University of Cape Town

STA5058W

Biostatistics Minor dissertation

Evaluating occupancy and the range dynamics of invasive bird species in South Africa

Author:
James Swingler

Student Number:
SWNJAM003

Dissertation submitted in partial fulfilment of the requirements for the degree of Masters in Biostatistics in the Department of Statistical Sciences

Supervisors:
Dr. Gregory Distiller
Mr. Allan Clark

Plagiarism Declaration

I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

I have used the Harvard convention for citation and referencing. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed and has been cited and referenced.

This report is my own work.

I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

Signed by candidate

James Swingler

October 10, 2022

Acknowledgements

First and foremost, I would like to give my sincerest thanks and appreciation to my supervisors Dr Greg Distiller and Mr Allan Clark, without whom this dissertation would not have been possible. The continued support, respect and commitment they provided helped me through an incredibly turbulent time caused by the coronavirus pandemic and made the transition from Stellenbosch to UCT a lot less daunting and challenging than was first anticipated. Although afflicted with their own set of academic deadlines and personal turmoil, they never once shunned their responsibilities as my supervisors, and I will forever be grateful to them. I would also like to thank my loving family and friends for the support they provided me. Through my lowest points they picked me up and proved time and time again that I would overcome whatever obstacle tried to hinder my progress. This journey was an incredibly enlightening experience and I cannot wait to put my newly acquired skills to good use in the near future.

Abstract

There is great interest in the distribution of invasive species that threaten indigenous wildlife. All effective conservation management decisions need to be based on sound inference and predictions so that these species can be controlled and the risk posed to the local ecosystem minimized. Thus, there is significant benefit in the study of invasive species as a means of aiding those charged with protecting indigenous wildlife. The occupancy and population range dynamics of the Myna and Mallard species are individually investigated in the South African region by fitting static and dynamic occupancy models to a set of citizen science data for a 10-year study period between 2010-2019. The occupancy and detectability of the respective species is analysed using static occupancy models for the 2010 study season. The covariates included in the best fitting static models are used to estimate the initial occupancy and detection parameters for the dynamic models which now include estimates for colonization and local extinction. A sensitivity analysis pertaining to the dynamic models is implemented by altering the data structures in terms of the number of analysed sites and length of the detection histories. The results find the Myna's proximity to urban environments to play a significant role on its occupancy in 2010, and yearly changes in climatic and anthropogenic factors influence its 10-year range dynamics. The models fitted to the Mallard are inconclusive possibly due to the violation of the closure assumption potentially caused by migratory behaviour. The results are limited by the presence of a potentially migratory species when using a poorly designed study and highlights the difficulties of conducting an occupancy analysis on a highly mobile avian species as opposed to their sedentary counterpart. The workings of this dissertation support previous claims that an increase in the quantity of sites, and thus the degree of overlapping sites over the different seasons, will improve the precision of the model estimates. However, caution must be exercised when increasing the length of the seasonal detection histories and should generally be set to no more than 10 repeated visits to a site.

Keywords — Invasive species, static occupancy, dynamic occupancy, sensitivity analysis, non-random movement patterns

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

An invasive species is best described as a non-native species being introduced to a foreign environment at some point in time. Whilst the introduction of some non-native species does not necessarily cause direct harm to the natural environment, there are negative consequences when the competing interests of a non-native species causes an imbalance in the ecosystem. This imbalance can result in substantial environmental, social and economic costs (Gormley et al., 2011). Hence, there is considerable interest in managing these invasive populations.

Increased understanding and knowledge of these species' movement patterns and general behaviour has the potential to mitigate any detrimental impact that could arise. An accurate assessment of the risk posed by these species to an area, through predicting and quantifying the current and potential distributions of established invasive species, allows for informed management decisions that can be made to either stem the expansion of these species or eradicate them from the environment entirely (Davis et al., 2018). Thus, effective conservation strategies are reliant on robust and accurate information pertaining to the distribution of the target species, which will in turn greatly influence the management decision-making process.

Statistical ecology is a relatively new field of research that develops statistical methods designed to draw inference and predictions to better understand ecological patterns and processes (Ludwig et al., 1988). There are various methods within statistical ecology with objectives ranging from estimating the abundance or density of a population in an area, investigating the movement patterns of tagged individuals, or performing species distribution modelling (SDMs) based on presence-only data. While these approaches each have their respective merits and disadvantages, the occupancy modelling framework initially developed by MacKenzie et al. (2002) is the one of the more effective methods of investigating the current and potential distributions of invasive species in a geographical area.

Occupancy models are the only approach to modelling the distribution of a species that explicitly accounts for one of the hallmarks of ecological data, namely imperfect detection (Kéry et al., 2013). This is the situation whereby a species of interest is recorded as absent when it was actually present and is referred to in statistics as a false negative.

The sampling situation used in the occupancy modelling framework ideally involves defining the area of interest and then studying a sample of sites within that area. These sampled sites are then visited multiple times in a given period to record the presence or absence of a target species. Data collected this way are commonly referred to as detection/non-detection data and the repeated visits to these sites, within a set period of time, allows one to estimate the detection probability and explicitly account for imperfect detection (MacKenzie et al., 2006).

This detection/non-detection data are then used to draw inferences on the likelihood that a certain site is occupied by the target species. This is accomplished by the simultaneous consideration of both the occupancy (state) and detection (observation) processes (MacKenzie et al., 2006). By using the occupancy modelling framework, the focus of this dissertation is not to quantify the abundance or density of a specific species, but rather evaluate its current and potential distribution, in addition to what factors may affect the occupancy and detection probability.

1.2 Focus and Scope

The focus of this study is to analyse the distributions of two invasive bird species in South Africa, namely, the Common Myna (*Acridotheres tristis*) and the Mallard (*Anas platyrhynchos*). The original study region included the entirety of South Africa using data collected through a citizen science project, namely the African Bird Atlas Project (ABAP), from which the actual study regions that were analysed were then constrained to where the different species occur. This project collects survey records of 16,689 sites set-up as a spatial grid superimposed over the country, whereby a site (pentad) is defined as a five minute by five minute latitude by longitude grid cell.

The distribution of these species are evaluated independently, and are initially assessed using static (single-season) occupancy models (MacKenzie et al., 2002) in 2010 as a means of inferring a snapshot of the occupancy status of the species and to inform the more complex dynamic occupancy models. Subsequently, the population range dynamics of these species, and the potential factors that influence changes in the occupancy of a site, are evaluated between 2010-2019 using dynamic (multi-season) occupancy models (MacKenzie et al., 2003).

1.3 Motivation

Invasive species are problematic for a variety of reasons, and the study of both the Common Myna and the Mallard allows for an investigation into two species that bring unique threats to indigenous wildlife.

The Common Myna is a world-renowned pest and is characterised as being highly aggressive, territorial in nature and negatively impacts the local avifauna through nest cavity displacement and providing competition for breeding resources. It is a commonly observed species in most major cities and has many recorded sightings in South Africa (Peacock et al., 2007).

In contrast, the Mallard is neither aggressive nor competitive, but the genetic pollution created through its interbreeding with indigenous waterfowl has the potential to entirely eradicate the identity of species native to a particular area (de Souza et al., 2019). It is a more rarely observed species which tends to be spotted around bodies of water and has significantly fewer sightings in South Africa than the Myna (Stephens et al., 2020).

There is great interest in the distribution of these respective species for these unique reasons. All effective conservation management decisions need to be based on sound inference and predictions, so that these species can be controlled, and the risk posed to the local ecosystem minimized. Thus, there is significant benefit in the study of these species as a means of aiding those charged with protecting indigenous wildlife (Davis et al., 2018).

1.4 Objectives

The aims and objectives of this dissertation are both practical and theoretical.

1. The practical objectives include:
    - Drawing inference from static models on the most important factors affecting occupancy and the most important factors affecting the detection probability of a species.
    - Using the results from the static models, fitting dynamic models to determine what factors play a role on the colonization and extinction probabilities of any given site.
    - Investigating whether there has been an apparent expansion or contraction in occupancy of these species over a ten-year period.
2. The structure of the data used in the dynamic occupancy modelling process is imbalanced since both the number of sites visited as well as the number of surveys conducted at each site vary during each survey period. Thus, the theoretical objectives include:
    - Evaluating how alterations to the number of sites visited in each primary sampling period (varying degrees of temporal overlap) impacts the results of the

dynamic models.

- Evaluating how changes to the maximum number of surveys conducted at a site within each primary sampling period (varying degrees of survey effort) impacts the results of the dynamic models.
- If these differences are significant, determining what causes these dissimilarities and discuss potential avenues of recourse.

1.5 Chapter Layout

The structure of this dissertation is as follows: Chapter 1 provides a general introduction, scope and focus, motivation and objectives for the research conducted in this dissertation. Chapter 2 provides a brief overview of the invasive species under investigation in this dissertation and provides a review of the literature concerning occupancy models and all significant developments from their first introduction to their modern-day implementation while Chapter 3 then provides a detailed description of the data and how the chosen statistical methods were applied to the data. In Chapter 4, the results of the modelling process are provided and analysed, with a specific focus on how the results are impacted when the structure of the imbalanced data is altered. Finally, Chapter 5 is devoted to: an assessment of whether the objectives of this dissertation are achieved; a presentation of the concluding results uncovered in the study; an acknowledgement of any pitfalls and potential limitations of the chosen method of analysis; and final remarks on how future work can use these results to aid and better their respective workings.

Chapter 2

Literature Review

This chapter begins with a brief overview of the invasive species under investigation in this dissertation and then provides a review of the relevant literature pertaining to occupancy models. There has been extensive research into occupancy models over recent years for which a large amount of literature can be cited. The formulation of this family of models began with the single-species, single-season approach and developed over time to account for more complex models such as those that include multiple species, and spatially and temporally complex range dynamics (MacKenzie et al., 2006).

However, for the purposes of this review, only methods applied in this minor dissertation will be discussed. This includes the literature of single-season (static) occupancy models and the approaches to multi-season (dynamic) occupancy models.

2.1 Invasive Species

2.1.1 Common (Indian) Myna

The Common Myna (*Acridotheres tristis*), henceforth referred to as the Myna, is a medium-sized, omnivorous woodland bird in the *Sturnidae* family native to most regions in Asia (Measey et al., 2020). It is a strongly territorial species which evolved in open woodland habitats in India which are characterised by widely spaced tall trees with little to no vegetation cover. It is believed that evolution in this type of habitat allowed for the pre-adaption of the species to similar environments encountered in urban areas (Pell and Tidemann, 1997).

The successful adaption to urban environments coupled with its aggressive and territorial nature has seen the Myna declared, by the International Union for Conservation of Nature (IUCN) in 2000, as only one of three birds among the 'World's 100 worst' invasive species (Peacock et al., 2007). Its distribution sees it occur on all continents, except Antarctica and South America, in addition to several islands in the Pacific, Indian and Atlantic oceans (Peacock et al., 2007).

Historically, the Myna has been deliberately introduced to foreign countries as a biological pest control agent to aid agricultural sectors, particularly for the management of locusts and grasshoppers. However, the introduction of this species to South Africa occurred through the accidental escape of captive birds in Durban, KwaZulu-Natal in 1902 (Peacock et al., 2007).

Over a century later the Myna has become a widespread species in the country where its distribution is more concentrated around larger human disturbances or a higher human population density. The sheer quantity of these birds inhabiting the country, coupled with its territorial nature, has resulted in the Myna being considered a major pest and disruptor of the natural ecosystem, posing a serious threat to indigenous biodiversity. Thus, it has been declared an invasive species and requires human management and control (Peacock et al., 2007; Pell and Tidemann, 1997).

The Myna is considered an invasive pest for a variety of reasons. Although deliberately introduced to aid farmers by eradicating grasshoppers and locusts, their diet includes agriculturally valuable insects in some regions, and thus they have become agricultural pests. Additionally, they are known to be carriers of invasive plants, diseases (avian malaria) and parasites (the mite) which are similarly detrimental to the natural environment, especially in regions where they congregate in close proximity to one another (Bomford and Sinclair, 2002).

Arguably, the greatest risk the Myna poses to the regions it has invaded is the effect they have had on indigenous species. A study conducted by Bomford and Sinclair (2002) found the Myna to pose a serious threat to several endangered native birds due to their aggressive nature and monopolisation of scarce nesting holes that would otherwise be occupied by these native birds. Another study, conducted by Pell and Tidemann (1997), investigated the specific impact of the Myna on hole-nesting native parrots in woodland and savannah areas of eastern Australia. Their results indicated that the Myna was the dominant users of these nesting holes and consequently had the potential to reduce the successful breeding of these native parrots.

The territorial dominance of the Myna has further ramifications for indigenous species beyond competition for nesting space. They are known to prey on indigenous eggs and chicks, in addition to committing direct attacks on other birds, particularly in island ecosystems where it has been noted that a strong relationship exists between the number of alien invasions and the number of local extinctions (Bomford and Sinclair, 2002; Peacock et al., 2007).

The impact of the Myna on human welfare is also a slight concern since their nests are known to block gutters and cause water damage to the exterior of buildings. Additionally, they are a risk to human safety at airports, and their communal roosts create a social nuisance through poor hygiene and noise pollution (Bomford and Sinclair, 2002).

2.1.2 Mallard Duck

The Mallard or wild duck (*Anas platyrhynchos*), henceforth referred to as the Mallard, is an omnivorous, sexually dimorphic dabbling duck in the *Anatinae* subfamily of the waterfowl family, *Anatidae* (de Souza et al., 2019). This species evolved predominantly in wetlands and grasslands, although it is also commonly sighted in sheltered estuaries and marine habitats, and is a native of North America, Eurasia and parts of Northern Africa.

The introduction of the Mallard to foreign environments is a result of hunting practices in the 20th century, but the species is also considered to have been deliberately introduced outside of its native region for purely aesthetic purposes. Currently, the species has a near global distribution and was initially introduced to South Africa in the 1940s. They have subsequently established their habitats within wetland, grassland, urban and peri-urban areas in the Western Cape and Gauteng provinces, but have also been identified and recorded in all provinces of the country (Stephens et al., 2020).

While the Mallard poses many of the same threats as the Myna in South Africa – specifically through competition with local waterfowl for food, roosting and nesting sites, as well as carrying harmful diseases such as avian cholera – this invasive species poses an additional and more acute threat to the country's indigenous biodiversity which is not experienced with the Myna. This is the problem of hybridisation, caused by the Mallard's ability to breed with closely related indigenous waterfowl, thus threatening the genetic integrity of native ducks. The Mallard is known to breed with approximately 60 *Anatidae* species, causing genetic pollution, population decline and sometimes the extinction of indigenous ducks (with which the Mallard breeds) as a whole (de Souza et al., 2019; Shivambu et al., 2020).

Mallards are more aggressive, larger, and more dominant than other duck species. It has further been suggested that the plumage of the Mallard drakes provides an additional stimulus to indigenous hens (Stephens et al., 2020). All of which consequently allows them to sexually outcompete indigenous species. The result of such interbreeding has seen the population of the American black duck substantially decline and has caused the extinction of the Mexican duck as a whole (de Souza et al., 2019). Similarly, interbreeding with the Mallard has seen the once widespread Hawaiian and New Zealand grey duck listed as endangered species (Shivambu et al., 2020).

In the context of South Africa, the indigenous waterfowl at risk of hybridisation include the Yellow-billed, Cape Teal and African Black ducks. While the offspring are likely to be fertile, the genetic diversity of these indigenous species are currently threatened. While

a study conducted by de Souza et al. (2019) has found minimal evidence in South Africa of such genetic pollution, the risk is still present and the interbreeding of the Mallard with these species needs to be monitored to minimise any potential loss of indigenous biodiversity.

2.2 Static Occupancy Models

The methods for single-species, static occupancy models range from the more naive and historic ad hoc approaches to occupancy estimation, to the more modern and flexible model-based approach. Moreover, the model-based approach makes an important consideration regarding the problem of imperfect detection by explicitly accounting for it as part of the model structure (MacKenzie et al., 2002).

2.2.1 Sampling situation

The general sampling situation pertaining to the static occupancy modelling framework aims to obtain suitable data, to achieve the method's objective of estimating the proportion of an area that is inhabited by a target species, whereby in this area a researcher samples a sub-selection of s sampling units (sites) from the collection of S sites about which inference is to be made. These sites represent either naturally occurring landmarks such as lakes or reserves, or arbitrarily defined spatial units such as grid cells of a specified size (MacKenzie et al., 2006). To aid in the understanding and following of the terminology used throughout the paper, Table 2.1 provides a point of reference and summarises the relevant abbreviations and indices.

Abbreviation	Description
K	Number of repeated surveys
s	Number of sites
i	Individual site
j	Individual survey
t	Single season/primary sampling period
T	Multiple seasons/secondary sampling period
h	Detection history
p	Detection probability
ψ	Occupancy probability
ϵ	Extinction probability
γ	Colonisation probability

Table 2.1: Table of abbreviations and indices used throughout the paper

Within the occupancy modelling framework, a standard sampling scheme involves performing repeated surveys (K) at a number of sites (s) within a sampling period (t) to establish where, whether and/or to what degree the target species inhabits the area of interest. These surveys are performed under the assumption that the sites are closed to changes in the state of occupancy within the survey sampling period (season).

This recorded sequence of detections (1) and non-detections (0) of the species during the K surveys of site i creates what is known as a detection history ($\mathbf{h}_i$[1]). As an example, assuming there are four surveys conducted within a season at a site where the target species was identified on the first and last visit, the detection history is expressed as $\mathbf{h}_i = 1001$.

There are two stochastic processes which generate the data, commonly known as the state and observation processes, respectively. The state process pertains to occupancy (ψ), the prime quantity of interest, which relates to the occupancy probability of the target species at the surveyed sites during the sampling period. Similarly, the observation process concerns detectability (p), which is generally regarded as a nuisance parameter and concerns how easily the species is detected (MacKenzie et al., 2006).

2.2.2 Detection Probability is 1 or known without error

While it is unlikely that a species of interest is so conspicuous that it will always be detected when present at a site, MacKenzie et al. (2006) suggest that it is still informative to determine how well an estimator performs in the case of perfect detection when the occupancy is known without error.

Assuming that all sites have a common probability of being occupied by the species, Ψ, MacKenzie et al. (2006) provide an estimate for the proportion of occupied sites under the assumption of perfect detection, when using the standard results for a binomial proportion, as:

$$\widehat{\Psi}_B = \frac{x}{s}$$

where x is the number of sites occupied by the target species and s is the total number of sites that are surveyed. The associated variance is formulated as:

$$\text{Var}\left(\widehat{\Psi}_B\right) = \frac{\Psi(1-\Psi)}{s}$$

which can be approximated by substituting the estimated value for Ψ.

Albeit of minimal practical value, some useful insight can be gleaned from the properties of the derived estimator and associated variance term. However, in most cases, both

[1]The subscript t is not required since this is the single-season case.

occupancy and detection probabilities are unknown and need to be estimated jointly from the collected data.

2.2.3 Two-step Ad Hoc Approaches

The initial class of methods for the estimation of unknown occupancy and detection is classed as a two-step ad hoc procedure. This approach estimates the probability of detecting the target species as a first step and subsequently uses this detection parameter to estimate the probability of occupancy as a second step. This class of ad hoc estimation procedures includes three different methods.

Geissler-Fuller Method

The earliest literature pertaining to occupancy estimation is the method put forward by Geissler and Fuller (1987). Their approach made use of a conditional probability of detection estimated for each site as the proportion of surveys in which the species was detected following the first sighting of the target species at the site of interest. This detection probability is formulated as:

$$\hat{p}_{\mathrm{GF},i} = \frac{\sum_{j=t_i+1}^{K_i} h_{ij}}{K_i - t_i}$$

where K_i denotes the number of surveys at a particular site, t_i is the survey at which first detection at a particular site occurred, and h_{ij} is a binary indicator of whether the target species was detected (1) or not (0) in survey j at site i.

The probability of detecting the species at least once during the K_i surveys can be calculated as:

$$\hat{p}^*_{\mathrm{GF},i} = 1 - \left(1 - \overline{\hat{p}}_{\mathrm{GF}}\right)^{K_i},$$

where $\overline{\hat{p}}_{\mathrm{GF}}$ is the simple average (across all sites) of the $\hat{p}_{\mathrm{GF};i}$'s.

This conditional probability is then used in the calculation of a Horvitz-Thompson-based-estimate for the occupancy probability as follows (Geissler and Fuller, 1987):

$$\hat{\psi}_{\mathrm{GF}} = \frac{\sum_{i=1}^{s} \frac{w_i}{\hat{p}^*_{\mathrm{GF},i}}}{s},$$

where w_i is a binary indicator variable of a species detection (1) or absence (0) at a particular site.

The authors proceed to suggest that the estimator should be the mean or median value obtained from a large number of nonparametric bootstraps from which the standard errors

could similarly be extracted. The purpose of such a procedure is to achieve a desirably negligible level of bias in the estimates (Hongyi Li and Maddala, 1996).

The premise of the Geissler-Fuller approach is based on two key assumptions. The first of which requires the probability of detection to be constant across each site for all surveys whilst the second requirement states that the probability of occupancy is similarly constant across each site for all surveys (Geissler and Fuller, 1987).

Azuma-Baldwin-Noon Method

The two-step approach developed by Azuma et al. (1990) took a different approach to occupancy estimation, whereby a site is not surveyed again after the presence of two individuals of the target species is confirmed. For this method they assume a constant detection probability and use a truncated geometric distribution (TGD) to then model the number of surveys required to detect the species of interest. The distribution is formulated as:

$$\begin{aligned}\Pr(Y = t \mid 0 < Y \leq K) &= \frac{p(1-p)^{t-1}}{1-(1-p)^t}, \quad t = 1, 2, 3, \ldots, K \\ &= 0, \text{ otherwise },\end{aligned}$$

where Y indicates the number of surveys required to detect the species, p denotes the constant detection probability and K is the number of repeated surveys. This provides a time-to-first-detection model for those sites at which detection of the target species occurred during the K surveys. The occupancy estimate is then calculated by dividing the number of sites at which the species was detected at least once by the product of the total number of sites and the estimated p.

Nichols-Karanth Method

A more extensive approach to occupancy estimation is proposed by Nichols and Karanth (2002). This approach generates an estimate of the number of occupied sites using closed-population capture-recapture methods. The application of capture-recapture methods to occupancy estimation is possible since an estimate of the number of occupied sites where the species was not detected is equivalent to estimating the number of individuals in the population that were never captured (MacKenzie et al., 2006). Once an estimate for the number of occupied sites has been determined using an appropriate technique, it is then possible to obtain an estimate of occupancy as follows:

$$\hat{\psi}_{\text{NK}} = \frac{\bar{x}}{s},$$

where $\bar{x}$ denotes the estimated number of occupied sites and s represents the number of available sites.

Whilst the two-step ad hoc approaches to occupancy estimation provide valid estimates, they are unable to provide the flexibility needed for the comparison of competing hypotheses about the nature of the system. Furthermore, they are not able to adapt to more complex sampling situations in which the assumptions of constant occupancy and detection probabilities do not hold. The more recent class of methods, which involve directly modelling the sampling process in a way that results in the simultaneous estimation of the occupancy and detection parameters, is provided by a single model-based framework.

2.2.4 Model-Based Approach

A more flexible approach to occupancy estimation was first introduced by the work of MacKenzie et al. (2002). Their workings are based on the grounds that a model-based framework, which directly models the probability of the observed outcomes resulting from the stochastic sampling process, should be implemented.

This in turn allows for the simultaneous estimation of both the occupancy and detection parameters in the case where detection probabilities are thought to be less than one, commonly known as the case of imperfect detection (MacKenzie et al., 2002).

As mentioned prior, the modelling framework includes two stochastic processes that govern whether the target species is detected at a site. Firstly, whether the site is occupied (with probability ψ) or unoccupied $(1 - \psi)$ by the target species. The probability of detecting the target species at each survey (j) will be 0 if the site is unoccupied and equal to some probability (p_j) if the site is occupied.

Using these specified occupancy and detection probabilities, a statement expressing the probability of observing the aforementioned detection history ($\mathbf{h}_i = 1001$) for site i could be expressed as:

$$\Pr(\mathbf{h}_i = 1001) = \psi p_1 (1 - p_2)(1 - p_3) p_4.$$

If the species was not observed at all during the survey period, the other two possibilities, for which the probability statements need to account, are the situations whereby the species was either not there at all, or it was present but missed during every visit. The corresponding probability of observing this detection history at site i would be:

$$\Pr(\mathbf{h}_i = 0000) = \psi (1 - p_1)(1 - p_2)(1 - p_3)(1 - p_4) + (1 - \psi).$$

The probability statements expressing the observed detection history for each surveyed site is then used to construct the model likelihood for the set of observed data, under the assumption of independent site surveys, as:

$$L\left(\boldsymbol{\psi}, \mathbf{p} \mid \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_s\right) = \prod_{i=1}^{s} \Pr\left(\mathbf{h}_i\right)$$

which can be simplified to:

$$L\left(\boldsymbol{\psi}, \mathbf{p} \mid \mathbf{h}_1, \ldots, \mathbf{h}_\mathrm{s}\right) = \left[\psi^{s_D} \prod_{j=1}^{K} p_j^{s_j} \left(1-p_j\right)^{s_D - s_j}\right] \left[\psi(\prod_{j=1}^{K}\left(1-p_j\right)) + (1-\psi)\right]^{s-s_D} \tag{2.1}$$

where $\mathbf{h}_1$ to $\mathbf{h}_s$ represents the independent detection histories for the s sites, p_j denotes the probability of detection at survey j (given occupancy), s_D is the number of sites where the species was detected at least once, and s_j is the number of sites where the species was detected during the jth survey.

The result of the above model is equivalent to modelling the number of detections at each site as a zero-inflated binomial random variable when the probability of detection is assumed to be constant for all occasions. This model for the detections is formulated as:

$$\begin{aligned} \Pr\left(Y = y_i\right) &= \psi \begin{pmatrix} K \\ y_i \end{pmatrix} p^{y_i} (1-p)^{K-y_i}, \quad y_i > 0 \\ &= \psi(1-p)^K + (1-\psi), \qquad y_i = 0 \end{aligned} \tag{2.2}$$

where y_i is the number of detections at site i.

Additionally, this model-based approach allows for the incorporation of missing observations and, unlike the two-step ad hoc approaches, allows for unequal survey effort at distinct sites which subsequently allows for flexibility of the model to suit realistic study designs (MacKenzie et al., 2006).

The central assumptions of this model are: the state of occupancy for all sites is fixed during the survey period (closure assumption, which assumes that no migration/emigration of the species occurs at the sampling site); the occupancy probability is equivalent across all sites of interest; conditional on its presence the detection probability of the target species is assumed to be constant across all sites and surveys/occasions; and observations made in each survey are independent within and across all sites.

