sim \text{Binomial}(N_{is}, p_s), \\ p_s &\sim \text{Beta}(\tau\bar{p}, \tau(1-\bar{p})), \end{aligned}$$

where $\bar{p}$ is the mean detection probability among species, and τ is a precision parameter that measures the similarity in detection probabilities. The use of the beta distribution to model the detection probability allows for sharing strength between species, using information for abundant species to inform detection probabilities for less abundant species. In turn, this can allow the estimation of abundance for species whose rarity may otherwise preclude them from examination.

Moral et al. (2018) developed a method that allows for the estimation of abundances of two species, as well as a measurement of the relationship between them. This is achieved through the inclusion of a parameter in the abundance of one species which links it to the other species. At site i and sampling occasion t, the abundance of one species is allowed to depend on the other species as follows:

$$\begin{aligned} Y_{it1} &\sim \text{Binomial}(N_{i1}), \\ N_{i1} &\sim \text{Poisson}(\lambda_{i1}), \\ Y_{it2} &\sim \text{Binomial}(N_{i2}), \\ N_{i2} &\sim \text{Poisson}(\psi_i + \lambda_{i2}N_{i1}), \end{aligned}$$

where $\lambda_{i2} > 0$ allows for a positive impact of the abundance of one species on the other, while $\lambda_{i2} = 0$ suggests that one species does not impact the other. In this case, the parameter ψ_i allows the species to be independently modelled. This modeling framework provides the opportunity to make inferences as to the relationship between species with the inclusion of the λ_{i2} parameter, though it focuses on pairs of species, rather than the community as a whole.

Mimnagh et al. (2022) proposed the multi-species N-Mixture (MNM) model, detailed in Chapter 3. This is a methodology that allows for the measurement of abundances of multiple species, and the relationships between them through the estimation of inter-species correlations, which are introduced using a multivariate normal random effect in the abundance. For species s at site i and sampling occasion t, this model may be summarised as:

$$\begin{aligned} Y_{its} &\sim \text{Binomial}(N_{is}, p_{its}), \\ N_{is} &\sim \text{Poisson}(\lambda_{is}), \\ \log(\lambda_{is}) &= a_{is} + \mathbf{x}_i^\top \boldsymbol{\beta}_s, \\ \mathbf{a}_i &\sim \text{MVN}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \end{aligned}$$

where $\text{MVN}(\cdot, \cdot)$ denotes a multivariate normal distribution and $\mathbf{x}_i$ represents a vector of covariates at the site level that can be used to better predict the latent abundance N_{is}. The unstructured covariance matrix $\boldsymbol{\Sigma}_a$ allows for the estimation of positive and negative inter-species correlations in abundance, and allows the user to make inferences as to the relationships that these species have with one another.

A disadvantage common to all of the models described in this Section is an inability to deal with species present in the study area which are not observed during the survey. This is an area of interest that has been examined in terms of occupancy modelling, but a solution remains to be found for abundance modelling.

4.3.2 N-Mixture Models for an Open Population

The N-mixture model introduced by Royle (2004) assumes that the population at each site is closed, since it considers that the latent population, N_i, does not change

over time. However, this assumption can be violated in studies where animals are observed during many years or even decades. For these cases, Kéry et al. (2009) and Dail and Madsen (2011) proposed extensions to deal with open populations. In this Section, we focus on the model proposed by Dail and Madsen (2011), as the model proposed by Kéry et al. (2009) can be viewed as a specific case.

The generalised N-mixture for open populations considers that the site abundance vary throughout the periods of observation. The counts $Y_{it} \sim \text{binomial}(N_{it}, p)$, where N_{it} denotes the population size at site i and time t. Hence, the integrated likelihood function is given in the form of

$$L(p, \boldsymbol{\theta}|\{Y_{it}\}) = \prod_{i=1}^{R} \left[\sum_{N_{i1}=Y_{i1}}^{\infty} \cdots \sum_{N_{it}=Y_{it}}^{\infty} \left\{ \left(\prod_{t=1}^{T} \binom{N_{it}}{Y_{it}} p^{Y_{it}} (1-p)^{N_i - Y_{it}} \right) \times \pi(N_{i1}, \ldots, N_{it}, \boldsymbol{\theta}) \right\} \right].$$

The estimation of the parameters is carried out assuming that the abundance at each site and time has a first-order Markovian structure, i.e., that N_{it} depends only on $N_{i(t-1)}$. Thus, the distribution for the abundance can be written as $\pi(N_{i1}, \ldots, N_{iT}, \boldsymbol{\theta}) = \pi(N_{i1}, \boldsymbol{\theta}) \prod_{t=2}^{T} \pi(N_{it}, N_{i(t-1)}, \boldsymbol{\theta})$, where $\pi(N_{i1}, \boldsymbol{\theta})$ is the distribution of the initial abundance at site i and time 1. In addition, $\pi(N_{it}, N_{i(t-1)}, \boldsymbol{\theta})$ is modelled through migration decomposition (Nichols et al., 2000) as a sum of two independent random variables, S_{it} and G_{it}, where S_{it} represents the animals at site i and time t who survived from $t-1$ and G_{it} denotes gains (new animals, e.g., due to births and/or immigration) at site i since time $t-1$. In probabilistic terms, these variables are represented as

$$\begin{aligned} G_{it}|N_{i(t-1)} &\sim \text{Poisson}(\gamma(N_{i(t-1)})), \\ S_{it}|N_{i(t-1)}, \omega &\sim \text{Binomial}(N_{i(t-1)}, \omega), \end{aligned}$$

where $\gamma(N_{i(t-1)})$ is the rate of the new arrivals at site i, which can be a function of the site abundance in the previous time, and ω represents the survival probability. In this case, the discrete convolution, $P_{a,b}$, that is used to represent the prior

distribution from state $N_{i(t-1)} = a$ to $N_{it} = b$, for $t > 1$, is given by

$$\begin{aligned} P_{a,b} &= \sum_{c=0}^{min\{a,b\}} \text{Binomial}(c; a, \omega) \times \text{Poisson}(b - c; \gamma(a)) \\ &= \sum_{c=0}^{min\{a,b\}} \left\{ \binom{a}{c} \omega^c (1-\omega)^{b-a} \times \frac{\gamma(a)^{b-c} e^{-\gamma(a)}}{(b-c)!} \right\}. \end{aligned}$$

Assuming that $\pi(N_{i1}; \boldsymbol{\theta})$ can be any discrete positive distribution, e.g., a Poisson(λ), the integrated likelihood in (4.4) can be rewritten as

$$L(p, \lambda, \omega, \alpha | \{Y_{it}\}) = \prod_{i=1}^{R} \left[\sum_{N_{i1}=Y_{i1}}^{\infty} \cdots \sum_{N_{it}=Y_{it}}^{\infty} \left\{ \left(\prod_{t=1}^{T} \binom{N_{it}}{Y_{it}} p^{Y_{it}} (1-p)^{N_i - Y_{it}} \right) \times \frac{\lambda^{N_{i1}} e^{-\lambda}}{N_{i1}!} \prod_{t=2}^{T} P_{N_{i(t-1)}, N_{it}} \right\} \right]. \quad (4.4)$$

Although the sum over N_{it} is infinite, in practice it is necessary to set an upper bound, L, large enough that the remainder sum does not impact significantly the parameter estimates. For the simulations and examples, Dail and Madsen (2011) set $L = 200$. However, they mention that the ideal choice may depend on the observed counts, and its choice varies according to the problem at hand. As in the N-mixture model for closed population, to estimate the parameters via classical inference, numerical optimisation methods are required, as it is not possible to find a closed-form expression for the integrated likelihood.

Additionally, Dail and Madsen (2011) proposed a closure test to verify whether the population under analysis is from a closed population or not. As the model proposed by Royle (2004) is a particular case when $\omega = 1$ and $\gamma = 0$, it is possible to utilise, for T sufficiently large, the asymptotic test introduced by Self and Liang (1987) to test $\{H_0 : \gamma = 0 \text{ and } \omega = 1\}$ versus $\{H_1 : \gamma \neq 0 \text{ and } 0 \leq \omega < 1\}$. As the asymptotic distribution of the test is based on mixtures of Chi-squared distributions and depends on the Fisher information matrix, they recommend the use of the observed information matrix, as the expected one cannot be obtained analytically. However, under the Bayesian perspective, the results of the asymptotic test are based only on the posterior distributions of ω and γ, and it is not necessary to obtain an information matrix.

Due to the Markovian structure in the estimation of the parameters, the abundance estimate at each time, assuming $\pi(N_{i1}; \boldsymbol{\theta}) = \text{Poisson}(\lambda)$, is obtained as

$$\hat{N}_{.1} = R\hat{\lambda},$$
$$\hat{N}_{.t} = \hat{\omega}\hat{N}_{.t-1} + R\hat{\gamma},$$

where R is the number of sites, $\hat{N}_{.1}$ is the initial abundance at time 1 (regardless of the site), $\hat{N}_{.t}$ is the abundance at time t, and $\hat{\lambda}$, $\hat{\omega}$ and $\hat{\gamma}$ are the estimates of the parameters λ, ω and γ, respectively. In addition, an estimate of trend in the abundances can be obtained dividing $\hat{N}_{.t-1}$ by $\hat{N}_{.t}$. If there is no interest in obtaining the abundance per period, the total abundance can be computed as $\hat{N} = R\hat{\lambda}$.

While Dail and Madsen (2011) partition the open-population model into survival S_{it} and recruitment G_{it}, Hostetler and Chandler (2015) explain that this partitioning is not always possible if sites are not closed with respect to movement. In this case, it is suggested that we might replace the mechanistic model of Dail and Madsen (2011) with a classical population growth model. In an unlimited environment, population growth may be simply modelled using an exponential growth model, where r is the maximum per capita rate of increase:

$$N_{it} \sim \text{Poisson}(e^r N_{i(t-1)}).$$

For scenarios in which a limit on population size exists, density-dependent versions of this model are also possible. If K is the stable equilibrium of a population and r is the population growth rate at low population density, abundance may be given by:

$$N_{it} \sim \text{Poisson}\left(e^{r\left(1-\frac{N_{i(t-1)}}{K}\right)} N_{i(t-1)}\right).$$

Further to this, immigration models that allow for population growth following extinction, may be implemented through the addition of a term ν that describes the average number of immigrants per year:

$$N_{it} \sim \text{Poisson}\left(e^r N_{i(t-1)} + \nu\right).$$

The density-dependent models may also incorporate immigration in the same way.

Mutshinda et al. (2009) proposed an approach which breaks down changes in species abundance over time into contributions from species interactions and environmental noise to determine which ones impose the greatest influence on community dynamics. This model proposes that the abundance of species i at time t may be modelled as follows.

$$N_{it}|N_{i(t-1)} = N_{i(t-1)}e^{r_i[1-\frac{\sum_{j=1}^{S}\alpha_{ij}\log N_{j(t-1)}}{k_i}]+e_{it}}$$

where r_i is the intrinsic growth rate and k_i is the natural logarithm of the carrying capacity of species i, α_{ij} represents the interaction between species i and j, and e_{it} is a Gaussian variable that introduces environmental noise to the framework.

The multi-species N-mixture model by Mimnagh et al. (2022) described in section 4.3.1 may also be extended to allow for a relaxation of the closure assumption through the introduction of an autoregressive component on the abundance. Abundance is collected at i sites, t sampling occasions, for s species over k years, and is modelled as:

$$\begin{aligned} Y_{itks} &\sim \text{Binomial}(N_{iks}, p_{itks}), \\ N_{iks} &\sim \text{Poisson}(\lambda_{iks}). \end{aligned}$$

If $k = 1$, then λ_{i1s} is defined as:

$$\log(\lambda_{i1s}) = a_{is} + \mathbf{x}_i^\top \boldsymbol{\beta}_s.$$

For $k > 1$, λ_{iks} is allowed to depend on the latent abundance at year $k-1$:

$$\log(\lambda_{iks}) = a_{is} + \mathbf{x}_i^\top \boldsymbol{\beta}_s + \phi_s \log(N_{i(k-1)s} + 1).$$

This model allows the estimation of inter-species correlations that vary by year, which can allow for inter-species relationships that change in time, though the current specification of ϕ_s is restricted to correlations whose sign does not change from one year to the next.

4.3.3 N-Mixture Models for Zero-Inflated Data

The original N-mixture model assumes that the latent abundance can be described using a Poisson distribution. This may not always be justified, particularly when data contains a large number of zero-counts; a common scenario to encounter when working with animal observation data which arises from surveying unoccupied sites. A negative binomial distribution allows for extra-Poisson variation by allowing the mean abundance to vary stochastically, and so a substitution of the Poisson distribution on abundance for a negative binomial distribution may accommodate a limited amount of zero-inflation. However, many datasets contain a larger number of zero-counts than may be modelled using the negative binomial distribution. If the negative binomial distribution proves unsuitable, distributions that accommodate zero-inflation in the data, including the zero-inflated Poisson distribution and zero-altered (hurdle) Poisson distribution, may be used as an alternative. The modelling frameworks described in this section accommodate zero-inflation by first determining whether a site is occupied, and subsequently estimating abundance of occupied sites.

Wenger and Freeman (2008) proposed a method that allows for the use of the N-mixture model to simultaneously model occurrence and abundance by specifying a zero-inflated distribution for the abundance. This is done by specifying a binomial distribution for the occupancy, O_i, and introducing a variable K_i, which is the realised abundance at site i, given presence, as:

$$\begin{aligned} O_i &\sim \text{Bernoulli}(\phi_i), \\ K_i &\sim \text{Poisson}(\lambda_i), \\ N_i &= O_i \times K_i. \end{aligned}$$

This modelling framework was not intended for use specifically with count data that contains a large number of zero-counts, but rather as an alternative to the original N-mixture model for any count dataset, with the aim of obtaining both occupancy and abundance estimates from a single model. This model retains the assumptions inherent under the original N-mixture model, and so may be used to estimate single-species abundance for a closed population.

Joseph et al. (2009) developed a framework that uses the zero-inflated Poisson and zero-inflated negative binomial models. These are both a combination of a Bernoulli process, which determines occupancy, and a negative binomial or Poisson process, which determines abundance. To implement the zero-inflated Poisson model, the site occupancy is given by:

$$O_i \sim \text{Bernoulli}(1-\theta),$$

where O_i represents the occupancy at site i, and θ is the probability of obtaining a zero-count. If the site is deemed to be occupied, the abundance is then estimated using a Poisson distribution as in the original N-mixture model.

Hostetler and Chandler (2015) also proposed an extension that uses a zero-inflated Poisson distribution to model excess zero-counts. This is achieved by specifying the true abundance at site i and sampling occasion t as follows.

$$N_{it} \sim \begin{cases} \text{Poisson}(0) & \text{with probability } \gamma \\ \text{Poisson}(\Lambda) & \text{with probability } (1-\gamma) \end{cases}$$

where γ represents the proportion of excess zero counts.

The MNM model proposed by Mimnagh et al. (2022) may also be used to model excess zeros in the observations through the use of a hurdle-Poisson (or zero-altered) model in the abundance. The observations Y_{its} are assigned a binomial distribution, as described in Section 4.3.1. The hurdle-Poisson distribution then consists of two separate processes. The first is a Bernoulli process, which determines whether a site is occupied (abundance is non-zero) or unoccupied (abundance is zero). If the abundance is non-zero, a second random variable with a zero-truncated Poisson distribution determines the value of the abundance, i.e.,

$$\begin{aligned} O_{is} &\sim \text{Bernoulli}(1-\theta), \\ C_{is} &\sim \text{zero-truncated Poisson}(\lambda_{is}). \end{aligned}$$

where O_{is} represents the occupancy of species s at site i, θ is the probability of obtaining a zero-count, and C_{is} represents the abundance of species s at site i. This abundance C_{is} is only estimated at sites that are occupied by a particular

species (i.e., $O_{is} = 1$). The abundance N_{is} is then defined as

$$N_{is} = \begin{cases} 0, & \text{if } O_{is} = 0 \\ C_{is}, & \text{if } O_{is} = 1 \end{cases},$$

This may be written as:

$$N_{is} \sim \text{hurdle-Poisson}(\lambda_{is}, \theta).$$

The hurdle-Poisson model described by Mimnagh et al. (2022) differs from the zero-inflated Poisson model described by Joseph et al. (2009) in the assumptions required for abundance estimation. The use of a *zero-truncated* Poisson distribution for abundance assumes that all zeros arising in the data arise from the occupancy process, while the *zero-inflated* Poisson distribution allows zeros to arise from both the occupancy and abundance processes. The decision as to which model to use will depend on the goal of the user. If the user is interested in examining true and false zeros (i.e., zeros produced because the site is unoccupied, and zeros produced because, though the site is occupied, no observations were made), then a zero-inflated model is an appropriate choice. If the analysis is concerned with whether a count is zero or non-zero, and the user is uninterested in the origin of the zero counts, then all zeros may be assigned to the occupancy process and the hurdle-Poisson model may be used.

We now employ some of the different extensions to N-mixture models to a real dataset in the case study described below.

4.4 Case Study: Bee Abundance

Wild bee species play major roles in pollination, increasing the yield of approximately 85% of all cultivated crops (Zattara and Aizen, 2021). Abundance and diversity of bee species are reported to be in decline on a near-global level (Theisen-Jones and Bienefeld, 2016; Pettis and Delaplane, 2010; Leonhardt et al., 2013), with economic and ecological repercussions inherent in this decline. In order to make decisions concerning management of bee populations (i.e., conservation, use, and monitoring, according to Caughley (1994)), it is beneficial for us to be able to estimate species abundance.

Here, we examine how bee population sizes may be estimated using the original N-mixture model (Royle, 2004) and the multi-species N-mixture model (Mimnagh et al., 2022). This analysis is motivated by data collected as part of the BeeWalk Survey Scheme (Comont et al., 2021), a programme established in 2008 by the Bumblebee Conservation Trust, which involves transects being surveyed by volunteers across the UK on a monthly basis. By the end of the 2019 data-collection period, data had been collected for approximately 70 bee species, at over $1,300$ sites in the UK. Here we examine bee observation data collected in 2016 and 2019. To ensure that we are comparing data collected from the same seasonal cycles, we examine data collected in June of both years.

The models described in this Section are implemented using a Bayesian framework. Each of the models were implemented in `R` (R Core Team, 2022) through the probabilistic programming software `JAGS` (Plummer, 2003, 2017) using four chains with $50,000$ iterations each, of which the first $10,000$ were discarded as burn-in, with a thinning of five to reduce autocorrelation in the MCMC samples. Parameter convergence was determined using the potential scale reduction factor ($\hat{R}$), a diagnostic criteria proposed by Gelman and Rubin (1992). An $\hat{R}$ value that is very close to one is an indication that the four chains have mixed well. If $\hat{R}$ value was less than 1.05, the chains were considered to have mixed properly, and the posterior estimates of the parameters were determined to be reliable.

Prior to modelling this data to estimate bee abundance, we check to confirm that we should not expect to encounter issues regarding infinite estimates of abundance, using the covariance diagnostic proposed by Dennis et al. (2015), and detailed in Section 4.2. A negative value for this diagnostic would suggest that problems with infinite parameter estimates may occur. Covariance diagnostic values obtained for the data utilised in this section were all positive, and so do not suggest that we should expect issues with parameter estimates.

The original N-mixture model (Royle, 2004) may be used to estimate abundance for a single species, and assumes that populations are closed. This analysis will focus on data collected in June of 2019 at 60 sites for the common carder bumblebee (*Bombus pascuorum*). Transects examined in this study range in length from 167

to 3,670 metres. It was thought that perhaps longer transects may provide more opportunity for bee observations, which may in turn affect abundance estimates. To account for possible effects on abundance due to transect length, the length in metres of each transect was included in the linear predictor for the abundance parameter. Additionally, to account for transect location, the latitude, longitude, and their interaction term latitude × longitude were examined. All covariates were scaled to have zero-mean and unit variance. The estimates for the effect of each of these covariates on carder bumblebee abundance are available in Table 4.1. The effect of longitude and latitude on carder bumblebee estimates may be difficult to discern, due to the use of the interaction term latitude × longitude, though we can see from Figure 4.1 that abundance estimates do appear larger in the south of the UK, and appear to decrease towards the north. Additionally, transect length appears to have a positive effect on estimates of abundance, with a mean estimate of 0.16 (i.e., abundance estimates associated with longer transects are greater than those associated with shorter transects).

Species	Intercept	Latitude	Longitude	Latitude × Longitude	Length
Carder Bumblebee	4.27 (4.12, 4.43)	-0.09 (-0.28, 0.08)	0.39 (0.18, 0.61)	0.12 (-0.05, 0.28)	0.16 (0.03, 0.29)

Table 4.1: Mean parameter estimates and 95% credible intervals for the original N-mixture model applied to Common Carder Bumblebee count data collected in June 2019.

The estimates of abundance obtained for the common carder bumblebee using this original N-mixture model are shown in Figure 4.1. The abundance estimates provided here may be considered as an estimate of the size of the population that is are currently foraging at this site. It cannot be viewed as representative of the full carder bee abundance in the area, or the abundance of the local carder bee colony as a whole, as only approximately 30% of a bee colony's population will forage at a certain time. Additionally, bees may travel several kilometres while foraging (Greenleaf et al., 2007), so it is not impossible that the bees observed at each site may not be local to the area, and may have travelled a distance from their colony to forage. For works on estimates of bee colony abundances, we refer

the reader to McGrady et al. (2021); Russo et al. (2015); Kuhlman et al. (2021)

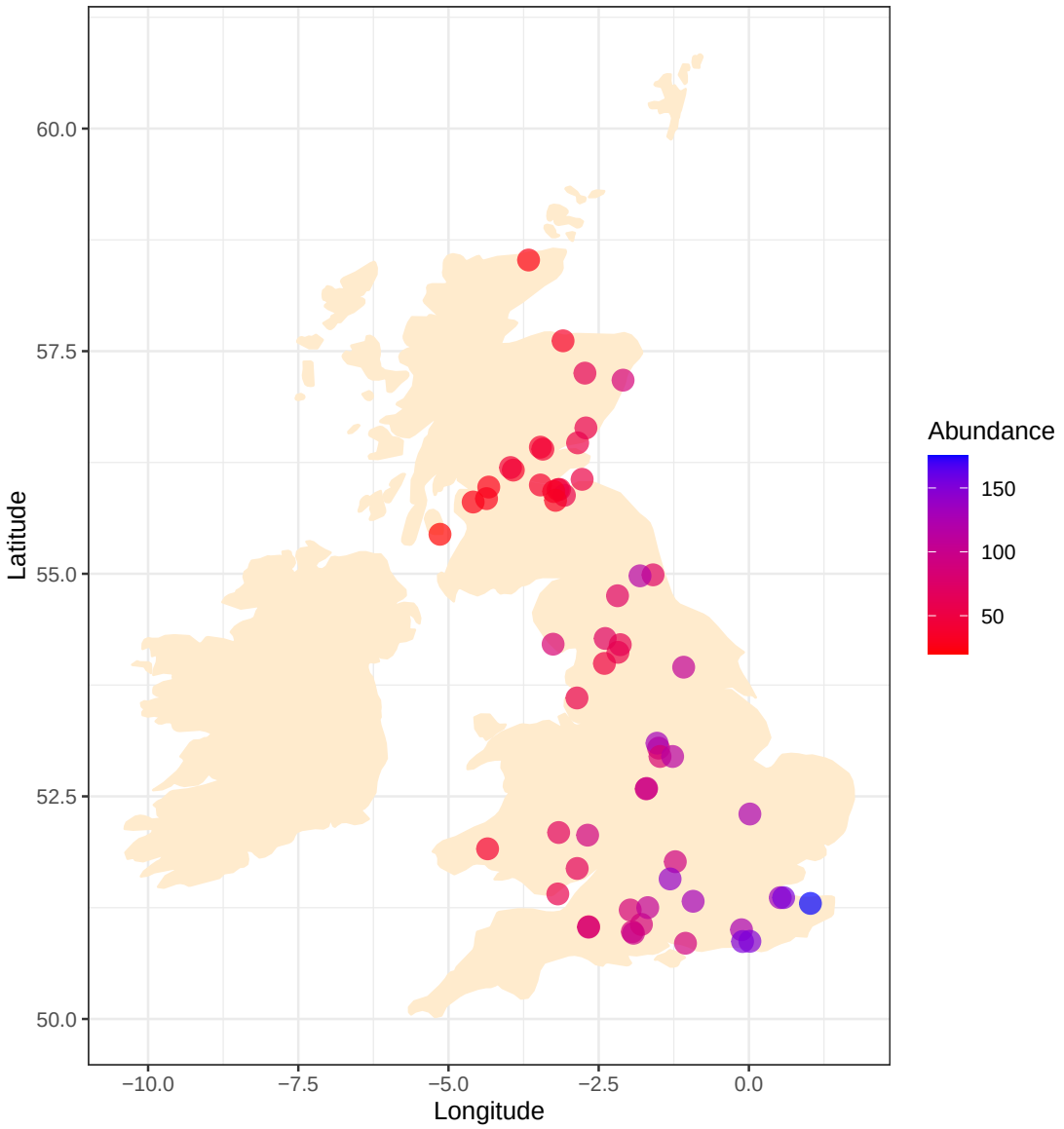


Figure 4.1: Common Carder Bumblebee (*Bombus pascuorum*) abundance estimates for June 2019 at 60 sites in the UK, obtained using the original N-mixture model.

The MNM model (Mimnagh et al., 2022) may be used to estimate abundances for multiple species over longer time frames. This allows us to examine differences in abundance estimates from June 2016 to June 2019 for the European

honeybee (*Apis mellifera*), and seven species of bumblebee: the white-tailed bumblebee (*Bombus lucorum*), the buff-tailed bumblebee (*Bombus terrestris*), the garden bumblebee (*Bombus hortorum*), the tree bumblebee (*Bombus hypnorum*), the early bumblebee (*Bombus pratorum*), the red-tailed bumblebee (*Bombus lapidarius*), and the common carder bumblebee (*Bombus pascuorum*).

Initial examination of this data reveals that 54% of observations (1,049 out of a total of 1,920 observations) are composed of zero-counts. For this reason, the MNM model with a hurdle component, described in section 4.3.3 is employed.

For appropriate comparison with the N-mixture model implementation detailed above, the same sites are examined here using the MNM model. The effect on abundance estimates of the covariates latitude, longitude, latitude $\times$ longitude, and transect length are provided in Table 4.2. As with the results in table 4.1, the presence of the interaction term latitude $\times$ longitude makes it difficult to interpret the effect of latitude and longitude on abundance estimates. For this reason, these results could be displayed in a similar manner to those displayed in Figure 4.1, with a separate abundance map per species, though we do not present those maps here. It appears that the length of the transect has a positive effect on abundance in the case of the buff-tailed bumblebee, the red-tailed bumblebee, the tree bumblebee, the early bumblebee, and the common carder bumblebee, as was demonstrated in Table 4.1. It may appear at first glance from the mean estimates that transect length also has a positive effect on abundance estimates for the European honeybee, and a negative effect for the white-tailed bumblebee, and the garden bumblebee. However, as the 95% credible intervals associated with the effect of transect length for these species contains 0, we cannot say with certainty that these mean estimates are reliable, and instead conclude that it seems that transect length does not have an effect on abundance estimates for these species.

Species	Intercept	Latitude	Longitude	Latitude × Longitude	Length
White-tailed Bumblebee	2.72 (2.01, 3.40)	1.02 (0.41, 1.59)	0.67 (0.07, 1.25)	-0.03 (-0.61, 0.51)	-0.13 (-0.27, 0.28)
Buff-tailed Bumblebee	3.50 (3.02, 3.93)	0.28 (-0.13, 0.69)	0.12 (-0.05, 0.62)	-0.13 (-0.29, 0.31)	0.29 (0.16, 0.65)
Garden Bumblebee	1.72 (1.03, 2.35)	0.04 (-0.62, 0.74)	-0.45 (-1.19, 0.27)	0.33 (-0.33, 1.01)	-0.04 (-0.56, 0.45)
Red-tailed Bumblebee	3.67 (3.16, 4.14)	-0.58 (-1.04, -0.17)	-0.03 (-0.49, 0.42)	0.26 (-0.18, 0.71)	0.64 (0.29, 1.01)
Tree Bumblebee	2.92 (2.36, 3.45)	0.04 (-0.54, 0.67)	0.38 (-0.25, 1.04)	0.83 (0.24, 1.41)	0.66 (0.26, 1.07)
Early Bumblebee	3.15 (2.67, 3.57)	0.30 (-0.12, 0.73)	0.24 (-0.23, 0.76)	0.74 (0.29, 1.21)	0.54 (0.20, 0.89)
Carder Bumblebee	4.03 (3.73, 4.31)	0.09 (-0.21, 0.41)	0.22 (-0.11, 0.55)	0.13 (-0.17, 0.42)	0.21 (0.01, 0.43)
European Honeybee	0.98 (-0.20, 1.99)	-0.42 (-1.45, 0.61)	0.87 (-0.11, 2.01)	0.45 (-0.44, 1.45)	0.60 (-0.15, 1.35)

Table 4.2: Mean parameter estimates and 95% credible intervals for the MNM model applied to count data for eight bee species, collected in June 2016 and 2019.

We can also examine how abundance estimates at each site change between 2016 and 2019. Table 4.3 shows the number of sites (out of the total 60 sites examined) at which abundances increased, decreased, or remained unchanged between 2016 and 2019. Species such as the European honeybee and the buff-tailed bumblebee appear to have experienced abundance increases at a large number of sites, while species such as the early bumblebee and common carder bumblebee appear to have experienced a decrease in abundance at a majority of sites.

Species	Increase	Decrease	No Change
White-tailed Bumblebee	32	22	6
Buff-tailed Bumblebee	34	21	5
Red-tailed Bumblebee	28	28	4
Garden Bumblebee	25	32	3
Tree Bumblebee	26	28	6
Early Bumblebee	17	34	9
Common Carder Bumblebee	18	29	13
European Honeybee	35	20	5

Table 4.3: The number of sites at which abundance estimates (obtained using the MNM model) for each bee species increased, decreased, or remained unchanged between 2016 and 2019.

Figure 4.2 displays inter-species abundance correlations, which may allow for inferences to be made as to the relationships that these species have with one another, or with their environments. For example, we can see that the early bumblebee has a strong positive correlation with the garden bumblebee, which suggests that the abundances of these species may be increasing or decreasing together, while the early bumblebee has a slightly negative correlation with the white-tailed bumblebee, which suggests that one of these species may be experiencing an increase in abundance while the other decreases. These correlations seem to correspond with the results obtained in Table 4.3, as both the early bumblebee and garden bumblebee experience a decrease at a large number of sites, while the white-tailed bumblebee experiences an increase in abundance at a majority of sites examined.

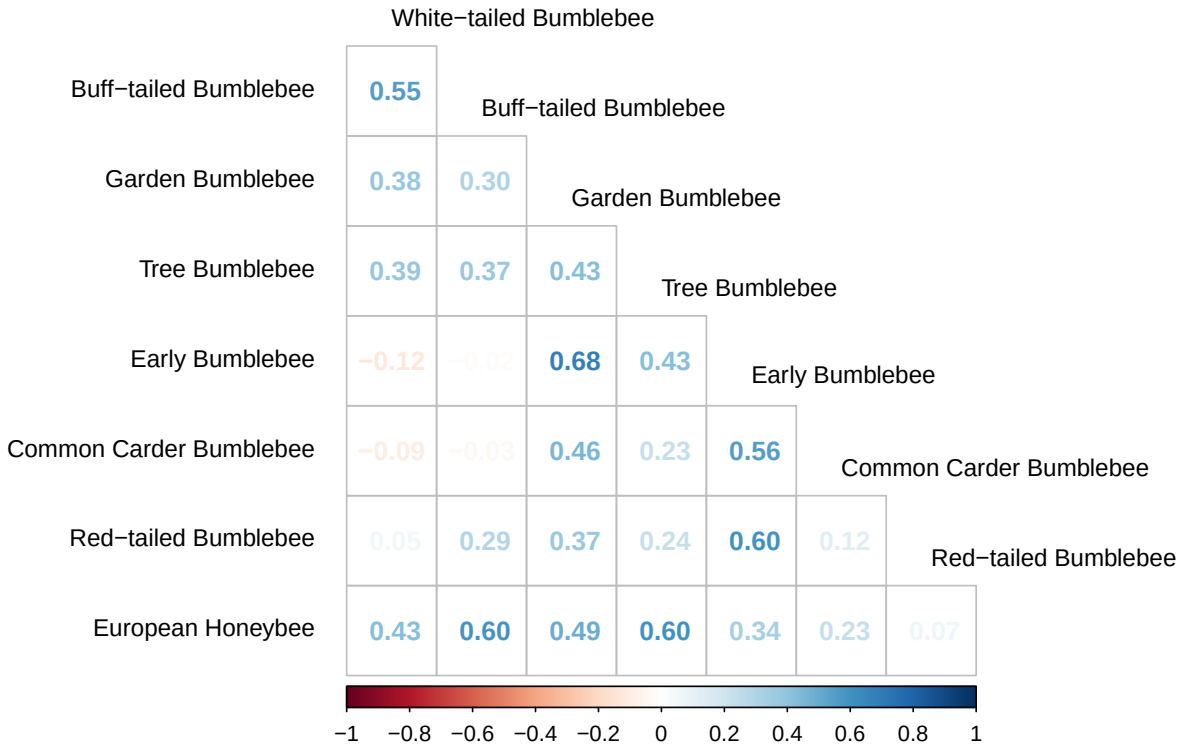


Figure 4.2: Correlation between abundance estimates for bee species obtained using the MNM model.

4.5 Discussion

In this chapter, we have examined several approaches to modelling abundance data, beginning with the original N-mixture model (Royle, 2004) and continuing to explore model extensions, which allow us to estimate animal abundance using data collected in a range of scenarios. The N-mixture family of models are widely used due to their ability to estimate both abundance and detection probability.

We have also addressed previous work, which has demonstrated that N-mixture models can sometimes suffer from issues with identifiability (Dennis et al., 2015), which lead to very small estimates for detection probability and very large estimates of abundance. This is an issue that must be kept in mind when using

N-mixture models, and the covariance diagnostic provided by Dennis et al. (2015) is a useful tool in assessing whether an N-mixture model is appropriate for use with a certain data set.

We finished by demonstrating how the N-mixture model by Royle (2004) and an N-mixture model for multiple species (Mimnagh et al., 2022) may be implemented to estimate foraging bee abundance using the software `R` and the probabilistic programming language `JAGS`. Results of this analysis (Figure 4.1) suggest that foraging bee populations may be larger in the South of England, and decrease as we travel through the North of England and into Scotland. As mentioned previously, due to bee colony dynamics, this abundance estimate does not represent the total bee abundance in the area, but rather the foraging abundance and can be thought of as an index of the local population size.

Appendix

4.A Bayesian N-Mixture Models for Closed Populations in JAGS

In this Section, we present an implementation of the Bayesian N-mixture model for closed population using the software `R` and the probabilistic programming language `JAGS`. Here, we present a simple Bayesian N-mixture model which assumes that

$$\begin{aligned} Y_i|N_i, p &\sim \text{Binomial}(N_i, p), \\ N_i &\sim \text{Poisson}(\lambda), \\ \lambda &\sim \text{Gamma}(\zeta = 1, \eta = 0.1), \\ p &\sim \text{Beta}(\nu = 1, \xi = 1). \end{aligned} \tag{4.5}$$

The prior on λ is set up such that it is non-informative as *a priori* $\text{E}(\lambda) = 10$ and $\text{Var}(\lambda) = 100$. In practice, that means that we believe that the true value of λ is around 10, but we are not sure about it, which is demonstrated by the large variance of the prior distribution. The prior on the detection probability, p, is also non-informative in the sense that it gives equal prior probability to any possible value in the range $[0, 1]$.