Under the assumption of a constant detection probability, the model likelihood presented

by Equation 2.1 is rewritten as:

$$L\left(\boldsymbol{\psi}, \mathbf{p} \mid \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_s\right) = \left[\psi^{s_D} p^{\sum_{j=1}^{K} s_j} (1-p)^{K s_D - \sum_{j=1}^{K} s_j}\right] \left[\psi(1-p)^K + (1-\psi)\right]^{s - s_D}$$

with p representing a constant detection probability for all survey occasions.

The maximum likelihood estimates are obtained by taking the first derivatives with respect to each parameter and equating to zero, giving the following estimating equations:

$$\hat{\psi}_{\text{MLE}} = \frac{s_D}{s\hat{p}^*_{\text{MLE}}},$$

where $\hat{p}^*_{\text{MLE}} = 1 - (1 - \hat{p}_{\text{MLE}})^K$ is the estimated constant probability of detecting the species at least once during the survey (given presence); and

$$\tilde{p}_{\text{MLE}} = \frac{\hat{p}_{\text{MLE}}}{\hat{p}^*_{\text{MLE}}} = \frac{\sum_{j=1}^{K} s_j}{K s_D},$$

where $\tilde{p}_{\text{MLE}}$ is the estimated conditional probability of detecting the species during the survey given the species was detected at least once at a site.

Similarly, the flexibility of the model-based approach allows for the parameter to be treated as temporally varying across surveys. In the case of survey-specific detection probabilities, the model likelihood presented by Equation 2.1 can be manipulated using the same techniques as before to obtain the following estimating equations:

$$\hat{\psi}_{\text{MLE}} = \frac{s_D}{s\hat{p}^*_{\text{MLE}}},$$

where now

$$p^*_{\text{MLE}} = 1 - \prod_{j=1}^{K} \left(1 - \hat{p}_{j,\text{MLE}}\right)^{2}$$

and

$$\tilde{p}_{j,\text{MLE}} = \frac{\hat{p}_{j,\text{MLE}}}{1 - \prod_{j=1}^{K} \left(1 - \hat{p}_{j,\text{MLE}}\right)} = \frac{s_j}{s_D}.$$

In these equations, s_D is now equivalent to the number of surveys conducted at time j at sites where at least one detection took place during the K surveys.

[2] $\hat{p}_{j,\text{MLE}}$ is now treated as a survey specific detection probability.

The likelihood of site occupancy, given the target species is not detected, can similarly be determined via Bayes' Theorem, which uses the parameter estimates to calculate a conditional occupancy probability. This is presented by MacKenzie et al. (2006) as:

$$\begin{aligned}\psi_{\text{condl}} &= \Pr(\text{ species present } \mid \text{ species not detected }) \\ &= \frac{\Pr(\text{ species present and not detected })}{\Pr(\text{ species not detected })} \\ &= \frac{\psi \prod_{j=1}^{K} (1 - p_j)}{(1 - \psi) + \psi \prod_{j=1}^{K} (1 - p_j)}.\end{aligned}$$

MacKenzie et al. (2002) conducted a simulation study as a means of evaluating the properties of their proposed method of estimating occupancy. The results of this simulation study found that the method was able to provide reasonable estimates of the proportion of occupied sites. The occupancy estimates are reasonably unbiased when the number of surveys is greater than four and the detection probability is greater than or equal to 0.3. Further, the occupancy estimates are only reasonable when the detection probability is at least 0.5 for scenarios in which there are only two conducted surveys.

The concluding remarks of the simulation study are: increasing the number of surveys improves both accuracy and precision of the estimated occupancy probability, although there is minimal benefit in conducting up to 10 (in comparison to five) total surveys; a higher quantity of sites included in the analysis will similarly enhance the estimated parameter's estimated accuracy and precision when detection probabilities are low; and the model is able to provide adequate estimates of the occupancy parameter which are generally unbiased for moderate detection probabilities (MacKenzie et al., 2002).

As previously mentioned, a main assumption of the model-based approach is that the likelihood of a site being occupied, as well as the likelihood of detecting the target species at a site, are constant across all sites of interest. However, most practical cases will find this assumption unlikely to hold and a study is likely expected to experience heterogeneous probabilities across sites. Thus, in many cases the objective of the study will be to characterize the dissimilarity across sites (MacKenzie et al., 2006).

The ability of the model-based procedure to incorporate environmental and survey covariates, which characterize these differences in site occupancy and detection, is thus a desirable property of the approach. Both the occupancy and detection parameters can be modelled as functions of measured covariates by fitting appropriate link functions, such as the logistic, probit, or logit-link.

As noted by MacKenzie et al. (2002), this consequently allows us to treat the procedure as a form of generalized logistic regression analysis in which some variability depends on whether the observed absence corresponds to a true absence of the target species from a particular site.

Using a logit-link function, the expressions for the probability of occupancy at site i as well as the probability of detection at site i during survey j are provided as follows:

$$\begin{aligned} \text{logit}\,(\psi_i) &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_U x_{iU} \\ \text{logit}\,(p_{ij}) &= \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_U x_{iU} + \alpha_{U+1} y_{ij1} + \ldots + \alpha_{U+V} y_{ijV} \end{aligned} \tag{2.3}$$

where $\beta_0, \ldots, \beta_U$ and $\alpha_0, \ldots, \alpha_U$ denote the corresponding effect parameters, $x_{i1}, \ldots, x_{iU}$ represent the U season-constant covariates associated with site i and $y_{ij1}, \ldots, y_{ijv}$ represent the V survey-specific covariates associated with survey j of site i.

When the assumptions of the model-based approach do not hold, it leads to unenviable results, as the estimators will have high bias and inferences regarding the factors affecting the parameters will be inaccurate. A study conducted by Kendall (1999) evaluated the consequences of violating the closure assumption. When the targeted species exhibits trends in migration, which result in movement patterns that are not random, then the occupancy estimator will have biased results due to changes in occupancy that the model cannot accommodate.

The biased estimator will tend to underestimate the occupancy parameter in the case of emigration and overestimate the parameter in the case of immigration of the target species. Kendall (1999) notes that the bias is easier to assess when a model with survey-specific detection probabilities has been fit to the data as opposed to a model with constant detection probabilities.

Heterogeneity of detection probabilities across sites, that cannot be accounted for through measured covariates, also results in the underestimation of the true occupancy parameter (MacKenzie et al., 2006). Alternative ways of modelling this heterogeneity using random effects is possible, but will not be discussed or applied in this dissertation.

Another problematic violation of the model assumptions occurs when detection is not independent among sites, due to potential spatial autocorrelation, which results in overstated precision of the occupancy estimate. This is a form of overdispersion, since the number of independent sites will be smaller than the number of sites surveyed, and results in underestimated standard errors (MacKenzie and Bailey, 2004).

The assessment of multiple fitted models has some power, to varying degrees, to detect models with relatively poor fit in breach of model assumptions. Ecological studies frequently apply the Akaike Information Criterion (AIC) model selection technique as a means of concluding the best fitting model (Akaike, 1973; MacKenzie and Bailey, 2004; Burnham and Anderson, 2002). Similarly, the use of the AIC statistic is connected to a hypothesis driven approach whereby a set of hypotheses are mapped on to model structures and then AIC is used to assess the relative support in the data for the different hypotheses.

The formulation of the AIC test statistic is as follows:

$$AIC = 2P - 2\ln(\hat{L}), \tag{2.4}$$

where P represents the number of estimated parameters in the model and $\hat{L}$ denotes the maximum value of the likelihood function for the model.

While the AIC statistic is a useful technique to choose from amongst a set of candidate models, it does not provide the power to infer if the fit of the chosen model is poor. However, as noted by MacKenzie and Bailey (2004), many ecological studies will have sample sizes that are too small to adequately detect poor model fit using conventional tests. Thus, they present a method of model assessment (that treats all individuals as a single cohort solely for the purpose of assessing model fit) where a Pearson chi-square statistic is calculated and a parametric bootstrap procedure subsequently is used to determine whether the observed statistic is unusually large. The formulation of the Pearson chi-square test statistic is provided in MacKenzie and Bailey (2004) as follows:

$$\chi^2 = \sum_{h=1}^{2^K} \frac{(O_h - E_h)^2}{E_h}, \tag{2.5}$$

where O_h and E_h represent the number of sites observed and expected to have detection history h out of 2^K possible detection histories defined by the fitted model. Using this result, and assuming the fitted model is correct, the parametric bootstrapping algorithm is implemented in the following manner:

1. Fit model to the observed data and estimate parameters $\hat{\psi}_i$ and $\hat{p}_{ij}$ (which may be functions of covariates).
2. Calculate the test statistic for the observed data, χ^2_{Obs}, using the model fit in Step 1.
3. For each site generate a pseudo-random number (r) between 0 and 1.

	If $r \leq \hat{\psi}_i$, then the site is occupied, hence generate K further pseudo-random numbers (r_j) between 0 and 1. If $r_j \leq \hat{p}_{ij}$, then the species was "detected" and the corresponding bootstrapped observation is a "1", otherwise "0". If $r > \hat{\psi}_i$, then the site is unoccupied and the bootstrapped observations will all be "0" for that site.
4.	Fit a model with the same structure as in Step 1 to the bootstrapped dataset.
5.	Calculate the test statistic for the bootstrapped data, χ^2_B, using the model fit in Step 4, and store the result.
6.	Repeat Steps 3-5 many times to approximate the distribution of the test statistic, given the fitted model is correct.
7.	Compare χ^2_{Obs} to the values of χ^2_B to determine the probability of observing a larger value (the p value).

Table 2.2: MacKenzie and Bailey (2004) model assessment algorithm

Additionally, a dispersion parameter can be calculated in the case of a poor model fit, and the quasi-likelihood version of AIC can be used for model selection. This dispersion parameter is constructed by MacKenzie and Bailey (2004) as:

$$\hat{c} = \chi^2_{\text{Obs}} / \bar{\chi}^2_B, \tag{2.6}$$

where $\bar{\chi}^2_B$ is the average of the test statistics obtained from the parametric bootstrap and $\hat{c}$ is used as a variance inflation factor to adjust the model selection procedures and standard errors. Based on this statistic, a model that fits the data adequately should result in a $\hat{c}$ of approximately 1, and values greater or less than one indicate more or less variation in the observed data, respectively.

The AIC selection criteria assumes that at least one of the fitted models provides an adequate fit to the data, by contrast, the method proposed by MacKenzie and Bailey (2004) enables both an evaluation of which model fits the data best, as well as whether this best fitting model is an adequate fit to the observed data. Furthermore, the method is flexible enough to include potential measured covariates that may vary across sites.

MacKenzie and Bailey (2004) conducted a simulation study to evaluate the power of their procedure in identifying poor model fit for a variety of hypothetical scenarios. Their results demonstrate that their procedure had relative (good) power when the number of surveys conducted was larger than five at a significance level of five percent. Further, their

method had greater power of identifying poor fit caused by the exclusion of an important site-specific detection covariate when the average detection parameter was high. However, when an inadequate model fit was caused by a lack of independence among sites, their assessment performed better when the average detection probabilities were lower. Regrettably, their approach had no power with respect to identifying an inadequate model due to heterogeneous occupancy probabilities across sites. The concluding remark made from the study suggests that in most cases model selection for static model-based occupancy models should be evaluated using QAIC values due to poor fitting global models, which include all measured covariates (MacKenzie and Bailey, 2004).

2.3 Dynamic Occupancy Models

The static occupancy model is a useful tool when the objective of a study is to infer some occupancy pattern as a snapshot of the target species' population for a region at a single point in time, however, some ecologists are more interested in the changes of an environmental system over a longer time frame (Bailey et al., 2014; Kéry et al., 2013; Weir et al., 2009).

The dynamic occupancy model allows for the inference of population (range) dynamics through the addition of two estimated (vital rate) parameters, namely colonization (γ) and (local) extinction (ϵ) probabilities, of a target species (MacKenzie et al., 2003). The inclusion of these parameters allows for a better understanding of how occupancy patterns change over time due to changes in the underlying population dynamics. The two general approaches to modelling these population dynamics comprise an implicit dynamics model and a model which estimates these vital rates explicitly (MacKenzie et al., 2006).

2.3.1 Sampling situation

The standard sampling scheme for the dynamic occupancy framework is similar to the static counterpart, however, the inference of whether the target species inhabits a site is now assessed over multiple points in time. The time frame of the study is split into two scales, namely the multiple seasons (T) over which the study is conducted and the season (t) in which the assumption of closure is made (MacKenzie et al., 2006). Thus, the state of occupancy (and the population dynamics driving the occupancy state) can change between seasons but not within seasons.

Figure 2.1 provides a useful representation of the sampling situation for a dynamic occupancy analysis. The recorded detection history for site i conducted within season t is denoted as $\mathbf{h}_{t,i}$, and the full detection history for site i over the entire study is now denoted $\mathbf{h}_i$ (MacKenzie et al., 2006). For example, an observed detection history for site i over a

three year study with three surveys per season is now represented as $\mathbf{h}_i = \ 011 \ \ 010 \ \ 001$.

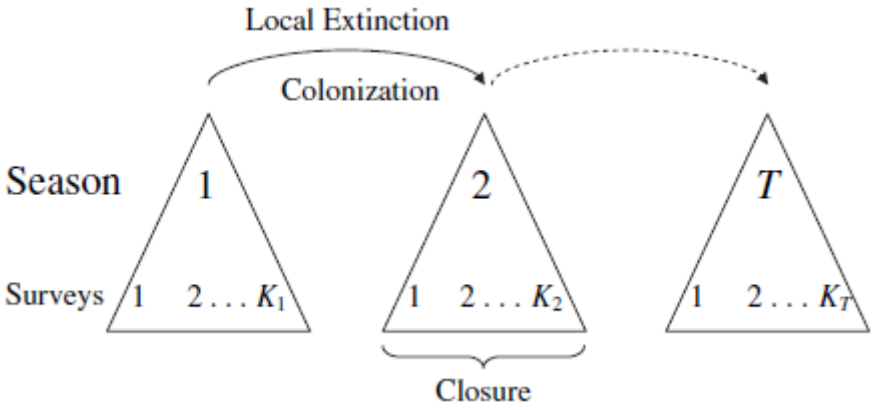


Figure 2.1: Sampling situation for the dynamic occupancy model with each triangle representing a season (t) with multiple surveys (K_t) conducted within seasons (MacKenzie et al., 2006)

2.3.2 Implicit Dynamics Model

The implicit dynamics model, applied to a target species over multiple seasons (T), is equivalent to applying multiple and separate static models to the data collected in each season's specified window. The occupancy between the current and previous season is considered an independent and random process and only the resulting occupancy estimates are modelled across the seasons, regardless of the underlying changes in population dynamics within a system (MacKenzie et al., 2006).

Since interest now lies over multiple seasons, the likelihood for the observed data (Eq. 2.1) for a given season (t) is modified to:

$$L_t\left(\psi_t, \mathbf{p}_t \mid \mathbf{h}_{t,1}, \mathbf{h}_{t,2}, \ldots, \mathbf{h}_{t,s}\right) = \prod_{i=1}^{s} \Pr\left(\mathbf{h}_{t,i}\right) \tag{2.7}$$

where ψ_t denotes the occupancy probability and p_t denotes the detection probability during season t, respectively. The full likelihood analysed for the seasons of interest is then defined as the product of all seasonal likelihoods, expressed as:

$$L\left(\psi, \mathbf{p} \mid \mathbf{h}_1, \ \mathbf{h}_2, \ldots, \mathbf{h}_s\right) = \prod_{t=1}^{T} L_t\left(\psi_t, \mathbf{p}_t \mid \mathbf{h}_{t,1}, \ \mathbf{h}_{t,2}, \ldots, \mathbf{h}_{t,s}\right). \tag{2.8}$$

The model does not clearly account for the extinction and colonization vital rates that dictate the changes in the range dynamics and makes a restrictive assumption regarding these

parameters in the sense that the local colonization probability of the species occupying a previously unoccupied site is equivalent to the local extinction probability of the species at the previously occupied site. This assumption generally does not hold in practice, thus, previous research has found it more useful to consider the more general approach which explicitly models the occupancy dynamics from which this implicit dynamics model can be derived (Kéry et al., 2013; Weir et al., 2009; MacKenzie et al., 2006).

2.3.3 Explicit Dynamics Model

Ecologists are often directly interested in the processes driving changes in population occupancy for the purposes of well-informed management decisions pertaining to ecological systems. Hence, the use of a model which explicitly estimates both colonization and extinction probabilities is the more desirable approach to dynamic occupancy models (MacKenzie et al., 2003).

This method considers the dynamic changes in occupancy as a first-order Markov process, which states that site occupancy in the current season is dependent on the state of site occupancy in the previous season, therefore accounting for a form of temporal autocorrelation. In essence, this Markovian process defines values close in time to be more similar than those separated by longer periods when these observations are positively correlated.

Furthermore, the changes in occupancy when modelled as a Markovian process can be regarded as providing a form of occupancy heterogeneity since the probability of a site being occupied by the target species in the current season will be different for sites that were either occupied or unoccupied the previous season (MacKenzie et al., 2006). There are three approaches to modelling dynamic occupancy models that explicitly account for these vital rate parameters.

Modelling Dynamic Processes when Detection Probability is 1

The historical, and most straightforward, approach to the modelling of dynamic processes assumes that the detection probability is set to one. The development of ecological studies pertaining to animal populations assessed over a multi-year time frame began in the late 1960s and was primarily influenced by the methods of Wilson and MacArthur (1967).

Their work suggested that species richness on islands reflected a situation of dynamic equilibrium between the rates of extinction and colonization. This situation was considered to have resulted from a stationary Markov process defined by equivalent vital rates (Simberloff, 1969; Diamond and May, 1977).

The estimation of equivalent vital rates for a stationary Markov process from detection/non-detection data were examined by Clark and Rosenzweig (1994) for which they provided

maximum likelihood estimates assuming both detection probabilities of one with constant rate parameters over time.

This approach was further expanded by Erwin et al. (1998) and allowed for time-specific vital rates in addition to non-Markovian models with reduced parameters which all assumed detection probabilities of one.

However, as with the static occupancy models, these historical models are of little practical value since the estimates are only reasonable when presence and absence of a target species is attainable with absolute certainty.

The Conditional Approach

The work of Barbraud et al. (2003) considered the same estimation problem as Erwin et al. (1998) but wanted to account for detection probabilities of less than one. They identified that site occupancy and capture-recapture studies with temporary emigration (specifically the robust design of Pollock (1982)) are analogous and considered the Markovian temporary emigration model, put forth by Kendall et al. (1997), as a useful method of estimating the occupancy and vital rate parameters.

Essentially, this is considered a conditional approach since the detection history is only modelled from the first detection (capture) of the species at a site and does not model the previous non-detections up until the point of capture, thus, it is only possible to derive estimates of the vital rate parameters and not the seasonal estimates of occupancy (MacKenzie et al., 2006).

The Unconditional Approach

The unconditional approach of MacKenzie et al. (2003) is the most commonly applied method of explicitly estimating the extinction and colonization probabilities and is a direct extension of the static occupancy model put forth by MacKenzie et al. (2002). This method of occupancy modelling allows for the assumptions of a stationary Markov process and perfect detection to be relaxed, thus creating a flexible framework which results in unbiased parameter estimates.

The estimated occupancy probability of a site is determined in the first season and any changes in site occupancy probability between seasons is derived using the vital rate parameters. Formally, these vital rate probabilities are defined as:

γ_t = the probability that an unoccupied site in season t is occupied by the species in season $t+1$; and

ε_t = the probability that a site occupied in season t is unoccupied by the species in season $t+1$.

The conditional approach using the Markovian temporary emigration model, developed by Barbraud et al. (2003), is a special case of the model described by MacKenzie et al. (2003). Additionally, the work of MacKenzie et al. (2003) is considered an unconditional approach to the modelling of dynamic processes since the non-detections of the target species from the start of the study up until its first observation (capture histories) are modelled. Thus, both the vital rate parameters and the seasonal occupancy parameter can be directly estimated.

Both approaches provide unbiased estimates of the colonization and extinction probabilities, but the unconditional approach is considered to be more efficient due to the ability to estimate site occupancy for each study season (MacKenzie et al., 2003, 2006).

Using matrix notation, the probability of a site transitioning between occupancy states from season t to $t+1$ is defined as:

$$\phi_t = \begin{bmatrix} 1-\varepsilon_t & \varepsilon_t \\ \gamma_t & 1-\gamma_t \end{bmatrix},$$

where the rows of ϕ_t indicate the occupancy state of the site in season t and the columns represent the occupancy state in season $t+1$. Additionally, ϕ_0 is defined as a row vector that models whether a site was occupied or unoccupied in the first surveyed season:

$$\phi_0 = \begin{bmatrix} \psi_1 & 1-\psi_1 \end{bmatrix},$$

with ϕ_1 representing the probability of site occupancy.

Furthermore, $p_{h,t}$ is defined as a column vector where each entry represents the probability of observing the detection history $\mathbf{h}_{t,i}$ in season t, conditional on the occupancy state. Two examples of this column vector are:

$$\mathbf{p}_{101,t} = \begin{bmatrix} p_{t,1}\,(1-p_{t,2})\,p_{t,3} \\ 0 \end{bmatrix} \quad \& \quad \mathbf{p}_{000,t} = \begin{bmatrix} \prod_{j=1}^{3}(1-p_{t,j}) \\ 1 \end{bmatrix}$$

where, whenever the species is detected at least once during season t, the second element of $\mathbf{p_{h,t}}$ will always equate to zero. This is because the chance of observing such a detection history if the site is in an unoccupied state is impossible. Similarly, if the species is never detected at a site during season t then the second element will always equate to 1, as this is the only observable detection history for an unoccupied site.

Using the results of this matrix notation, an observed detection history probability can be defined (MacKenzie et al., 2006) as:

$$\Pr\left(\mathbf{h}_i\right) = \phi_0 \prod_{t=1}^{T-1} D\left(\mathbf{p}_{\mathrm{h},t}\right) \phi_t \mathbf{p}_{\mathrm{h},T},$$

where $D\left(\mathbf{p}_{\mathrm{h},t}\right)$ denotes a diagonal matrix with the elements of $p_{h,t}$ along the main diagonal. Assuming independence of detection histories, the model likelihood can be calculated using the results of the probability statements for each observed detection history, expressed as:

$$L\left(\psi_1, \gamma, \varepsilon, \mathbf{p} \mid \mathbf{h}_1, \ldots, \mathbf{h}_s\right) = \prod_{i=1}^{s} \Pr\left(\mathbf{h}_i\right) \tag{2.9}$$

In the same manner as the static occupancy modelling of MacKenzie et al. (2002), the dynamic occupancy model can assume constant parameters across all sites at every instant of time. Failure to meet this assumption introduces heterogeneity and results in biased parameter estimates. Similarly, this model can allow for the addition of covariate information to account for potential heterogeneity present due to these measured factors, albeit inaccurate and biased estimates still present themselves when this heterogeneity is created by some unknown quantity.

Additionally, the model of MacKenzie et al. (2003) allows for the presence of missing information and handles these missing observations by simply omitting them from the detection history which results in unbiased parameter estimates.

A simulation study generated data, under the assumption that all four parameter estimates were constant over time, to assess the model's general performance. MacKenzie et al. (2003) found through their simulations that their model generally produced unbiased estimates in all scenarios except for the case when the values of the detection probability in a survey (0.2) and the number of surveys (2) in a season were low.

This resulted in overestimated occupancy and colonization parameters in the case of a low to moderate (0.2-0.5) true occupancy value derived at the first season of the study and underestimated otherwise. Additionally, the extinction parameter tended to be overesti-

mated when the number of surveys and detection probability were low.

The general findings of MacKenzie et al. (2003) were that the model's ability to estimate all relevant parameters precisely and accurately was improved when the number of sites, seasons, surveys within a season and detection probability within a survey were increased, respectively.

A study may not be directly interested in the estimates of occupancy, colonization and extinction, but may be interested in quantities such as the rate of change in the occupancy status of an invasive species in an area, or seasonal occupancy estimates over time (MacKenzie et al., 2003, 2006; Bailey et al., 2014). The dynamic occupancy model allows for the reparameterization of the four estimated parameters to estimate these occupancy-related quantities directly.

2.3.4 Extensions of the Unconditional Model

The dynamic occupancy model makes several fundamental assumptions that need to be met for unbiased and efficient maximum likelihood estimates. Namely: the closure assumption within seasons must hold, as was the case with the static model of MacKenzie et al. (2002); the initial occupancy probability and subsequent vital rate parameters are either constant across sites, or any introduced heterogeneity is accounted for using measurable covariates; there is no heterogeneity in the detection probabilities that hasn't been captured by appropriate covariates; survey results are independent of each other; and no false-positives exist (MacKenzie et al., 2003, 2006; Bailey et al., 2014; Sepulveda, 2018).

The model development of Miller et al. (2011) made special note of the problem of false-positives and extended the standard static and dynamic models to allow for these false-positive errors when a subset of the total detections is known with absolute certainty (Miller et al., 2013; Bailey et al., 2014). Models that account for false-positives are fairly new and in the stages of improvement. However, given the severe bias caused by ignoring the problem of false-positive detections, it is recommended to use these false-positives models when a study suspects such errors to test whether the detection history for any particular site within a given season is truly equivalent to zero (Bailey et al., 2014).

Random movement patterns of a species that cause differences in occupancy probabilities across all sites within a season will still lead to unbiased results. However, when non-random movement (migration) patterns are present then the closure assumption is violated and leads to biased results (Kendall, 1999; MacKenzie et al., 2006). There are two main approaches to eliminate the bias induced from this violation. The first involves the recommendations of Kendall (1999) which state that, for the case of immigration or

emigration-only movement patterns, the survey data can be pooled into two surveys per season and subsequently modelled with survey-specific detection probabilities. Alternatively, the problem is addressed by restricting the data to include surveys within the time period when it is known that the availability of the target species is uninterrupted (such as within a breeding season) (MacKenzie et al., 2006).

In situations when there is an apparent correlation among survey outcomes then the assumption of independent surveys will be violated. This occurs when there are differences in the observation process from survey to survey within a site, and is caused when there is evidence of individual animals being detected at different rates due to differences in the detection ability of observers (Bailey et al., 2014).

Additionally, this model violation can occur when the study area has set the survey sites close enough that the same individual will be detected at different sites within the same survey. This particular violation is directly addressed through the development of multi-scale occupancy models (Hines et al., 2010). These models make use of this observer dependence by allowing for inference regarding species occurrence to take place at multiple hierarchical scales.