This model can be specified as follows.

```
library(rjags)
library(R2jags)
# Simulate synthetic data --------------------
set.seed(1234)
```

```
T <- 10   # Number of periods of observation
R <- 100  # Number of sites
p <- 0.7  # Detection probability
y <- matrix(data = NA,ncol = T,nrow = R)
lambda <- 20
N <- rpois(n = R, lambda)
for (t in 1:T){
  y[,t] <- rbinom(n = R,size = N, prob = p)
}
# Specify the model -------------------------
model_code = '
model{
for (i in 1:R) {
    N[i] ~ dpois(lambda)
        for (t in 1:T) {
            y[i,t] ~ dbin(p, N[i])
        }
}
# Priors
p~dunif(0, 1)
lambda~dgamma(1, 0.1)
}'
# Package data for R2jags
data_list <- list(R = R, T = T, y = y)
# Initial values
initial_values <- function(){list(N = apply(y, 1, max))}
# parameters to monitor
par_save <- c("p", "lambda")
# run model
model_run <- jags(data               = data_list,
                  inits              = initial_values,
                  parameters.to.save = par_save,
                  model.file         = textConnection(model_code),
                  n.chains           = 4,
                  n.iter             = 10000,
```

```
n.burn               = 5000)
```

4.B Bayesian N-mixture Models for Open Populations in JAGS

In this Section, we present an implementation of the multi-species N-mixture model (Mimnagh et al., 2022) for an open population with a large proportion of zero-counts, using the software `R` and the probabilistic programming language `JAGS`. Here, we present a model which assumes that

$$
\begin{aligned}
Y_{itsk}|N_{isk}, p &\sim \text{Binomial}(N_{isk}, p),\\
N_{isk} &\sim \text{Hurdle-Poisson}(\lambda_{isk}, \theta),\\
\log(\lambda_{isk}) &= \begin{cases} a_{is} & \text{for } k = 1,\\ a_{is} + \phi_s \log\left(N_{is(k-1)} + 1\right) & \text{for } k > 1, \end{cases}\\
a_i &\sim \text{MVN}(\boldsymbol{\mu} = \mathbf{0}_s, \boldsymbol{\Sigma}),\\
\boldsymbol{\Sigma} &\sim \text{Inverse-Wishart}\left(\Omega = \mathbf{I}_s, \omega = S + 1\right),\\
p &\sim \text{Beta}(\nu = 1, \xi = 1),\\
\theta &\sim \text{Beta}(\nu = 1, \xi = 1),\\
\phi_s &\sim \text{Normal}(\mu = 0, \sigma^2 = 100).
\end{aligned} \tag{4.6}
$$

The $Y_{i,t,s,k} = y_{i,t,s,k}$ denote the observed counts at site $i = 1, \cdots, R$, sampling occasion $t = 1, \cdots, T$, species $s = 1, \cdots, S$ and year $k = 1, \cdots, K$. We denote $\mathbf{0}_s$ a vector of zeros of dimension S, $\mathbf{I}_s$ an identity matrix of dimension S and $\mathbf{a}_i = (a_{i,1}, \cdots, a_{i,S})$. The model may be specified as follows in `JAGS`.

```
library(R2jags)
library(clusterGeneration)
library(mvtnorm)
library(extraDistr)

# Simulate synthetic data ------------------
set.seed(1234)
R <- 40                    # number of sites
T <- 10                    # number of sampling occasions
S <- 4                     # number of species
```

```
K <- 2                      # number of years
theta <- 0.3                # probability of a site being unoccupied
p <- 0.8                    # probability of detection
phi <- runif(S, -0.5, 0.5) # autocorrelation coefficient

# Empty arrays to store the observed and latent counts
y <- array(NA, dim=c(R,T,S,K))
N <- lambda<-array(NA, dim=c(R,S,K))
# Occupancy array
Occ <- array(rbinom(R*S*K, size=1, prob=1-theta), dim=c(R,S,K))
# Scale matrix for wishart distribution
Omega <- diag(1, nrow=S, ncol=S)

covariance <- genPositiveDefMat(S, rangeVar=c(0.2, 1),
                                covMethod="unifcorrmat")[["Sigma"]]
correlation <- cov2cor(covariance)
# species-level MVN random effect
a <- rmvnorm(R, mean=rep(0,S), sigma=covariance)

# Generate the latent abundance, N[i,s,k]
for(i in 1:R){
    for(s in 1:S){
        # for year K = 1
        lambda[i,s,1]<-exp(a[i,s])
        N[i,s,1] <- ifelse(Occ[i,s,1]==0, 0,
                           rtpois(1, lambda=lambda[i,s,1], a=0))
        # for year K > 1
        for(k in 2:K){
           lambda[i,s,k] <- exp(a[i,s]+ phi[s]*log(N[i,s,k-1]+1))
           N[i,s,k] <- ifelse(Occ[i,s,k]==0, 0,
                              rtpois(n=1, lambda=lambda[i,s,k], a=0))
        }
    }
}
```

```
# Generate the observed abundance, y[i,s,k]
  for(i in 1:R){
    for(t in 1:T){
      for(s in 1:S){
        for(k in 1:K){
          y[i,t,s,k] <- rbinom(1, size=N[i,s,k], prob=p)
        }
      }
    }
  }

# Specify the model -------------------------
model_code = 'model {
for(s in 1:S){
    for (i in 1:R) {
        Occ[i,s,1] ~ dbern(1-theta)
        log(lambda[i,s,1]) <- a[i,s]
        C[i,s,1] ~ dpois(lambda[i,s,1])T(1,)
        N[i,s,1] <- ifelse(Occ[i,s,1]==0, 0, C[i,s,1])
        for(k in 2:K){
            Occ[i,s,k] ~ dbern(1-theta)
            log(lambda[i,s,k]) <- a[i,s] + phi[s]*log(N[i,s,k-1]+1)
            C[i,s,k] ~ dpois(lambda[i,s,k])T(1,)
            N[i,s,k] <- ifelse(Occ[i,s,k]==0, 0, C[i,s,k])
            }
        }
    }
for(i in 1:R){
    for(s in 1:S){
        for(t in 1:T){
            for(k in 1:K){
                y[i,t,s,k] ~ dbin(p, N[i,s,k])
            }
        }
    }
```

```
}

# species-level random effect
for(i in 1:R){
    a[i,1:S] ~ dmnorm(rep(0,S), precision[,])
}

# Inter-species correlations
precision[1:S,1:S] ~ dwish(Omega[,], df)
covariance[1:S,1:S] <- inverse(precision[,])
for (s in 1:S){
    for (s1 in 1:S){
        correlation[s,s1] <- covariance[s,s1]/sqrt(covariance[s,s]*covariance[s1, s1])
    }
}

# Priors
theta~dbeta(1,1)
p~dunif(0,1)
for(s in 1:S){
    phi[s] ~ dnorm(0,0.01)
}
}
' # end of model specification

# Package data for R2jags
data_list <- list(R=R, y=y, T=T, S=S, K=K, Omega=diag(1, S), df=S+1)
# Initial values
initial_values <- function(){
 list(C = apply(y,c(1,3,4), max)+1,
     Occ = apply(y, c(1,3,4), function(z) ifelse(any(z>0), 1, 0)))
}
# parameters to monitor
par_save=c("correlation", "N", "p", "theta", "Occ")
```

```
# run model
model_run <- jags(data               = data_list,
                  inits              = initial_values,
                  parameters.to.save = par_save,
                  model.file         = textConnection(model_code),
                  n.chains           = 4,
                  n.iter             = 25000,
                  n.burn             = 10000)
```

4.C Covariance Diagnostic

In this Section, we provide an implementation of the covariance diagnostic proposed by Dennis et al. (2015) using the software R. To obtain this covariance diagnostic, we need the number of sampling occasions in our data, T, and the observed data, Y.

A negative value for this diagnostic indicates that there may be an issue with estimates of infinite abundance. In the example below, the simulated data produces a negative value for the covariance diagnostic, which suggests that the use of the N-mixture model on this data may produce estimates of abundance that are very large and estimates of detection probability that are very small.

```
# Simulate synthetic data -------------------
set.seed(1234)
T <- 5  # Number of periods of observation
R <- 20  # Number of sites
p <- 0.1  # Detection probability
y <- matrix(data = NA,ncol = T,nrow = R)
lambda <- 2

N <- rpois(n = R, lambda)
for (t in 1:T){
  y[,t] <- rbinom(n = R,size = N, prob = p)
}

# Diagnostic for the N-mixture model --------
CovarianceDiagnostic<-function(y, T){
  ninj <- 0
  for(i in 1:(T-1)){
    for(j in (i+1):T){
      ninj<-cbind(ninj,y[,i]*y[,j])
    }
  }
  covDiag <- sum(colMeans(ninj))*2/(T*(T-1))-((sum(colMeans(y)))/T)^2
```

```
  print(paste0("Covariance Diagnostic: ", round(covDiag,4)))
}

# Run diagnostic
CovarianceDiagnostic(y, T)
```

CHAPTER 5

A Triple Poisson Model for Scarce Vestige Data

We propose a new class of models for the estimation of animal abundance using animal vestige data. We demonstrate that our approach is competitive to similar models, and is particularly useful when data is very scarce, using both simulation studies and real-world datasets. For interested readers, `R` *code that implements this approach is available at* `https://github.com/niamhmimnagh/triple_poisson`.

5.1 Introduction

Estimating wildlife abundance may be relatively expensive and time-consuming. Therefore, whenever possible monitoring population fluctuation by an abundance index tends to be more cost-effective (Nichols, 2014). However, in order to understand evolutionary-ecological processes and make decisions concerning wildlife management (i.e., conservation, use, coexistence, and monitoring, according to Caughley (1994)), one might need to estimate actual species abundance (Verdade et al., 2014). In addition, methods that involve capturing or even direct sightings of animals can be invasive and pose risks to both wildlife and humans (Verdade et al., 2013).

In this chapter, we propose a modelling framework based on a triple Poisson hierarchy, which allows for the estimation of animal abundance from animal vestige count data, where a vestige may include any trace that an animal leaves behind as it moves through a study area. In the simulation studies and case studies examined in this chapter, the vestiges examined are scats, that we assume are produced at a constant rate. Advantages associated with the use of vestige data to estimate animal abundance include the reduced cost and labour required to carry out the survey, as well as the reduced risk of disturbance caused to the animal, when compared with direct methods of estimating animal abundance (Verdade et al., 2014). In addition, we show that the modelling framework proposed here can be useful for estimating animal abundance from small-scale monitoring programmes, which may have very few transects and so produce data that is scarce.

The remainder of this chapter is organised as follows. In Section 5.3 we introduce the modelling framework to estimate animal abundance from vestige count data. In Section 5.4 we present simulation studies, which were carried out to assess the estimates of abundance under a range of scenarios and with varying levels of prior information. Finally, in Section 5.5 we present a number of case studies which we use to illustrate how this modelling approach may have real-world applications.

5.2 Related Works

Several methods have been previously developed for estimating animal relative abundance based on vestige count data.

Distance sampling (Buckland et al., 1993) is a methodology that while originally proposed to estimate animal density using animal count data, may also be used to estimate animal abundance using vestige data by making some small modifications (Marques et al., 2001). Distance sampling using vestige data involves modelling, with a detection function, the assumption that the detectability of vestiges decreases with increasing distance from a transect. This is done to obtain an estimate of vestige density, which may then be used to obtain an estimate of animal density and finally animal abundance. Marques et al. (2001) estimate the abundance of sika deer using this distance sampling methodology. In Section 5.4

we provide details of a simulation study which involves the comparison of abundance estimates obtained using a distance sampling model with those obtained using the novel triple Poisson model that we propose in this chapter.

Becker (1991), Becker et al. (1998), and Patterson et al. (2004) propose a model for estimating abundance based on the observation of animal tracks in snow. This model assumes that the number of animal groups in the area may be obtained by following tracks and locating each group. For this reason, this methodology could prove time- and resource-intensive. The triple Poisson model that we propose in this chapter allows us to assign a prior distribution to group size, which means that we may estimate an unknown number of groups, and true values do not need to be obtained.

The Formozov–Malyshev–Pereleshin formula for estimating animal abundance using animal track data is described by Stephens et al. (2006). This formula, originally proposed and published through Russian in Chelintsev (1995), involves estimating the probability that a transect will intersect an animal's track. This probability is then used to estimate the total number of track crossings, which can be used to estimate animal density. The successful implementation of this formula requires both estimates of animal daily travel distances and counts of animal tracks whose age is known.

5.3 Methods

5.3.1 Model Formulation

This model assumes that we are examining closed populations of mammals which move around randomly in G groups of size N_i, $i = 1, \ldots, G$, within a study area which is homogeneous in terms of habitat use. We assume that

$$G \sim \text{Poisson}(\lambda_G),$$

and therefore the total number of animals in the area is

$$T = \sum_{i=1}^{G} N_i.$$

By assuming the group sizes N_i are independent and distributed as

$$N_i \sim \text{Poisson}(\lambda_N),$$

we can conclude that the conditional distribution of $T|G$ is

$$T|G \sim \text{Poisson}(G\lambda_N).$$

However, the data that we examine is not composed of realisations of the abundance T. Instead, we observe a fraction of the number of vestiges left in the study area by a certain species. Let V_t represent the total vestige count in the area at time t. We assume that the number of vestiges left at time $t = 1$ has mean βT, i.e. depends on the individual vestige production rate β. By letting $V_1 \sim \text{Poisson}(\beta T)$ and assuming vestiges left in the environment disappear exponentially over time with some constant rate, we may write

$$V_t|V_1, \ldots, V_{t-1} \sim \text{Poisson}\left(\beta T + \sum_{j=1}^{t-1} \beta T e^{-\delta(t-j)}\right).$$

where δ is a vestige decay parameter.

We are interested, however, in the limiting distribution of V_t, and for that we assume that after a short period of time, the total number of vestiges produced plus old vestiges that remain in the environment from previous time points will become constant. We next examine the rate term in the Poisson distribution above, which can be written as follows:

$$\beta T + \beta T \sum_{j=1}^{t} e^{-j\delta} = \beta T e^{-(0)\delta} + \beta T \sum_{j=1}^{t} e^{-j\delta} = \beta T \sum_{j=0}^{t} e^{-j\delta}.$$

We may then say that

$$\lim_{t\to\infty} \beta T \sum_{j=0}^{t} e^{-j\delta} = \frac{e^{\delta}\beta T}{e^{\delta} - 1}.$$

We may then define a new parameter $\alpha = \left(\frac{e^{\delta}\beta}{e^{\delta}-1}\right)$ which we call the individual vestige surplus, given that it accounts for the new vestiges produced per individual, plus vestiges that were produced at previous time points, minus those

that have decayed. We can then write the marginal distribution of $V = V_t$ as $V \sim \text{Poisson}(\alpha T)$.

By assuming random deposition of vestiges, the distribution of the observed number of vestiges will depend on the coverage level of the sampling method (e.g. transects). We refer to this coverage as $\nu \in (0, 1)$. Given that study area and transect length are generally known, we assume that ν can be calculated and is thus a known parameter, and that every vestige within the covered area is detected. Therefore, the distribution of the observed number of vestiges Y, alongside the full modelling hierarchy, can be written using the following triple Poisson hierarchical model:

$$Y|T, G, \nu \sim \text{Poisson}(\alpha T \nu) \tag{5.1}$$

$$T|G \sim \text{Poisson}(G\lambda_N) \tag{5.2}$$

$$G \sim \text{Poisson}(\lambda_G) \tag{5.3}$$

The above expression for observed vestiges Y may involve only a single visit per transect (i.e., Y as a vector of length R), but may also involve multiple visits per transect (i.e., Y as an $R \times S$ matrix, where S is the number of visits to each transect). In each case, the true abundance T is assumed not to change between transect visits. Abundance estimates obtained from single-visit data are compared to those obtained from data involving temporal replication in Appendix 5.A.

The performance of the model is highly dependent on how well α, the individual vestige surplus, is estimated. Therefore we may either opt to set up an informative prior for α, or simply fix it as a "known" value. To estimate λ_G and λ_N we may use informative or non-informative priors, depending on the level of prior information that we possess.

It would also be reasonable to assume an aggregated process of resource allocation and/or aggregated animal behaviour, which may result in vestiges which are produced at a non-constant rate. This would, in turn, affect the number of vestiges present at each transect. A simple extension that would accommodate this assumption would be to treat the top tier of the hierarchy as an over-dispersed

process, e.g.

$$Y|T, G, \nu \sim \text{Negative Binomial}(\alpha T \nu, \phi) \tag{5.4}$$

and use a non-informative gamma prior to estimate ϕ, the over-dispersion parameter. Analogously, if the group sizes are believed to be highly variable, a reasonable modification to the model above would use $N_i \sim \text{Negative Binomial}(\lambda_N, \phi_N)$, which results in $T \sim \text{Negative Binomial}(G\lambda_N, G\phi_N)$. Finally, if numbers of groups are highly variable when comparing across different regions, we may assume $G \sim \text{Negative Binomial}(\lambda_G, \phi_G)$. This introduces a family of models that can accommodate different scenarios, depending on the species being monitored.

Gallant et al. (2007), Murray et al. (2005) , and Barnes (2001) examine the limitations associated with the use of scat surveys to estimate abundance. Among these limitations is the possibility for vestiges to be concentrated at certain sites. As described in Equation 5.4, the use of a negative binomial distribution for vestige counts provides a possible solution to this issue. Another of these limitations is the relatively small amount of a sample area that tends to be covered by transects. The triple Poisson model as detailed in Equation 5.2 incorporates the transect coverage rate in the vestige count Y, which allows us to take into account the fact that very small sections of the study area are visible during transect surveys, though as we will discover in section 5.5.2, very small transect coverage may lead to inflated estimates of abundance, and this should be taken into account at the experimental design phase. A final limitation is the possibility for vestiges produced at certain times of the year to decompose more quickly than those produced at other times. It is true that the triple Poisson model as described in this chapter assumes that vestiges decay at a constant rate. A subject of planned future work will allow vestiges to decay at different rates, possibly incorporating the effect of covariates into the rate of decay, which may allow us to model the transient phase of the population.

5.4 Simulation Studies

In this section we describe simulation studies which were conducted with the aim of determining the effect on abundance estimates of varying the choice of prior

distribution used for individual vestige surplus α. To this end, vestige count data was simulated from a triple Poisson model, and models were fitted in `JAGS` with different options for α. We wished to compare abundance estimate from scenarios when α is given a weakly-informative Uniform prior to scenarios when α is known and supplied correctly to the model as data, and when α is "known" but is incorrect, and is supplied as data.

The results are presented in Figure 5.1, in which the true value for α is 35 vestiges and the true abundance is 86. The small crossbar in Figure 5.1 represents the 95% Credible Interval for the estimate of abundance when α is known correctly. From this we can see that when the true value for α is known, estimates of abundance are highly accurate and precise. The large crossbar represents the 95% Credible interval for the estimate of abundance when a Uniform(10, 50) prior is supplied for α. When α is unknown and is given a Uniform prior, the estimate for abundance has still got a high degree of accuracy, but is now imprecise. The ribbon represents the 95% Credible Interval when incorrect values in the range $\{20, 50\}$ are supplied for α. In this case, the estimate for abundance is still precise (i.e., the credible interval is still quite narrow), but is now highly inaccurate.

Another possible scenario that we have assessed is that in which no prior information is available for the value of α, and a non-informative prior must be used. Results of these simulations are available in Appendix 5.A, and from these results we conclude that if no information is available for the value of α, then informative prior distributions are required for λ_N and λ_G.

Further to this, simulation studies wherein data is simulated from a triple Poison model were run to assess model estimation for different prior distributions on λ_G and λ_N, the effects of using a Poisson distribution on vestige count versus using a negative binomial distribution, an the effect on abundance estimation of using temporally replicated vestige counts (i.e. transects that are visited on multiple occasions). The full details and results of all simulation studies can be found in Appendix 5.A.

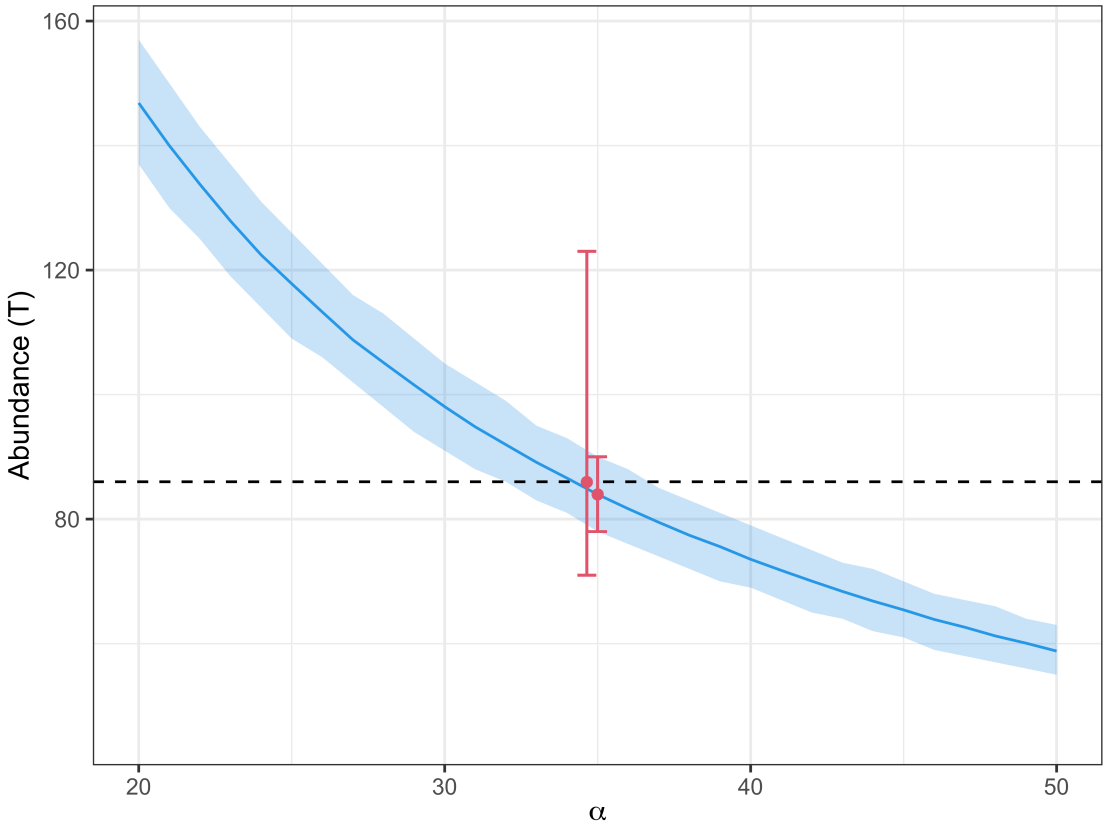


Figure 5.1: Abundance mean estimates and 95% Credible intervals when the true value of α is known correctly (short crossbar), when α is known incorrectly (ribbon) and when α is supplied a Uniform prior (long crossbar), with the true abundance denoted by the dashed horizontal line.

In addition to the simulation studies described above, studies were run to compare the accuracy of abundance estimates produced by the triple Poisson model to those produced by a distance sampling model (both in a frequentist and Bayesian framework), when data is scarce.

The Bayesian distance sampling model was implemented following the example of Kéry and Royle (2015). The distance sampling model involves employing counts of observed vestiges Y to estimate the true number of vestiges per transect z, and subsequently estimating vestige density, animal density, and finally animal abundance.

Data used for the distance sampling model is composed of J total vestiges, each collected from one of R transects, along with the distance between each vestige and the transect.

The observed vestige count at transect i ($i = 1, ..., R$) is given by

$$\begin{aligned} Y_i &\sim \text{Binomial}(z_i, p_i) \\ z_i &\sim \text{Poisson}(\beta_i) \end{aligned}$$

where z_i us the actual number of vestiges present, and p_i is the vestige detection probability at transect i.

The maximum distance from each transect at which vestiges might be observed is denoted as B, and this distance B is divided into $k = 1, ..., K$ intervals. $\pi_{i,k}$ is the probability of detecting a vestige within each of these K intervals around transect i, and is as follows.

$$\pi_{i,k} = \left(-\frac{m_k^2}{2\sigma_i^2}\right)\left(\frac{d}{B}\right)$$

where m is the midpoint of each interval, σ is the half-normal scale parameter, $-\frac{m_k^2}{2\sigma_i^2}$ is the half-normal detection function evaluated at m, and d is the length of each interval. The overall probability of detection for transect i is then:

$$p_i = \sum_{k=1}^{K} \pi_{i,k}$$

Because we are working with K distance intervals (labelled $1, \ldots, K$) the original continuous distance values must be converted into the corresponding distance interval. For observation j, we model this distance interval C_j using a categorical distribution as follows:

$$C_j \sim \text{Categorical}\left(\hat{\pi}_{*_j}\right)$$

where $\hat{\pi}_{*_j}$ is a vector of length K that contains the probabilities associated with vestige j being observed within each interval. Given that we know which transect each vestige is observed at, this parameter is calculated for all R transects as the $R \times K$ matrix $\frac{\pi_{i,k}}{p_i}$, and the probabilities associated with the appropriate transect are then utilised in the categorical distribution above.

The vestige density can be calculated as $D_{\text{vestiges}} = \frac{\sum_i z_i}{A}$, where A is the area of the sample space. This vestige density can now be used to determine animal abundance (Marques et al., 2001). The total number of vestiges produced per day is calculated as

$$T_{\text{daily vestiges}} = \frac{D_{\text{vestiges}}}{\delta},$$

where δ is the time until vestige decay. The animal density D can then be given as:

$$D = \frac{T_{\text{daily vestiges}}}{\lambda},$$

where λ is the vestige production rate. This density can then be used to calculate abundance as

$$T = D \times A,$$

Data was simulated from this distance sampling model, which required the specification of the size of the study area, the number of vestiges within the study area, the distance between transects, the truncation distance (the distance from a transect within which vestiges may be observed), the detection function (which is used to model the distribution of vestiges given their distance from a transect), the vestige production rate (which was specified as 15 vestiges per day), and the time to vestige decay (which was specified as 10 days). Data was simulated using the `DSsim` package (Marshall, 2020) using the `R` statistical software version 4.0.2 (R Core Team, 2022). Each simulation had a total of 5000 vestiges contained within a 2×5km area, with vestige density constant across the study area. A truncation distance of 10m was chosen, and transects were specified at a distance of 1km from each other, which ensured that each simulation contained only two transects, each 5km long. As a result, the data used in the triple Poisson model

is comprised of just two values, i.e., the total number of vestiges observed at each transect. A half-normal detection function was used to model the probability of observing vestiges, given their distance from the transect.

Triple Poisson models, with combinations of informative and non-informative λ_G and λ_N were fitted to these simulations. Simulation studies typically involve the true abundance given as a fixed value across all simulated datasets. However, because we are simulating from a distance sampling model, we first simulate 5000 total vestiges, and then use a probability detection function to determine which of these are visible near the transect, and finally calculate the true abundance. The result is that the true abundance T was slightly different for each simulated dataset ($T \in \{25, 45\}$). For this reason, in order to provide informative priors for λ_N and λ_G, it was assumed that the mean number of animals per group might be between three and seven, and the mean number of groups in the area might be between one and ten. This allowed us to use the informative priors $\lambda_N \sim \text{gamma}(5, 1)$ and $\lambda_G \sim \text{gamma}(10, 1)$.

The non-informative priors for λ_N or λ_G were specified using a $\text{gamma}(0.01, 0.01)$ distribution. Data was simulated with 5000 vestiges present in the study area, so the individual vestige surplus of the triple Poisson model α was given a weakly-informative $\text{Uniform}(10, 10000)$ prior, where α represents the number of new vestiges produced per individual per day plus the number of vestiges still present in the area from previous days. These models were implemented through JAGS (Plummer, 2003), using the `R2jags` package (Su and Yajima, 2020).

Subsequently, distance sampling models were fitted to these simulations, using both frequentist and Bayesian frameworks. The frequentist model was fitted using the `Distance` (Miller et al., 2019) package, with combinations of δ and λ specified correctly and incorrectly. The aim of the frequentist application of the distance sampling model was to determine the effect on abundance estimates of slight inaccuracies in the values supplied for these parameters. The Bayesian implementation of the distance sampling model was then performed so that the distance sampling model might be fairly compared with the Bayesian implementation of the triple Poisson model, and this involved the use of informative and non-informative

gamma prior distributions for δ and λ.

The full list of models fitted are given in Table 5.1. The accuracy of abundance estimates was assessed using relative bias for the true abundance T, averaged over all simulations, calculated as

$$\text{Relative mean bias} = \frac{\hat{T} - T}{T}.$$

The smaller the value for relative bias, the closer to the true value our estimated abundance values were.

(a) Triple Poisson Models			
Model	λ_N	λ_G	Abundance Relative Bias
TP1	Informative	Informative	0.109
TP2	Informative	Non-Informative	0.142
TP3	Non-Informative	Informative	0.186
TP4	Non-Informative	Non-Informative	0.104

(b)(i) Frequentist Distance Sampling Models			
Model	λ	δ	Abundance Relative Bias
DS1	15	10	0.000
DS2	16	10	0.091
DS3	15	11	0.062
DS4	16	11	0.147
DS5	17	12	0.265
DS6	18	13	0.359
DS7	19	14	0.436

(b)(ii) Bayesian Distance Sampling Models			
Model	λ	δ	Abundance Relative Bias
DS8	Known	Known	0.003
DS9	Informative	Informative	0.022
DS10	Informative	Non-Informative	1.592
DS11	Non-Informative	Informative	2.523
DS12	Non-Informative	Non-Informative	36.146

Table 5.1: (a) Triple Poisson models fit with informative and non-informative gamma priors on mean group size λ_N and mean number of groups λ_G (b)(i) Distance Sampling models fit using a frequentist framework, with individual vestige production (λ) and time to vestige decay (δ) supplied correctly ($\lambda = 15, \delta = 10$) and incorrectly, with incorrect values for $\lambda \in (16, 19)$ and incorrect values for $\delta \in (11, 14)$ (b)(ii) Distance Sampling models fit using a Bayesian framework fit with informative and non-informative gamma priors on individual vestige production λ and time to vestige decay δ.

In Figures 5.2, 5.3 and 5.4, we present the results of the comparison of true abundances with estimates obtained using triple Poisson models and distance sampling models. In the case of the Bayesian implementation of the distance sampling model (Figure 5.4), the model in which vestige production rate (λ) and time to vestige decay (δ) are assigned non-informative priors is not displayed. This is because - with estimates of abundance of approximately 12,000 individuals - it produces

estimates of abundance that are on a much larger scale than the other Bayesian distance sampling models, which makes results from the other models difficult to read.

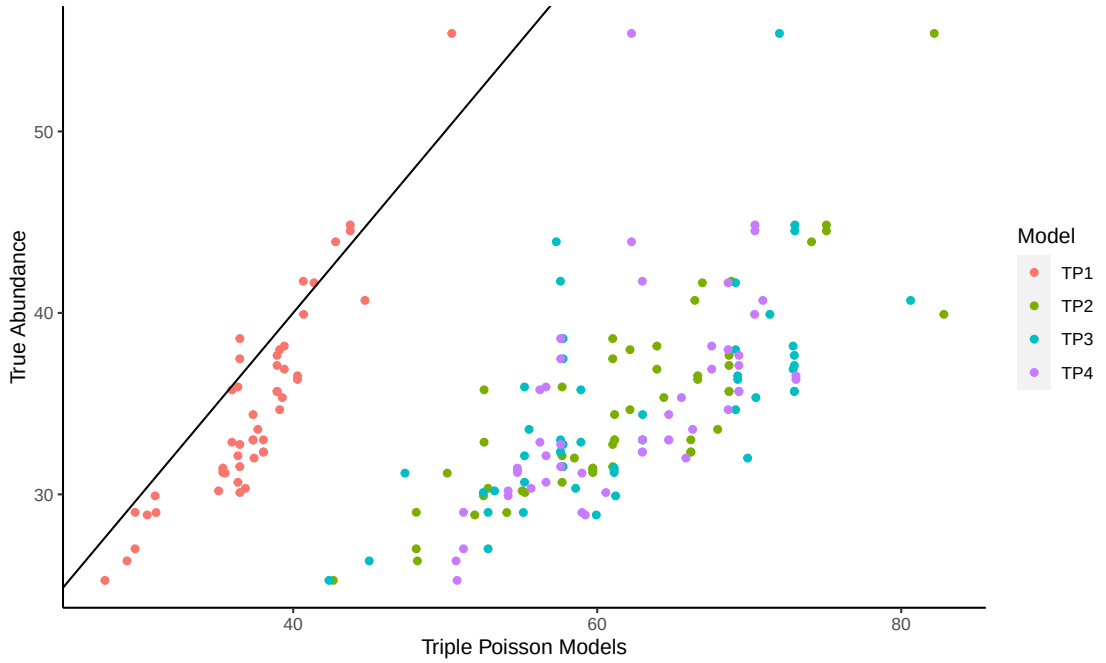


Figure 5.2: A scatter plot comparing the abundance estimates from triple Poisson models (TP1 – TP4) with varying priors on λ_N and λ_G to the true abundances.

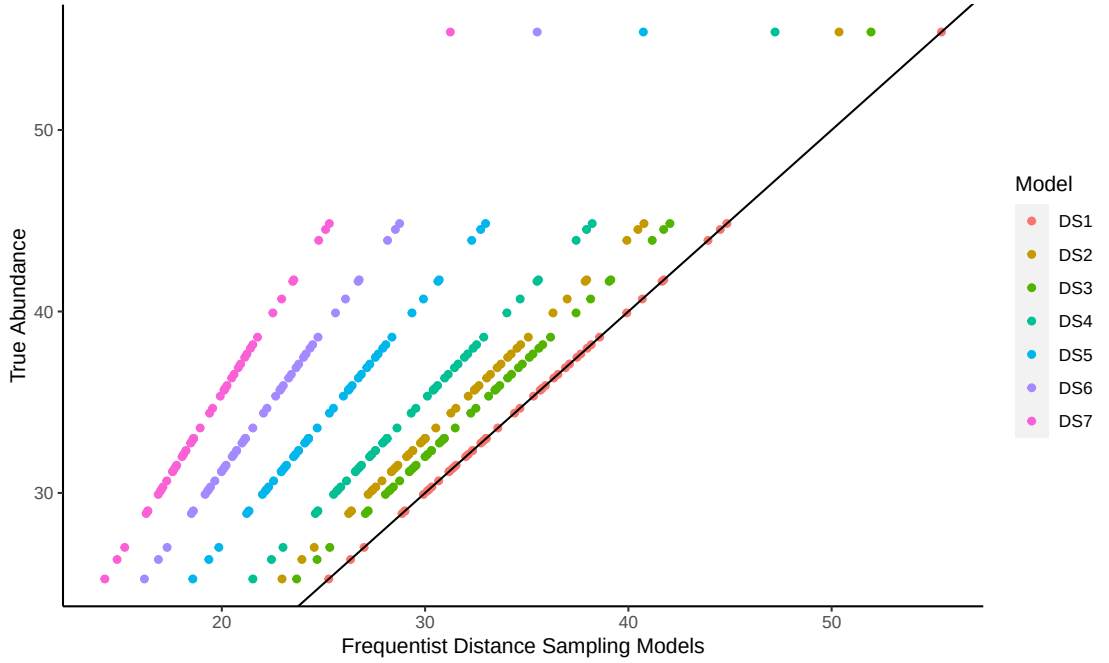


Figure 5.3: A scatter plot comparing the abundance estimates from frequentist implementations of the distance sampling model (DS1 – DS7, as described in Table 5.1) to the true abundances.