The standard dynamic model has also been extended to account for heterogeneous detection probabilities that are still unaccounted for after the inclusion of measured covariates. This problem is initially addressed by extending the finite mixture approach of Norris and Pollock (1996). Alternatively, an approach which fits a model with a random effect assigned to each parameter for each specific site from some defined distribution could be implemented; albeit this extension is computationally difficult to execute using a maximum likelihood approach and is generally reserved for use within the Bayesian framework (MacKenzie et al., 2006).

This problem could alternatively be addressed within the maximum likelihood framework by extending the abundance distribution modelling of Royle and Nichols (2003), in which the probability of detecting the target species at a given site is modelled as a function of the detection probability of individual animals.

Chapter 3

Data and Methodology

A recap of the research questions are as follows: the practical objectives center around the apparent and potential distributions of the two invasive species of interest, the factors affecting both occupancy and detectability, as well as whether these distributions are expanding or contracting over time; while the theoretical objectives focus on how the observed results of the fitted models, specifically pertaining to the dynamic framework, change when the structure of the data was altered in terms of the sites surveyed as well as the total number of surveys conducted at every site within each sampling period .

This chapter discusses the methodological approach taken in answering these research questions. It contains a step-by-step explanation of the data collection process including the extraction of the detection/non-detection data, the observational covariates, and the environmental covariates. The chapter continues with a brief explanation of the data exploration process and finishes with a full description of the method of data analysis.

All analysis was conducted using the R programming language (R Core Team, 2021) which applies a variety of packages needed to conduct statistical analysis, including the **unmarked**, **ABAP** and **AICcmodavg** libraries (Fiske and Chandler, 2011; BIRDIE Development Team, 2022; Mazerolle, 2020). When necessary, computationally intensive code was implemented with the support offered by the University of Cape Town's High Performance Computing (HPC) facility (hpc.uct.ac.za).

3.1 Data

This section discusses the steps that were taken in the data collection process. This collection process was achieved through the functionality provided by the **ABAP** package (BIRDIE Development Team, 2022), which allowed for the extraction of the data needed to fit both static and dynamic occupancy models.

3.1.1 ABAP

The African Bird Atlas Project (ABAP) is a citizen science database of bird checklists collected by volunteers (citizen scientists) which aims to enable the mapping of different bird distributions across multiple Southern African countries. This database contains the recorded detection and non-detection data taken by these volunteers, who follow a strict protocol, across a grid of locations superimposed across Southern Africa from 2007 to present (Bled et al., 2013). This project has been expanded over the last seven years to

include coverage of West and East African countries such as Nigeria and Uganda; however, for the purposes of this dissertation only sites located in South Africa were of interest.

These sites are known as *pentads* and represent 5-minute x 5-minute latitude and longitude rectangular cells (Greenwood, 2007; Broms et al., 2014). Each submitted checklist represents one survey at a particular site with the submission protocol requiring at least two hours of intensive birding over a maximum period of five days (Harebottle et al., 2007). These checklists include a recorded measure of effort, namely the total active number of hours spent bird watching. With the help of citizen scientists, the assessment of spatial distributions for wide ranging elusive species is possible where more intensive sampling schemes would incur high costs and are considered largely inefficient (Karanth et al., 2011; Johnson et al., 2013).

A problematic trait of citizen science data is the lack of a rigid design, which was portrayed by the varying number of surveys per site and the number of surveyed sites per year, resulting in an imbalanced set of data. The inconsistent levels of effort, varying levels of birding expertise, lack of quality control, and possible species misidentification all negatively impact the reliability of the data (Altwegg and Nichols, 2019). This trait, in conjunction with the lack of literature guiding the level of temporal overlap necessary for the models to be reliable, motivated the assessment of model sensitivity through altering the structure of the data.

3.1.2 Environmental Data

The covariate data were extracted using the functionality provided by the **ABAP** package (BIRDIE Development Team, 2022) which extracts the data from the Google Earth Engine (GEE) catalogue at the same spatial and temporal resolution as the detection/non-detection data. A full list of the data catalogues, the environmental covariates (extracted as the recorded mean per pentad per year), and a brief description of these variables is provided in the Appendices by Table A.1.

The process of extracting and cleaning from the GEE data catalogue was conducted according to the following steps:

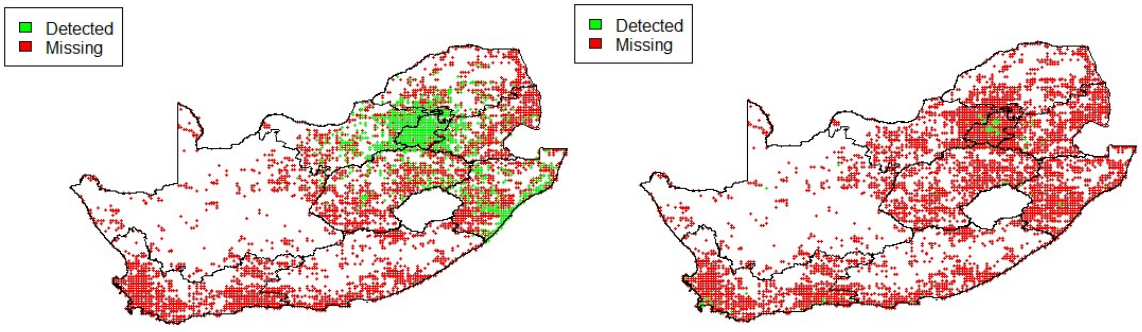
1. Pentads were mapped and those falling outside the borders of South Africa were excluded.
2. Extracted covariate data at the pentad spatial resolution for the years 2010-2019.
3. Created yearly environmental dataframes that include fixed and year-specific covariates.

4. Conducted kNN ($k = 5$) imputation for missing values where necessary. For every observation to be imputed, it identified the k closest observations based on the Euclidean distance and computed the weighted average (weighted based on distance) of k (Malarvizhi and Thanamani, 2012).

Once both the detection/non-detection and environmental data were collected, a function to clean and manipulate the data into the structure needed to fit occupancy models was written. This included, but was not limited to, the removal of erroneous checklists; formatting the site-level and observation-level covariates; the removal of sites falling outside of the study region; and the creation of the Season and Julian date variables to form part of the structure needed to generate the observation-level covariates. The full algorithm describing the workings of this function is provided in the Appendices by Table A.2.

3.2 Data Exploration

This section focused on preliminary analysis of the data by performing a naive exploration of occupancy. The years of interest for this study were 2010 through to 2019. Thus, the yearly presence and absences of the target species at the surveyed sites for the years 2010, 2013, 2016 and 2019 were mapped as a means of determining which areas were to be included in the study region of the occupancy models. In addition, tabulation of the yearly number of sites with at least one detection per province were produced to aid the decision regarding the restriction of the respective study regions. The full results can similarly be found in Table 4.2.



(a) Myna **(b)** Mallard

Figure 3.1: Detection/non-detection data for the invasive species in 2010

Figure 3.1 provides insight into the decision-making process regarding the definition of the study regions for the respective species. This map represents the detections (green) and non-detections (red) of the two invasive species for the year 2010, whereby the species was recorded as detected at a site if it was observed at least once, regardless of the number

of surveys needed to record it. The same maps for the years 2013, 2016 and 2019 were produced and can be found in the Appendices by Figure A.1.

Investigation of these spatial maps (Figures 3.1 and A.1) in conjunction with the yearly detections (Table 4.2) led to the decision that the study region of the Myna excluded the Western, Eastern and Northern Cape, while the study region of the Mallard excluded Limpopo, Eastern Cape and the Northern Cape. This was based on the fact that these species were (relatively) infrequently recorded as present at sites in the respective provinces, and thus it was considered uninformative to predict occupancy in these areas.

Multicollinearity of the environmental covariates is an issue that requires consideration prior to the model fitting process (De Marco and Nóbrega, 2018) and was explored using the **usdm** package (Naimi et al., 2014). The variance inflation factor (VIF) threshold was set to six based on previous literature (De Marco and Nóbrega, 2018) and, based on the results, three covariates (`AET`, `Urban_Cover` and `Seasonal_Water`) were excluded from any analysis of occupancy.

3.3 Data Analysis

This section describes how the data were structured and analysed for both the static and dynamic occupancy models which were fit using the functionality available in the **unmarked** package.

3.3.1 Static Occupancy Models

The hierarchical model implemented in the **unmarked** package for the modelling of static occupancy expresses the latent occupancy state and detection processes defined by Fiske and Chandler (2011) as follows:

$$\begin{aligned} Z_i &\sim \text{Bernoulli}(\psi) \quad \text{for } i = 1, 2, \ldots, S \\ Y_{ij} \mid Z_i &\sim \text{Bernoulli}\left(Z_i p\right) \quad \text{for } j = 1, 2, \ldots, K_i \end{aligned} \tag{3.1}$$

where Y_{ij} represents species detection(1) or non-detection(0) at a surveyed site and Z_i is a binary indicator representing species occurrence such that $Z_i = 1$ if site i is occupied by the target species and 0 if it is absent. The marginalization of the latent Z variables yields the likelihood as provided in Chapter 2 by Equation 2.1 (Fiske and Chandler, 2011).

The predictors that can be fitted to the probability of occupancy or detectability are either site- or observation-level covariates. The site-level covariates are those that are site-specific and do not change based on the conducted survey, these included covariates such as elevation above sea level and are generally used to explain changes in occupancy.

Observation-level covariates are survey-dependent and generally will vary with every survey conducted at a given site. Covariates with this structure include factors such as season, total hours spent bird watching or any covariate that is likely to have an influence on detectability (Bailey et al., 2014).

The individual static occupancy models were fitted to the Myna and Mallard data corresponding to the year (season) 2010 and the pool of potential covariates that were included in the model-fitting process is provided by Table 3.1.

Covariate structure	Variable	Description
Site-level	`Latitude`	Geographic coordinate system (Y)
	`Longitude`	Geographic coordinate system (X)
	`Elevation`	Height above sea level
	`TreeCanopy`	Tree height
	`SoilPH`	pH of the surface soil
	`SurfaceTemp`	Temperature of the surface
	`Permanent_Water`	Permanent water covering (%) pentad
	`Population.Density`	Population (density)
	`Avg.Temp`	Air temperature
	`Precipitation`	Rainfall
	`Soil.Moisture`	Soil moisture
	`Wind.Speed`	Wind speed
	`Pressure`	Air pressure
	`NDVI`	Normalized difference vegetation index
	`Population`	Population (count)
Observation-level	`lhours`	Log of total hours
	`ltotspp`	Log of total species
	`Season`	Season of survey

Table 3.1: Covariates included in the static occupancy model fitting process

The static occupancy modelling process was conducted in this dissertation using a two-stage approach. The first stage comprised a purely statistical-based approach to modelling the detection parameter. The pool of factors, that were considered potentially influential to the changes in the probability of detection amongst sites, were fitted as a set of univariate occupancy models with no covariates modelling occupancy. A set of factors that were individually statistically significant (at a significance level of $\alpha = 0.05$) were then combined to investigate the significance of adjusted effects and, after removing those insignificant parameters, was then applied as the structure that best described the detection process. Additionally, the potential effect of nonlinear relationships between the covariates and the

detection probability were accounted for through the inclusion of quadratic effects where it made sense, namely the `lhours` and `ltotspp` covariates.

It is worth noting that most studies tend to only use site-level covariates to model differences in occupancy among sites and observation-level covariates to model changes in the detection probability among sites. However, it is possible to use site-level covariates to model the probability of detection (and vice versa) when the use of such a covariate could be considered intuitive (MacKenzie et al., 2006).

With this in mind, the `Population.Density` and `NDVI` site-covariates were included in the modelling of the detection process. The `NDVI` covariate was included since it is considered logical that areas with greener areas are likely to have thicker shrubbery which will make it more difficult to observe certain species, while the `Population.Density` covariate was included since it is assumed by the author that more people in a region are likely to result in a larger proportion of birders searching for the species, and thus a higher likelihood of detection.

The second stage comprised a hypothesis-driven approach to occupancy modelling. A set of hypotheses was formulated and various site-level covariates were used to reflect the hypothesis of interest. This approach compared various anthropogenic, climatic and environmental factors, and explicitly accounted for potential nonlinear effects. These hypotheses were tabulated and provided for both the Myna (Table 3.2) and the Mallard (Table 3.3), respectively.

#	Name	Hypothesis	Covariates [1]
1	Climate	Mynas prefer temperate and subtropical climate regions	`Pressure`, `SurfaceTemp`, `Precipitation`, `Wind.Speed`
2	Weather	Mynas prefer windless areas with high rainfall and high temperatures	`SurfaceTemp`, `Precipitation`, `Wind.Speed`
3	Urban	Mynas prefer urban areas	`Population.Density`
4	Env	Mynas prefer environments with high tree canopies and vibrant insect life	`Elevation`, `TreeCanopy`, `SoilPH`

[1]The static occupancy model hypotheses for both species include only linear terms for the

5	Urban.Clim.Env	Mynas prefer areas that are urban, areas that have high tree canopies, and areas with temperate or subtropical climates	`Pressure`, `SurfaceTemp`, `Precipitation`, `Wind.Speed`, `TreeCanopy`, `Population.Density`
7	Urban.Clim.NoWind	Mynas prefer areas that are urban, and areas with temperate or subtropical climates regardless of average wind speeds	`Pressure`, `SurfaceTemp`, `Precipitation`, `Population.Density`
8	Urban.Clim.Windy	Mynas prefer areas that are urban, and areas with temperate or subtropical climates	`Pressure`, `SurfaceTemp`, `Precipitation`, `Wind.Speed`, `Population.Density`

Table 3.2: Static occupancy model hypotheses for the Myna

#	Name	Hypothesis	Covariates
1	Low.Water	Mallards prefer low-land areas, areas with large bodies of water, and areas with slightly alkaline soil	`Elevation`, `SoilPH`, `Permanent_Water`
2	Coastal	Mallards prefer coastal regions	`Elevation`, `Wind.Speed`, `Permanent_Water`
3	Climate	Mallards prefer Mediterranean and temperate climates	`Pressure`, `SurfaceTemp`, `Precipitation`, `Wind.Speed`
4	Urban	Mallards prefer urban areas	`Population.Density`

`Population.Density` and `Permanent_Water` covariates, and both linear and quadratic terms for the `Pressure`, `SurfaceTemp`, `Precipitation`, `Wind.Speed`, `Elevation`, `TreeCanopy` and `SoilPH` covariates.

5	Urban. Water. Env. Clim	Mallards prefer urban areas, areas near large bodies of water, areas with rich soil, high temperature areas, and areas with high rainfall	SoilPH, SurfaceTemp, Precipitation, Permanent_Water, Population.Density
6	Urban. Env. Clim	Mallards prefer urban areas, areas with high annual rainfall, areas with rich soil, and areas with high tree canopies	SoilPH,SurfaceTemp, Precipitation,TreeCanopy, Population.Density
7	Water. Climate	Mallards prefer areas with large bodies of water, and areas with Mediterranean or temperate climates.	Pressure, SurfaceTemp, Precipitation, Wind.Speed, Permanent_Water
8	Env	Mallards prefer a natural environment with large bodies of water, tree cover and plant/insect rich soil	Elevation, SoilPH, TreeCanopy, Permanent_Water
9	Urban. Climate	Mallards prefer urban areas, and areas with Mediterranean or temperate climates	Pressure, SurfaceTemp, Precipitation, Wind.Speed, Population.Density
10	Urban.Env	Mallards prefer urban areas, and areas with a natural environment with low elevation, large bodies of water, tree cover, and rich soil	Elevation, SoilPH, TreeCanopy, Permanent_Water, Population.Density

Table 3.3: Static occupancy model hypotheses for the Mallard

Assessment of model fit

The selection of the best fitting model out of the pool of hypothesized models was determined using the **AICcmodavg** package (Mazerolle, 2020) by an evaluation of the AIC statistics in conjunction with the AIC weights. When the difference between the AIC values of the best models are low, the acceptance of one model and exclusion of the others can lead to a misplaced sense of confidence and a form of selection bias (Wagenmakers and Farrell, 2004). The inclusion of the AIC weights, best described as the probability that the chosen model is the best model given the data and the candidate set of models

(Burnham and Anderson, 2002), allowed for the consideration that more than one model out of the candidate set may adequately describe the data.

The goodness-of-fit (GOF) of the selected model was not computed using the chi-square and parametric bootstrapping technique, as presented by MacKenzie and Bailey (2004), since this first stage involving the static models was considered exploratory and used primarily to inform the dynamic models. Further, the number of surveys was not constrained at this stage, and the benefit of determining the goodness-of-fit at this stage of the analysis did not warrant the major computational burden of this technique.

Although the parametric bootstrapping approach was not implemented in the static modelling analysis, the assessment of model fit was purely visual and performed via the mapping of the predicted occupancy superimposed by the observed occupancy state, derived from the data corresponding to the year 2010.

3.3.2 Dynamic Occupancy Models

The dynamic occupancy models were similarly applied using the functionality provided by the **unmarked** package, namely the `occu` and `colext` functions. Using the hierarchical model implemented in this package: the study period of interest for both species from 2010 to 2019; colonization and local extinction probabilities were estimated based on both the initial occupancy state derived from the first study season and changes in occupancy between seasons; and the occupancy state of any given site was assumed to develop according to a Markovian process. Hence, the data are described using a two-state hidden Markov model (Fiske and Chandler, 2011). The state transitions were controlled by the probability of local colonization and extinction and were dependent on the occupancy status at the previous time step (Kéry et al., 2013). A visual representation of the occupancy dynamics and transitions between states using a two-state hidden Markov model is provided by Figure 3.2.

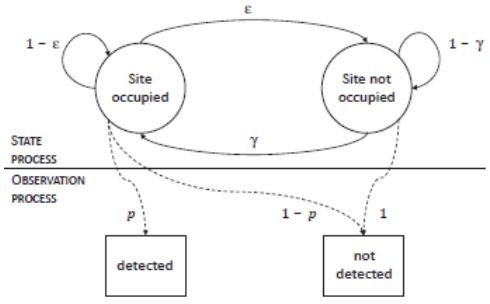


Figure 3.2: Hidden Markov model for occupancy dynamics under imperfect detection, extracted from Kéry et al. (2013).

The figure depicts both the latent occupancy (state) and detection (observation) processes divided by a horizontal line. The state process is governed by the occupancy state of a particular site, represented by the circles, and the solid arrows which denote the state transitions characterizing the occupancy dynamics. Additionally, the observation process is governed by possible observations, denoted by the squares, and the dashed arrows represent the observation process pertaining to each survey replicate (Kéry et al., 2013).

The latent state and observation processes describing the dynamic occupancy model as implemented by Fiske and Chandler (2011) are defined as:

$$\begin{gathered} Z_{i1} \sim \text{Bernoulli}(\psi) \\ Z_{it} \sim \begin{cases} \text{Bernoulli}\,(\gamma_{t-1}) & \text{if } Z_{i(t-1)} = 0 \\ \text{Bernoulli}\;\,(1-\epsilon_{t-1}) & \text{if } Z_{i(t-1)} = 1 \end{cases}, \\ \text{for } t = 2, 3, \ldots, T \\ Y_{ijt} \mid Z_{it} \sim \text{Bernoulli}\,(Z_{it}p)\,, \end{gathered} \tag{3.2}$$

where Z_{it} denotes a binary indicator of the latent occupancy state of the species at site i during season t as present (1) or absent (0) and Y_{ijt} represents the observed species occurrence status of site i at survey j during season t. The static occupancy model is generalised by this Markov model, through the relaxation of the between-season closure assumption, and the likelihood for this model is provided in Chapter 2 by Equation 2.9.

While there is not much research pertaining to the optimal design of dynamic occupancy models, it has been noted that important consideration must be made for: the time interval between two seasons; the number of sites surveyed each season; whether the same or different sites are surveyed each season; and how many surveys are conducted within each season (MacKenzie et al., 2006). This dissertation placed an emphasis on analysing the data using different combinations of surveys and sites studied. A summary of the data structures (the number of sites and surveys in the study design) is provided by Table 3.4.

Species	**(Seasonal) Surveys**	**Common Sites**	**Intermediate Sites**	**All Sites**
Myna	10	775	1918	6911
	50	775	1918	6911
Mallard	10	858	2240	7041
	50	858	2240	7041

Table 3.4: Summary of the six different data structures per species

The investigation of the different study designs began by defining a survey season as one full year, from the start of January to the end of December, with no breaks between seasons. The number of sites surveyed in a given season was split into three categories based on the following criteria: "Common Sites" included those sites surveyed every year in the 10-year study period; "Intermediate Sites" included those sites surveyed at least six times in the 10 year study period; "All Sites" included those sites surveyed at least once in the 10-year study period. These three categories were further split into designs that included seasonal detection histories with a maximum length of 10 or 50 per site.

Table 3.5 depicts the different covariate structures that were available for the dynamic occupancy model fitting process.

Covariate structure	Variable	Description
Site-level	`Latitude`	Geographic coordinate system (Y)
	`Longitude`	Geographic coordinate system (X)
	`Elevation`	Height above sea level
	`TreeCanopy`	Tree height
	`SoilPH`	pH of the surface soil
	`SurfaceTemp`	Temperature of the surface
	`Permanent_Water`	Permanent water covering (%) pentad
	`Population.Density`	Population density
	`Avg.Temp`	Air temperature
	`Precipitation`	Rainfall
	`Soil.Moisture`	Soil moisture
	`Wind.Speed`	Wind speeds
	`Pressure`	Air pressure
	`NDVI`	Normalized difference vegetation index
	`Population`	Population count
Observation-level	`lhours`	Log of total hours
	`ltotspp`	Log of total species
	`Season`	Season of survey
Yearly-site-level	`Yearly_Temp`	Yearly air temperature
	`Yearly_Rain`	Yearly rainfall
	`Yearly_Soil`	Yearly soil moisture
	`Yearly_Wind`	Yearly wind speeds
	`Yearly_Pressure`	Yearly air pressure
	`Yearly_NDVI`	Yearly NDVI
	`Yearly_Pop`	Yearly population count

Table 3.5: Covariates included in the dynamic occupancy model fitting process

While the site-level and observation-level covariates were the same as those shown in Table 3.1, there was a new structure of interest when modelling dynamic processes. This was the yearly-site-level covariate structure, which represented a set of measured covariates that were recorded in each primary period for the 10 year study. While initial occupancy was estimated as a function of the site-level covariates, the yearly-site-covariates were used to estimate the vital rate parameters as well as the detectability of the target species.

The approach to modelling the dynamic occupancy models was implemented as a single-stage approach based on the following steps. The covariates that described the occupancy and detection processes, of the best fitting models produced from the static occupancy models, were used to estimate the initial occupancy and detection parameters for the dynamic models. The choice of colonization and extinction parameters were then evaluated using a hypothesis-driven modelling approach. These hypotheses were symmetric in the sense that a potential covariate was considered to influence both colonization and extinction and was expected to display opposing effects if the covariate was statistically significant. The set of formulated hypotheses is provided for both the Myna (Tables 3.6) and the Mallard (Table 3.7), respectively.

#	Name	Hypothesis	Covariates [2]
1	Temp.Rain.Urban	Dynamic population rates of the Myna are primarily governed by changes in temperature, rainfall and population	`Yearly_Temp`, `Yearly_Rain`, `Yearly_Pop`
2	Wet	Dynamic population rates of the Myna are primarily governed by changes in wet areas	`Yearly_Rain`, `Yearly_Soil`
3	Environment	Dynamic population rates of the Myna are primarily governed by changes in environmental factors	`Yearly_Soil`, `Yearly_NDVI`
4	Humid.Urban	Dynamic population rates of the Myna are primarily governed by changes in urban areas in humid climates	`Yearly_Pressure`, `Yearly_Pop`

[2]The relevant covariates were fitted to estimate both vital rate parameters and did not include any quadratic effects.

5	Clim.Urban	Dynamic population rates of the Myna are primarily governed by changes in a combination of climatic and anthropogenic factors	`Yearly_Pressure` `Yearly_Temp`, `Yearly_Rain`, `Yearly_Pop`
6	Clim.Urban.Wind	Dynamic population rates of the Myna are primarily governed by changes in a combination of climatic and anthropogenic factors	`Yearly_Pressure` `Yearly_Temp`, `Yearly_Rain`, `Yearly_Wind`, `Yearly_Pop`
7	Env.Clim.Urban	Dynamic population rates of the Myna are primarily governed by changes in a combination of environmental, climatic, and anthropogenic factors	`Yearly_Pressure` `Yearly_Temp`, `Yearly_Rain`, `Yearly_Wind`, `Yearly_Pop`, `Yearly_Soil`, `Yearly_NDVI`

Table 3.6: Dynamic occupancy model hypotheses for the Myna

#	Name	Hypothesis	Covariates
1	Humid.Env	Dynamic population rates of the Mallard are primarily governed by changes in soil moisture, temperature and rainfall	`Yearly_Soil`, `Yearly_Temp`, `Yearly_Rain`
2	Weather	Dynamic population rates of the Mallard are primarily governed by changes in weather patterns	`Yearly_Temp`, `Yearly_Rain`, `Yearly_Wind`
3	Anthropogenic	Dynamic population rates of the Mallard are primarily governed by changes in anthropogenic factors	`Yearly_Pop`

4	Wetlands	Dynamic population rates of the Mallard are primarily governed by changes in wetland biomes	`Yearly_Soil`, `Yearly_Pressure`
5	Arid	Dynamic population rates of the Mallard are primarily governed by changes in hot and dry climates	`Yearly_Temp`, `Yearly_NDVI`
6	Wet.Urban	Dynamic population rates of the Mallard are primarily governed by changes in urban areas with wet weather conditions	`Yearly_Pop`, `Yearly_Soil`, `Yearly_Rain`
7	Clim.Urban	Dynamic population rates of the Mallard are primarily governed by changes in climatic and anthropogenic factors	`Yearly_Temp`, `Yearly_Rain`, `Yearly_Pressure`, `Yearly_Wind`, `Yearly_Pop`,

Table 3.7: Dynamic occupancy model hypotheses for the Mallard

Model selection and goodness-of-fit

The assessment and validation of the dynamic model fit now utilised the method developed by MacKenzie and Bailey (2004) in addition to the AIC statistic used for the static model selection procedure. It is noted by Kéry and Chandler (2012) that the AIC statistic cannot be used for model selection if missing values for yearly-site-covariates are present due to the manner in which **unmarked** removes missing data. However, in this dissertation there were no missing observations in the yearly-site-covariates, thus the AIC along with the AIC weights were used to compute model fit.

The goodness-of-fit (GOF) of the selected best model was then computed using the chi-square and parametric bootstrapping technique, as presented by MacKenzie and Bailey (2004), using the algorithm described in Chapter 2 by Table 2.2. This approach allowed for a formal statistical hypothesis test through the chi-square statistic in addition to the calculation of an overdispersion parameter as a means of assessing the model fit.

Chapter 4

Results

4.1 Exploratory data analysis

4.1.1 Exploration of the presence/absence data

The exploration of the data began with a simple summation of the applicable pentads in the study region in terms of how they were distributed between provinces, provided by Table 4.1.

Province	**Provincial Code**	**Pentads**
Eastern Cape	EC	2255
Free State	FS	1817
Gauteng	GP	258
KwaZulu-Natal	KZN	1293
Limpopo	LP	1639
Mpumalanga	MP	1004
North West	NW	1411
Northern Cape	NC	5182
Western Cape	WC	1830
Total		**16689**

Table 4.1: Total number of applicable pentads in the study region by province.

This table indicates that there were $16,689$ pentads included in the study region that were not evenly distributed across the relevant provinces. This was to be expected since the pentads are of a uniform size, but the provinces differ in geographic size. As mentioned in Chapter 3, the study regions for the different species were constrained based on an investigation of a set of detection/non-detection maps provided by Figures 3.1 and A.1 in conjunction with the yearly detections presented by Table 4.2.

An evaluation of this table suggested that while the Myna's detections allowed for the creation of a definitive provincial exclusion threshold, of at least 25 detections per province per year, the Mallard's threshold was not as easily defined. It could have been argued that additional provinces may be excluded, however, the potential loss of information by

excluding more regions than necessary was a cause for concern. It was thus assumed that the presence of an area with comparatively infrequent detections was not expected to create any problems with the observed results.

Detected Mynas		Province									Total
		EC	FS	G	KZN	L	M	NW	NC	WC	
Year	2010	3	170	216	259	127	165	274	8	0	1222
	2011	6	152	221	330	146	165	274	22	0	1316
	2012	2	129	216	302	141	152	265	15	0	1222
	2013	6	109	219	313	148	143	199	23	0	1160
	2014	2	145	232	300	194	158	224	17	0	1272
	2015	7	135	230	185	156	189	285	14	0	1201
	2016	8	114	230	154	170	173	255	15	0	1119
	2017	5	119	223	163	204	138	222	19	0	1093
	2018	7	111	224	156	255	170	198	15	0	1136
	2019	7	97	220	144	237	149	207	11	0	1072
Detected Mallards		**Province**									**Total**
		EC	FS	G	KZN	L	M	NW	NC	WC	
Year	2010	3	8	27	9	2	10	10	2	38	109
	2011	8	8	33	5	1	7	3	2	41	108
	2012	8	6	27	7	1	4	7	0	41	101
	2013	6	5	29	3	0	5	9	3	40	100
	2014	4	4	33	6	0	6	13	2	36	104
	2015	8	5	43	5	1	3	13	0	36	114
	2016	6	5	36	4	1	3	14	0	34	103
	2017	5	4	34	1	1	6	16	0	30	97
	2018	7	4	34	1	0	3	12	0	28	89
	2019	7	1	35	3	2	5	11	1	27	92

Table 4.2: Exploration of surveyed pentads and pentads with at least one observation by province

Once the study regions were finalised, a naive evaluation of the yearly site-occupancy per species was conducted to give some surface-level insight into the occupancy rates of these species and is provided by Table 4.3.

Year	Myna		Mallard	
	Total	At least one detection(%)	Total	At least one detection(%)
2010	3331	1211 (36.36%)	3725	102(2.74%)
2011	3100	1288 (41.55%)	3461	97 (2.80%)
2012	2963	1205 (40.67%)	3293	92 (2.79%)
2013	2663	1131 (42.47%)	2750	91 (3.31%)
2014	2833	1253 (44.23%)	2862	98 (3.42%)
2015	2757	1180 (42.80%)	2953	105(3.56%)
2016	2571	1096 (42.63%)	2966	96 (3.24%)
2017	2539	1069 (42.10%)	2864	91 (3.18%)
2018	2534	1114 (43.96%)	2614	82 (3.14%)
2019	2392	1054 (44.06%)	2594	82 (3.16%)

Table 4.3: Annually surveyed pentads per species

This table provides insight into the structure of the data for each species, in terms of the total number of sites surveyed annually as well as the yearly proportion of sites with at least one detection. It is evident that the number of sites surveyed varies by year and there is no indication as to whether these surveyed sites were the same in each season.

This imbalance in the data, caused by differing levels of sampling effort per site in conjunction with a lack of a rigid study design by not sampling the same sites every season, will not have an effect on a study pertaining to just one season. However, these differences (in terms of both the number of sites sampled each season as well as the sampling effort per site) will cause problems in the dynamic models if the data structure is not transformed into a sensible study design (MacKenzie et al., 2017). Thus, the effect on the results caused by changes to the extent of site overlap and the maximum number of surveys in a season are explored later in this chapter.

Table 4.3 allows for a simple estimation of the species' respective occupancy rates, by dividing the number of sites at which at least one detection occurred by the total number of sites surveyed in the corresponding year. These estimates suggest increasing and fairly stable changes in the overall occupancy of the Mallard and also suggest that occupancy has been increasing for the Myna since 2010, but it is hard to determine if these changes are just stochastic variability. However, this type of occupancy estimate is naive and inadequate, since it would provide no inference as to how occupancy varies within the study region, ignores imperfect detection, nor would it provide any information relating

to the factors affecting occupancy or the detectability of the species.

4.1.2 Exploration of the environmental covariates

The environmental covariates that were to be included in the model fitting process were first investigated for potential multicollinearity through both an investigation of correlation plots as well as a calculation of variance inflation factors (VIF) with a threshold value set to six. Figure 4.1 provides the correlations for the set of measured covariates.

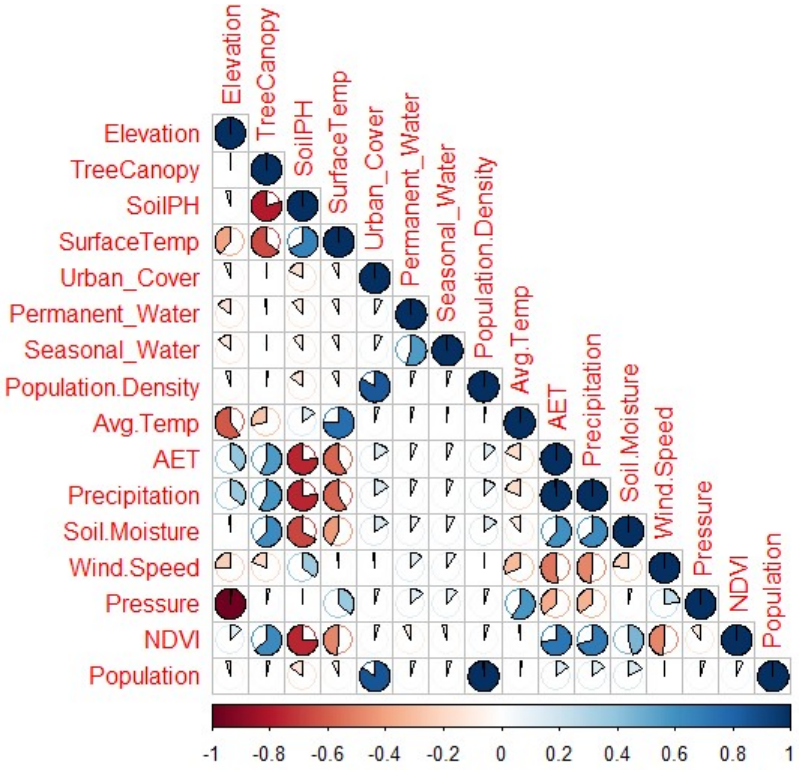
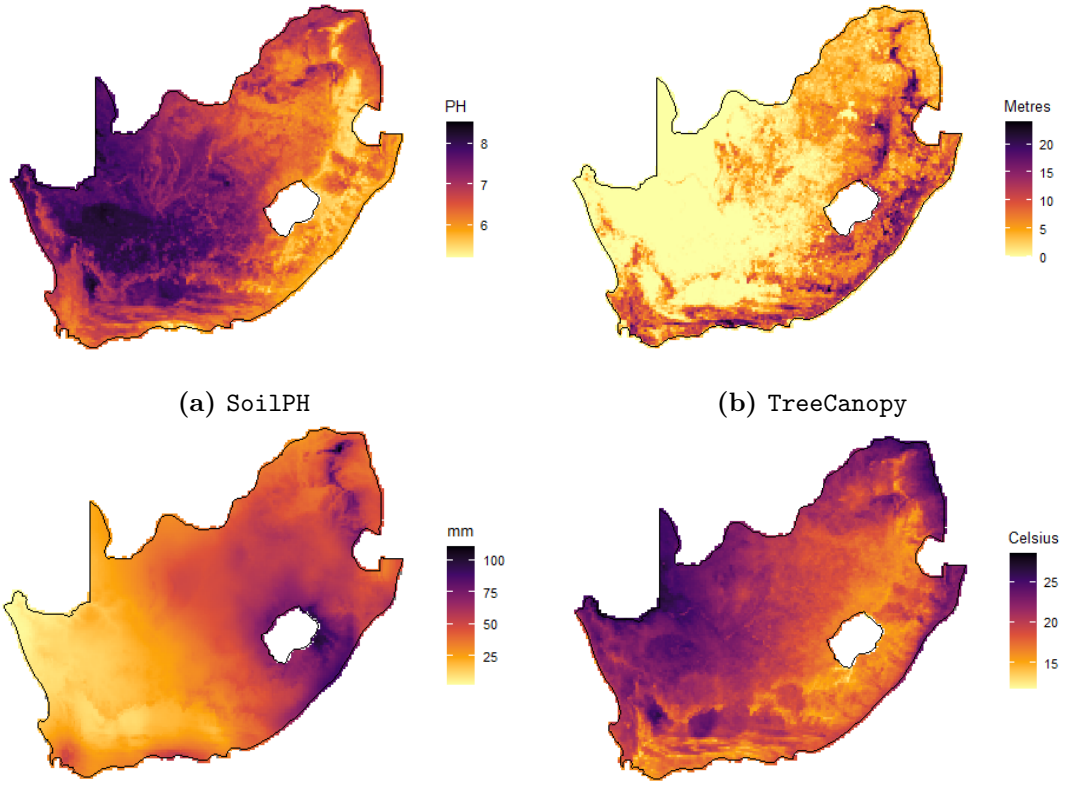


Figure 4.1: Correlation plot for the 2010 environmental covariates for which the filled proportion of the pie charts denote the absolute correlation coefficient between 0 and 1.

This graphic depicts a few highly correlated variables. Those covariates displaying an almost one-to-one correlation include: `Pressure` and `Elevation`; `Population.Density` and `Urban_Cover`; `Population` and `Population.Density`; in addition to `Precipitation` and `AET`. An evaluation of these correlations together with the VIF scores, using a threshold value of six, for the covariates corresponding to the years 2010, 2013, 2016 and 2019 suggested that the `AET` and `Urban_Cover` variables were to be excluded from the study since they had VIF scores exceeding the specified threshold. `Seasonal_Water` was also excluded from the study since it was more intuitive to use the `Permanent_Water` covariate

to represent bodies of water when the study was conducted at an annual rather than seasonal scale.

`Elevation` displayed VIF scores exceeding six and displayed a high negative correlation to `Pressure`, however, it was still included in the set of potential covariates. This is because `Elevation` was a factor representing the natural environment of the site, while `Pressure` was a factor representing the climate of a site. They were thus used in the formulation of separate hypotheses for the occupancy models, but were never used in the same model. Additionally, the `Population.Density` covariate was fitted to the static occupancy models, since it was only recorded for the year 2010, while the `Population` covariate was used for the dynamic occupancy models. These covariates both acted as proxy variables representing anthropogenic factors but were implemented in separate analyses to avoid problems with multicollinearity. The correlations of all measured covariates did not significantly differ between seasons, and the correlation plot corresponding to the final year of the study is provided in the Appendices by Figure B.2 to substantiate this point.



(a) `SoilPH` **(b)** `TreeCanopy`

(c) `Precipitation` **(d)** `SurfaceTemp`

Figure 4.2: Mapped environmental covariates (2010)

After the collinear predictors were removed, there were 13 potential covariates that were included at some stage of the static or dynamic occupancy modelling process. Figure 4.2 provides four of these covariates from the candidate pool (`SoilPH`, `TreeCanopy`, `Precipitation` and `SurfaceTemp`) for which the inconsistent measurement scales would indicate different contributions to the estimated parameters if not transformed to a comparable scale, and highlights the importance of standardising the predictors before conducting any occupancy analysis. The remaining predictors not shown by this graphic similarly depict contrasts in both spatial intensities and the measurement scales and an additional six covariates are provided in the Appendices by Figure B.1 to support this statement.

4.2 Static occupancy models

4.2.1 Myna

Estimating detection

The results of the purely data-driven approach to modelling the covariates with the potential to affect the Myna's detectability, while no measured covariates are used for the estimation of the occupancy probability, are tabulated and provided by Table 4.4 for which the coefficients are given on the logit scale. Inference of the estimate of occupancy at this stage is irrelevant but instructive points regarding the covariates can be made.

Parameter	**Estimate**	**SE**	**z**	**P(>|z|)**	**p-value**
Occupancy (Ψ)					
Intercept	0.31	0.06	5.45	5e-08	0.00
Detection (p)					
Intercept	0.01	0.07	0.18	8.57e-01	0.86
lhours	0.11	0.04	2.57	1.00e-02	0.01
lhours^2	-0.08	0.02	-3.16	1.57e-03	0.00
ltotspp	0.54	0.04	15.38	2.14e-53	0.00
ltotspp^2	0.05	0.02	3.64	2.73e-04	0.00
SeasonSpring	0.38	0.08	4.98	6.20e-07	0.00
SeasonSummer	0.11	0.08	1.45	1.47e-01	0.15
SeasonWinter	0.26	0.08	3.24	1.21e-03	0.00
Population.Density	0.40	0.02	23.63	1.91e-123	0.00
NDVI	-0.12	0.03	-3.66	2.52e-04	0.00

Table 4.4: Factors driving the Myna's detectability

The modelling process found all the potential covariates as significant for the estimation of the detection probability. The relationships were mostly linear, but included quadratic effects for both `lhours` and `ltotspp`. The inference pertaining to the `lhours` covariates suggests that an increase in hours spent actively bird watching increases the likelihood of observing the Myna but at a decreasing rate, while the estimates corresponding to the `ltotspp` covariates suggest that an increase in the number of other species detected at a given site during a sampling survey will increase the probability of detecting the Myna at an increasing rate.

The `Season` factor variable set Autumn as the reference category, and found that surveys conducted in the Spring and Winter seasons resulted in significantly different detection probabilities compared to Autumn, while there was little difference between detections in the Summer and Autumn seasons. In short, the `Season` variable suggests that the detectability of the Myna differs based on the season during which a survey took place.

The `Population.Density` covariate acted as a proxy for urban environments and was highly influential on the detectability of the Myna. The estimate suggests that an increase in the population density at a site will increase the likelihood of detecting the Myna. This finding supports the prior belief that an increase in the population density leads to a larger proportion of potential birders, and thus a higher likelihood of species detection.

The final result pertaining to the detectability of the Myna relates to the `NDVI` covariate. This covariate is an index of greenness where high values (1) represent areas with high natural obscurity and low values (-1) represents bodies of water. The inference made from this covariate indicates that the likelihood of detecting the Myna will decrease as the NDVI value increases. This estimate is reasonable from an intuitive outlook since a higher greenness index could potentially act as a proxy for thicker shrubbery which leads to higher obscurity and lower visibility in an area.

Model selection procedure

The hypothesis-driven approach to modelling the occupancy of the Myna included eight hypothesized models that utilised a combination of anthropogenic, environmental and climatic factors. The model selection process considered all fitted models, however only the top five candidate models are illustrated by Table 4.5.

The selection of the best fitting model was based on the AIC statistic, in conjunction with the AIC weights, and found that the hypothesized model which best described the occupancy of the Myna was `Urban.Clim.Env`. The best fitting hypothesized model included linear and quadratic terms for: `Pressure`; `SoilPH`; `SurfaceTemp`; `Precipitation`;

`Wind.Speed` and `TreeCanopy`, in addition to a linear term for the `Population.Density` covariate.

Hypothesis	K	AIC	ΔAIC	AICWt	Log-Likelihood
Urban.Clim.Env	24	10026.20	0.00	1.00	-4989.10
Urban.Clim.Windy	20	10178.31	152.12	0.00	-5069.16
Urban.Clim.NoWind	19	10253.69	227.50	0.00	-5108.85
Urban	12	10477.83	451.63	0.00	-5226.91
Climate	19	10623.72	597.53	0.00	-5292.86

Table 4.5: Static model selection procedure: Myna

Model estimates

The estimates of the model which best describes the occupancy and detection probability of the Myna is provided by Table 4.6. This hypothesized model included a few insignificant terms, namely the quadratic term of `Pressure`, both terms relating to `Precipitation`, the linear term of `Wind.Speed` as well as both terms relating to `TreeCanopy`. The results of this table indicate that the estimated occupancy parameter is significantly influenced by the linear effects of the `Pressure` and `Population.Density` covariates, as well as the nonlinear effects of the `SoilPH`, `SurfaceTemp` and `Wind.Speed` covariates.

Parameter	Estimate	SE	z	P(> |z|)	p-value
Occupancy (ψ)					
Intercept	1.10	0.28	3.97	7.10e-05	0.00
Pressure	-0.82	0.18	-4.72	2.35e-06	0.00
Pressure^2	0.11	0.08	1.47	1.42e-01	0.14
SoilPH	-2.42	0.24	-10.04	1.04e-23	0.00
SoilPH^2	-0.96	0.14	-6.75	1.52e-11	0.00
SurfaceTemp	0.90	0.29	3.13	1.72e-03	0.00
SurfaceTemp^2	-0.49	0.14	-3.63	2.83e-04	0.00
Precipitation	-0.06	0.38	-0.16	8.75e-01	0.88
Precipitation^2	-0.18	0.13	-1.39	1.66e-01	0.17
Wind.Speed	-0.38	0.27	-1.43	1.52e-01	0.15
Wind.Speed^2	-0.76	0.16	-4.68	2.88e-06	0.00
TreeCanopy	0.31	0.20	1.54	1.24e-01	0.12
TreeCanopy^2	-0.12	0.08	-1.63	1.03e-01	0.10
Population.Density	9.50	1.05	9.06	1.30e-19	0.00

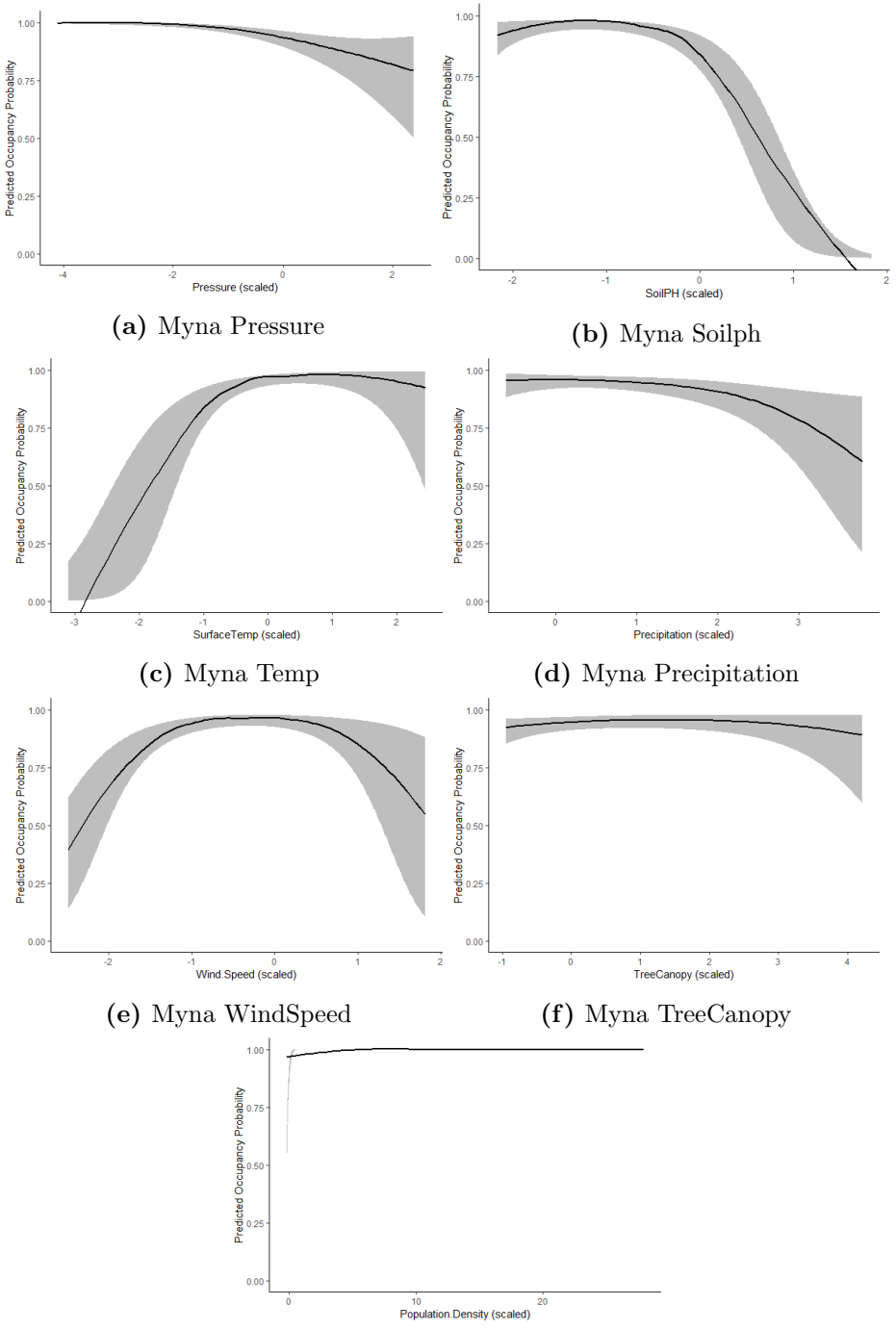
Detection (p)					
Intercept	0.09	0.07	1.38	1.68e-01	0.17
lhours	0.07	0.04	1.65	9.93e-02	0.10
lhours^2	-0.05	0.02	-2.29	2.22e-02	0.02
ltotspp	0.53	0.03	15.23	2.35e-52	0.00
ltotspp^2	0.05	0.01	3.20	1.39e-03	0.00
SeasonSpring	0.40	0.08	5.22	1.79e-07	0.00
SeasonSummer	0.14	0.08	1.80	7.23e-02	0.07
SeasonWinter	0.26	0.08	3.31	9.38e-04	0.00
NDVI	-0.15	0.03	-4.45	8.67e-06	0.00
Population.Density	0.36	0.02	22.87	8.70e-116	0.00

Table 4.6: Static occupancy model estimates: Myna

The majority of these estimated coefficients are reasonable on the logit scale, however, the coefficient corresponding to the `Population.Density` term used to estimate the occupancy probability is notably large. This coefficient would suggest that the Myna's occupancy probability is essentially equal to one at a site if there is a value for this covariate larger than zero, regardless of the scale. This possibly represents a situation of complete separation, which occurs commonly in logistic regression, where the continuous predictor variable separates the dichotomous response completely (Mansournia et al., 2018). This coefficient depicts a comparatively large standard error and is likely to be inflated due to the selection of the study region, yet, the result supports the literature which would suggest that Mynas are highly adaptable to urban environments. Thus, although this coefficient is questionably high (Szumilas, 2010), it is still a significant result which follows from what was expected based on the literature. This covariate indicates that, within the study region, any measure of `Population.Density` will result in an almost certainty that the Myna occupies the site.

A set of graphics that illustrate the fitted relationships between these covariates and the probability of occupancy is provided by Figure 4.3. Albeit conservative inference must be made for those insignificant fitted covariates, these relationships indicate that an increase in the measured `SurfaceTemp` and `Population.Density` covariates lead to increases (of varying degrees) in the likelihood of Myna occupancy, while an increase in the measured `Precipitation` covariate results in a potential decrease in this occupancy probability. Further, an increase in the measured `SoilPH`, `Wind.Speed` and `TreeCanopy` covariates will initially increase the probability of occupancy until a point of inflection, which occurs at the means of both the `SoilPH` and `Wind.Speed` covariates and between 1 to 2 standard

deviations above the mean of the `TreeCanopy` covariate respectively, beyond which an increase in the measured covariate leads to a decrease in the Myna's occupancy probability.



(a) Myna Pressure

(b) Myna Soilph

(c) Myna Temp

(d) Myna Precipitation

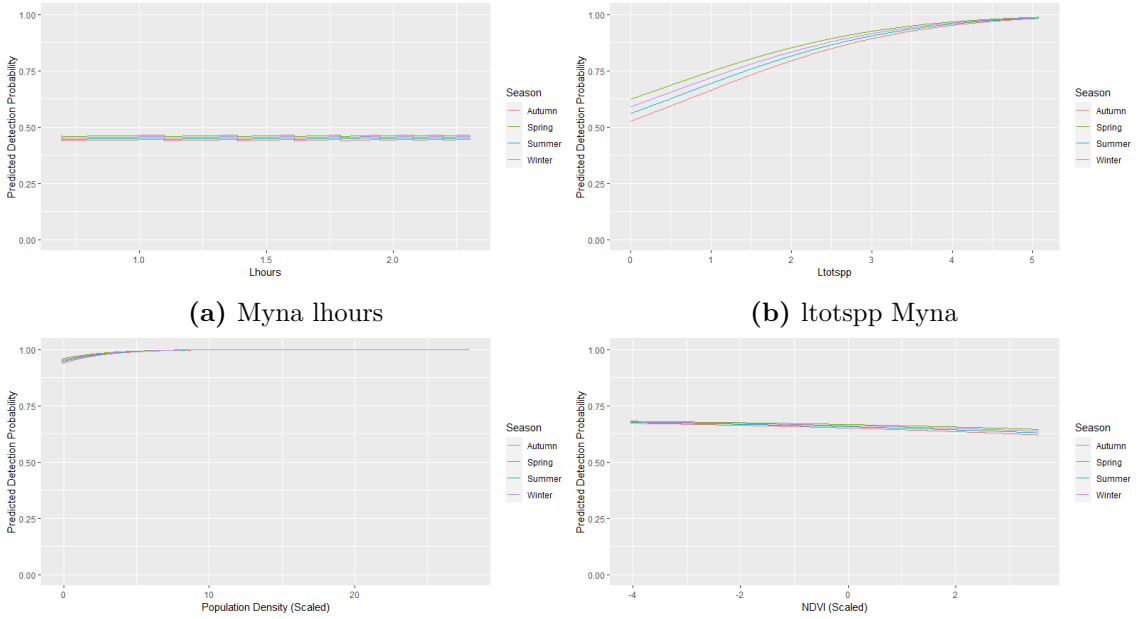
(e) Myna WindSpeed

(f) Myna TreeCanopy

(g) Myna PopDens

Figure 4.3: Plot of static occupancy fitted relationships: Myna

The covariates used to estimate the detection probability portray the same signs and similar inference as that described using the results of Table 4.4. A key difference is that the `lhours` linear term is now statistically insignificant, although its quadratic term is still significant, which is reflected by the horizontal fitted relationship observed between the `lhours` covariate and the estimated probability of detection. Additionally, these fitted relationships are congruent to the estimated detection probabilities of Table 4.6 and show that there is no observable difference between detectability and `Season`.