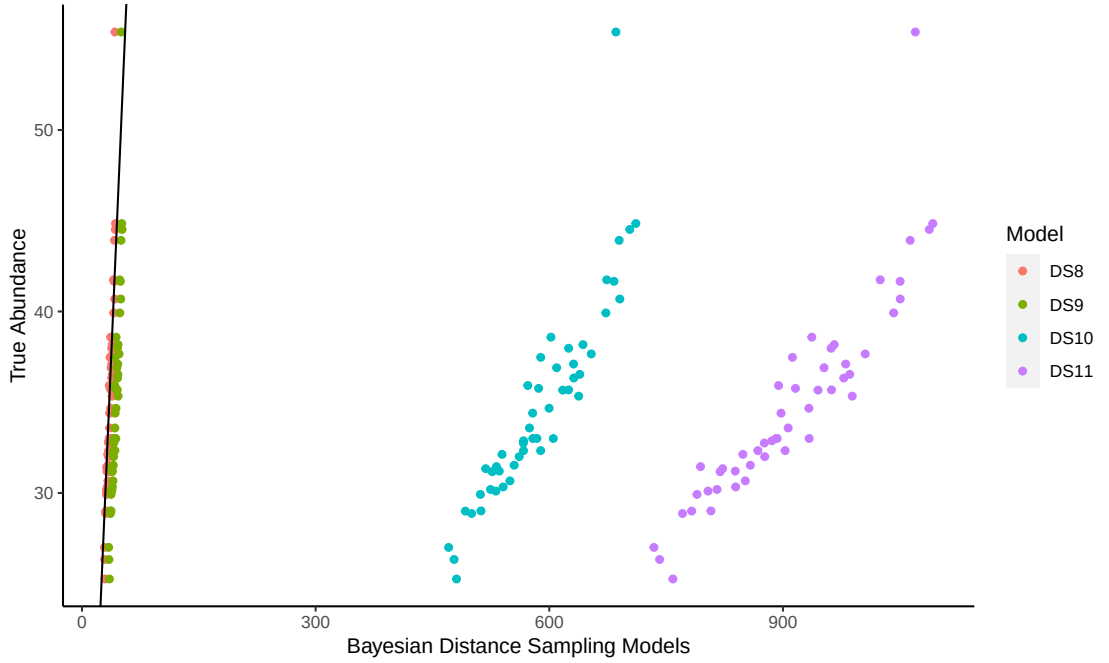


Figure 5.4: A scatter plot comparing the abundance estimates from Bayesian implementations of the distance sampling model (DS8 – DS11, as described in Table 5.1) to the true abundances.

5.5 Case Studies

As discussed in Chapter 2, we have three different case studies with which we demonstrate various aspects of the triple Poisson model. Initially this modelling framework was applied to each case study, assuming that the observed vestige count Y could be appropriately described using a Poisson distribution. Subsequently the models were re-run, assuming a negative binomial distribution for Y. For each case study the results of each of these models was compared using BIC and DIC values, and this comparison is presented in Appendix 5.C.

5.5.1 Collared Peccary

As described in Chapter 2, we first examine a case study regarding the collared peccary, (*Dicotyles tajacu*), using vestige data collected in southeast Brazil (Assis, 2012). The data collected as part of this case study is comprised of vestige counts

collected on a single sampling occasion from only two transects. As a result, the sample size is very small, as it contains only two counts: $(7, 1)$.

In order to implement the triple Poisson model on this dataset, we must first determine the prior distributions that we will assign to the mean group size λ_N and mean number of groups λ_G. In order to do this, we take any prior information available to us regarding these parameters into account. Collared peccary are estimated to live in groups of between seven and nine individuals (Sowls, 1997), and the number of groups in the area is estimated at between three and five. This information allowed us to place informative priors on λ_G and λ_N. However, for this case study, information is not available on the individual vestige surplus, so we are unable to place an informative prior on α in this case. For this reason we place a non-informative prior on vestige surplus: $\alpha \sim \text{gamma}(0.01, 0.01)$. As is demonstrated in Appendix 5.A, if a non-informative prior distribution must be used for vestige surplus α, and particularly if sample size is small, it is required that we specify informative prior distributions for λ_N and λ_G.

The transect along which seven vestiges were found was 8km long, while the other transect was 12km long. In order to estimate v, the transect coverage rate, we require the distance around the transect within which vestiges might be observed. This distance was not measured, so we make a conservative estimate that vestiges within 2m on either side of the transect are visible. This results in transects whose area are $32,000\text{m}^2$ and $48,000\text{m}^2$ respectively. The study area is 43.65km^2, and we can therefore estimate the transect coverage rate as $v \in \{0.00073, 0.0011\}$.

Our model was fitted with first a Poisson distribution on the vestige count Y, and then a negative binomial distribution, and the results were compared using DIC values. The model with the lower DIC value was the one with a negative binomial distribution on the observed number of vestiges Y. This model was chosen as the best fit, of the models available. The results obtained from this model were a mean estimate of a population of 44 collared peccaries within the study area, with a 95% Credible Interval of $(16, 87)$.

5.5.2 Sika Deer

As described in Chapter 2, the data analysed in this section (Marques et al., 2001) is available as part of the `Distance R` package (Miller et al., 2019). This data contains counts of sika deer (*Cervus nippon*) vestiges located in eight different regions in Scotland, as well as the distance from each vestige to the transect, measured in centimetres. Because this distance information is available, this case study provides an ideal opportunity to compare estimates produced using a triple Poisson model to those produced using a modification of a distance sampling model, as described by Marques et al. (2001).

As detailed briefly in Section 5.2, the implementation of this modified distance sampling model to estimate abundance using vestige data requires estimates of the rate at which vestiges are produced, and time to vestige decay. Our implementation of this modelling framework follows one described by Marques et al. (2001) and so we utilise their provided estimate that sika deer produce approximately 25 vestiges per day. The paper by Marques et al. (2001) provides estimates and standard errors for the time to vestige decay for this data. However, they provide these estimates based on the habitat groups for the data, and the month during which data is collected, neither of which we have access to within the dataset provided as part of the `Distance R` package (Miller et al., 2019). For this reason, we base this value upon analysis presented in Rexstad (2022) and obtain an estimate of 163.4 days (with a standard error of 14.2) for vestiges to decay.

We then implement a triple Poisson model on this data, and can compare estimates provided by these two methodologies. In order to implement our triple Poisson model, we must determine the prior distributions that should be assigned to estimate the mean group size λ_N and the mean number of groups λ_G. The prior information that we possess is that sika deer may live in groups of up to ten animals (Ratcliffe, 1987), and that territory size of sika deer is in the range of $0.02-0.12\text{km}^2$. For this reason, we place an informative prior on $\lambda_N \sim \text{gamma}(5, 1)$ which can allow for group sizes of up to approximately 15 individuals. Our knowledge of the typical territory size allows us to also place an informative prior on λ_G. Each of the eight regions has a different area, and so we can use each area and

the known territory size to provide a different prior distribution for λ_G for each region. The details of model implementation per region are available in Appendix 5.B.

From the estimates provided for the distance sampling implementation, we know that sika deer produce approximately 25 vestiges per day. We also assume that vestiges may take several months to decompose, and so we place a weakly informative prior distribution on α, i.e. $\alpha \sim \text{Uniform}(0, 4000)$. The use of this prior distribution implies that we assume that there are between 0 and 4000 vestiges present in the study area due to each individual, which, at a rate of 25 vestiges produced per day, allows vestiges to remain for up to approximately 5 months before becoming close to fully decomposed.

We know that the largest distance at which vestiges were observed from the transect was two metres. For this reason, the triple Poisson models are run assuming that vestiges are visible within two metres on either side of the transect, which allows us to estimate the transect coverage rate ν for each area.

Initially, each model was run assuming that the vestige count is adequately described using a Poisson distribution. This was then compared to a model in which the vestige count is assigned a negative binomial distribution. The model of best fit for each area was chosen using DIC values (the details of which are available in Appendix 5.C), and a comparison of abundance estimates produced using these triple Poisson models of best fit and the distance sampling model can be found in Table 5.1.

In this table we see that point estimates obtained using a triple Poisson model for areas A and B are very close to those obtained using a distance sampling model, while there is a greater difference in point estimates obtained for other areas. In particular, we see a large difference in abundance estimates for areas H and J. For these areas, the distance sampling model produces quite small estimates, while the triple Poisson model produces larger point estimates. This can be explained by a combination of factors. The first is the amount of data available for these areas. In each of areas H and J, only one transect is surveyed. This means that the triple Poisson models fitted to these regions are fitted to data that contains only

one observation. The simulation studies that we have performed for scarce data contain data collected from two transects, and so the models fitted to areas H and J are run using data that is more scarce than even the most scarce simulations performed. A further reason for the inflated point estimates obtained for these areas is the very small transect coverage rate associated with them. Areas A and B, for which triple Poisson point estimates are closest to distance sampling point estimates, have a coverage rate of 0.00048 and 0.00042 respectively. The transect coverage rates associates with areas A and B are twice as large as the next largest transect coverage rate (associated with area C), and over 10 times larger than those associated with areas H and J, which have coverage rates of 0.000071 and 0.000041 respectively. Due to the formulation of this model, the abundance T is confounded with transect coverage ν, and very small values for ν may lead to inflated estimates for abundance T.

Area	Distance Sampling		Triple Poisson	
	Estimate	95% CI	Estimate	95% CI
A	1027	(690, 1528)	1028	(584, 1952)
B	382	(219, 667)	409	(167,982)
C	33	(15, 74)	106	(20, 451)
E	29	(8, 99)	71	(19,312)
F	209	(173, 252)	340	(128, 1036)
G	125	(18, 856)	281	(78, 834)
H	17	(14, 21)	95	(11, 431)
J	69	(57, 83)	165	(35, 622)

Table 5.1: Mean estimates and their 95% confidence (distance sampling) or credible (triple Poisson) intervals for sika deer abundance per area from a distance sampling model and the triple Poisson model.

5.5.3 Red Foxes

As detailed in Chapter 2, the final case study examined as part of this chapter involves vestige data collected from the red fox (*Vulpes vulpes*) by Cavallini (1994). This data was collected from nine transects in central Italy. In the paper by Cavallini (1994), an index of abundance is obtained, estimated as the number of vestiges per kilometre. This data was collected once every month over the course of a year, and so it is a good candidate for our triple Poisson modelling framework

that takes into account temporal replication. As detailed in Section 5.3, here the observed vestige count Y is allowed to vary by transect and month. However, in our model implementation we assume that the true abundance T varies only by transect.

Before we can implement the triple Poisson model on this case study, we must first determine the prior distributions that to be assigned to the mean group size λ_N and mean number of groups λ_G. We possess prior knowledge on common group sizes for red foxes, as well as their territory size, which will allow us to place informative priors on λ_G and λ_N.

Red foxes may have as many as nine cubs in a litter, which means that a group of red foxes may contain as many as 11 individuals Baker and Harris (2004). Their territory ranges from 5 to 12km^2. This case study is contained within a study area of size 2448km^2, and so in this area there could be as many as 490 groups of red foxes.

Transect coverage in this case was again calculated using transect length and study area, assuming that vestiges within 2m of the transect are visible. To inform our prior distribution of vestige surplus α, we know that red foxes may produce eight vestiges per day (Webbon et al., 2004). This data was collected monthly over a 12 month period, so we remove the initial observations (April 1992) for each region, as we do not know how long these vestiges may have been present in the environment. The original paper (Cavallini, 1994) mentions that at each sampling occasion the vestiges observed were collected and taken for laboratory analysis. For this reason, counts collected disregard counts initially collected in previous months, and so we assume that the longest time that the vestiges observed may have been present in the environment was approximately 30 days. The prior distribution used for vestige surplus α reflects this belief that each individual might produce eight vestiges per day, and these vestiges are only present for approximately one month until the following sampling occasion: $\alpha \sim \text{Uniform}(0, 240)$.

The model was fitted initially with a Poisson distribution on the vestige count, and then a negative binomial distribution. The results were compared using DIC values, and the model with the lower DIC value was the one with a negative

binomial distribution on the vestige count. This model was chosen as the best fit for this data. The details of the comparison via DIC values are available in Appendix 5.C. The result obtained from this model was a mean estimate of 3304 red foxes in the study area, with a 95% Credible Interval of $(2603, 4442)$.

5.6 Discussion

In this chapter, we have presented a novel modelling framework that allows for the estimation of animal abundance using vestige data. We have presented the results of an extensive simulation study, which demonstrates the performance of this model under a range of conditions. Full details of these simulations are available in Appendix 5.A.

As part of the simulation study, we simulated scarce data from a distance sampling model, and fitted various distance sampling and triple Poisson models to this data to compare abundance estimates. While distance sampling models consistently produce accurate estimates of abundance when vestige production rate λ and time to vestige decay δ are known precisely (Table 5.1 (b)), small inaccuracies in λ and δ produce estimates of abundance with relative bias that increases quite rapidly. Triple Poisson models were fitted with priors on λ_N and λ_G both informative and non-informative. These models produced estimates of abundance with relative bias $\in (0.1, 0.2)$ regardless of prior choice for λ_N or λ_G. In Figure 5.2, the triple Poisson model with informative priors on λ_G and λ_N better reflects the true abundances, while models for which information is unavailable to inform λ_N and λ_G do not perform quite as well.

While these scarce data simulation studies (detailed in Section 5.4) reveal that the triple Poisson model can produce unbiased abundance estimates when even very few transects are available for surveying, having more data is always preferable. Simulation studies designed to examine the effect of sample size on abundance estimates reveal that scenarios in which a greater number of sites are available for survey produce less biased abundance estimates, as might be expected. Model formulation suggests that abundance estimates T may be confounded with transect coverage rate ν. The results obtained using the sika deer case study in Section

5.5 demonstrate this. When the amount of the study area covered by transects is very small, estimates of abundance will be inflated, as can be seen in Table 5.1. This highlights the need at the experimental design phase to ensure the use of appropriately sized study areas for which adequate transect coverage can be obtained.

Simulations reveal that vestige surplus α is confounded with abundance T, which is also to be expected given model formulation. If α is known, estimates for abundance are not dependent on prior distributions used for λ_G and λ_N. When α is not known precisely, but enough information is available to provide a prior distribution that is weakly informative, dependence on λ_G and λ_N increases. Finally, if no information is available and a non-informative prior must be supplied for α, to avoid producing abundance estimates that are highly biased, informative priors are required for both λ_G and λ_N.

The assumption that vestige count Y can be adequately described using a Poisson distribution relies on vestige deposition constant across the study area. In many cases, this can prove to be an unrealistic assumption. In such cases, a simple alternative is the use of a negative binomial distribution to describe the vestige count. This allows for overdispersion in vestige counts, which could be attributed to the occurrence of vestiges at higher rates in some areas, due to factors such as resource allocation, animal behaviour, or non-constant rates of vestige production. The use of a negative binomial distribution for vestige counts was examined by simulation study. A notable finding of this study was that even a small level of overdispersion in vestige counts can cause quite a large increase in relative bias of abundance estimates, when compared to the Poisson model for vestige count. This impact is seen most strongly when both sample size is small and non-informative prior distributions are used for both λ_G and λ_N. This occurs even when α is known. Consequently, if overdispersion is expected in vestige count, and prior knowledge that may inform λ_G and λ_N is not available, it is advisable that larger sample sizes are collected, either by increasing numbers of transects established, or by visiting transects on multiple occasions.

In this study we assumed vestiges experience an exponential rate of decay. How-

ever, in the future we plan to examine alternative scenarios, in which vestiges may decay according to different distributions, to allow us to model the transient phase of the population, as mentioned in Section 5.3. In the future, we intend to perform agent-based modelling simulations of animal movement, while incorporating vestige deposition and decay probabilities. By assuming different sampling designs, we will then be able to estimate the abundance T and compare with the true number of animals in the environment. This might facilitate field work on data collection and improve cost-efficiency of animal counting.

The decision-making process concerning wildlife management is usually based on the cost-efficiency relations of survey/monitoring methods available (Nichols, 2014). Traditional methods involving capturing and/or direct sight seen of animals tend to be invasive, time-consuming, and relatively expensive (Verdade et al., 2013, 2014). In addition, traditional methods of animal counting tend to have low precision and unknown accuracy (Verdade et al., 2014). The use of vestiges on the estimation of actual abundance or population density simplifies the process of animal counting; therefore, improving the decision-making process concerning wildlife management.

Appendix

5.A Simulation Studies

In this section, we present the results of simulations that were run to examine the effect on abundance estimates of varying parameter values and prior distributions. To assess abundance estimates for different population sizes, we assigned true values for group size $\lambda_N \in \{5, 10\}$, and number of groups $\lambda_G \in \{5, 10\}$. We also examine the effect of our level of prior knowledge of vestige surplus α, using scenarios in which α is presumed somehow "known", and can be supplied to the triple Poisson model as data, and comparing this to scenarios in which a level of prior information allows us to place weakly informative uniform priors on α, and scenarios in which we possess no prior information for α and must assign it a non-informative gamma prior. To examine the effect of sample size, we present abundance estimates when data is collected on a single visit to 10 transects, and compare these to estimates obtained using data collected on a single visit to 100 transects. We then briefly examine the effect of temporal replication by estimating abundance using data collected on 10 visits to 10 transects.

The effect on abundance estimates of prior information on mean group size λ_N and mean number of groups λ_G is assessed using combinations of informative Gamma(7,1) and non-informative Gamma(0.01, 0.01) prior distributions, where the gamma distributions are parameterised in terms of their shape and rate.

To assess abundance estimates when overdispersion is present in our data, we first perform simulations which allow for overdispersion in the number of vestiges observed at each transect Y as follows.

$$Y|T, G, \nu \sim \text{Negative Binomial}(\alpha T\nu, \phi)$$

where $\phi \in \{0.2, 2\}$ is the overdispersion parameter. The smaller the value of ϕ, the greater the degree of overdispersion present in the vestige counts. This allows us to account for situations in which an aggregated process of resource allocation or animal behaviour cause vestiges to be produced at a non-constant rate.

Finally, we present simulation results for a "triple negative binomial" model, formulated as follows.

$$\begin{aligned} Y|T, G, \nu &\sim \text{Negative Binomial}(\alpha T\nu, \phi) \\ G &\sim \text{Negative Binomial}(\lambda_G, \phi_G) \\ T &\sim \text{Negative Binomial}(G\lambda_N, G\phi_N) \end{aligned}$$

This formulation allows for overdispersion in all layers of the model hierarchy, which allows for a non-constant vestige production, and allows for variability in group size and numbers of groups.

Overall, the simulation study examined for the triple Poisson model contained a total of 288 variable combinations, each of which was simulated 100 times. In each case we estimated mean group size λ_N, mean number of groups λ_G, abundance T and the vestige surplus α.

The accuracy of abundance estimates was assessed using relative bias for the true abundance T, averaged over all simulations, calculated as follows.

$$\text{Relative bias} = \frac{\hat{T} - T}{T}.$$

The smaller the value for relative bias, the closer to the true value our estimated abundance values were. Figures 5.A.1 to 5.A.7 present results for each of the previously described simulations. Each is divided into four sub-figures, (a) to (d) corresponding to varying prior distributions assigned to λ_N and λ_G.

(a) This sub-figure corresponds to a scenario in which prior information is available for group size ($\lambda_N \sim \text{Gamma}(7, 1)$) but not number of groups ($\lambda_G \sim \text{Gamma}(0.01, 0.01)$)

(b) This sub-figure corresponds to a scenario in which prior information is available for number of groups ($\lambda_G \sim \text{Gamma}(7, 1)$) but not group size ($\lambda_N \sim \text{Gamma}(0.01, 0.01)$)

(c) This sub-figure corresponds to a scenario in which prior information is available for both group size ($\lambda_N \sim \text{Gamma}(7, 1)$) and number of groups ($\lambda_G \sim \text{Gamma}(7, 1)$)

(d) This sub-figure corresponds to a scenario in which prior information is not available for group size ($\lambda_N \sim \text{Gamma}(0.01, 0.01)$) or number of groups ($\lambda_G \sim \text{Gamma}(0.01, 0.01)$)

Within each sub-figure we present relative biases for the mean estimate estimate and the 95% credible interval for the abundance, using data collected on a single visit to 10 transects, data collected on a single visit to 100 transects, and data collected on 10 visits to 10 transects. The true value for vestige surplus α in all cases is 20 vestiges.

In Figure 5.A.1, we present results when observed vestige count $Y \sim \text{Poisson}(\alpha T \nu)$, and vestige surplus α is known and correctly supplied to the model as data. Relative bias in abundance estimates in each case are very small, and there are no obvious differences in relative bias across sub-figures (a) to (d). This suggests that when α is known, prior distributions used for λ_N and λ_G have little effect on abundance estimation. In each case, relative bias is largest when using data collected on a single visit to 10 transects, and results for data collected on a single visit to 100 transects are comparable to those collected on 10 visits to 10 transects.

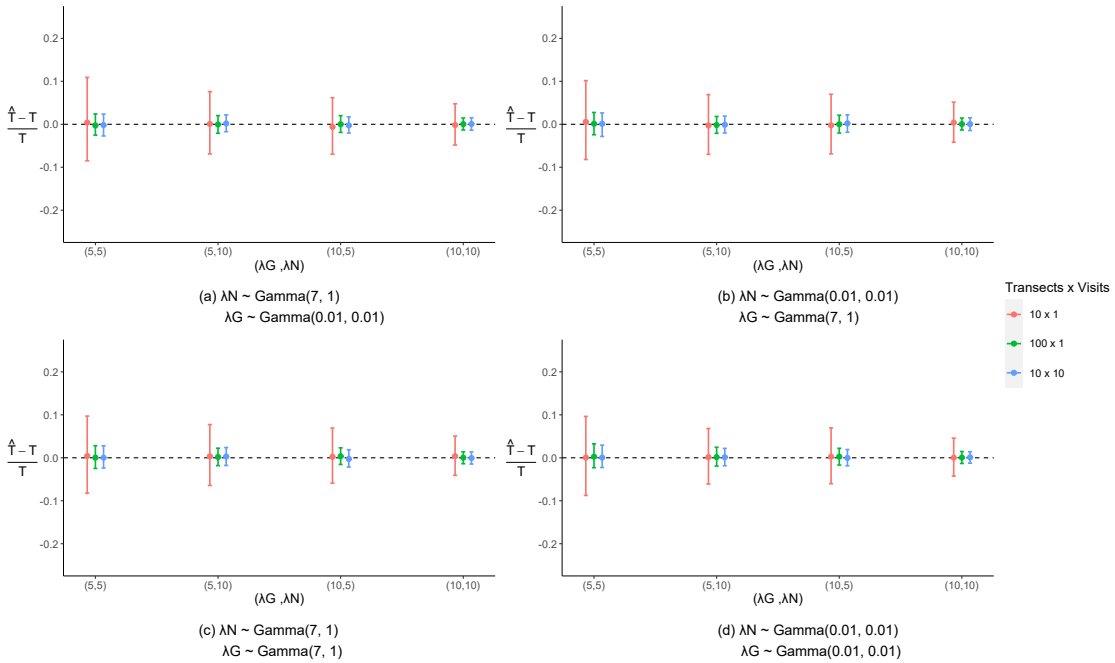


Figure 5.A.1: Relative bias in mean estimate and 95% credible intervals for abundance T, for Poisson Y and α known, with true values for λ_N and $\lambda_G \in \{5, 10\}$, at varying sample sizes, and for different prior distributions on λ_N and λ_G.

In Figure 5.A.2, we present results when observed vestige count $Y \sim \text{Poisson}(\alpha T \nu)$, and vague prior knowledge of vestige surplus allows us to specify $\alpha \sim \text{Uniform}(1, 100)$. Relative biases are larger when α must be estimated, when compared with those in Figure 5.A.1. Relative biases for Figures 5.A.2(a), 5.A.2(b) and 5.A.2(c) are very similar to each other, and display very little difference due to sample size. However, a difference in relative bias is obvious in Figure 5.A.2(d), when non-informative gamma prior distributions are used for both λ_N and λ_G. In this case, relative biases are larger, and appear dependent on sample size, with the largest relative bias associated with the smallest sample size, which contains data collected on a single visit to 10 transects, and the smallest relative bias associated with data collected over 10 visits to 10 transects. This suggests that when α must be estimated with a weakly-informative uniform prior, best results are obtained when information is available for at least one of λ_G and λ_N.

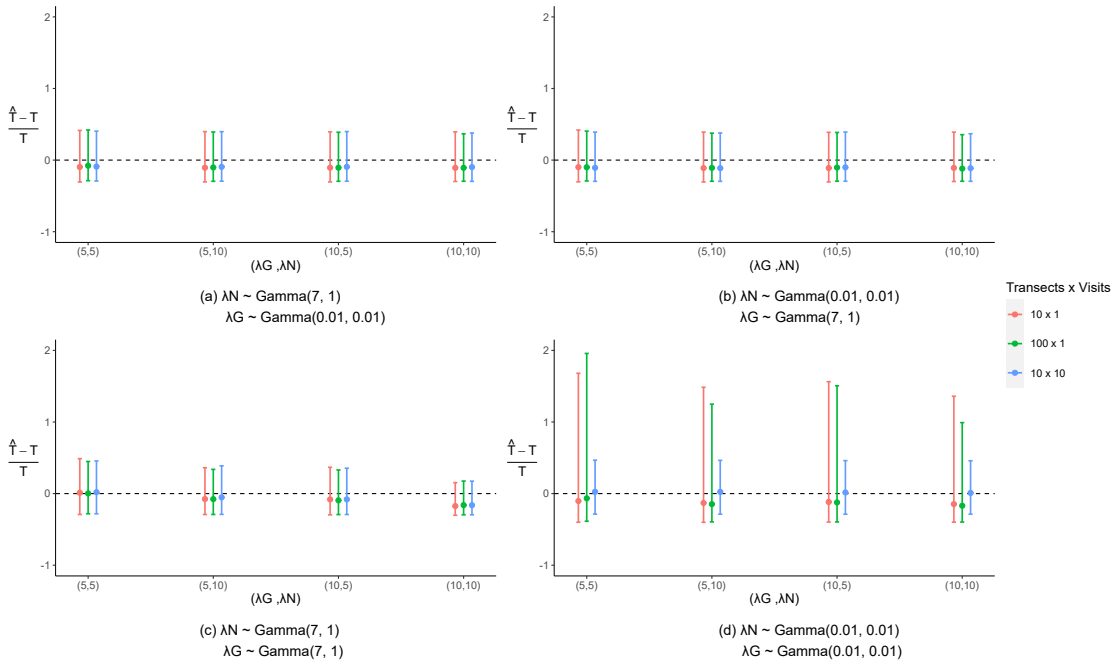


Figure 5.A.2: Relative bias in mean estimate and 95% credible intervals for abundance T, for Poisson Y and uniform α, with true values for λ_N and $\lambda_G \in \{5, 10\}$, at varying sample sizes, and for different prior distributions on λ_N and λ_G.

In Figure 5.A.3 we present results when observed vestige count $Y \sim \text{Poisson}(\alpha T \nu)$, and no prior information is available for vestige surplus, and so it must be specified as $\alpha \sim \text{Gamma}(0.01, 0.01)$. In this case we can see that there is large variability present in relative bias for abundance estimates,and that relative biases are much larger than the scenarios presented in Figure 5.A.1 and Figure 5.A.2. This is particularly evident in Figure 5.A.3(d), which corresponds to a situation in which no information is available for α, λ_N or λ_G, and all are estimated using non-informative gamma priors. In this case,relative biases of up to approximately 110 tell us that abundance estimates are unreliable when no information is available for α, λ_N or λ_G, and the triple Poisson model should not be used to produce abundance estimates in this case. Relative biases are also very high in Figure 5.A.3(a) and Figure 5.A.3(b), when information is only available for either λ_N or λ_G, but not both. Figure 5.A.3(c), which corresponds to a situation in which prior information is available for both λ_N and λ_G produces much smaller relative biases, and is the only scenario in which abundance estimates may be relied upon. From

this we may conclude that if no information is available for α, the triple Poisson model may be used to estimate abundance, provided that information is available for λ_N and λ_G, and both can be specified by informative prior distributions.

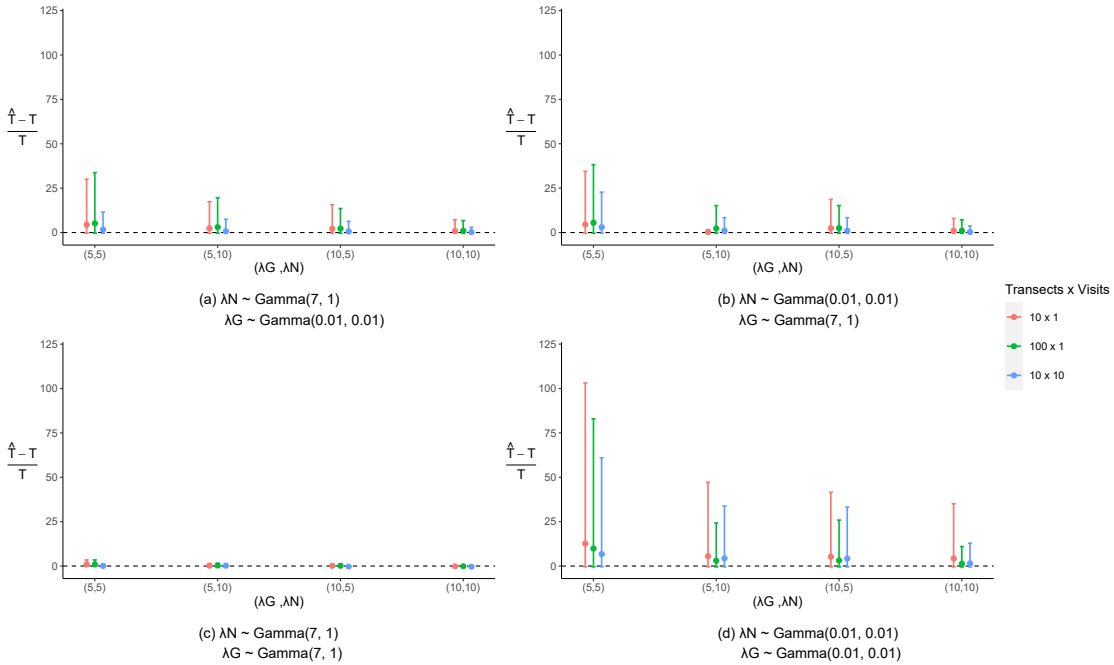


Figure 5.A.3: Relative bias in mean estimate and 95% credible intervals for abundance T, for Poisson Y and α estimated using a non-informative gamma distribution, with true values for λ_N and $\lambda_G \in \{5, 10\}$, at varying sample sizes, and for different prior distributions on λ_N and λ_G.

In Figure 5.A.4, we present results when $Y \sim$ Negative Binomial$(\alpha T\nu, \phi)$, overdispersion levels in vestige counts are small ($\phi = 2$), and vestige surplus α is known and correctly supplied to the model as data. Results in Figures 5.A.4(a)-(c) are very similar to those observed in Figure 5.A.1, when $Y \sim$ Poisson$(\alpha T\nu)$ and α is known. However, relative bias presented in Figure 5.A.4(d) is much larger than relative bias presented in the corresponding Figure 5.A.1(d). This is most evident when data collected on a single visit to 10 transects means that sample size is small. This suggests that when overdispersion is present in vestige counts, even if levels of overdispersion are small, that larger sample sizes are required for reliable abundance estimation.

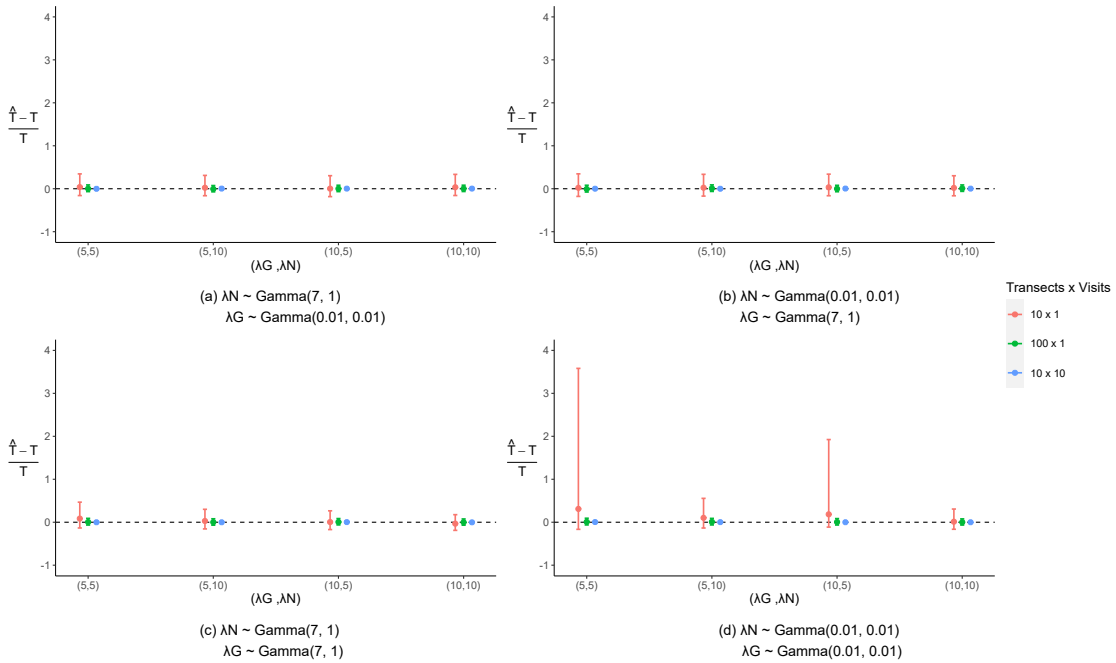


Figure 5.A.4: Relative bias in mean estimate and 95% credible intervals for abundance T, for negative binomial Y with small levels of overdispersion and α known, with true values for λ_N and $\lambda_G \in \{5, 10\}$, at varying sample sizes, and for different prior distributions on λ_N and λ_G.

In Figure 5.A.5, we present results obtained when $Y \sim$ Negative Binomial$(\alpha T\nu, \phi)$, overdispersion levels in vestige counts are large ($\phi = 0.2$), and vestige surplus α is known and correctly supplied to the model as data. Results are similar to those presented in Figure 5.A.4, though the issues with relative bias encountered in Figure 5.A.4(d) are worsened in Figure 5.A.5 by the greater levels of overdispersion present in this data. The conclusions that we can draw from this simulation are similar to those from Figure 5.A.4. When overdispersion is present in vestige counts Y, due to animal behaviour or non-constant vestige production, the ideal situation involves prior information available to inform prior distributions for λ_N and/or λ_G, and larger samples are preferable to obtain accurate estimates for abundance.

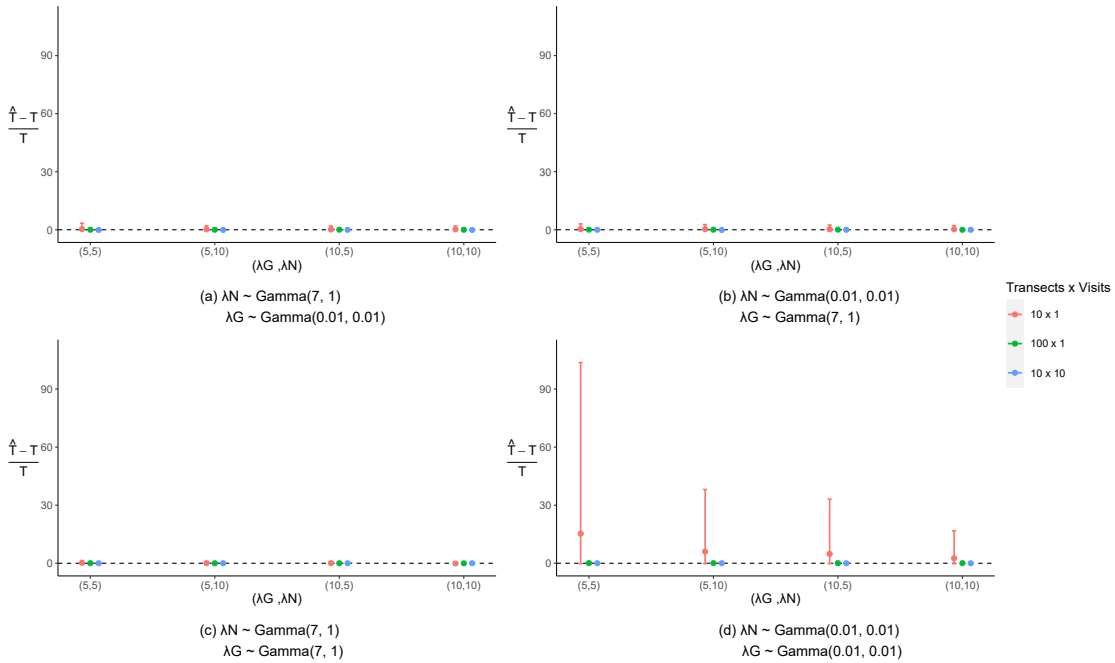


Figure 5.A.5: Relative bias in mean estimate and 95% credible intervals for abundance T, for negative binomial Y with large levels of overdispersion and α known, with true values for λ_N and $\lambda_G \in \{5, 10\}$, at varying sample sizes, and for different prior distributions on λ_N and λ_G.