(a) Myna lhours **(b)** ltotspp Myna

(c) Myna PopDens **(d)** Myna NDVI

Figure 4.4: Plot of static detection fitted relationships: Myna

The equations to formally define the static occupancy model in 2010 for the Myna are expressed in terms of the estimated occupancy and detection parameters as follows:

$$\begin{aligned}
\text{logit}\left(\psi_i\right) = 1.10 &- 0.82\left(\texttt{Pressure}_i\right) + 0.11\left(\texttt{Pressure}_i^2\right) \\
&- 2.42\left(\texttt{SoilPH}_i\right) - 0.96\left(\texttt{SoilPH}_i^2\right) \\
&+ 0.90\left(\texttt{SurfaceTemp}_i\right) - 0.49\left(\texttt{SurfaceTemp}_i^2\right) \\
&- 0.06\left(\texttt{Precipitation}_i\right) - 0.18\left(\texttt{Precipitation}_i^2\right) \\
&- 0.38\left(\texttt{Wind.Speed}_i\right) - 0.76\left(\texttt{Wind.Speed}_i^2\right) \\
&+ 0.31\left(\texttt{TreeCanopy}_i\right) - 0.12\left(\texttt{TreeCanopy}_i^2\right) \\
&+ 9.50\left(\texttt{Population.Density}_i\right)
\end{aligned} \tag{4.1}$$

$$\begin{aligned}\text{logit}\left(p_{ij}\right) = 0.09 &+ 0.07\left(\texttt{lhours}_{ij}\right) - 0.05\left(\texttt{lhours}_{ij}^2\right) \\ &+ 0.53\left(\texttt{ltotspp}_{ij}\right) + 0.05\left(\texttt{ltotspp}_{ij}^2\right) \\ &+ 0.40\left(\texttt{SeasonSpring}_{ij}\right) + 0.14\left(\texttt{SeasonSummer}_{ij}\right) \\ &+ 0.26\left(\texttt{SeasonWinter}_{ij}\right) - 0.15\left(\texttt{NDVI}_{i}\right) \\ &+ 0.36\left(\texttt{Population.Density}_{i}\right)\end{aligned} \tag{4.2}$$

$$\begin{aligned}\text{for } i &= 1, 2, \ldots, 3331 \\ j &= 1, 2, \ldots, 146.\end{aligned}$$

Assessment of model fit

The evaluation of the static occupancy model's goodness-of-fit was assessed via the plotting of the observed and predicted occupancy maps. These graphics are provided by Figure 4.5 for which the top image illustrates the observed occupancy, which was extracted using the results of the presence/absence data based on the 2010 surveys, where the orange crosses denote no detections and the red crosses denote at least one detection of the Myna at a particular site. The predicted occupancy for the Myna is provided by the lower image, derived from the estimates of the best fitting model (Table 4.6), where navy blue denotes low predicted occupancy while light blue denotes high predicted occupancy.

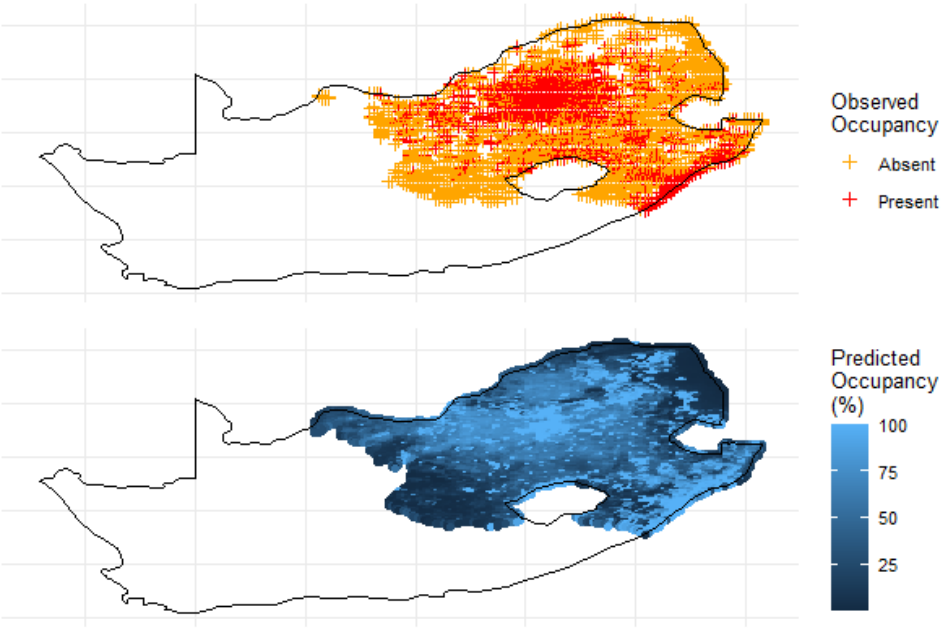


Figure 4.5: Predicted and observed occupancy maps: Myna

This graphic illustrates that the chosen model provides an adequate fit to the data since detections of the Myna are superimposed on areas with high predicted occupancies. Similarly, areas with no detections are generally superimposed on low predicted occupancies.

An important aspect of this map that is worth discussion is the effect on the predicted occupancies due to the large `Population.Density` coefficient. The isolated areas with high predictions of occupancy, that are surrounded by low occupancy predictions, are the result of this large coefficient. These light blue points represent areas that are unlikely to provide natural habitats suitable for the Myna but the presence of people in these areas, and by proxy urban environments with houses and buildings in which the Myna may create dwellings, result in a high predicted occupancy of the Myna.

However, it is also possible that these sites do provide a suitable habitat for the Myna to occupy, and these isolated areas with high predictions of occupancy may simply not be visited in the sampling period or depict instances of false negatives in the cases where they are surveyed at least once during the year. Despite this somewhat undesirable trait, it is reasonable to conclude that the model fitted to represent the occupancy of the Myna provides a decent description of the data at this stage of the analysis.

4.2.2 Mallard

Estimating detection

The same data-driven approach, that was applied to determine the significant covariates affecting the Myna's detection probability, was applied to ascertain the factors affecting the Mallard. This approach was identical in the sense that no covariates were fitted to model the occupancy parameter and the same set of candidate variables were included to model the detection probability on the logit scale, however, the findings were not the same and are presented by Table 4.7.

The modelling process found the linear term of `lhours` to be insignificant in driving the detection probability of the Mallard. However, since the quadratic term is significant, the linear `lhours` term is not removed from the model. The significant and positive `ltotspp` linear term indicates that the probability of detecting the Mallard increases if there is an increased number of other species observed at the surveyed site. The quadratic term is insignificant and was removed before conducting the hypothesis-driven approach to modelling occupancy. Therefore, it can be concluded that the relationship between `ltotspp` and the Mallard is a linear one.

Parameter	Estimate	SE	z	P($>|z|$)	p-value
Occupancy process					
Intercept	-1.97	0.13	-14.8	9.28e-50	0.00
Detection process					
Intercept	-2.18	0.16	-13.46	2.67e-41	0.00
lhours	0.11	0.09	1.26	2.08e-01	0.21
lhours^2	0.14	0.05	3.11	1.90e-03	0.00
ltotspp	0.44	0.08	5.32	1.02e-07	0.00
totspp^2	-0.01	0.05	-0.12	9.07e-01	0.91
SeasonSpring	-0.15	0.16	-0.99	3.59e-01	0.36
SeasonSummer	-0.42	0.17	-2.44	1.46e-02	0.02
SeasonWinter	0.26	0.16	1.63	1.03e-01	0.10
Population.Density	0.05	0.05	1.14	1.27e-01	0.13
NDVI	-0.19	0.08	-2.40	1.66e-02	0.02

Table 4.7: Factors driving the Mallard's detectability

The `Season` factor variable would suggest that there is a significant difference in the detectability of the Mallard between the Autumn and Summer seasons, albeit there is no difference in the Mallard's detectability between Autumn and Winter or Autumn and Spring. Since there is still some significance between the levels of this factor variable, it was not removed from the model prior to the modelling of the occupancy parameter.

Although the `Population.Density` covariate would indicate that an increase in the population density at a site potentially increases the likelihood of detecting the Mallard, it was removed from the model due to its statistical insignificance. However the `NDVI` estimate was found to be significant and suggests that an increase in the greenness of an area leads to a decrease in the detectability of the Mallard.

Model selection procedure

The hypothesis-driven approach to modelling the occupancy of the Mallard included a set of 10 hypothesized models each representing a combination of anthropogenic, climatic and environmental factors. The top five candidate models selected using the AIC statistic, together with the AIC weights, are provided by Table 4.8.

Hypothesis	K	AIC	ΔAIC	AICWt	Log-Likelihood
Urban.Water.Env.Clim	17	2604.96	0.00	0.77	-1285.48
Urban.Climate	18	2608.38	3.41	0.14	-1286.19
Urban.Env.Clim	18	2609.29	4.33	0.09	-1286.64
Urban.Env	17	2647.21	42.25	0.00	-1306.60
Urban	10	2669.69	64.73	0.00	-1324.85

Table 4.8: Static model selection procedure: Mallard

This table indicates that the difference between the AIC scores were smaller and the AIC weights were more widely dispersed between the top five hypothesized models in comparison to the best models selected for the Myna. However, the model reflecting the `Urban.Water.Env.Clim` hypothesis was selected as the best fitting since it has the lowest AIC statistic, highest AIC weighting, and the fewest parameters. This hypothesis included linear and quadratic terms for `SoilPH`, `SurfaceTemp` and `Precipitation`, and linear terms for the `Permanent_Water` and `Population.Density` covariates.

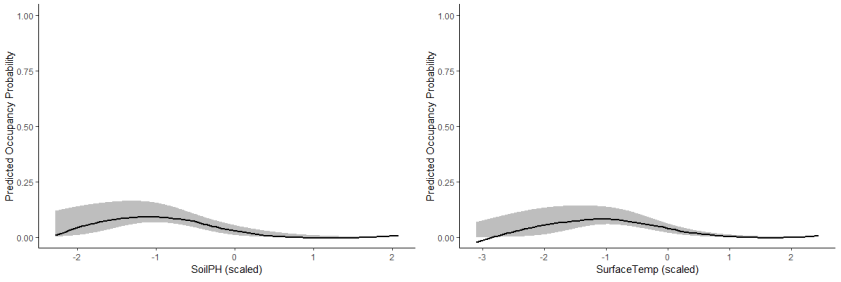
Model estimates

The estimates for the model that best describes the occupancy and detection probabilities of the Mallard is provided by Table 4.9. The fitted covariates are all statistically significant barring the linear term corresponding to the `lhours` covariate and the quadratic term relating to `Precipitation`. The results of this table suggest that the Mallard's estimated occupancy is influenced by the nonlinear effects of `SoilPH` and `SurfaceTemp` as well as the linear effects of the `Precipitation`, `Permanent_Water` and `Population.Density` covariates.

Parameter	Estimate	SE	z	P($>\|z\|$)	p-value
Occupancy (ψ)					
Intercept	-4.20	0.46	-9.23	2.59e-20	0.00
SoilPH	-2.97	0.88	-3.38	7.28e-04	0.00
SoilPH^2	-1.36	0.46	-2.96	3.05e-03	0.00
SurfaceTemp	-1.92	0.42	-4.59	4.35e-06	0.00
SurfaceTemp^2	-0.91	0.28	-3.32	9.02e-04	0.00
Precipitation	-0.71	0.24	-2.91	3.63e-03	0.00
Precipitation^2	0.15	0.13	1.16	2.46e-01	0.25
Permanent_Water	0.14	0.06	2.40	1.63e-02	0.02
Population.Density	0.45	0.06	7.68	1.60e-14	0.00

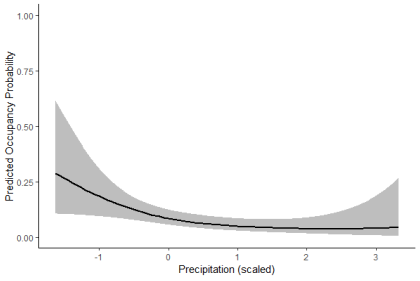
Detection (p)					
Intercept	-1.82	0.13	-14.19	1.00e-45	0.00
lhours	0.04	0.09	0.48	6.30e-01	0.63
lhours^2	0.17	0.05	3.64	2.68e-04	0.00
ltotspp	0.45	0.08	5.67	1.44e-08	0.00
SeasonSpring	-0.18	0.16	-1.14	2.54e-01	0.25
SeasonSummer	-0.45	0.17	-2.65	8.08e-03	0.00
SeasonWinter	0.22	0.16	1.40	1.61e-01	0.16
NDVI	-0.20	0.08	-2.63	8.58e-03	0.00

Table 4.9: Static occupancy model estimates: Mallard

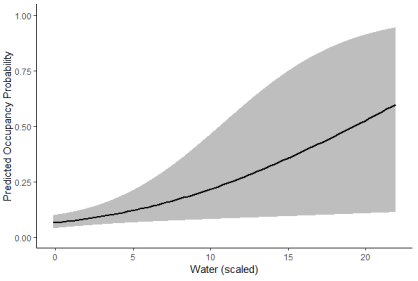


(a) Mallard SoilPH

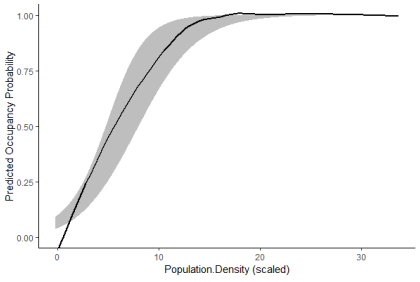
(b) Mallard SurfaceTemp



(c) Mallard Precipitation



(d) Mallard Water

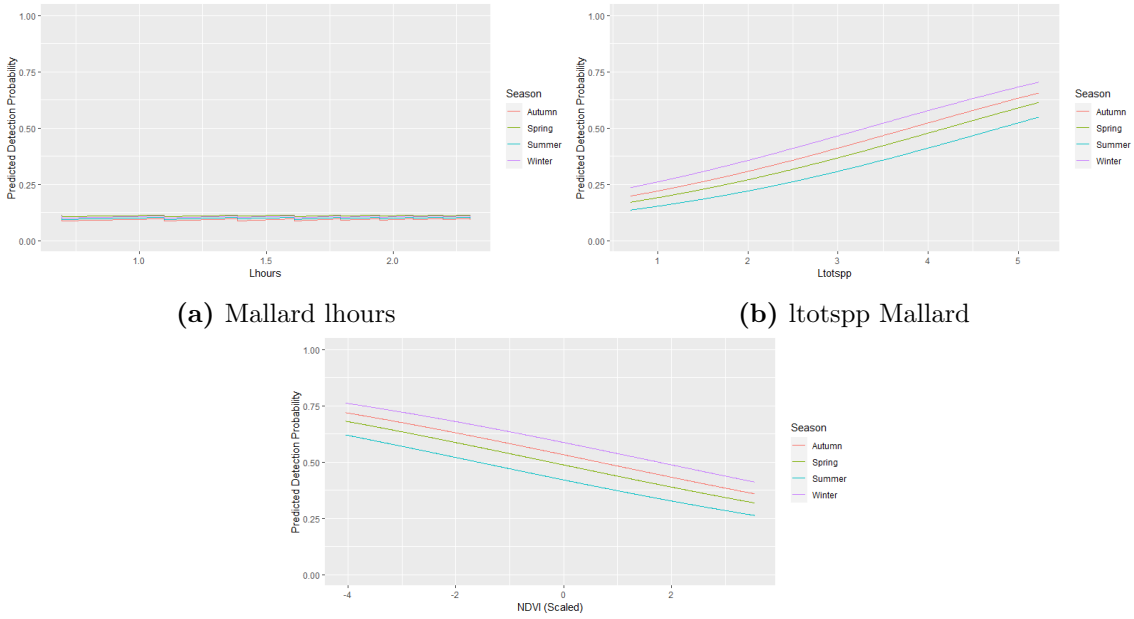


(e) Mallard PopDens

Figure 4.6: Plot of static occupancy fitted relationships: Mallard

These estimates (in conjunction with the fitted relationships provided by Figure 4.6) are indicative that an increase in either `Permanent_Water` or `Population.Density` will result in increases of varying degrees to the occupancy probability of the Mallard while an increase in `Precipitation` will lead to a decrease in the probability of Mallard occupancy. Further, considering the significant quadratic terms, it is apparent that an increase in either `SurfaceTemp` or `SoilPH` will initially result in increases in the occupancy probability, until a point beyond which an increase in either of these covariates will decrease the Mallard's occupancy probability.

An important aspect of this model worth discussion is that the observed coefficient for the `Population.Density` covariate is not as extreme as that observed in Table 4.6. Since the estimate is considered reasonable (Szumilas, 2010) for the Mallard's model but unusually large for the Myna's model, this would suggest that nearly all urban areas within the Myna's study region were occupied by the Myna and resulted in a case of complete separation (Mansournia et al., 2018). Therefore, if the Western Cape or Eastern Cape were included in the Myna's study region, hence introducing metropolitan areas in the form of Cape Town and Gqeberha where there are little to no Myna sightings, then it is expected that the `Population.Density` coefficient would shrink drastically.



(a) Mallard lhours **(b)** ltotspp Mallard

(c) Mallard NDVI

Figure 4.7: Plot of static detection fitted relationships: Mallard

The detection covariates in the full model present the same signs and result in the same inference as drawn from Table 4.7. These graphics support the findings that: `lhours` has no influence in the estimated detectability of the Mallard; the relationship between the Mallard's detectability and `ltotspp` is positive; while the detectability of the Mallard will decrease as the average greenness (`NDVI`) of a given pentad increases.

With these results in mind, the equations to formally define the Mallard's static occupancy model for 2010 are expressed in terms of the occupancy and detection probabilities as:

$$\begin{aligned}\text{logit}\left(\psi_i\right) = -4.20 &- 2.97\left(\texttt{SoilPH}_i\right) - 1.36\left(\texttt{SoilPH}_i^2\right) \\ &- 1.92\left(\texttt{SurfaceTemp}_i\right) - 0.91\left(\texttt{SurfaceTemp}_i^2\right) \\ &- 0.71\left(\texttt{Precipitation}_i\right) + 0.15\left(\texttt{Precipitation}_i^2\right) \\ &+ 0.14\left(\texttt{Permanent\_Water}_i\right) + 0.45\left(\texttt{Population.Density}_i\right)\end{aligned} \tag{4.3}$$

$$\begin{aligned}\text{logit}\left(p_{ij}\right) = -1.82 &+ 0.04\left(\texttt{lhours}_{ij}\right) + 0.17\left(\texttt{lhours}_{ij}^2\right) \\ &+ 0.45\left(\texttt{ltotspp}_{ij}\right) - 0.18\left(\texttt{SeasonSpring}_{ij}\right) \\ &- 0.45\left(\texttt{SeasonSummer}_{ij}\right) + 0.22\left(\texttt{SeasonWinter}_{ij}\right) \\ &- 0.20\left(\texttt{NDVI}_i\right) \\ \text{for } i &= 1, 2, \ldots, 3725 \\ j &= 1, 2, \ldots, 146.\end{aligned} \tag{4.4}$$

Assessment of model fit

The assessment of the model goodness-of-fit through the evaluation of the observed and predicted occupancy maps for the Mallard in 2010 is provided by Figure 4.8. This graphic includes the same features as that described for the plotted map of the Myna's occupancy (Figure 4.5) and suggests that the chosen model provides an adequate fit to the data since detections of the Mallard are at the pentads corresponding to high predicted occupancies. Similarly, areas with no detections correspond to areas with low predicted occupancies.

These maps would suggest that the Mallard is an infrequently observed species and the only site-level covariates, from the pool of potential variables included in the analysis, that seem to have an influence on occupancy are `Permanent_Water` and `Population.Density`. This implies one of two things: either the Mallard's occupancy is only positively influenced by these two covariates; or there is one or more key variables missing from the model that would be significant in predicting their occupancy. Nonetheless, the results of these maps provide evidence that the fitted static occupancy model is able to adequately describe the

data in terms of both the occupancy and detection parameters of the Mallard.

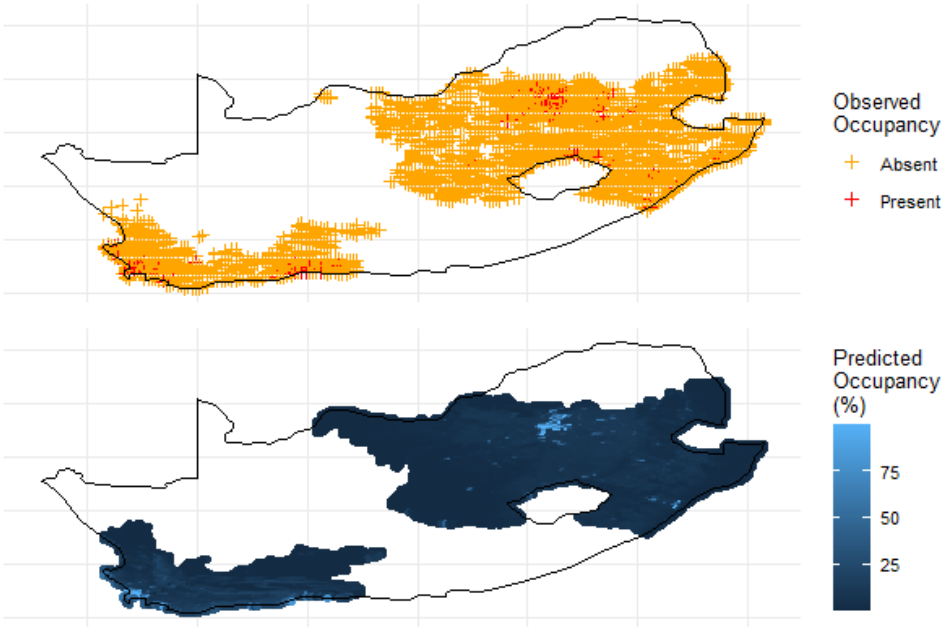


Figure 4.8: Predicted and observed occupancy map: Mallard

4.3 Dynamic occupancy models

4.3.1 Myna

Model selection procedure

The dynamic occupancy modelling of the Myna made use of the results extracted from the static occupancy modelling process, via the covariates fitted to describe occupancy and detectability in the best fitting static model (Table 4.6), to model the occupancy and detection components of the dynamic models. The only noticeable difference in these covariates is that the static measurements for the `Population` and `NDVI` variables were substituted for their yearly counterparts.

The hypothesis-driven approach to modelling the vital rate parameters included a candidate set of seven formulated hypotheses that were symmetric in the sense that a covariate was used to estimate both the colonization and extinction parameters in the same dynamic model. These hypotheses were fitted to the six different data structures and the five best fitting models for each structure, based on the AIC statistic and the AIC weights, are provided by Table 4.10.

Structure	Hypothesis	K	AIC	ΔAIC	AICWt
Common10 (Common50)	Env.Clim.Urban (Env.Clim.Urban)	40	38149.66 (73107.00)	0.00 (0.00)	1 (1)
	Clim.Urban.Wind (Clim.Urban.Wind)	36	38174.07 (73132.68)	24.41 (25.68)	0.00 (0.00)
	Clim.Urban (Clim.Urban)	34	38288.31 (73231.83)	138.65 (124.83)	0.00 (0.00)
	Humid.Urban (Humid.Urban)	30	38302.28 (73247.81)	152.62 (140.81)	0.00 (0.00)
	Temp.Rain.Urban (Temp.Rain.Urban)	32	38327.12 (73272.52)	177.46 (165.52)	0.00 (0.00)
Intermediate10 (Intermediate50)	Env.Clim.Urban (Env.Clim.Urban)	40	57985.38 (94861.05)	0.00 (0.00)	1 (1)
	Clim.Urban.Wind (Clim.Urban.Wind)	36	58049.45 (94915.91)	64.08 (54.86)	0.00 (0.00)
	Clim.Urban (Clim.Urban)	34	58147.74 (95062.46)	162.37 (201.42)	0.00 (0.00)
	Humid.Urban (Humid.Urban)	30	58256.17 (95167.08)	270.79 (306.03)	0.00 (0.00)
	Temp.Rain.Urban (Temp.Rain.Urban)	32	58316.31 (95226.60)	330.93 (365.55)	0.00 (0.00)
All10 (All50)	Env.Clim.Urban (Env.Clim.Urban)	40	72811.30 (110430.60)	0.00 (0.00)	1 (1)
	Clim.Urban.Wind (Clim.Urban.Wind)	36	72899.43 (110520.50)	88.14 (89.91)	0.00 (0.00)
	Clim.Urban (Clim.Urban)	34	72979.98 (110570.40)	168.68 (139.75)	0.00 (0.00)
	Humid.Urban (Humid.Urban)	30	73198.62 (110814.8)	387.32 (384.20)	0.00 (0.00)
	Temp.Rain.Urban (Temp.Rain.Urban)	32	73237.00 (110858.1)	425.70 (427.48)	0.00 (0.00)

Table 4.10: Dynamic model selection procedure: Myna

These results suggest that, from the set of hypothesized models, the best fitting model (`Env.Clim.Urban`) does not change if the structure of the data is altered in terms of the total number of visited sites. Similarly, for the structures with the same number of visited

sites, an increase in the maximum length of the detection history from 10 to 50 does not affect the model selection outcome.

Assessment of model fit

The goodness-of-fit of the `Env.Clim.Urban` model was then evaluated via the chi-square statistic described by Table 2.2, using each of the six data structures, for which the summarized results are provided below by Table 4.11. The mean bootstrapped chi-square statistic ($\tilde{\chi}^2_{\mathbf{B}}$) was calculated for each structure using 2000 simulations (**nsim**) with run times varying from 27 to 330 hours. The All50 structure was only bootstrapped for 100 simulations, since the computational burden to run this test was extremely high, and extrapolating the run time of the All10 structure would indicate that it would take significantly longer than 330 hours to complete the All50 test. The results pertaining to the All50 structure are simply provided to indicate that from no more than 100 simulations the test starts to indicate that the model does not provide a good fit to the data.

The null hypothesis corresponding to the chi-square statistic (the model provides an adequate fit to the data) is rejected for every data structure when implementing a significance level of 5%, however, the overdispersion estimates for the three data structures with seasonal detection histories of max length 10 per site are not unreasonable as they are estimated to be 2.29, 2.26 and 2.29, respectively. The estimates associated with seasonal detection histories of maximum length 50 are significantly higher and show that these best fitting models do not provide a good fit to the observed data.

Structure	$\chi^2_{\mathbf{O}}$	$\tilde{\chi}^2_{\mathbf{B}}$	$\hat{\mathbf{c}}$	**p-value**	**nsim**	**Run time (hours)**
Common10	46557.35	20300.92	2.29	0.01	2000	27
Intermediate10	45924.63	20282.86	2.26	0.01	2000	45
All10	46452.02	20267.29	2.29	0.02	2000	330
Common50	2.66e+18	1.68e+16	158.26	0.00	2000	48
Intermediate50	1.09e+18	1.80e+16	60.49	0.00	2000	123
All50	6.96e+17	1.66e+16	42	0.00	100	68

Table 4.11: Dynamic model goodness-of-fit: Myna

The results of Table 4.11 suggest that, if the length of the detection history is increased, the same hypothesized model is unable to capture the variability in the detections and results in markedly different conclusions about the model's goodness-of-fit. Further analysis was

conducted for the All10 data structure since it included the most sites (thus the most information) from the three structures with reasonable dispersion parameters. The model selection procedure was repeated using the Quasi-AIC (QAIC) statistic to account for the overdispersed data but did not chose a model different to the one previously selected. The results corresponding to this procedure can be found in the Appendices by Table B.1.

Model estimates

The standard errors corresponding to the estimates of the `Env.Clim.Urban` model were adjusted upwards to account for the measured overdispersion. For the purposes of further evaluating the potential difference in estimated coefficients, caused by altering the data structures, the parameter coefficients for the Common10 (C10) and Intermediate10 (I10) are included adjacent to the coefficients, standard errors, z-scores and p-values relating to the All10 (A10) data structure, provided by Table 4.12.

The (initial) occupancy parameter is estimated for the year 2010 and, although additional sites that were not surveyed at all in the first study year were included in the estimation of occupancy, the same covariates were found to be statistically significant. The only notable exception was the `SurfaceTemp` covariate, which was previously significant but found to be insignificant in the dynamic modelling process, but could likely just be due to the inflation of the standard errors. The increase in the number of sites visited gives inconclusive results pertaining to the occupancy parameter as no definitive pattern can be observed in the coefficients.

Parameter	**Estimate**			**SE**	**z**	**p-value**
	C10	**I10**	**A10**			
Occupancy (ψ_1)						
Intercept	(1.29)	(1.49)	1.39	0.40	3.46	0.00
Pressure	(-1.15)	(-0.85)	-0.69	0.24	-2.93	0.00
Pressure^2	(0.07)	(-0.03)	0.20	0.11	1.85	0.06
SoilPH	(-3.22)	(-2.58)	-2.39	0.32	-7.53	0.00
SoilPH^2	(-1.87)	(-1.17)	-1.11	0.19	-5.97	0.00
SurfaceTemp	(1.17)	(0.88)	0.59	0.38	1.57	0.12
SurfaceTemp^2	(-0.33)	(-0.45)	-0.65	0.20	-3.31	0.00
Precipitation	(2.19)	(0.56)	0.16	0.50	0.33	0.74
Precipitation^2	(-0.73)	(-0.35)	-0.34	0.18	-1.88	0.06
Wind.Speed	(-0.34)	(0.07)	-0.75	0.38	-1.98	0.05
Wind.Speed^2	(-0.94)	(-0.77)	-0.99	0.24	-4.17	0.00
TreeCanopy	(0.35)	(0.49)	0.63	0.25	2.54	0.01

TreeCanopy^2	(-0.03)	(-0.19)	-0.25	0.09	-2.71	0.01
Population.Density	(14.31)	(7.77)	13.16	2.08	6.32	0.00
Colonization (γ)						
Intercept	(-1.41)	(-2.31)	-2.94	0.16	-18.68	0.00
Yearly_Pressure	(-0.89)	(-1.02)	-0.79	0.15	-5.31	0.00
Yearly_Temp	(1.04)	(1.03)	0.97	0.14	6.73	0.00
Yearly_Rain	(0.04)	(0.35)	0.43	0.12	3.69	0.00
Yearly_Wind	(0.88)	(0.42)	0.37	0.10	3.66	0.00
Yearly_Pop	(0.35)	(0.42)	0.58	0.21	2.81	0.00
Yearly_Soil	(-0.06)	(-0.11)	-0.03	0.07	-0.47	0.63
Yearly_NDVI	(0.41)	(0.20)	0.39	0.12	3.38	0.00
Extinction (ϵ)						
Intercept	(-2.73)	(-2.11)	-2.17	0.19	-11.64	0.00
Yearly_Pressure	(0.83)	(0.90)	0.99	0.16	6.34	0.00
Yearly_Temp	(-0.40)	(-0.47)	-0.70	0.17	-4.14	0.00
Yearly_Rain	(0.05)	(-0.24)	-0.31	0.15	-2.10	0.04
Yearly_Wind	(-0.63)	(-0.13)	-0.28	0.13	-2.18	0.03
Yearly_Pop	(-1.81)	(-3.00)	-2.45	0.40	-6.21	0.00
Yearly_Soil	(0.29)	(0.40)	0.38	0.08	5.00	0.00
Yearly_NDVI	(0.00)	(0.12)	0.12	0.11	1.14	0.25
Detection (p)						
Intercept	(0.42)	(0.39)	0.33	0.04	9.24	0.00
lhours	(-0.07)	(-0.06)	-0.07	0.02	-3.44	0.00
lhours^2	(-0.06)	(-0.06)	-0.06	0.01	-5.09	0.00
ltotspp	(0.19)	(0.24)	0.27	0.02	12.22	0.00
ltotspp^2	(0.02)	(0.01)	0.01	0.01	1.86	0.07
SeasonSpring	(-0.05)	(-0.03)	0.01	0.04	0.18	0.84
SeasonSummer	(0.02)	(0.02)	0.04	0.04	1.00	0.31
SeasonWinter	(-0.02)	(0.01)	0.02	0.04	0.37	0.73
Yearly_NDVI	(-0.04)	(-0.06)	-0.05	0.02	-2.67	0.01
Yearly_Pop	(0.24)	(0.25)	0.26	0.01	25.44	0.00

Table 4.12: Dynamic occupancy model estimates: Myna

The increase in total sites does not result in an observable difference in the estimates relating to the detection probability. This is to be expected since the detection histories

are the same length for the three data structures and an increase in surveyed sites is unlikely to alter the estimated detection probability of the Myna at a given site but could possibly affect its precision.

Regarding the estimation of the vital rate parameters, all covariates used to estimate the probability of colonization in this model are statistically significant apart from `Yearly_Soil` and the same covariates used to estimate (local) extinction are statistically significant barring `Yearly_NDVI`. Since these covariates are symmetric it would be reasonable to assume that the signs of the estimates will oppose one another for the different processes if the covariate is truly significant. This assumption is based on the rationale that, for a hypothetical increase in a covariate, if the probability of colonizing a site increases, it should decrease the probability of extinction at that same site. The results of the table would suggest that this is indeed the case for the `Yearly_Pressure`, `Yearly_Temp`, `Yearly_Rain`, `Yearly_Wind`, and `Yearly_Pop` covariates since these estimates are both statistically significant and have opposing signs for the different processes.

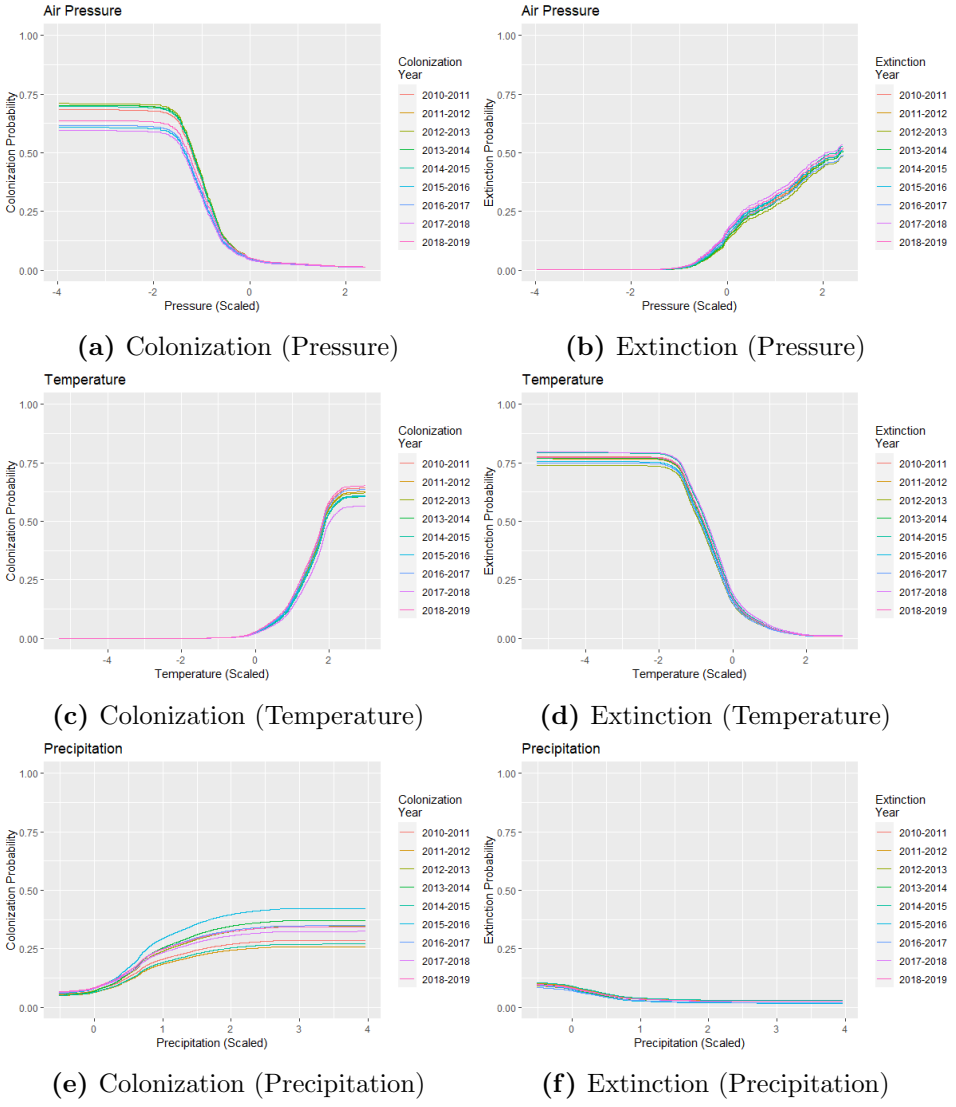
These results indicate that an increase in the annual air pressure (`Yearly_Pressure`) at a particular site is likely to decrease the likelihood of the Myna colonizing that site and increase its probability of extinction. By contrast, an increase in the annual temperature (`Yearly_Temp`), precipitation (`Yearly_Rain`), wind speed (`Yearly_Wind`) or population (`Yearly_Pop`) at a given site is expected to increase the probability of the Myna colonizing that site and decrease its probability of extinction. The fitted covariate relationships, which are generated by holding all other variables constant and show the yearly average probability across all sites for both the colonization and extinction parameters, can be found as Figures 4.9 and 4.10 on the subsequent pages.

There seems to be little variation in the estimates between the different data structures which suggests that, assuming the length of the detection histories remain constant, an increase in the number of visited sites does not markedly influence the model estimates. The estimates for the vital rate parameters relating to the A10 data structure can be formally expressed as:

$$\begin{aligned}\text{logit}\left(\gamma_{it-1}\right) = -2.94 &- 0.79\left(\texttt{Yearly\_Pressure}_{it}\right) + 0.97\left(\texttt{Yearly\_Temp}_{it}\right) \\ &+ 0.43\left(\texttt{Yearly\_Rain}_{it}\right) + 0.37\left(\texttt{Yearly\_Wind}_{it}\right) \\ &+ 0.58\left(\texttt{Yearly\_Pop}_{it}\right) - 0.03\left(\texttt{Yearly\_Soil}_{it}\right) \\ &+ 0.39\left(\texttt{Yearly\_NDVI}_{it}\right)\end{aligned} \tag{4.5}$$

$$\begin{aligned}\text{logit}\left(\epsilon_{it-1}\right) = &-2.17 + 0.99\left(\texttt{Yearly\_Pressure}_{it}\right) - 0.70\left(\texttt{Yearly\_Temp}_{it}\right) \\ &- 0.31\left(\texttt{Yearly\_Rain}_{it}\right) - 0.28\left(\texttt{Yearly\_Wind}_{it}\right) \\ &- 2.45\left(\texttt{Yearly\_Pop}_{it}\right) + 0.38\left(\texttt{Yearly\_Soil}_{it}\right) \\ &+ 0.12\left(\texttt{Yearly\_NDVI}_{it}\right) \end{aligned} \tag{4.6}$$

$$\begin{aligned}\text{for } i &= 1, 2, \ldots, 6911 \\ t &= 2, 3, \ldots, 10.\end{aligned}$$

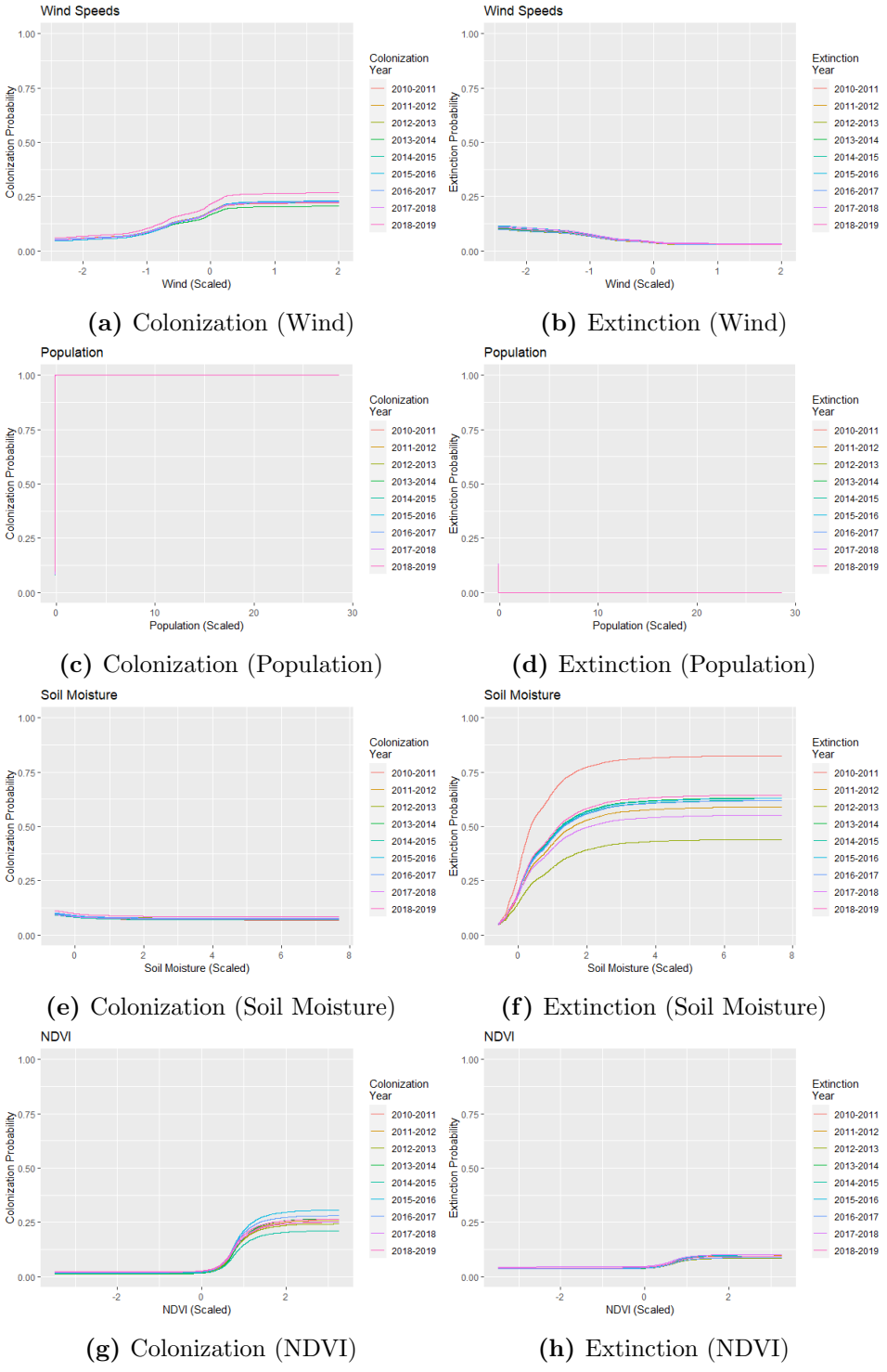


(a) Colonization (Pressure) **(b)** Extinction (Pressure)

(c) Colonization (Temperature) **(d)** Extinction (Temperature)

(e) Colonization (Precipitation) **(f)** Extinction (Precipitation)

Figure 4.9: Fitted relationships of dynamic occupancy vital rates: Myna (Part 1)



(a) Colonization (Wind) **(b)** Extinction (Wind)

(c) Colonization (Population) **(d)** Extinction (Population)

(e) Colonization (Soil Moisture) **(f)** Extinction (Soil Moisture)

(g) Colonization (NDVI) **(h)** Extinction (NDVI)

Figure 4.10: Fitted relationships of dynamic occupancy vital rates: Myna (Part 2)

In addition to evaluating the estimated model coefficients, which do not display distinct trends between the various data structures, it is worth investigating how the precision of the estimate is influenced when the degree of temporal overlap (total number of surveyed sites) is altered and the maximum length of the detection history is set to 10.

Figure 4.11 illustrates the (logit) standard errors of every covariate used in the estimation of the four model parameters whereby the index of 1-14, 15-22, 23-30 and 31-40 represent the covariates used to estimate the initial occupancy (ψ_1), colonization (γ), local extinction (ε) and detection parameter (p), respectively.

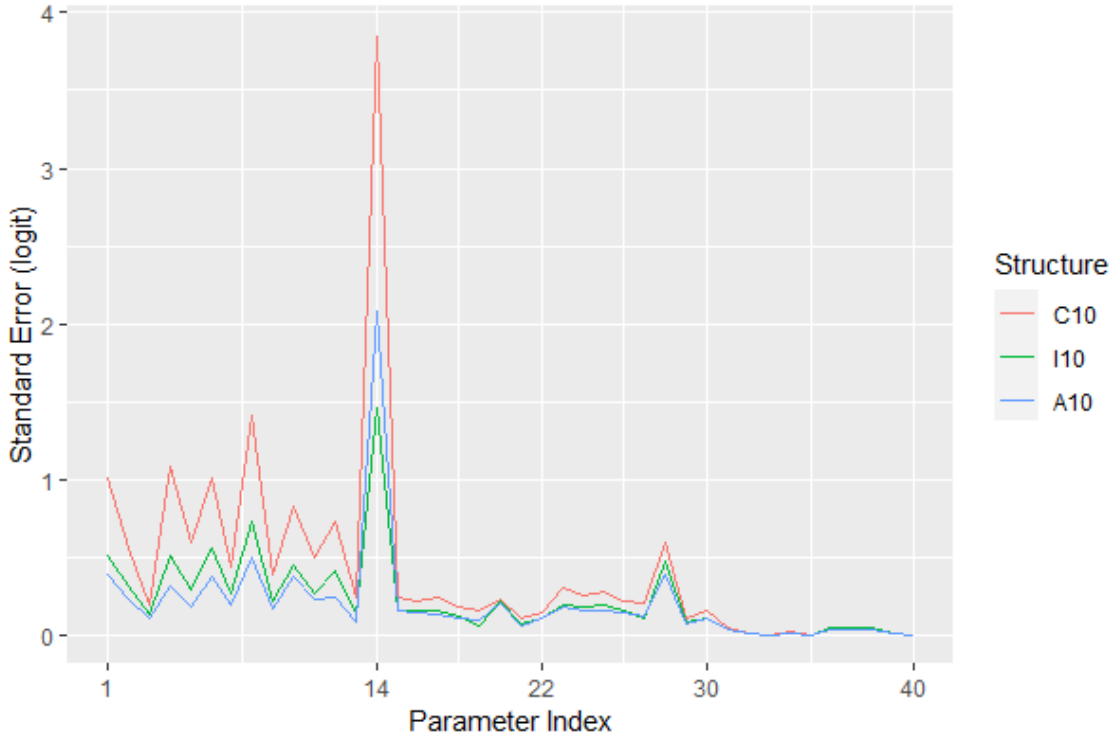


Figure 4.11: Precision of the estimate (Myna)

The trend observed within these standard errors across each estimated parameter is that the precision of the estimate is higher when the degree of temporal overlap is larger. Based on this result it can be concluded that, while holding the maximum length of the detection history constant, an increase in the number of sites included in the study will result in an increase in the precision of the estimate.

The variation between the different structure's estimated standard errors is largest for the initial occupancy parameter and smallest for the detection parameter. Further, the

large spike observed in the standard error corresponding to the 14th fitted covariate (`Population.Density`) would suggest that, albeit the model indicates that the covariate is important in the estimation of initial occupancy, the high variation within this estimate advocates for conservative inference if the estimated coefficient is to be used in any real decision-making process.

The Myna's predicted (derived) occupancy across the study region over the 10-year period is provided by Figure 4.12 and indicates that the target species' distribution is expanding over time. This graphic would suggest that the probability of the Myna's occupancy within the study region is predicted to increase from approximately 47% in 2010 to 50% in 2019, and that the variation in the estimated yearly occupancy (vertical lines) fluctuates by roughly 1-2%.

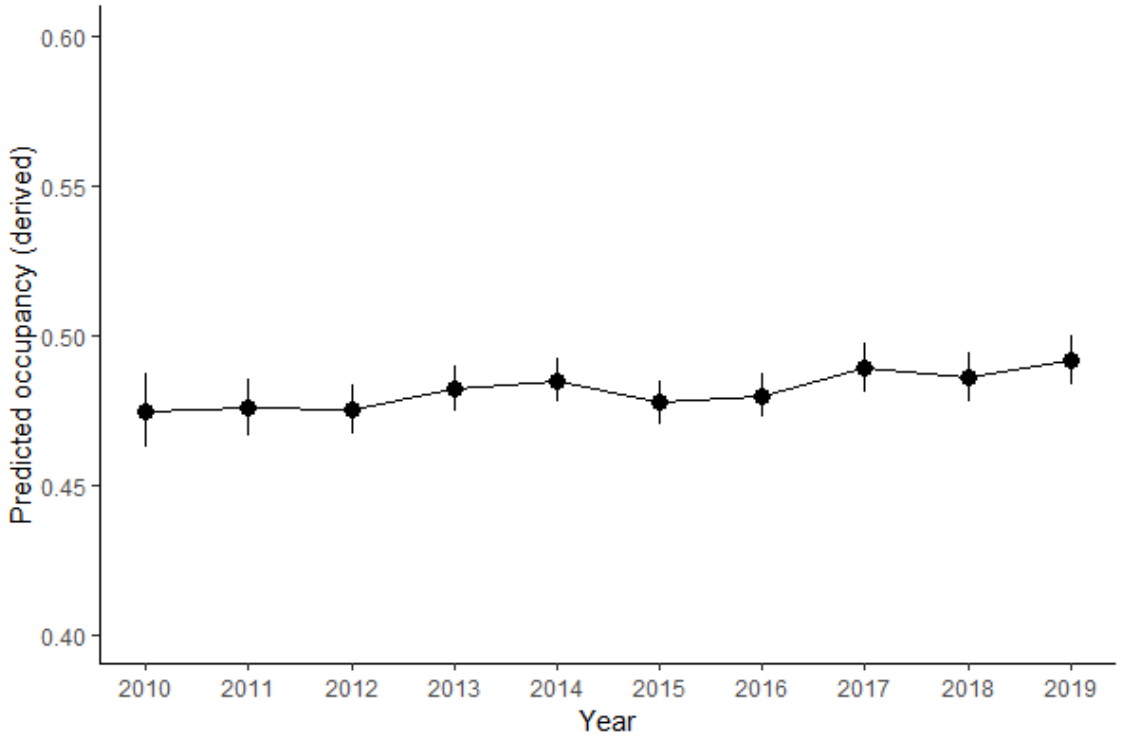


Figure 4.12: Myna's predicted occupancy for the study period 2010-2019

A comparison of this plot to the naive occupancy rates provided by Table 4.3 would suggest that the fitted dynamic model predicts approximately the same occupancy each year for the study region. The naive estimates showed a general increase in the Myna's yearly predicted occupancy, and the dynamic model was also able to uncover a slight expansion in the species range with the help of the information provided by the fitted covariates.

4.3.2 Mallard

Model selection procedure

The dynamic occupancy modelling of the Mallard similarly made use of the results extracted from the static occupancy models via the covariates fitted to describe the occupancy and detection parameters in the best fitting static model (Table 4.8). Further, the static `Population` and `NDVI` variables were treated the same as the Myna's dynamic models and were substituted for their yearly counterparts.

Structure	Hypothesis	K	AIC	ΔAIC	AICWt
Common10 (Common50)	Clim.Urban (Clim.Urban)	29	7564.26 (19678.87)	0.00 (0.00)	1 (1)
	Wet.Urban (Wet.Urban)	25	7583.47 (19704.51)	19.21 (25.64)	0.00 (0.00)
	Anthropogenic (Anthropogenic)	21	7584.82 (19712.08)	20.56 (33.21)	0.00 (0.00)
	Humid.Env (Humid.Env)	25	7614.30 (19729.59)	50.04 (50.72)	0.00 (0.00)
	Weather (Weather)	25	7614.53 (19732.65)	50.27 (53.78)	0.00 (0.00)
Intermediate10 (Intermediate50)	Clim.Urban (Clim.Urban)	29	9201.22 (21543.37)	0.00 (0.00)	1 (1)
	Wet.Urban (Wet.Urban)	25	9215.39 (21565.65)	14.16 (22.28)	0.00 (0.00)
	Anthropogenic (Anthropogenic)	21	9217.22 (21570.25)	16.00 (26.88)	0.00 (0.00)
	Wetlands (Weather)	23 (25)	9257.75 (21601.23)	56.52 (57.85)	0.00 (0.00)
	Weather (Humid.Env)	25 (25)	9258.11 (21607.80)	56.89 (64.43)	0.00 (0.00)
	Clim.Urban (Clim.Urban)	29 (29)	9680.83 (22031.78)	0.00 (0.00)	1 (1)
	Anthropogenic (Wet.Urban)	21 (25)	9700.44 (22057.17)	19.60 (25.39)	0.00 (0.00)
	Wet.Urban (Anthropogenic)	25 (21)	9701.03 (22058.31)	20.19 (26.54)	0.00 (0.00)

All10 (All50)	Weather (Weather)	25	9740.21 (22094.46)	59.38 (62.68)	0.00 (0.00)
	Humid.Env (Humid.Env)	25	9747.22 (22108.76)	66.39 (76.98)	0.00 (0.00)

Table 4.13: Dynamic occupancy model selection procedure: Mallard

The approach to modelling the vital rate parameters was also symmetric and implemented using a hypothesis-driven approach, for which seven hypotheses were formulated and fitted to the same six data structures. The five best fitting models for each structure, based on the AIC and weighted AIC selection criteria, is provided by Table 4.13.