In Figures 5.A.6 and 5.A.7, we present results obtained when $Y \sim$ Negative Binomial$(\alpha T\nu, \phi)$, and vestige surplus is estimated as $\alpha \sim$ Uniform$(1, 100)$. Figure 5.A.6 contains small overdispersion in vestige counts ($\phi = 2$) and 5.A.7 contains vestige counts with greater overdispersion ($\phi = 0.2$). The scenarios presented here are similar to the previous negative binomial simulations presented in Figures 5.A.4 and 5.A.5. The issue of very large relative bias when non-informative gamma prior distributions are used for λ_N and λ_G are again worsened by both the estimation of α and the increase in the degree of overdispersion present in the counts.

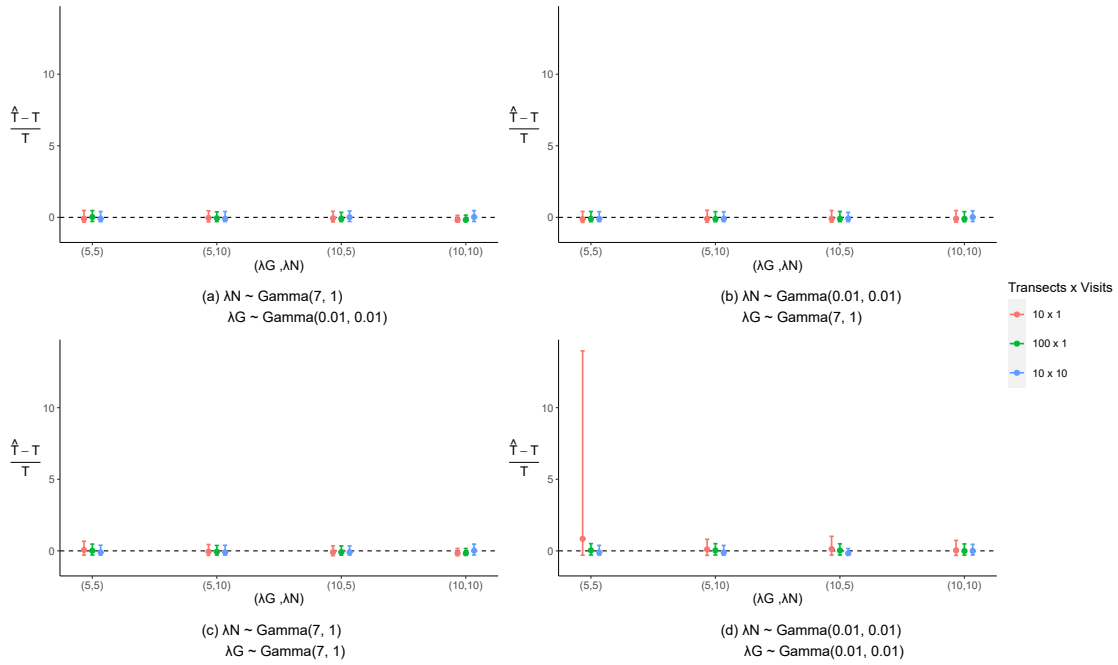


Figure 5.A.6: Relative bias in mean estimate and 95% credible intervals for abundance T, for negative binomial Y with small levels of overdispersion and α estimated with a uniform distribution, with true values for λ_N and $\lambda_G \in \{5, 10\}$, at varying sample sizes, and for different prior distributions on λ_N and λ_G.

Finally, simulation studies were performed to assess estimates obtained using a triple negative binomial model, in which the observed vestige count Y, the number of groups G and the abundance T are all assumed to contain overdispersion. Data was simulated from three different models: a triple Poisson model (TP), a triple negative binomial model with small levels of overdispersion in Y, T and G (TNB-1) and finally a triple negative binomial model with large levels of overdispersion in Y, T and G (TNB-2). Six different triple Poisson and triple negative binomial models were then fitted to each of these sets of simulated data. These models were fitted with varying specifications in terms of the vestige surplus α and the mean number of groups λ_G. The models fitted were as follows:

(A) A triple negative binomial model, for which α is known, and a non-informative Gamma(0.01, 0.01) prior distribution is provided for λ_G;

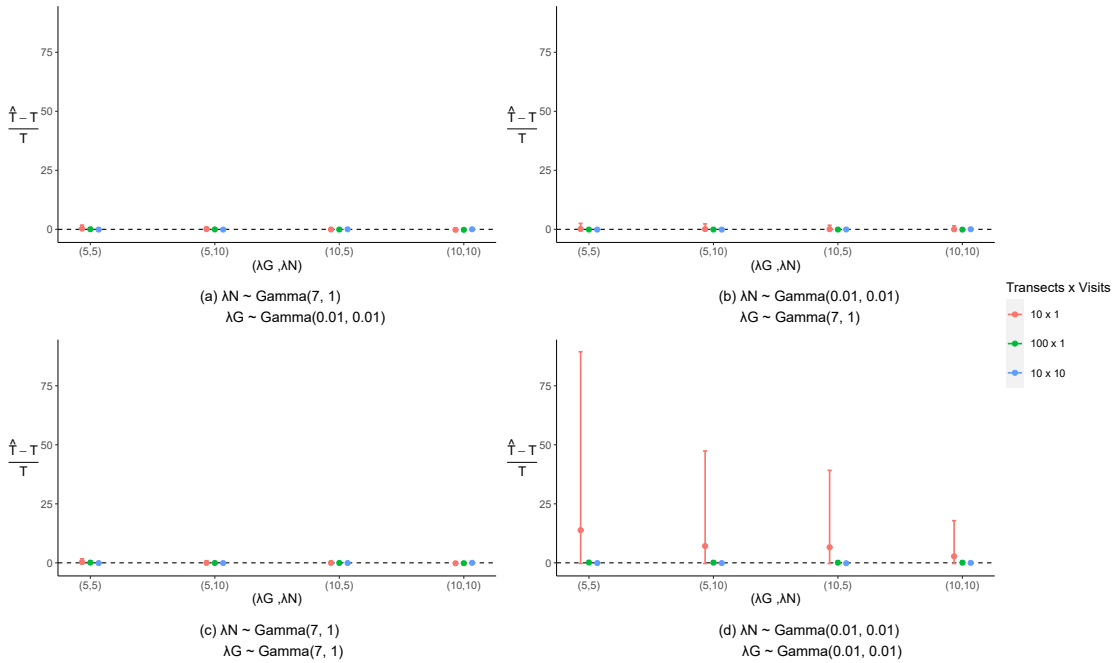


Figure 5.A.7: Relative bias in mean estimate and 95% credible intervals for abundance T, for negative binomial Y with large levels of overdispersion and information available to inform a uniform α, with true values for λ_N and $\lambda_G \in \{5, 10\}$, at varying sample sizes, and for different prior distributions on λ_N and λ_G.

(B) A triple negative binomial model, for which α is known, and an informative Gamma(7, 1) prior distribution is provided for λ_G;

(C) A triple negative binomial model, for which α is estimated using a uniform distribution, and a non-informative Gamma(0.01, 0.01) prior distribution is provided for λ_G;

(D) A triple negative binomial model, for which α is estimated using a uniform distribution, and an informative Gamma(7, 1) prior distribution is provided for λ_G;

(E) A triple Poisson model, for which α is known, and a non-informative Gamma(0.01, 0.01) prior is provided for λ_G;

(F) A triple Poisson model, for which α is estimated using a uniform distribution, and a non-informative Gamma(0.01, 0.01) prior is provided for λ_G.

In Figure 5.A.8, results for data simulated from a triple Poisson model (TP) are very similar to results for data simulated from a triple negative binomial model with small overdispersion (TNB-1). This indicates that if the true data had small amounts of overdispersion, that fitting a triple Poisson model would suffice, and that if the true data were triple Poisson, that fitting a triple negative binomial model would provide similar results. As we have seen in simulation studies described above, relative bias in abundance estimates is larger when α is estimated (models C and D) than when it is known and provided as data. When α is estimated using a Uniform distribution, an informative gamma prior on λ_G (model D) produces less uncertainty in abundance estimates than a non-informative gamma prior on λ_G (model C). For both of these simulations, when α is known and provided as data, relative bias in abundance estimates is very small regardless of the model fitted or the prior distribution provided to λ_G.

When data is simulated from a triple negative binomial model with higher levels of overdispersion (TNB-2), relative bias in abundance estimates are larger in almost all scenarios. When α must be estimated (models C, D and F), relative bias in abundance estimates are considerably higher for this data. When α is known and an informative gamma prior is supplied for λ_G (model B), relative bias in abundance estimates are similar to those associated with the triple Poisson and triple negative binomial with small overdispersion simulations. However, when α is known and a non-informative prior is supplied for λ_G (model A), relative bias in abundance estimates increases again. From this we can conclude that when overdispersion levels in our data are high, that an informative prior for λ_G and prior information on α are required to reduce uncertainty in abundance estimates.

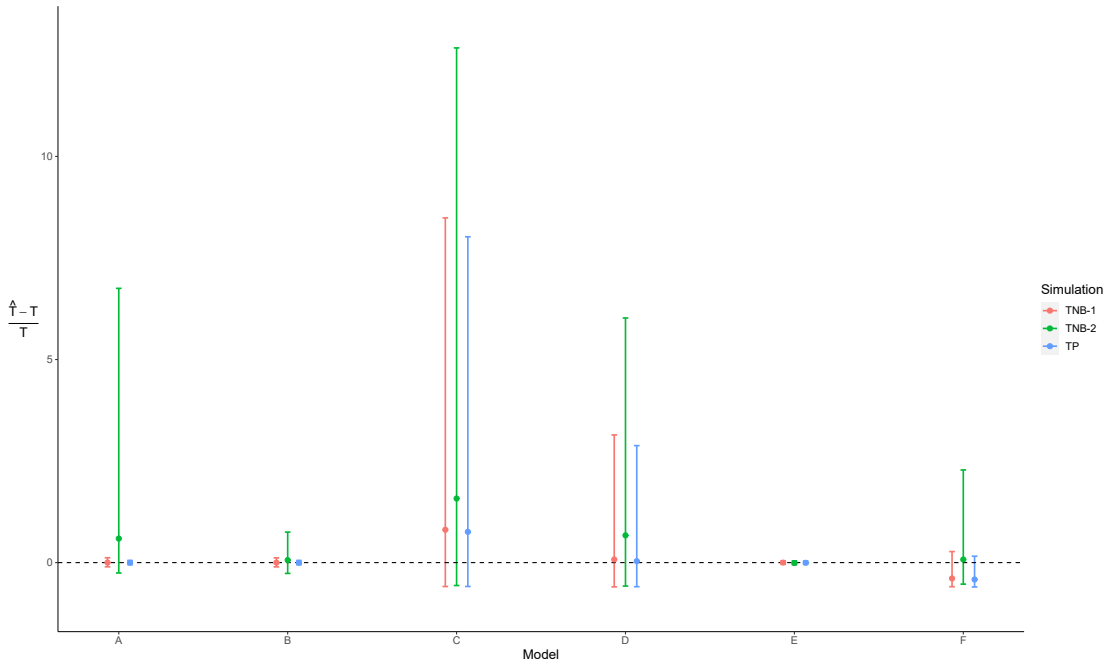


Figure 5.A.8: Relative mean bias in abundance estimates for triple negative binomial models (A-D) and triple Poisson models (E-F), fitted to data simulated from triple negative binomial models (denoted by the green and red lines) and triple Poisson models (denoted by the blue line).

5.B Case Studies

In this section we present the details of model implementation for each of our three case studies. In order to choose a prior distribution for the number of groups of animals in an area, in each case study we use the size of the study area and prior information regarding the species' territory size to obtain a theoretical maximum number of groups that might be in the area.

5.B.1 Sika Deer

The sika deer dataset contains vestige counts collected from eight different areas, each of a different size with sizes ranging from 8km^2 to 15.2km^2. For this reason, we obtain a separate theoretical group maximums per area. In order to determine the prior to use for λ_G, we use the size of the region, and prior knowledge that the territory of sika deer lies between two and twelve hectares. Table 5.B.1 contains the

area of each region, the theoretical maximum number of groups it might contain, and the gamma prior chosen for our analysis.

Region	Area	Max. Groups	Prior
A	13.9km^2	700	$\text{gamma}(1, 72)$
B	10.3km^2	500	$\text{gamma}(1, 45)$
C	8.6km^2	430	$\text{gamma}(1, 40)$
E	8km^2	400	$\text{gamma}(1, 40)$
F	14km^2	700	$\text{gamma}(1, 72)$
G	15.2km^2	760	$\text{gamma}(1, 75)$
H	11.3km^2	565	$\text{gamma}(1, 52)$
J	9.6km^2	480	$\text{gamma}(1, 50)$

Table 5.B.1: This table presents the area of each of the eight regions in the sika deer dataset, as well as the theoretical maximum number of groups of sika deer that each area could contain, and the prior assigned for λG during our analyses

5.B.2 Red Foxes

Similar to the sika deer dataset, we know that the size of the study area in this case is 2448km^2. We also possess prior information on the average territory size of a red fox, which ranges from 5km^2 to 12km^2. For this reason we estimate that the theoretical maximum number of groups of red foxes within this study area is 490. We can now specify that $\lambda_G \sim \text{gamma}(1, 50)$, which allows a maximum of between 400 and 500 groups.

5.C DIC Values

Table 5.C.1 provides the DIC and BIC values associated with a model with a Poisson vestige count and a negative binomial vestige count for each of the case studies examined as part of this chapter. In each comparison, the DIC values and BIC values were in agreement as to the model of best fit. For each case study, the model with the lower DIC/BIC value was chosen as the optimal fit.

	DIC		BIC	
Model	Poisson	Negative Binomial	Poisson	Negative Binomial
Collared Peccary	16.12	15.14	15.54	14.44
Red Fox	1195.94	629.33	1427.36	759.74
Sika Deer A	555.63	145.10	561.37	151.39
Sika Deer B	283.9	94.40	288.82	99.69
Sika Deer C	14.92	15.52	16.37	17.62
Sika Deer E	46.25	36.34	48.41	34.94
Sika Deer F	7.29	8.41	5.32	6.26
Sika Deer G	34.839	27.52	36.22	25.72
Sika Deer H	5.07	5.50	3.34	3.52
Sika Deer J	5.82	6.34	3.96	4.35

Table 5.C.1: DIC and BIC values for models fitted to case studies in which we compare a Poisson vestige count to a Negative binomial vestige count.

CHAPTER 6

Final Remarks

In this chapter, we review and summarise the work presented in this manuscript, examine any obstacles or difficulties encountered, and provide recommendations for possible extensions to the work described.

In this thesis, we have introduced modelling frameworks that allow us to estimate animal abundance using different types of data and in a range of scenarios. In this final chapter, we briefly re-examine the work presented in previous chapters, examining some advantages and limitations to these modelling frameworks, providing final remarks, and discussing further work and possible extensions in each case.

Methods such as those detailed in this thesis that allow us to utilise indirect animal data to estimate abundance facilitate the implementation of wide-scale, long-term animal monitoring programmes. This is due to the relative affordability with which these indirect surveys can be carried out, the reduced risk of harm or disturbance posed to both animals and humans, and the fact that these indirect surveys may be carried out by individuals with very little training requirements.

The ability to estimate animal abundance using this data can consequently be of great value to those working within the wildlife monitoring space, including

statisticians, ecologists and policy-makers. Wildlife monitoring programmes are vitally important if we wish to determine when species are beginning to decline towards a possible extinction, and when species are beginning to increase to levels which could cause issues, either economical or societal. The aim of the work presented as part of this thesis was to provide modelling frameworks that might contribute to the literature in regard to estimating animal abundance and improve upon this ability to monitor wildlife populations. Taking Chapter 3 as a starting point, here we provide an extension the N-mixture model developed by Royle (2004) by considering observational counts collected for multiple species, through the addition of a species-level random effect. This random effect in turn allows us to estimate inter-species correlations, which may allow for inferences to be made as to the relationships that species have with one another.

We then examine further extensions to this model. The first of these extensions allows us to estimate abundances using data that has a large proportion of zero counts. This is a scenario that is very commonly encountered when working with data composed of counts of animal sightings, and so is an important aspect to consider when building methodologies that might be used with this data. A second extension allows us to estimate abundances using data collected over long time periods through the addition of a first-order autoregressive term on the abundance. The ability to utilise data collected over long time frames to estimate abundance is a vital component in the establishment of long-term monitoring programmes.

Finally we provide a straightforward combination of these two extensions, which allows us to examine data which is collected over long time periods and also has a large proportion of zero counts. Through an extensive simulation study, we show that estimates consistently demonstrate high levels of accuracy. We also explore the performance of this family of MNM models on a real-world application which contains both zero-inflated data and data collected over long time frames, using data collected as part of the North American Breeding Bird programme.

A limitation of note associated with the MNM model is the computational intensity involved in running these models. It is possible that in future, improvements may be made which could significantly increase the speed of our implementations, which

we believe would prove invaluable in facilitating wider-scale use of this modelling framework.