The power of the model selection procedure to choose a common hypothesized model (`Clim.Urban`) as the best fitting model for each structure was the same as observed for the Myna. This indicates that the model selection procedure is unlikely to be affected by an alteration to the degree of temporal overlap (in terms of both the number of sites as well as the maximum length of the detection history) used in the study.

Assessment of model fit

The `Clim.Urban` model's goodness-of-fit was subsequently assessed via the same methods described for the Myna's best fitting model but, as illustrated by Table 4.14, produces fewer desirable results.

Structure	$\tilde{\chi}^2_{\mathbf{O}}$	$\tilde{\chi}^2_{\mathbf{B}}$	$\hat{\mathbf{c}}$	**P-value**	**nsim**	**Run time (Hours)**
Common10	324089.3	20012.16	16.2	0	2000	3
Intermediate10	474614.7	20015.41	23.71	0	2000	40
All10	700496.6	20146.45	34.77	0	2000	300
Common50	1.99e+30	1.17e+16	1.70e+14	0	2000	50
Intermediate50	2.89e+30	5.25e+15	5.50e+14	0	2000	136
All50	3.57e+30	8.17e+14	4.37e+15	0	100	62

Table 4.14: Dynamic model goodness-of-fit: Mallard

This table depicts a situation whereby, regardless of the implemented structure, the model's goodness-of-fit is poor. The dispersion parameters are unreasonably large and indicates that even the models corresponding to the data structures which restrict the

maximum length of the detection histories to 10 seasonal surveys do not provide a reasonable fit to the data. The All50 structure was again run using only 100 simulations as a means of simply highlighting that, even for a relatively small number of simulations, this structure's model provides a poor fit to the data.

The p-values corresponding to the chi-square statistic indicate that the null hypothesis (of an adequate model fit) for each structure is rejected and the lowest estimated dispersion parameter is 16.2. This possibly signifies that one or more important predictors of occupancy are missing from the model, or the closure assumption associated with the model has been violated, and no firm conclusions can be made regarding the occupancy and range dynamics of the Mallard. Such an ill-fitting model does not allow for informative inference of the model estimates, yet the interpretations pertaining to the best fitting model are still discussed after adjustment for the dispersion parameter is made. Although the caveat of these interpretations is that they are extremely conservative and firm conclusions cannot be drawn from the results.

Model estimates

The standard errors corresponding to the estimates of the hypothesized model were adjusted upwards to account for the estimated overdispersion. The estimates, standard errors, z-scores and p-values associated with the All10 data structure's best fitting model (`Clim.Urban`) adjacent to the Common10 (C10) and Intermediate10 (I10) estimates, are presented by Table 4.15 with the motivation of assessing the contrast in estimated coefficients when it is known that the model provides an acutely poor fit to the data.

Parameter	**Estimate**			**SE**	**z**	**p-value**
	C10	**I10**	**A10**			
Occupancy (ψ_1)						
Intercept	(-4.44)	(-5.19)	-5.10	2.86	-1.79	0.07
SoilPH	(-4.52)	(-5.85)	-5.01	5.67	-0.88	0.38
SoilPH^2	(-1.84)	(-2.53)	-2.14	2.86	-0.75	0.45
SurfaceTemp	(-1.70)	(-1.82)	-2.29	2.82	-0.81	0.42
SurfaceTemp^2	(-1.35)	(-1.22)	-1.33	1.79	-0.74	0.46
Precipitation	(-0.94)	(-1.03)	-1.04	1.22	-0.85	0.39
Precipitation^2	(0.19)	(0.15)	0.01	0.83	0.01	0.99
Permanent_Water	(0.21)	(0.08)	0.08	0.26	0.32	0.75
Population.Density	(0.28)	(0.32)	0.36	0.34	1.06	0.29

Colonization (γ)						
Intercept	(-3.74)	(-4.22)	-4.65	0.78	-5.97	0.00
Yearly_Temp	(0.23)	(0.20)	0.13	1.05	0.13	0.90
Yearly_Rain	(-0.11)	(-0.03)	0.13	0.63	0.20	0.84
Yearly_Pressure	(-0.55)	(-0.49)	-0.46	1.03	-0.44	0.66
Yearly_Wind	(0.40)	(0.37)	0.37	0.81	0.46	0.65
Yearly_Pop	(0.12)	(0.12)	0.14	0.12	1.13	0.26
Extinction (ε)						
Intercept	(-1.54)	(-1.43)	-1.45	1.04	-1.39	0.16
Yearly_Temp	(-0.68)	(-0.49)	-0.63	1.30	-0.49	0.63
Yearly_Rain	(0.29)	(0.21)	0.24	1.05	0.23	0.82
Yearly_Pressure	(0.34)	(0.18)	0.19	0.90	0.21	0.83
Yearly_Wind	(-0.36)	(-0.23)	-0.25	0.77	-0.33	0.74
Yearly_Pop	(0.07)	(0.06)	0.05	0.12	0.42	0.68
Detection (p)						
Intercept	(-1.37)	(-1.41)	-1.46	0.41	-3.56	0.00
lhours	(0.16)	(0.14)	0.15	0.24	0.62	0.53
lhours^2	(0.02)	(0.03)	0.02	0.13	0.17	0.86
ltotspp	(0.00)	(0.04)	0.06	0.25	0.25	0.80
Yearly_NDVI	(-0.18)	(-0.19)	-0.19	0.24	-0.80	0.42
SeasonSpring	(-0.02)	(-0.03)	-0.02	0.54	-0.05	0.96
SeasonSummer	(-0.03)	(-0.02)	-0.03	0.49	-0.06	0.95
SeasonWinter	(-0.04)	(-0.05)	-0.03	0.53	-0.07	0.95

Table 4.15: Mallard parameter estimates for structures capped at 10

The initial occupancy of the Mallard was estimated for the year 2010, representing the first study season, and where the static model (Table 4.6) indicated that the covariates fitted to the occupancy parameter were all significant, the dynamic counterpart suggests that none of the fitted covariates are significant in the estimation of the Mallard's initial occupancy probability. This is to be expected, based on the goodness-of-fit results (Table 4.14), as the estimated standard errors were adjusted by a factor of 34.77 to account for the overdispersion present in the data.

Similarly, the covariates used to estimate the detection probability are all statistically

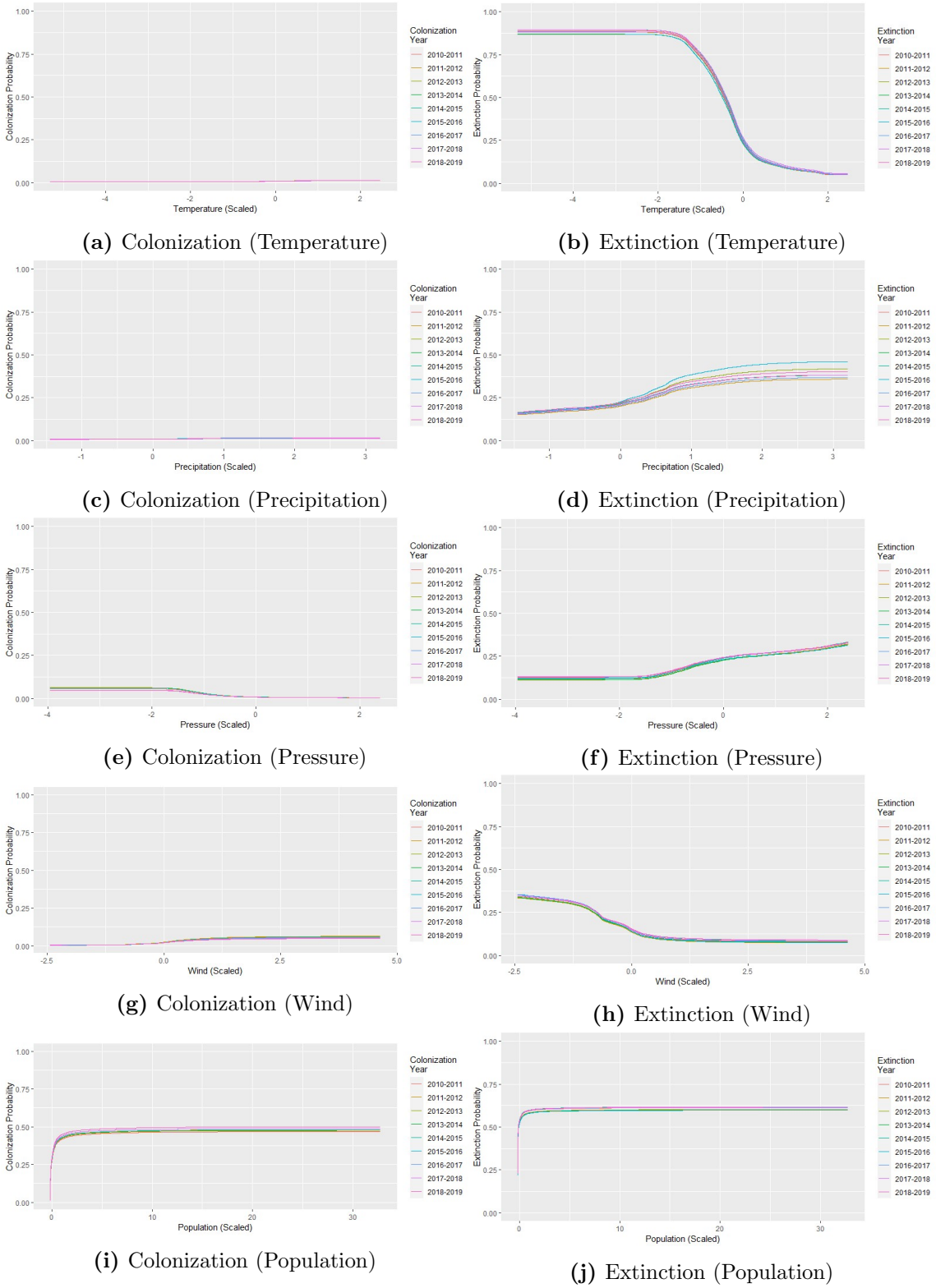
insignificant, with the estimated p-values ranging from 0.42 to 0.95. Little can be gleaned from these coefficients due to the model's poor fit, and the coefficients do not markedly vary between the data structures since the maximum length of the detection history is the same.

Analysis of these results in conjunction with Table 4.12 implies that, whether or not the fitted model provides a reasonably adequate (Myna) or markedly poor fit (Mallard) to the data, the effect of increasing the number of total sites visited each season does not affect the detection probability when the maximum length of the detection history is relatively small and kept constant.

The results of the colonization and extinction parameters indicate that, after the necessary adjustment to the standard errors, none of the fitted covariates are statistically significant. The coefficients of the three data structures are almost identical and (by scrutinising these results in conjunction with Table 4.12) a conservative conclusion would be that, regardless of the model's goodness-of-fit, there is little variation in the estimated coefficients when the detection histories are of the same length and the total number of surveyed sites is altered.

Although the covariates are all insignificant, the signs of these coefficients (with the exception of `Yearly_Pop`) oppose one another depending on whether they are used to estimate the colonization or extinction probability. These estimates depict the same pattern observed in the Myna's dynamic model for almost all effects and could possibly suggest that, as seen from the results of the significant vital rate parameters of Table 4.12, the measured covariates are able to account for variability in the Mallard's range dynamics but there is another unseen fundamental problem with the study design that prevents the model from providing an adequate fit to the data.

Some conservative inference can be made regarding the range dynamics by consulting the fitted relationships of the vital rate parameters provided by Figure 4.13. It is apparent that the probability lines are flat, and increases in the annual temperature (`Yearly_Temp`), precipitation (`Yearly_Rain`), pressure (`Yearly_Pressure`) or wind speed (`Yearly_Wind`) variables will have no effect on the colonization probability at a site. By contrast, an increase in `Yearly_Temp` or `Yearly_Wind` will decrease the probability of extinction and increases in either of `Yearly_Rain` or `Yearly_Pressure` result in an increased probability of extinction at a site.



(a) Colonization (Temperature)

(b) Extinction (Temperature)

(c) Colonization (Precipitation)

(d) Extinction (Precipitation)

(e) Colonization (Pressure)

(f) Extinction (Pressure)

(g) Colonization (Wind)

(h) Extinction (Wind)

(i) Colonization (Population)

(j) Extinction (Population)

Figure 4.13: Fitted relationships of dynamic occupancy vital rates: Mallard

The equations for the vital rate parameters corresponding to the A10 data structure can be mathematically expressed as:

$$\begin{aligned}\text{logit}\left(\gamma_{it-1}\right) = &-4.65 + 0.13\left(\texttt{Yearly\_Temp}_{it}\right) + 0.13\left(\texttt{Yearly\_Rain}_{it}\right) \\ &- 0.46\left(\texttt{Yearly\_Pressure}_{it}\right) + 0.37\left(\texttt{Yearly\_Wind}_{it}\right) \\ &+ 0.14\left(\texttt{Yearly\_Pop}_{it}\right)\end{aligned} \tag{4.7}$$

$$\begin{aligned}\text{logit}\left(\epsilon_{it-1}\right) = &-1.45 - 0.63\left(\texttt{Yearly\_Temp}_{it}\right) + 0.24\left(\texttt{Yearly\_Rain}_{it}\right) \\ &+ 0.19\left(\texttt{Yearly\_Pressure}_{it}\right) - 0.25\left(\texttt{Yearly\_Wind}_{it}\right) \\ &+ 0.05\left(\texttt{Yearly\_Pop}_{it}\right)\end{aligned} \tag{4.8}$$

$$\begin{aligned}&\text{for } i = 1, 2, \ldots, 7041 \\ &\quad t = 2, 3, \ldots, 10.\end{aligned}$$

Investigation of how the precision of the estimate is influenced by an alteration to the number of sites included in the study, when the maximum length of the detection history is set to 10, is provided by Figure 4.14.

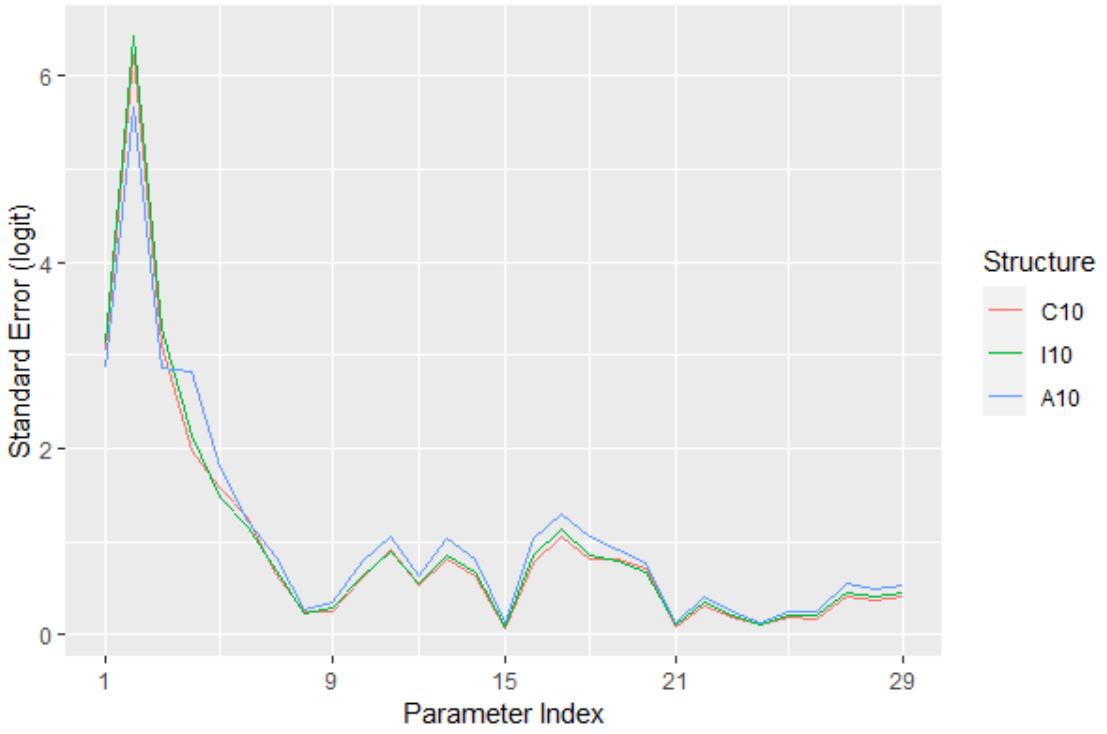


Figure 4.14: Precision of the estimate (Mallard)

Similar to the case of the Myna (Figure 4.11), this graphic shows the (logit) standard errors of every covariate used in the estimation of the four model parameters whereby the index of 1-9, 10-15, 16-21 and 22-29 represent the covariates used to estimate the initial occupancy (ψ_1), colonization (γ), local extinction (ε) and detection parameter (p), respectively.

The trend observed within these standard errors across each estimated parameter indicates that the precision of the estimate is higher when the number of sites included in the study is decreased. This is the opposite result to what was seen for the Myna and is likely due to the Mallard's much higher dispersion parameters corresponding to these three data structures (16.2, 23.71 and 34.77). The large standard errors observed for all fitted covariates, especially those used to estimate initial occupancy, support the prior results which would suggest the dynamic model provides a poor fit to the Mallard data.

The Mallard's predicted occupancy across the study region over the 10-year period is provided by Figure 4.15. The figure depicts a slight decrease in the predicted occupancy over the years, however, it is important to note that the y-axis only ranges from 0-0.1.

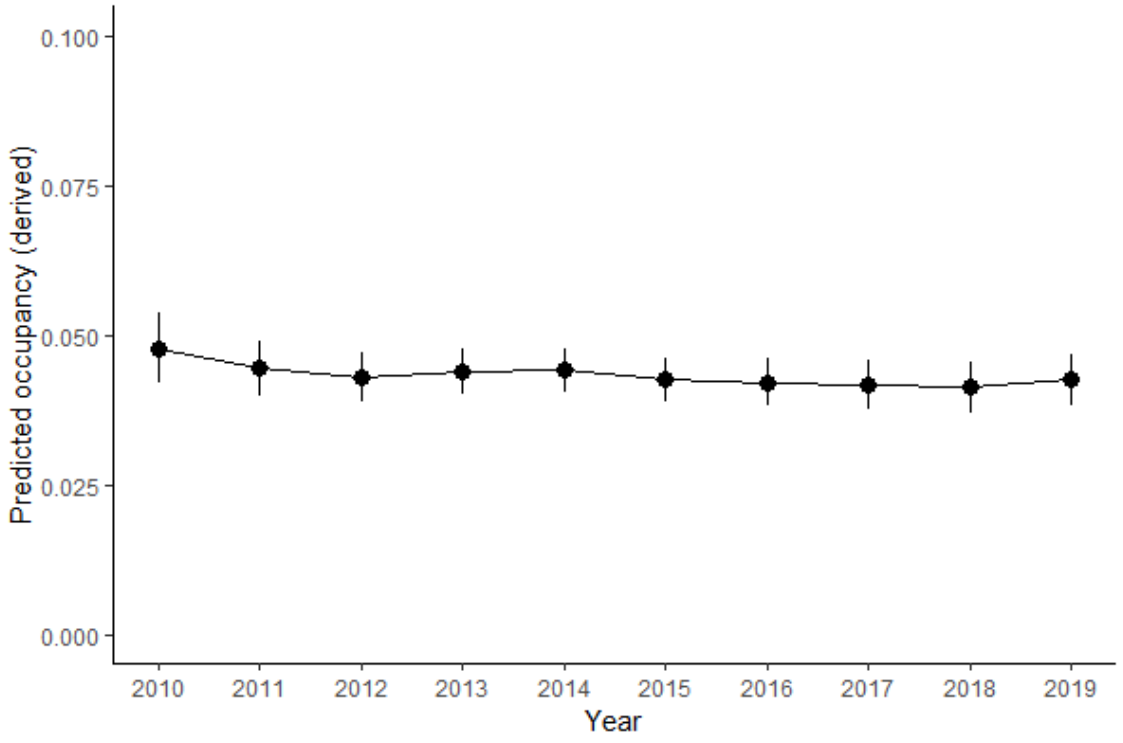


Figure 4.15: Mallard's predicted occupancy for the study period 2010-2019

A comparison of this plot with the naive occupancy rates provided by Table 4.3 would suggest that the fitted dynamic model shows a slightly higher (approximately 2%) predicted occupancy each year in comparison to the naive occupancy rates. Although the observed trends are slight and arguably negligible the results of Table 4.3 indicate that the distribution of the Mallard is expanding, but the dynamic model would suggest that the Mallard's distribution is slightly contracting during this 10-year period. Since the predicted occupancy is questionable due to the dynamic model's markedly poor fit to the data, and the naive occupancy rates are likely to be inaccurate due to stochastic volatility, it is difficult to ascertain how the distribution of the Mallard is changing within the study region during this 10-year period.

Chapter 5

Summary and Conclusions

The practical objectives of this dissertation were fulfilled through the fitting of static and dynamic occupancy models to the observed data. The factors which influenced the occupancy of the Myna (using the static modelling framework) included proximity to urban environments, atmospheric pressure, pH of the soil, temperature of the surface, tree canopy heights, wind speeds and rainfall. While the detection probability was influenced by the number of hours spent actively bird-watching, the total number of other species observed at the same site (which acts as a proxy for birding expertise or general birding conditions), the season in which the visit took place, the greenness of a site and the proximity to urban environments.

Similarly, the static models were indicative that the Mallard's occupancy was determined by the pH of the soil, temperature of the surface, rainfall, proximity to urban environments and access to permanent bodies of water, respectively. While its detectability was determined by the number of hours spent actively bird-watching, the total number of other species observed at the same site, the season in which the visit took place and the greenness of a site.

The results of the dynamic occupancy modelling were indicative that reasonable results could be obtained for the Myna, but there was likely a fundamental problem with the study design that inhibited the ability to fit a reasonable model to the observed Mallard data. The Myna's range dynamics over the 10-year study period were significantly influenced by yearly changes at a site in recorded atmospheric pressure, air temperature, rainfall, wind speeds and measured population. Although the results of the Mallard were insignificant and not conclusive, conservative inference pertaining to the Mallard's best fitting dynamic model suggested that the same yearly-site covariates fitted to the Myna's best model have the potential to influence the Mallard's range dynamics, but the poor fitting model did not allow for an in-depth evaluation and conclusive inference concerning these results.

The final practical objective concerned the distribution in the target species' occupancy over the study period for which the dynamic model predictions of yearly occupancy did not suggest there was a marked expansion or contraction in either of the invasive species over the 10-year period in their respective study regions. This is somewhat intuitive since the time scale at which analysis took place is comparatively negligible when considering the evolutionary rate at which species spread throughout an area (Du Plessis et al., 1995).

The theoretical objectives of this study were fulfilled through the fitting of the dynamic occupancy models. The objectives of interest focused on the potential impact of altering the imbalanced data structure on the results of the dynamic models, and how the estimated results changed if these alterations significantly impacted the observed analysis. The observed results were indicative that increasing the total number of sites included in the study (thus altering the degree of temporal overlap) had no apparent impact on the final results, however, an increase in the maximum length of the observed detection histories played a significant role on the assessment of model fit and the parameter estimates of the final model. The precision of the model estimates was also influenced by an alteration to the degree of temporal overlap, for which the inclusion of more surveyed sites in a study lead to more accurate estimates.

Regardless of the alterations to the data structures, the dynamic modelling framework did not have the power to identify the cause of the Mallard's poor model fit, and is assumed to be influenced by dominant large-scale nomadic movement patterns driven by rainfall that the framework was unable to accurately model (Measey et al., 2020). Thus, the final takeaway from this study is that the results of occupancy model's goodness-of-fit were greatly influenced by the presence of large-scale nomadic movement patterns of a species in combination with an ill-defined primary sampling period, the estimated occupancy probability was empowered by the addition of as many visited sites that the design will allow, and the length of the seasonal detection histories had a major impact on the model goodness-of-fit that cannot be ignored during the formation of the study design. Additionally, when using a large amount of sampling the occupancy models tended to suffer when modelling a species that was inconspicuous and more rarely sighted, as was observed in the case of the Mallard.

5.1 Static models

Previous literature pertaining to static occupancy models (MacKenzie et al., 2002) suggested that increasing the number of surveys improves both accuracy and precision of the estimated occupancy probability, although there is minimal utility in conducting up to 10 in comparison to five total surveys. Further, the work of MacKenzie et al. (2002, 2006) found that a higher quantity of analysed sites will similarly enhance the estimated parameter's estimated accuracy when the fitted model is known to provide a good fit to the data.

The results pertaining to the simulation study conducted by MacKenzie and Bailey (2004), as a means of evaluating the power of their procedure in identifying poor model fit, gave evidence that their approach (implemented in this study) had good power when the number

of surveys conducted was larger than five. Their simulation study was also able to identify poor fit caused by the exclusion of an important site-specific detection covariate when the estimated detection probability was low, however, the procedure was not able to identify an inadequate model due to heterogeneous occupancy probabilities across sites.

Previous work found that violation of the closure assumption, caused by non-random (migratory) movement patterns, leads to high bias and inconclusive inference regarding the factors affecting the estimated parameters (Kendall, 1999). Further, this study found that these biased estimates will tend to underestimate the occupancy parameter in the case of emigration and overestimate the parameter in the case of immigration of the target species.

Since the results pertaining to the dynamic occupancy of the Mallard gave substantial evidence that the models provided a poor fit to the observed data, it is likely that this outcome was the consequence of a violation of the closure assumption. This is based on the recent understanding that the Mallard is a migratory bird in the South African region (Measey et al., 2020). It is thus reasonable to conclude that the seemingly adequate results of the static occupancy model pertaining to the Mallard were biased and altogether inconclusive.

The choice, relating to the static occupancy models, of leaving the length of the detection histories unconstrained (maximum of 146 at one specific site) was likely an erroneous one, as the probability of occupancy was guided by these histories and the covariates selected from this static analysis were then used to estimate the detection and initial occupancy probabilities when the detection histories were restrained to much shorter lengths. It is recommended that further research should set the length of seasonal detection histories to match in both the static and dynamic occupancy modelling frameworks to remove the concern of potential biases or model selection results caused by differences in the seasonal detection history length.