We note also that in the models that contain an autoregressive component, while we obtain separate $\text{Corr}(N_s, N_{s'})$ per year, a feature of model formulation means that the correlation between two species cannot change sign from year to year. We can accommodate a change in sign by allowing for an unstructured covariance matrix of the autocorrelation coefficient $\boldsymbol{\Sigma}_{\phi}$, and this particular extension is subject of ongoing work. Furthermore, the models presented in this paper assume that sites are independent of one another. A further extension we are currently working on is the incorporation of spatial dependence in the abundances estimated by the MNM model.

We then present in Chapter 4 the original N-mixture model that our multi-species N-mixture model is based on, along with a number of alternative N-mixture model extensions that have been developed since the publication of the original. We then present an implementation of the N-mixture model by Royle (2004) and a further implementation of our MNM model detailed in Chapter 3, using a dataset of bee sighting collected as part of the BeeWalk Survey. This chapter is written with the intention of providing instruction in the use of various N-mixture models, to an audience who may not have a background in Bayesian statistics. To this end, several appendices are provided which contain `R` code that may be used to reproduce this analysis. It is hoped that providing ecologists and other practitioners with the information required to perform their own data analysis will further facilitate the implementation of wildlife monitoring programmes.

Finally in Chapter 5, we propose a new class of models which use animal vestige data to estimate animal abundance. Through simulation studies which involve data simulated from a distance sampling model, and using a dataset collected by Marques et al. (2001) composed of sika deer vestiges, we demonstrate that the performance of the triple Poisson model is comparable to the performance of a distance sampling model, the only other modelling framework that we have encountered that has been used to estimate abundance using this type of scat data. Via these simulation studies which saw datasets simulated that contained only

two counts, we also discover that the triple Poisson model is capable of estimating abundance even when data is very scarce. This method of data collection (the use of vestiges rather than animal sightings) is of particular use for many animal species which may otherwise go un-monitored for various reasons that might include the animal's reclusive behaviour, or the animals habitat being one that does not lend itself easily to obtaining animal counts (e.g. a dense woodland). For this reason we hope that modelling frameworks such as the one detailed in Chapter 5 might open up new opportunities for wildlife monitoring.

We have thus far assumed for this modelling framework that vestiges experience an exponential rate of decay. However, in the future we plan to examine alternative scenarios which will involve vestige decay occurring according to different distributions, to allow us to model the transient phase of the population. In the future, we also plan to examine the effect on abundance estimates of using convenience sampling rather than randomly placed transects. It is anticipated that this development might improve the cost-efficiency associated with carrying out this type of vestige survey.

In addition to the work presented in this thesis, we have explored other areas in estimating animal abundance and occupancy. For example, we have spent time exploring a spatial multi-species occupancy model, which might allow for the estimation of occupancy for multiple species, while taking spatial dependence into account through the incorporation of a site-level random effect with a covariance matrix that has a multivariate Matérn structure, as proposed by Gneiting et al. (2010). However, our implementation of this approach has thus far failed to converge to the posterior distribution in a manner that we find satisfactory. In particular, we noticed that while the spatial correlation of each species converges well, the inter-species spatial correlation struggles to converge.

Finally, we note that the implementation of the methods presented in this work are freely available at `https://github.com/niamhmimnagh` in public repositories named `insect_populations_ch11`, `mnm`, and `triple_poisson`. Thus, all analyses presented as part of this thesis are reproducible and methodologies are freely available to interested parties.

Bibliography

Alvarez, I., Niemi, J., and Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*. 31

Assis, C. K. (2012). Uso do espaço e densidade populacional de catetos (*Pecari Tajacu*) em paisagem silvicultural do centro-sul do estado de sã o paulo. monografia para obtenção de grau de bacharel em ciências biológicas pela escola superior de agricultura "luiz de queiroz". 16, 132

Baker, P. J. and Harris, S. (2004). Red foxes. *The behavioral ecology of red foxes in urban Bristol*, pages 207–216. 137

Barnes, R. F. W. (2001). How reliable are dung counts for estimating elephant numbers? *African Journal of Ecology*, 39(1):1–9. 122

Becker, E. F. (1991). A terrestrial furbearer estimator based on probability sampling. *The Journal of wildlife management*, pages 730–737. 119

Becker, E. F., Spindler, M. A., and Osborne, T. O. (1998). A population estimator based on network sampling of tracks in the snow. *The Journal of wildlife management*, pages 968–977. 119

Berlow, E. L., Neutel, A.-M., Cohen, J. E., De Ruiter, P. C., Ebenman, B., Emmerson, M., Fox, J. W., Jansen, V. A., Jones, J. I., Kokkoris, G. D., et al. (2004). Interaction strengths in food webs: issues and opportunities. *Journal of Animal Ecology*, 73(3):585–598. 37

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (1993). *Distance Sampling: Estimating abundance of biological populations*. Springer Netherlands. 118

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32. 86

Caughley, G. (1994). Directions in conservation biology. *Journal of animal ecology*, pages 215–244. ix, 1, 98, 117

Cavallini, P. (1994). Faeces count as an index of fox abundance. *Acta theriologica*, 39(4):417–424. 5, 19, 20, 136, 137

Chelintsev, N. G. (1995). Mathematical principles of winter censuses of mammals. *Byulleten Moskovskogo Obschestva Ispytatelei Prirody*, 100:3–19. 119

Comont, R., Luker, S., and Dickinson, H. (2021). Beewalk annual report 2021. *Bumblebee Conservation Trust, Stirling, Scotland UK*. 11, 99

Dail, D. and Madsen, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics*, 67(2):577–587. 86, 92, 93, 94

Delattre, M., Lavielle, M., and Poursat, M.-A. (2014). A note on bic in mixed-effects models. *Electronic journal of statistics*, 8(1):456–475. 34

Dennis, E. B., Morgan, B. J., and Ridout, M. S. (2015). Computational aspects of n-mixture models. *Biometrics*, 71(1):237–246. 4, 12, 38, 39, 80, 87, 88, 99, 105, 106, 115

Dorazio, R. M. and Connor, E. F. (2014). Estimating abundances of interacting species using morphological traits, foraging guilds, and habitat. *PloS one*, 9(4):e94323. 25, 26

Dorazio, R. M. and Royle, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, 100(470):389–398. 24

Dunstan, T. C. and Harper, J. F. (1975). Food habits of bald eagles in north-central minnesota. *The Journal of Wildlife Management*, pages 140–143. 8, 34

Fahrmeir, L., Tutz, G., Hennevogl, W., and Salem, E. (1994). *Multivariate statistical modelling based on generalized linear models*, volume 425. Springer. 29

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-valued garch process. *Journal of time series analysis*, 27(6):923–942. 29

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of eugenics*, 6(1):13–25. 24

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58. 24

Fiske, I. and Chandler, R. (2011). Unmarked: an r package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of statistical software*, 43:1–23. 85

Fokianos, K. and Tjøstheim, D. (2011). Log-linear poisson autoregression. *Journal of Multivariate Analysis*. 29

Gallant, D., Vasseur, L., and Bérubé, C. H. (2007). Unveiling the limitations of scat surveys to monitor social species: a case study on river otters. *The Journal of Wildlife Management*, 71(1):258–265. 122

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472. 30, 99

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177. 158

Golding, J. D., Nowak, J. J., and Dreitz, V. J. (2017). A multispecies dependent double-observer model: a new method for estimating multispecies abundance. *Ecology and evolution*, 7(10):3425–3435. 24, 25, 89

Gomez, J. P., Robinson, S. K., Blackburn, J. K., and Ponciano, J. M. (2018). An efficient extension of n-mixture models for multi-species abundance estimation. *Methods in Ecology and Evolution*, 9(2):340–353. 24, 25, 90

Greenleaf, S. S., Williams, N. M., Winfree, R., and Kremen, C. (2007). Bee foraging ranges and their relationship to body size. *Oecologia*, 153(3):589–596. 100

Haines, L. M. (2016). Maximum likelihood estimation for n-mixture models. *Biometrics*, 72(4):1235–1245. 87

Hansen, A. J. (1987). Regulation of bald eagle reproductive rates in southeast alaska. *Ecology*, 68(5):1387–1392. 7, 34

Herdin, M., Czink, N., Ozcelik, H., and Bonek, E. (2005). Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels. In *2005 IEEE 61st Vehicular Technology Conference*, volume 1, pages 146–140. 43

Hodges, J. I. (2011). Bald eagle population surveys of the north pacific ocean, 1967–2010. *Northwestern Naturalist*, 92(1):7–12. 7, 34

Hostetler, J. A. and Chandler, R. B. (2015). Improved state-space models for inference about spatial and temporal variation in abundance from count data. *Ecology*, 96(6):1713–1723. 86, 94, 97

Jenkins, M., Green, R. E., and Madden, J. (2003). The challenge of measuring global change in wild nature: are things getting better or worse? *Conservation Biology*, 17(1):20–23. 84

Joseph, L. N., Elkin, C., Martin, T. G., and Possingham, H. P. (2009). Modeling abundance using n-mixture models: the importance of considering ecological mechanisms. *Ecological Applications*, 19(3):631–642. 96, 98

Kendall, M. G. (1948). *Rank correlation methods*. Charles Griffin and Co. Ltd. 39

Kéry, M. and Royle, J. A. (2015). Applied hierarchical modeling in ecology: analysis of distribution, abundance and species richness in r and bugs. 124

King, J. G., Robards, F. C., and Lensink, C. J. (1972). Census of the bald eagle breeding population in southeast alaska. *The Journal of Wildlife Management*, pages 1292–1295. 7, 34

Krebs, C. J. (1972). The experimental analysis of distribution and abundance. *Ecology. New York: Harper and Row*. 83

Kuhlman, M. P., Burrows, S., Mummey, D. L., Ramsey, P. W., and Hahn, P. G. (2021). Relative bee abundance varies by collection method and flowering richness: Implications for understanding patterns in bee community data. *Ecological Solutions and Evidence*, 2(2):e12071. 101

Kéry, M., Dorazio, R. M., Soldaat, L., Van Strien, A., Zuiderwijk, A., and Royle, J. A. (2009). Trend estimation in populations with imperfect detection. *Journal of Applied Ecology*, 46(6):1163–1172. 87, 92

Leonhardt, S. D., Gallai, N., Garibaldi, L. A., Kuhlmann, M., and Klein, A.-M. (2013). Economic gain, stability of pollination and bee diversity decrease from southern to northern europe. *Basic and Applied Ecology*, 14(6):461–471. 98

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268. 42

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, pages 245–259. 39

Marques, F. F. C., Buckland, S. T., Goffin, D., Dixon, C. E., Borchers, D. L., Mayle, B. A., and Peace, A. J. (2001). Estimating deer abundance from line transect surveys of dung: sika deer in southern scotland. *Journal of Applied Ecology*, pages 349–363. 5, 18, 19, 118, 126, 134, 157

Marshall, L. (2020). *DSsim: Distance Sampling Simulations*. R package version 1.1.5. 126

Martin, J., Royle, J. A., Mackenzie, D. I., Edwards, H. H., Kery, M., and Gardner, B. (2011). Accounting for non-independent detection when estimating abundance of organisms with a bayesian approach. *Methods in Ecology and Evolution*, 2(6):595–601. 86

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall. 3, 23

McEwan, L. C. and Hirth, D. H. (1980). Food habits of the bald eagle in north-central florida. *The Condor*, 82(2):229–231. 8, 34

McGrady, C., Strange, J., López-Uribe, M., and Fleischer, S. (2021). Wild bumble bee colony abundance, scaled by field size, predicts pollination services. *Ecosphere*, 12(9):e03735. 101

Miller, D. L., Rexstad, E., Thomas, L., Marshall, L., and Laake, J. L. (2019). Distance sampling in r. *Journal of Statistical Software*, 89(1):1–28. 18, 127, 134

Mimnagh, N., Parnell, A., Prado, E., and Moral, R. A. (2022). Bayesian multi-species n-mixture models for unmarked animal communities. *Environmental and Ecological Statistics*, pages 1–24. 86, 91, 95, 97, 98, 99, 101, 106, 110

Moral, R. A., Hinde, J., Demétrio, C. G., Reigada, C., and Godoy, W. A. (2018). Models for jointly estimating abundances of two unmarked site-associated species subject to imperfect detection. *Journal of Agricultural, Biological and Environmental Statistics*, 23(1):20–38. 25, 37, 90

Murray, D., Ellsworth, E., and Zack, A. (2005). Assessment of potential bias with snowshoe hare fecal pellet-plot counts. *The Journal of wildlife management*, 69(1):385–395. 122

Mutshinda, C. M., O'Hara, R. B., and Woiwod, I. P. (2009). What drives community dynamics? *Proceedings of the Royal Society B: Biological Sciences*, 276(1669):2923–2929. 95

Nichols, J. D. (2014). The role of abundance estimates in conservation decision-making. In *Applied ecology and human dimensions in biological conservation*, pages 117–131. Springer. 117, 140

Nichols, J. D., Hines, J. E., Lebreton, J.-D., and Pradel, R. (2000). Estimation of contributions to population growth: a reverse-time capture–recapture approach. *Ecology*, 81(12):3362–3376. 92

Nichols, J. D. and MacKenzie, D. I. (2004). Abundance estimation and conservation biology. *Animal biodiversity and conservation*, 27(1):437–439. 1, 22

Niku, J., Hui, F. K., Taskinen, S., and Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, 10(12):2173–2182. 26

Norris, K. (2004). Managing threatened species: the ecological toolbox, evolutionary theory and declining-population paradigm. *Journal of Applied Ecology*, 41(3):413–426. 83

Novacek, M. J. and Cleland, E. E. (2001). The current biodiversity extinction event: scenarios for mitigation and recovery. *Proceedings of the National Academy of Sciences*, 98(10):5466–5470. 1

Pardieck, K. L., Ziolkowski, D., Lutmerding, M., Aponte, V., and Hudson, M.-A. R. (2020). North american breeding bird survey dataset 1966-2019. U.S. Geological Survey data release. 6, 24, 32

Patterson, B. R., Quinn, N. W. S., Becker, E. F., and Meier, D. B. (2004). Estimating wolf densities in forested areas using network sampling of tracks in snow. *Wildlife Society Bulletin*, 32(3):938–947. 119

Pettis, J. S. and Delaplane, K. S. (2010). Coordinated responses to honey bee decline in the usa. *Apidologie*, 41(3):256–263. 98

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. 30, 86, 99, 127

Plummer, M. (2017). *JAGS Version 4.3.0 user manual*. 30, 31, 99

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 5, 30, 43, 85, 99, 126

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 311–324. 24

Ratcliffe, P. (1987). Distribution and current status of sika deer, cervus nippon, in great britain. *Mammal Review*, 17(1):39–58. 134

Rexstad, E. (2022). Multipliers and indirect surveys: Dung surveys including estimates of production and decay rates. https://examples.distancesampling.org/Distance-mult/multipliers-distill.html. Accessed: 05-02-2023. 134

Royle, J. A. (2004). N–mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115. ix, 4, 5, 11, 22, 23, 26, 84, 87, 89, 91, 93, 99, 105, 106, 156, 157

Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence–absence data or point counts. *Ecology*, 84(3):777–790. 25

Russo, L., Park, M., Gibbs, J., and Danforth, B. (2015). The challenge of accurately documenting bee species richness in agroecosystems: bee diversity in eastern apple orchards. *Ecology and Evolution*, 5(17):3531–3540. 101

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610. 93

Sowls, L. K. (1997). *Javelinas and Other Peccaries: Their Biology, Management and Use*. Texas A&M University Press, 2 edition. 133

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639. 34

Stephens, P. A., Zaumyslova, O. Y., Miquelle, D. G., Myslenkov, A. I., and Hayward, G. D. (2006). Estimating population density from indirect sign: track counts and the formozov–malyshev–pereleshin formula. *Animal Conservation*, 9(3):339–348. 119

Su, Y.-S. and Yajima, M. (2020). *R2jags: Using R to Run 'JAGS'*. R package version 0.6-1. 43, 86, 127

Theisen-Jones, H. and Bienefeld, K. (2016). The asian honey bee (apis cerana) is significantly in decline. *Bee World*, 93(4):90–97. 98

Thomas, L., Buckland, S. T., Burnham, K. P., Anderson, D. R., Laake, J. L., Borchers, D. L., and Strindberg, S. (2006). Distance sampling. *Encyclopedia of Environmetrics*. 5, 18

Todd, C. S., Young, L., Owen Jr, R. B., and Gramlich, F. J. (1982). Food habits of bald eagles in maine. *The Journal of Wildlife Management*, pages 636–645. 8, 34

Verdade, L. M., Lyra-Jorge, M. C., and Pina, C. I. (2014). *Applied ecology and human dimensions in biological conservation*. Springer. ix, 117, 118, 140

Verdade, L. M., Moreira, J. R., and Ferraz, K. M. P. (2013). Counting capybaras. In *Capybara*, pages 357–370. Springer. ix, 2, 23, 117, 140

Webbon, C. C., Baker, P. J., and Harris, S. (2004). Faecal density counts for monitoring changes in red fox numbers in rural britain. *Journal of Applied Ecology*, 41(4):768–779. 137

Wenger, S. J. and Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, 89(10):2953–2959. 96

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic Press. 84

Witmer, G. W. (2005). Wildlife population monitoring: some practical considerations. *Wildlife Research*, 32(3):259–263. 1, 22

Witmer, G. W. (2007). The ecology of vertebrate pests and integrated pest management (ipm). *USDA National Wildlife Research Center-Staff Publications*, page 730. 2

Yamaura, Y., Royle, J. A., Kuboi, K., Tada, T., Ikeno, S., and Makino, S. (2011). Modelling community dynamics based on species-level abundance models from detection/nondetection data. *Journal of applied ecology*, 48(1):67–75. 24, 25

Yamaura, Y., Royle, J. A., Shimada, N., Asanuma, S., Sato, T., Taki, H., and Makino, S. (2012). Biodiversity of man-made open habitats in an underused country: a class of multispecies abundance models for count data. *Biodiversity and Conservation*, 21(6):1365–1380. 24, 25

Zattara, E. E. and Aizen, M. A. (2021). Worldwide occurrence records suggest a global decline in bee species richness. *One Earth*, 4(1):114–123. 98

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, pages 1019–1031. 29