5.2 Dynamic models

Myna

The Myna's dynamic model selection procedure, when the hypothesized models are able to provide a relatively adequate fit to the data, was not influenced by alterations to the total number of visited sites or the length of the detection histories. However, the results suggested that there could possibly be differences in the best selected model when the number of sites is fewer than 775 and the length of the detection histories is greater than 10. Possible future studies could create an experimental design using less than the number

of sites studied in the "Common Sites" data structure to determine at which point the model selection process is affected by alterations to the length of the detection histories.

The assessment of the Myna's dynamic occupancy model fit suggested that an increase in the length of the detection histories from 10 to 50 results in a drastic change to the observed chi-square statistic and dispersion parameter. Further, depending on the time constraints of the researcher, the high computational burden of the MacKenzie and Bailey (2004) parametric bootstrapping technique will result in restricted analysis due to an increase in the detection history length to 50 seasonal surveys per site.

The analysis of these dynamic models would suggest that the coefficients of the model estimates do not display marked differences when there is an increase in the number of visited sites, given the detection histories are restricted to 10. However, based on the large dispersion parameters corresponding to those data structures with detection history lengths of 50, it can be inferred that all the estimates would be insignificant, and the insufficient power to model the variability introduced by increasing the length of the detection histories would lead to inconclusive results.

Mallard

The Mallard's dynamic model selection procedure, when the hypothesized models cannot provide an adequate fit to the data, indicated that the model selection procedure was unlikely to be affected by an alteration to the total number of sites included in the study. However, the slight differences in the third to fifth best fitting models for the structures assuming the same number of sites, but altering the length of the detection histories, implies that an increase in the length of the detection histories may have the power to alter the model selection procedure when it is known that none of the hypothesized models provide an adequate fit to the data.

The assessment of the Mallard's dynamic model fit suggested that, regardless of increasing the length of the detection histories from 10 to 50, the observed chi-square statistic and dispersion parameter imply that the fitted model, regardless of the data structures, did not provide a reasonable fit to the observed data. Further, after adjusting for the dispersion parameter, none of the fitted covariates were significant.

There are a few potential reasons which could give rise to the poor model fitting relating to the Mallard's observed data. Based on the aforementioned literature (Stephens et al., 2020), it is highly likely that the closure assumption has been violated due to unmeasured migration of the Mallard within the specified study season and thus it is probable that the temporal window used to define a season was erroneous. The assumption of no false

positives may also have been violated, since the consequence of the Mallard's hybridisation with indigenous species makes the identification of purebred Mallards all the more difficult, especially to the untrained eye (Measey et al., 2020). It is also plausible that a set of important predictors of Mallard occupancy were omitted from the set of covariates.

The most logical approach to address the bias induced from the violation of the closure assumption would be to restrict the data to include surveys within the temporal window when it is known that the availability of the target species is uninterrupted (such as within a breeding season) (MacKenzie et al., 2006). Alternatively, a different approach would be to implement the suggestions of Kendall (1999), which state that the survey data can be pooled into two surveys per season and subsequently modelled with survey-specific detection probabilities. A potential avenue of recourse for the problem of species misidentification would be to train the citizen scientists to detect harder to observe species. In addition to the above avenues of recourse, a general recommendation, especially for the case of citizen science data using a grid-based design, is that future studies should implement spatial occupancy models as a means of accounting for intuitive correlation between neighbouring pentads (Johnson et al., 2013).

This dissertation has emphasized the important decisions that need to be made pertaining to the design of an occupancy study, specifically for studies which utilise citizen science data. The consequences of an ill-fitting model based on a poor design has been observed and discussed and has brought to the fore the difficulties of conducting an occupancy analysis on a migratory avian species as opposed to their sedentary counterpart. MacKenzie et al. (2003) suggested that the dynamic model's ability to estimate all relevant parameters precisely and accurately was improved when the number of sites, seasons, surveys within a season and detection probability within a survey were increased. This dissertation has presented results which support the claim that an increase in the quantity of sites will improve the model estimates, however, caution must be exercised when increasing the length of the detection histories and should generally be kept to 10 repeated visits to a given site within the study window.

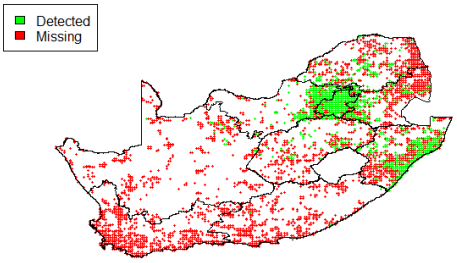
Appendix A

Data collection and extraction supplement

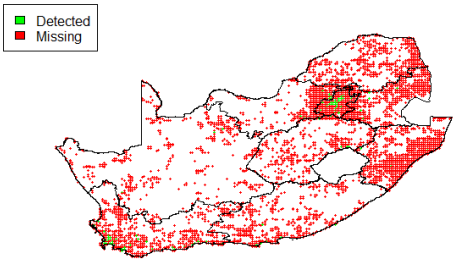
Catalogue (Years)	Variable	Units	Description
TerraClimate (1958-2020)	Avg.Temp	° C	Average air temperature [1]
	AET	mm	Actual evapotranspiration
	Precipitation	mm	Precipitation accumulation
	Soil.Moisture	mm	Soil moisture
	Wind.Speed	m/s	Wind speed
GCOM (2018-2021)	SurfaceTemp	K	Temperature of land surface
USGS (2010)	Elevation	m	Height above sea level
NASA/JPL (2005)	TreeCanopy	m	Tree heights
Copernicus (2015, 2019)	Urban_Cover	%	Percent of pentad classified as urban land cover
	Permanent_Water	%	Percent of pentad classified as permanent water cover
	Seasonal_Water	%	Percent of pentad classified as seasonal water cover
GLDAS (2000-2022)	Pressure	Pa	Atmospheric pressure
iSDA (2001-2016)	SoilPH	pH	pH of the soil
NOAA (1981-2022)	NDVI	Index	Normalized difference vegetation index
GPWv411 (2010, 2015)	Population.Density	km^2	Number of persons per square kilometre
WorldPop (2000-2020)	Population	Count	Estimated number of people residing in each grid cell

Table A.1: Summary of extracted covariates and the catalogues from which they were acquired

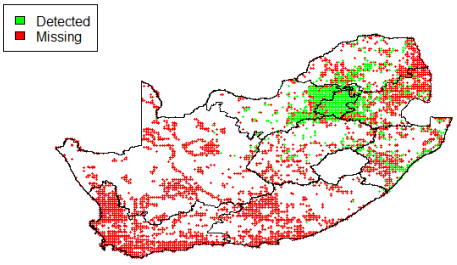
[1] This variable was derived from summing and averaging the minimum and maximum air temperatures



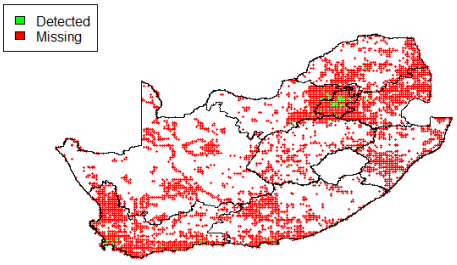
(a) Myna (2013)



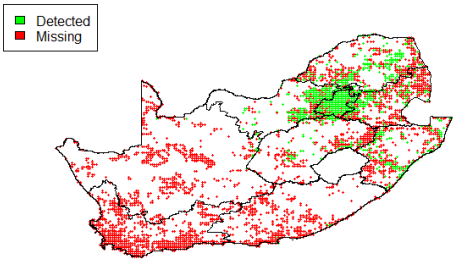
(b) Mallard (2013)



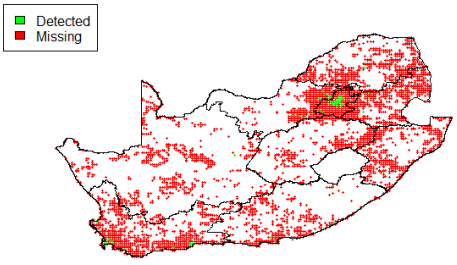
(c) Myna (2016)



(d) Mallard (2016)



(e) Myna (2019)



(f) Mallard (2019)

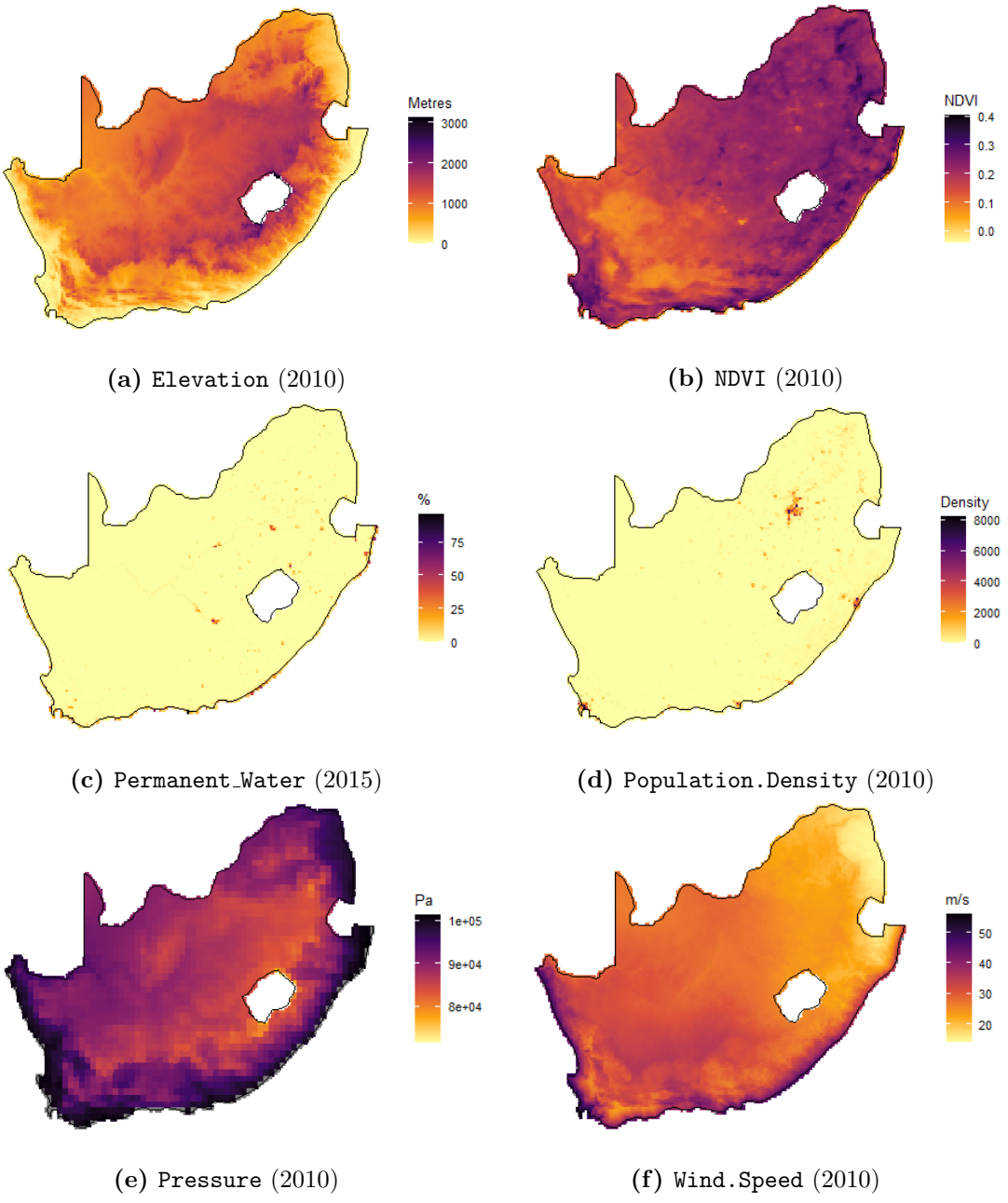
Figure A.1: Presence/absence data for the invasive species in 2013, 2016 & 2019

Data_Cleaner algorithm	
1.	Extract the presence/absence data for the target species for the whole of South Africa in a particular year.
2.	Remove submitted cards with no hours spent actively bird watching.
3.	Correct for erroneous records within the total hours variable.
4.	Correct for erroneous records within the total species variable.
5.	Import the relevant year-specific environmental data.
6.	Standardize the environmental data to remove problems with regard to measurement scale.
7.	Extract environmental data for provinces of interest.
8.	Remove pentads from presence/absence data that fall outside of the region of interest.
9.	Convert total hours and total species variables to their logarithmic form.
10.	Create Julian date variable.
11.	Create a categorical variable called Season using meteorological date of observation.
12.	Scale the observation-level covariates.
13.	Join the environmental, observational and presence/absence data to create the structure needed to conduct occupancy analysis.

Table A.2: Guide to the data cleaning algorithm

Appendix B

Supplementary Results



(a) `Elevation` (2010)

(b) `NDVI` (2010)

(c) `Permanent_Water` (2015)

(d) `Population.Density` (2010)

(e) `Pressure` (2010)

(f) `Wind.Speed` (2010)

Figure B.1: Additional environmental covariates

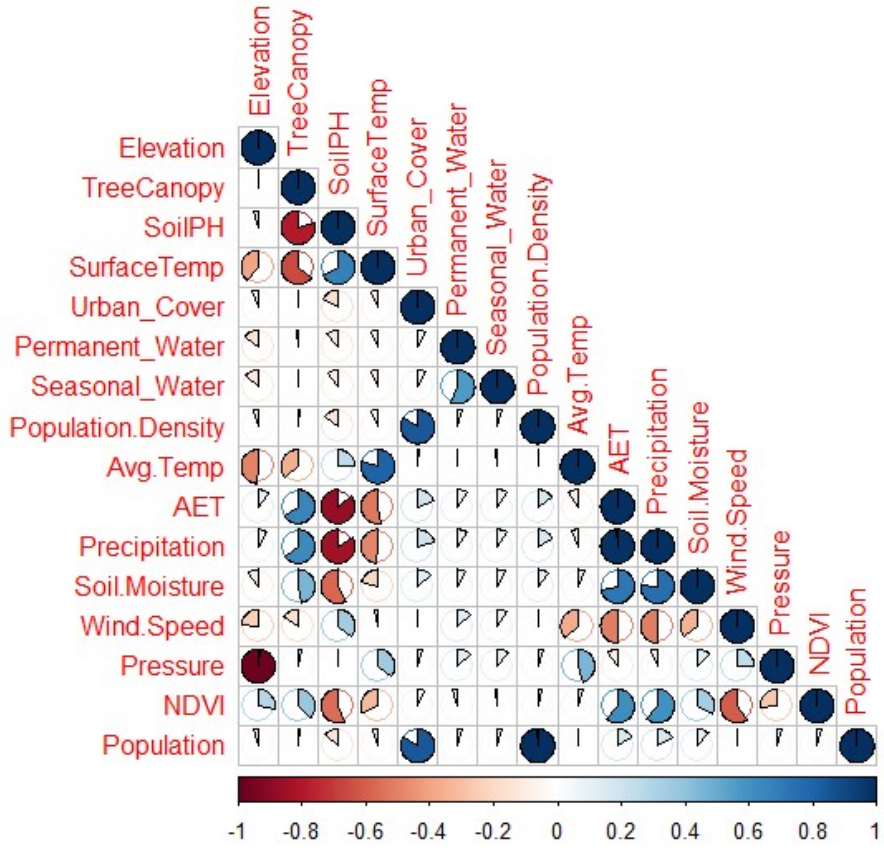


Figure B.2: Correlation plot for the 2019 environmental covariates for which the filled proportion of the pie charts denote the absolute correlation coefficient between 0 and 1.

Structure	Hypothesis	K	QAIC	ΔQAIC	QAICWt
Common10	Env.Clim.Urban	39	16706.20	0.00	0.96
	Clim.Urban.Wind	35	16712.35	6.16	0.04
	Clim.Urban	33	16759.93	53.73	0.00
	Humid.Urban	29	16772.29	66.09	0.00
	Positive	31	16774.60	68.40	0.00
Intermediate10	Env.Clim.Urban	39	25702.33	0.00	1.00
	Clim.Urban.Wind	35	25726.19	23.86	0.00
	Clim.Urban	33	25767.17	64.84	0.00
	Humid.Urban	29	25810.49	108.16	0.00
	Positive	31	25839.61	137.28	0.00
All10	Env.Clim.Urban	39	31709.22	0.00	1.00
	Clim.Urban.Wind	35	31743.00	33.77	0.00
	Clim.Urban	33	31772.61	63.39	0.00
	Humid.Urban	29	31861.16	151.93	0.00
	Positive	31	31880.69	171.47	0.00

[3]

Table B.1: Updated dynamic occupancy model selection procedure: Myna

[3]The updated model selection procedure could only be run for the Myna since the dispersion estimates were considered too large (> 4) to run by the **AICcmodavg** package.

Bibliography

Akaike, H. (1973), 'Maximum likelihood identification of gaussian autoregressive moving average models', *Biometrika* **60**(2), 255–265.

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

Azuma, D. L., Baldwin, J. A. and Noon, B. R. (1990), *Estimating the occupancy of spotted owl habitat areas by sampling and adjusting for bias*, Vol. 124, US Department of Agriculture, Forest Service, Pacific Southwest Research Station.

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

Barbraud, C., Nichols, J. D., Hines, J. E. and Hafner, H. (2003), 'Estimating rates of local extinction and colonization in colonial species and an extension to the metapopulation and community levels', *Oikos* **101**(1), 113–126.

BIRDIE Development Team (2022), *ABAP: Access to African Bird Atlas Project data*. R package version 0.0.4.
URL: *https://github.com/AfricaBirdData/ABAP*

Bled, F., Nichols, J. D. and Altwegg, R. (2013), 'Dynamic occupancy models for analyzing species' range dynamics across large geographic scales', *Ecology and evolution* **3**(15), 4896–4909.

Bomford, M. and Sinclair, R. (2002), 'Australian research on bird pests: impact, management and future directions', *Emu* **102**(1), 29–45.

Broms, K. M., Johnson, D. S., Altwegg, R. and Conquest, L. L. (2014), 'Spatial occupancy models applied to atlas data show southern ground hornbills strongly depend on protected areas', *Ecological Applications* **24**(2), 363–374.

Burnham, K. P. and Anderson, D. R. (2002), 'A practical information-theoretic approach', *Model selection and multimodel inference* **2**.

Clark, C. W. and Rosenzweig, M. L. (1994), 'Extinction and colonization processes: parameter estimates from sporadic surveys', *The American Naturalist* **143**(4), 583–596.

Davis, A. J., McCreary, R., Psiropoulos, J., Brennan, G., Cox, T., Partin, A. and Pepin, K. M. (2018), 'Quantifying site-level usage and certainty of absence for an in-

vasive species through occupancy analysis of camera-trap data', *Biological Invasions* **20**(4), 877–890.

De Marco, P. and Nóbrega, C. C. (2018), 'Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation', *PloS one* **13**(9), e0202403.

de Souza, S. G., Symes, C. T., Smit-Robinson, H. and Mollett, J. M. (2019), 'Minimal evidence of interspecific hybridisation between the yellow-billed duck and introduced mallard in central and northwestern south africa', *Ostrich* **90**(4), 285–301.

Diamond, J. M. and May, R. M. (1977), 'Species turnover rates on islands: dependence on census interval', *Science* **197**(4300), 266–270.

Du Plessis, M. A., Siegfried, W. R. and Armstrong, A. J. (1995), 'Ecological and life-history correlates of cooperative breeding in south african birds', *Oecologia* **102**(2), 180–188.

Erwin, R. M., Nichols, J. D., Eyler, T. B., Stotts, D. B. and Truitt, B. R. (1998), 'Modeling colony-site dynamics: a case study of gull-billed terns (sterna nilotica) in coastal virginia', *The Auk* **115**(4), 970–978.

Fiske, I. and Chandler, R. (2011), 'unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance', *Journal of Statistical Software* **43**(10), 1–23. **URL:** *https://www.jstatsoft.org/v43/i10/*

Geissler, P. and Fuller, M. (1987), 'Estimation of the proportion of an area occupied by an animal species', *Proceedings of the Section on Survey Research Methods of the American Statistical Association* **1986**, 533–538.

Gormley, A. M., Forsyth, D. M., Griffioen, P., Lindeman, M., Ramsey, D. S., Scroggie, M. P. and Woodford, L. (2011), 'Using presence-only and presence–absence data to estimate the current and potential distributions of established invasive species', *Journal of Applied Ecology* **48**(1), 25–34.

Greenwood, J. J. (2007), 'Citizens, science and bird conservation', *Journal of Ornithology* **148**(1), 77–124.

Harebottle, D., Smith, N., Underhill, L. and Brooks, M. (2007), 'Southern african bird atlas project 2: instruction manual', *Animal Demography Unit, University of Cape Town, Cape Town* .

Hines, J. E., Nichols, J. D., Royle, J. A., MacKenzie, D. I., Gopalaswamy, A., Kumar, N. S.

and Karanth, K. (2010), 'Tigers on trails: occupancy modeling for cluster sampling', *Ecological Applications* **20**(5), 1456–1466.

Hongyi Li, G. and Maddala (1996), 'Bootstrapping time series models', *Econometric reviews* **15**(2), 115–158.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C. and Pond, B. A. (2013), 'Spatial occupancy models for large data sets', *Ecology* **94**(4), 801–808.

Karanth, K. U., Gopalaswamy, A. M., Kumar, N. S., Vaidyanathan, S., Nichols, J. D. and MacKenzie, D. I. (2011), 'Monitoring carnivore populations at the landscape scale: occupancy modelling of tigers from sign surveys', *Journal of Applied Ecology* **48**(4), 1048–1056.

Kendall, W. L. (1999), 'Robustness of closed capture–recapture methods to violations of the closure assumption', *Ecology* **80**(8), 2517–2525.

Kendall, W. L., Nichols, J. D. and Hines, J. E. (1997), 'Estimating temporary emigration using capture–recapture data with pollock's robust design', *Ecology* **78**(2), 563–578.

Kéry, M. and Chandler, R. (2012), 'Dynamic occupancy models in unmarked', *Availableat: http://cran. r-project. org/web/packages/unmarked/vignettes/colext. pdf (Accessed 20 April 2015)* .

Kéry, M., Guillera-Arroita, G. and Lahoz-Monfort, J. J. (2013), 'Analysing and mapping species range dynamics using occupancy models', *Journal of Biogeography* **40**(8), 1463–1474.

Ludwig, J. A., Reynolds, J. F., QUARTET, L. and Reynolds, J. (1988), *Statistical ecology: a primer in methods and computing*, Vol. 1, John Wiley & Sons.

MacKenzie, D. I. and Bailey, L. L. (2004), 'Assessing the fit of site-occupancy models', *Journal of Agricultural, Biological, and Environmental Statistics* **9**(3), 300–318.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G. and Franklin, A. B. (2003), 'Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly', *Ecology* **84**(8), 2200–2207.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J. and Langtimm, C. A. (2002), 'Estimating site occupancy rates when detection probabilities are less than one', *Ecology* **83**(8), 2248–2255.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. and Hines, J. E.

(2006), *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*, Elsevier.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L. and Hines, J. E. (2017), *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*, Elsevier.

Malarvizhi, R. and Thanamani, A. S. (2012), 'K-nearest neighbor in missing data imputation', *International Journal of Engineering Research and Development* **5**(1), 5–7.

Mansournia, M. A., Geroldinger, A., Greenland, S. and Heinze, G. (2018), 'Separation in logistic regression: causes, consequences, and control', *American journal of epidemiology* **187**(4), 864–870.

Mazerolle, M. J. (2020), *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*. R package version 2.3-1.
URL: *https://cran.r-project.org/package=AICcmodavg*

Measey, J., Hui, C. and Somers, M. J. (2020), 'Terrestrial vertebrate invasions in south africa', *Biological Invasions in South Africa* pp. 787–830.

Miller, D. A., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E. and Mitchell, M. S. (2013), 'Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data', *PLoS one* **8**(6), e65808.

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L. and Weir, L. A. (2011), 'Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification', *Ecology* **92**(7), 1422–1428.

Naimi, B., a.s. Hamm, N., Groen, T. A., Skidmore, A. K. and Toxopeus, A. G. (2014), 'Where is positional uncertainty a problem for species distribution modelling', *Ecography* **37**, 191–203.

Nichols, J. and Karanth, K. (2002), *Statistical concepts: assessing spatial distributions*, Centre for Wildlife Studies, Bangalore, India.

Norris, J. L. and Pollock, K. H. (1996), 'Nonparametric mle under two closed capture-recapture models with heterogeneity', *Biometrics* pp. 639–649.

Peacock, D. S., J. van Rensburg, B. and Robertson, M. P. (2007), 'The distribution and spread of the invasive alien common myna, acridotheres tristis l.(aves: Sturnidae), in southern africa', *South African Journal of Science* **103**(11), 465–473.

Pell, A. and Tidemann, C. (1997), 'The ecology of the common myna in urban nature reserves in the australian capital territory', *Emu-Austral Ornithology* **97**(2), 141–149.

Pollock, K. H. (1982), 'A capture-recapture design robust to unequal probability of capture', *The Journal of Wildlife Management* **46**(3), 752–757.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Royle, J. A. and Nichols, J. D. (2003), 'Estimating abundance from repeated presence–absence data or point counts', *Ecology* **84**(3), 777–790.

Sepulveda, A. J. (2018), 'Novel application of explicit dynamics occupancy models to ongoing aquatic invasions', *Journal of applied ecology* **55**(2), 917–925.

Shivambu, T. C., Shivambu, N. and Downs, C. T. (2020), 'Impact assessment of seven alien invasive bird species already introduced to south africa', *Biological Invasions* pp. 1–19.

Simberloff, D. S. (1969), 'Experimental zoogeography of islands: a model for insular colonization', *Ecology* **50**(2), 296–314.

Stephens, K., Measey, J., Reynolds, C. and Le Roux, J. J. (2020), 'Occurrence and extent of hybridisation between the invasive mallard duck and native yellow-billed duck in south africa', *Biological Invasions* **22**(2), 693–707.

Szumilas, M. (2010), 'Explaining odds ratios', *Journal of the Canadian academy of child and adolescent psychiatry* **19**(3), 227.

Wagenmakers, E.-J. and Farrell, S. (2004), 'Aic model selection using akaike weights', *Psychonomic bulletin & review* **11**(1), 192–196.

Weir, L., Fiske, I. J. and Royle, J. A. (2009), 'Trends in anuran occupancy from northeastern states of the north american amphibian monitoring program', *Herpetological Conservation and Biology* **4**(3), 389–402.

Wilson, E. O. and MacArthur, R. H. (1967), *The theory of island biogeography*, Princeton University Press.